ANALYTICAL MODEL FOR DATA MINING PROCESSES IN HIGHER LEARNING SYSTEMS

HARIDAN BIN ABDUL RAHIM

NOVEMBER 2005

"I admit that I have read this masterpiece and in my point of view this masterpiece achieved the scope and quality for the purpose of graduation in Bachelor of Electrical Engineering (Power Industry)"

Signature:

Name Of Supervisor: Encik Saifulza B Alwi@Suhaimi

Date : 18 th November 2005

ANALYTICAL MODEL FOR DATA MINING PROCESSES IN HIGHER LEARNING SYSTEMS

HARIDAN BIN ABDUL RAHIM

This Report Is Submitted In Partial Fulfillment Of Requirements For The Degree of Bachelor In Electrical Engineering (Industry Power)

Fakulti Kejuruteraan Elektrik
Kolej Universiti Teknikal Kebangsaan Malaysia

November 2005

"I admit that this project is written by me and is my own effort and that no part has been plagiarized without citations"

Signature :

Name : Haridan Bin Abdul Rahim

Date : 18th November 2005

Specially dedicated to my beloved family

ACKNOWLEDGEMENT

First of all, syukur alhamdulillah to ALLAH S.W.T on his blessing to make this project successful. I would like to express my deepest gratitude to ex supervisor, Mr. Abu M Wahidullah for his invaluable knowledge, suggestion, advise and most importantly his guidance towards the completion of this project. Also to Encik Saifulza for his guiding through the completion of this thesis. I would also like to thank my lecturers and friends who had helped me a lot during the completion of this project. Last but not least, to my family who had sacrificed their time guiding and giving me all the support I needed to complete this project.

ABSTRACT

Data mining is a process of discovering various models, summaries, and derived values from a given collection of data. Data mining is not a random application of statistical, machine learning, and other methods and tools. It is not a random walk through the space of analytic techniques but a carefully planned and considered process of deciding what will be most useful, promising and revealing. One of the main worries of higher learning system, is evaluating and enhancing the educational structure of an Organization. This project will contribute through an analytical model the aspects of data mining processes and technologies. This could be utilized as a decision support tool for a higher learning system to develop a data mining system to assist and improve the traditional processes.

ABSTRAK

Data mining adalah suatu proses penghasilan berbagai-bagai model, kesimpulan daripada jumlah bilangan data yang banyak. Data mining bukanlah suatu aplikasi secara rawak dalam bidang statistik atau bidang-bidang yang lain. Ia adalah suatu proses secara automatik untuk memilih data yang berguna daripada sejumlah data yang banyak. Ia juga merupakan proses yang terancang dan teratur untuk memilih apa yang paling berguna untuk menghasilkan sesuatu projek yang berkesan. Salah satu kebimbangan dalam sistem pembelajaran tinggi adalah penilaian dan peningkatan stuktur pendidikan dalam sesuatu organisasi. Projek ini memberi sumbangan terhadap aspek proses dan teknologi dalam data mining menerusi model analitikal. Ini dapat digunakan sebagai alat untuk membuat keputusan dalam sistem pembelajaran tinggi untuk menghasilkan sistem data mining yang boleh membantu dan memperbaiki proses pembelajaran tradisional.

CONTENTS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing huge sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make useful decisions. Data mining tools can make decision faster than the traditional way by improving the system and can save lots of time and cost. They analyze through databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining the data to result a useful outcome. Both processes require either sifting through an immense amount of data, or intelligently probing it to find where the value resides.

1.1 Objectives

- To understand the data mining processes
- To improve the traditional way of learning system
- To create a data mining processes that can be useful in higher learning system

1.2 Problem Statement

- Understanding and applying data mining system in the right perspective
- Abundance of free flow of information and choosing the right sources such as journals, books and websites.

CHAPTER 2

LITERATURE REVIEW

Many researches have been done before I completed my project. These researches are important because through these processes, I gained a lot of knowledge that are very useful to develop my data mining system.

2.1 Data Mining Definition

The definition from Gartner Group seems to be most comprehensive, as they define data mining as .the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques.

Data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data.

2.2 Applications of Data Mining in Higher Education

There are many transferable techniques that can be applied in higher education system. Algorithms are similar in concept to stored procedures in object related programming in that they are universally applicable. Almost all algorithms or models currently used in the business sector are directly usable for research in higher education, especially in institutional research.

2.3 Graphical Support for a Data Mining Process.

The effectiveness of data mining as a business intelligence tool has been demonstrated with a large number of successful applications. However, in order to give data mining a wider appeal it has become apparent that a methodology or process is required to allow non data mining specialists to achieve the same degree of success as seasoned practitioners. Such a systematic and repeatable process will allow data mining to be successfully deployed by many people across organizations.

It is reassuring to see a common data mining process (methodology) starting to emerge. There is broad agreement on the main tasks within such a process which are data preparation, data exploration, pattern discovery, pattern validation and pattern deployment. XpertRule Miner provides a graphical environment for supporting all the stages of the data mining process. The click, drag and drop environment allows non programmers to carry out complex data preparation, mining and deployment processes.

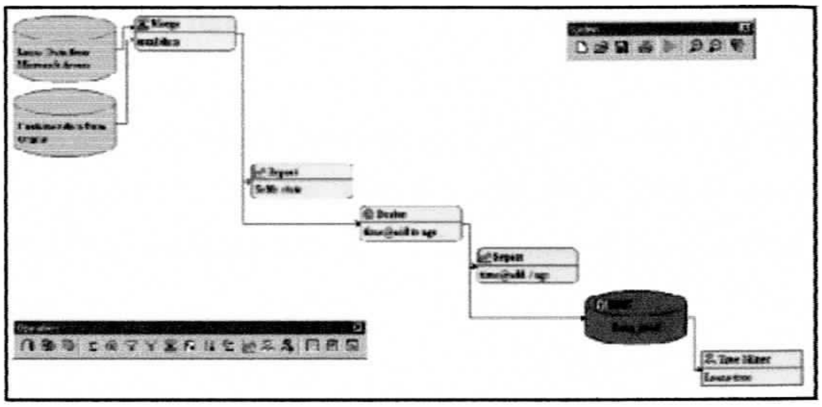

Figure 2.3: Graphical data transformation

2.4 Data Mining in a Scientific Environment.

The computer has brought the ability to generate and store huge amounts of data. For example, it is not unusual for power users to have the equivalent of three or four encyclopedias worth of data online.

When we add the data generated by government and other organizations, such as the recently completed census or the data collected every time we make a purchase at any modern supermarket, the volume of data available is almost incomprehensible. The problem is in how to turn this data into usable information.

This is not a new phenomenon. Scientists, especially experimentalists, have long had to tackle this problem. While Isaac Newton may have formulated his theory of gravity when an apple fell on his head, it was still followed by hundreds, if not thousands, of experiments demonstrating, validating and/or refining the original equation.

Over the centuries, various methods have been developed to deal with this volume of data, many of which were seen as major steps forward for mathematics at the time. Some of these methods include Fast Fourier Transforms, Multivariate Regression Analysis, as well as a whole range of statistical methods. More recently, Visualization has been widely adopted by scientists as a means of studying the ever-growing masses of data.

2.5 Verification Driven Data Mining

The most common use of Data Mining is verification driven, and is primarily aimed at confirmation of an idea. Basically, the mechanism is to propose some association or pattern and then to study the data to find support, or otherwise, for the proposal.

There are a number of standard techniques used in verification driven mining; these include the most basic form of query and reporting, presenting the output in graphical, tabular and textual forms, through to multi-dimensional analysis and on to statistical analysis.

2.6 Discovery Driven Data Mining.

The discovery driven approach depends on a much more sophisticated and structured search of the data for associations, patterns, rules or functions, and then having the analyst review them for value.

There are four common techniques for data mining approaches such as predictive modeling including neural nets, link-analysis technique which attempts to establish links between records, database segmentation which partitions the data into collections of related records, and finally deviation detection which identifies point that do not fit in a segment

Table 2.6: Example of How Data Mining Can Improve the Traditional Analysis Approach

Traditional analysis approach	Data mining discovery
An analyst may want to study the buying behavior of known classes of customers (e.g., retired school teachers or young urban professionals), to help design targeted marketing programs. First, the analyst would use known characteristics about those classes of customers and try to sort them into groups. Second, he or she would study the buying behaviour common to that group. The analyst would repeat this process until he or she was satisfied with the final customer groupings	The data mining tool would study the database to identify all groups of customers with distinct buying patterns. After the data is mined, the analyst could use various query, reporting, and multi-dimensional analysis tools to work with the results.

2.7 Software for the data mining course

The following software packages are recommended to use for the data mining projects. It is not advisable to use too many different tools together, as each of them will need its own data format, and a lot of time and effort will be wasted doing data conversions.

2.7.1 Weka

General Description:

Weka is open source data mining software. It does not only support machine learning algorithms, but also data preparation and meta-learners like bagging and boosting. The whole suite is written in java, so it can be run on any platform.

Functionalities

Decision trees: ID3 and C4.5 are implemented, and M5': a model tree induction algorithm for predicting numeric values (each leaf node has a regression model). PART is a rule-learner that makes rules by building different decision trees and each time keeping the leaf with the largest coverage.

Memory-based methods: kNN and locally weighted regression.

Neural Networks: only back propagation with momentum is supported. Simpler methods: naive Bayes (for numeric values, a normal distribution is used, but also 'kernel density estimation' can be used to avoid assuming a normal distribution) and linear regression are useful simple methods. Two-class logistic regression is also supported. The algorithm uses a 'ridge estimator'.

Advantages:

The obvious advantage of a package like Weka is that a whole range of data preparation, feature selection and data mining algorithms are integrated. This means that only one data format is needed, and trying out and comparing different approaches becomes really easy.

Disadvantages:

Probably the most important disadvantage of data mining suites like this is that they do not implement the newest techniques.

2.7.2 Netlab

General description:
Netlab is a toolbox of matlab functions and scripts, originally based on Bishop's Neural networks for pattern recognition book. They provide matlab implementations of some of the newest machine learning algorithms.

Functionalities:
This is a list of the functionalities supported by Netlab.
Gaussian mixture model with EM training algorithm
Linear and logistic regression with IRLS training algorithm
Multi-layer perceptron with linear, logistic and softmax outputs and appropriate error functions.

Advantages:
Working with netlab surely offers some important advantages. Compared to a package like Weka, the greatest advantage is probably that the algorithms implemented are up to date with the newest developments in the field. Also, using the functions assumes quite some knowledge of the theory behind the algorithms, so at least the data mining analyst knows what is going on. Finally, the integration in matlab gives some important advantages: all the matrix calculation and visualisation functions are available, which can be very useful for data preparation and exploration, and analysis of results afterwards. Also, Matlab makes scripting is possible, which should increase the efficiency of the analysis.

Disadvantages:
Disadvantages are that some important data preprocessing functionalities are missing: dealing with missing data, feature selection. Another problem is that matlab can only work with numeric data, so categorical data will have to be converted in advance, and techniques like apriority and decision trees cannot be implemented.

2.7.3 Bow

General Description:
Bow is a library of C code for statistical text analysis, language modeling and information retrieval. Together with this library, four executable programs based on it are distributed. They are Rainbow (for document classification), Arrow and Archer (for document retrieval) and Crossbow (for document clustering).

Functionalities:
Rainbow: Rainbow is the front-end to the library that supports text classification. It is the best documented of the four, and is the most useful for the projects. It has the following functionalities:
Text preprocessing: This supports the conversion of texts into bag-of-words format. Different options allow to choose a sparse or a full matrix model, to display a count or a binary present/absent flag, and to include the word names or not. Also, feature selection methods can be used, based on word frequency, word-document counts or information gain.

Evaluation: Especially the text preprocessing functionalities are very useful for text related projects. Once the conversion to bag-of-word format is done, you have the option to use the classification and clustering features of the bow package, or to export the matrix to other data mining programs. The classification functionalities of Rainbow give very fast and accurate results. The documentation for Rainbow is quite good.

CHAPTER 3

METHODOLOGY

Data Mining consists of three components such as the captured data that must be integrated into organization-wide views and often in a Data Warehouse, the mining of this warehouse and the organization and presentation of this mined information to enable understanding. The data capture is a standard process of gathering, organizing and cleaning up; for example, removing duplicates, deriving missing values where possible, establishing derived attributes and validation of the data. Much of the following processes assume the validity and integrity of the data within the warehouse.

3.1 Conducting Data Mining As a Process.

The following are the major stages in organizations for conducting data mining as a process:

1. Define the problem, including accurately describing it, to determine the appropriateness of using data mining, to decide the form of input and output, to decides cost effectiveness and so on
2. Collect and select data, such as deciding which data to collect and how to collect them
3. Prepare data, such as transform data to a certain format, data cleansing or integrate data from different sources.