2006

Web Data Mining: A Case Study

Samia Jones
Galveston College

Omprakash K. Gupta
Prairie View A&M University

Follow this and additional works at: http://scholarworks.lib.csusb.edu/ciima

Part of the Management Information Systems Commons

Recommended Citation

Web Data Mining: A Case Study

Samia Jones
Galveston College,
Galveston, TX 77550

Omprakash K. Gupta
Prairie View A&M,
Prairie View, TX 77446
okgupta@pvamu.edu

Abstract

With an enormous amount of data stored in databases and data warehouses, it is increasingly important to develop powerful tools for analysis of such data and mining interesting knowledge from it. Data mining is a process of inferring knowledge from such huge data. The main problem related to the retrieval of information from the World Wide Web is the enormous number of unstructured documents and resources, i.e., the difficulty of locating and tracking appropriate sources. In this article, a survey of the research in the area of web mining and suggest web mining categories and techniques. Furthermore, a presentation of a web mining environment generator that allows naive users to generate a web mining environment specific to a given domain by providing a set of specifications.

Introduction

In business today, companies are working fast to gain a valuable competitive advantage over other businesses. A fast-growing and popular technology, which can help to gain this advantage, is data mining. Data mining technology allows a company to use the mass quantities of data that it has compiled, and develop correlations and relationships among this data to help businesses improve efficiency, learn more about its customers, make better decisions, and help in planning. Data Mining has three major components *Clustering* or *Classification,* Association *Rules* and *Sequence Analysis.* This technology can develop these analyses on its own, using a blend of statistics, artificial intelligence, machine learning algorithms, and data stores.

Data Mining

Data mining is a tool that can extract predictive information from large quantities of data, and is data driven. It uses mathematical and statistical calculations to uncover trends and correlations among the large quantities of data stored in a database. It is a blend of artificial intelligence technology, statistics, data warehousing, and machine learning.

Data mining started with statistics. Statistical functions such as standard deviation, regression analysis, and variance are all valuable tools that allow people to study the reliability and relationships between data. Much of what data mining does is rooted in statistics, making it one of the cornerstones of data mining technology.

In the 1970's data was stored using large mainframe systems and COBOL programming techniques. These simplistic beginnings gave way to very large databases called "data warehouses", which store data in one standard format. The dictionary definition of a data warehouse is "a generic term for storing, retrieving, and managing large amounts of data." (dictionary.com).These data warehouses "can now store and query terabytes and megabytes of data in sophisticated database management systems." (Carbone,2000) These data stores are an essential part of data mining, because a cornerstone of the technology is that it needs very large amounts of organized data to manipulate.

In addition to basic statistics and large data warehouses, a major part of data mining technology is artificial intelligence (AI). Artificial intelligence started in the 1980's with a set of algorithms that were designed to teach a computer how to "learn" by itself. As they developed, these algorithms became valuable data manipulation tools and were applied to large sets of data. Instead of entering a set of pre-defined hypothesis, the data mining software, combined with AI technology was able to generate its own relationships between the data. It was even able to analyze data and discover correlations between the data on its own, and develop models to help the developers interpret the relationships that were found.

AI gave way to machine learning. Machine learning is defined as "the ability of a machine to improve its performance based on previous results." (dictionary.com) Machine learning is the next step in artificial intelligence technology because it blends trial and error learning by the system with statistical analysis. This lets the software learn on its own and allows it to make decisions regarding the data it is trying to analyze.

Later in the 1990's data mining became wildly popular. Many companies began to use the data mining technology and found that it was much easier than having actual people work with such large amounts of data and attributes. This technology allows the systems to "think" for themselves and run analysis that would provide trend and correlation information for the data in the tables. In 2001, the use of data warehouses grew by over a third to 77%. (Hardison, 2002).

Data mining is a very important tool for business and as time goes on, business is becoming more and more competitive and everyone is scrambling for a competitive edge (Hardison, 2002). Businesses need to gain a competitive edge, and can get it from the increased awareness they can get from data mining software that is available on the market right now (Montana, 2001) .

Data Mining Evolutionary Chart

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query	Oracle, Sybase, Informix,	Retrospective, dynamic data delivery at

		Language (SQL), ODBC	IBM, Microsoft	record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

*This chart is from: http://www.thearling.com/text/dmwhite/dmwhite.htm

Web Mining

Web Mining rapidly collect and integrate information from multiple Web sites. This approach relies on computing power, rather than programmer brain power, to handle many of the complex issues that arise when gathering information from disparate sources. One solution is the power by advanced artificial intelligence algorithms. For instance, machine learning techniques would be used for extracting information from Web sites. Based on a few examples, the software automatically determines how to extract different types of information from data sources.

Association Rule Algorithms

An association rule is a rule which implies certain association relationships among a set of objects (such as ``occur together" or ``one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y.

Classification Algorithms

In Data classification one develops a description or model for each class in a database, based on the features present in a set of class-labeled training data. There have been many data classification methods studied, including decision-tree methods, such as statistical methods, neural networks, rough sets, database-oriented methods etc.

Sequential Analysis

Here we are looking for a Sequential Patterns, called data-sequences. Each data sequence is an ordered list of transactions (or item sets), where each transaction is a sets of items (literals). Typically there is a transaction-time associated with each transaction. A sequential pattern also

consists of a list of sets of items. The problem is to find all sequential patterns with a user-specified minimum support, where the support of a sequential pattern is the percentage of data sequences that contain the pattern.

This article explains how to build a mining model and to solve a few concern problems to best teach online class using the web and get to know the students to serve them better.

The Study

This article is organized in two parts. The first part of the paper provides a brief presentation of the clustering algorithm, and explains a few basic data mining terms.
The second part of the paper focuses on the performance study of two classes that are using online course over the web and the web generating a mining algorithm using a variety of parameters. A number of experiments were conducted, and their results are presented in this part. The experiments were based on different parameter of interest. The parameters varied form the number of input attributes, the sample size of the class, and so on. The results of these experiments prove that the data mining algorithm are very efficient and scalable.

The Clustering Algorithm used

The Clustering algorithm is based on the Expectation and Maximization algorithm (Microsoft, 2003). This algorithm iterates between two steps. In the first step, called the "expectation" step, the cluster membership of each case is calculated. In the second step, called the "Maximization" step, the parameters of the models are re-estimated using these cluster memberships, which has the following major steps:

1. Assign initial means
2. Assign cases to each mean using some distance measure
3. Compute new means based on members of each cluster
4. Cycle until convergence.
5.

A case is assigned to each cluster with a certain probability and the means of each cluster is shifted based on that iteration. The following table shows a set of data that could be used to predict best achievement. In this study, information was generated on users that included the following list of measurements:

1. Most requested pages
2. Least requested pages
3. Top exit pages
4. Most accessed directories
5. Most downloaded files
6. New versus returning visitors
7. Summary of activity for exam period
8. Summary of activity by time increment
9. Number of views per each page
10. Page not found

The relevant measurements are viewed which implied the following:

Table 1: Statistics on Web site visits

Statistics Report		
Hits	Entire Site (Successful)	2390
	Average per day	72
	Home page	256
Page Views	Page views	213
	Average per day	97
	Document views	368
Visitor Sessions	Visitor sessions	823
	Average per day	65
	Average visitor session length	00:31:24
Visitors	Visitors who visited once	32
	Visitors who visited more than once	75

RESULTS

This brief case study gives a look at what statistics are commonly measured on web sites. The results of these statistics can be used to alter the web site, thereby altering the next user's experience. Table 1 displays some basic statistics that relate to frequency, length and kind of visitor. In Table 1 additional insights are gained with a breakdown of visitors by each day. This behavior might reflect variation in activity related to exams time or other issues. Monitoring and understanding visitor behavior is the first step in evaluating and improving the web site. Another relevant measurement is how many pages are viewed. This can reflect content as well as navigability. If a majority of visitors viewed only one page, it may imply that they did not find it easy to determine how to take the next step.

CONCLUSION

The web offers prospecting and user relationship management opportunities that are limited only by the imagination. Data mining is a tool that can extract predictive information from large quantities of data, and is data driven. It uses mathematical and statistical calculations to uncover trends and correlations among the large quantities of data stored in a database. It is a blend of artificial intelligence technology, statistics, data warehousing, and machine learning. This data mining technology is becoming more and more popular, and is one of the fastest growing technologies in information systems today. The future of data mining is wide open, and it will be exciting to see how far this technology will go.

REFERENCES

Carbone, P. (August, 2000). What is the Origin of Data Mining?
www.mitre.org/pubs/edge/august_00/carbone.htm.

Evolutionary Chart http://www.thearling.com/text/dmwhite/dmwhite.htm

"Data Mining Overview." http//:www.data-mine.com/

"Data Mining" Def. www. Dictionary.Com

"Data Warehouse" Def. www.Dictionary.com

Hardison,. (2002). Data Mining: The New Gold Rush. *Pharmaceutical Executive*. March, 26, 28, 30.

Microsoft. (2002). Performance Study of Microsoft Data Mining Algorithms. March 25.

Montana, J. (2001). Data Mining: A Slippery Slope. *Information Management Journal.* October, 50-54.