StockSim and JTrader: Two Differing Applications of Business Data-Mining

Final Research Paper

Andrew Baldinger

June 3, 2004

Abstract

The stock market is an intricate network of buying and selling. Every second counts, and every number, no matter how small, must be analyzed to ensure financially stable decisions are being made. Because of the large amounts of data that are required to make these analyzations, data mining can be used to make the process both speedier and more efficient, by "weeding-out" what is needed from the superfluous information. Similarly, when we visit an online seller (ie. Amazon.com) and login to our account, a "profile" is constructed about us using data from our buying history, our personal information, and other sources. Both of these applications in the real world are examples of the use of data mining. Data mining is being used increasingly today to make intelligent decisions based on data acquired from different sources. It is the source of the data that this project examines, as the distinction is becoming ever more important in our connected world.

1 Introduction

This project examines two different applications of data mining which are becoming ever more common in today's business world. The first, an application which I have termed "StockSim" is a real-world simulation of the stock market, allowing teachers to track their students' progress as they navigate the ever-changing markets. This application gathers real- time stock data from the internet, the first type of data mining that I examine. The second application that I examine is a program named JTrader, which was created under a business plan constructed by the Future Business Leaders of America, a student organization at Thomas Jefferson High School. This application examines a different type of data mining; that of gathering data from users. Comparison of these two different applications of data mining will become even more important in years to come, as internet companies strive to integrate data with consumer-friendly interfaces.

1.1 Data Mining the Web

The exponential growth of the Web has made it into a huge inter-connected source of information. It is estimated that the Web today has close to 7 billion static pages, and millions of databases at various web sites that can generates many more billions of dynamic pages. Today there are three main methods of accessing information from the Web:

1. Using a search engine like Google to find pages containing certain keywords

2. Using a Web index like Yahoo to find pages relevant to certain concepts Randomly browsing or surfing the web, going from page to page, using browsers like Internet Explorer.

These methods of accessing data on the internet are all based on using the web in an exploratory manner. This may be useful for individuals looking for information, especially for their private purposes, but the needs of organizations or corporations accessing the Web as an information source is turning out to be quite different. In addition, the need to gather updated information for company websites is also necessitating different data extraction techniques. For example, financial analysts in the equity research department of Morgan Stanley do not care (as much) about being able to search through the 1.4 billion web pages that Google indexes, but are very interested in full details that can be obtained from the top 100 financial sites, e.g. MorningStar, Wall Street Journal, etc. While data mining in this case does not remove all the need for exploration of the internet to take place, it does allow for the establishment of ”information channels” that provide the equity researchers high-quality and detailed information from the top 100 financial sites on a continuous basis.

1.2 Why Data Mining?

Information provided is usually much more valuable if it in the form of business intelligence, rather than just text found on web pages. This automatic extraction of business intelligence from relevant web sites is being called Web Intelligence. The past couple of years have seen the introduction of a set of technologies that are making Web Intelligence possible. First of these is Web Content Acquisition, which provides the capability of not only searching static pages at a site, but also of crawling entire sites, and executing search engines of web sites to extract dynamic pages. There has been considerable work on developing tools that can be used for rapid development of wrappers and crawlers used for these tasks. The second technology is Web Mining, which is the application of data mining techniques to content, structure, and usage data from the Web to extract higher order knowledge. Web Mining has seen rapid growth in the past few years. The third and final technology is the creation management of user profiles, which is being used for personalization and relevance analysis. This type of personalization is what my program design feeds off of, and is the basis of my techlab project.

2 Background on Data Mining

Data mining refers to the automated process of searching data for relationships and patterns. It is often done under the supervision of a human agent. The computer using data mining algorithms will identify possible patterns in data and the human agent will visually check to see if the patterns are relevant.

Applications Applications of data mining include credit risk assessment in financial institutions, fraud detection in credit card companies, sales analysis for retail industries and any company that collects large amounts of historical data. It is quite useful in that it can improve revenue and reduce costs in a company. E.g. in a company that uses postal mail to market products to customers, data mining can select the attributes of customers that best predict if they will buy a product from a mailing campaign. Traditionally the company may only be using customers income to target the mailings. However the data mining process may discover that marital status, age and income are much better predictors of the chances of a customer buying a product from a mailing campaign than income alone. Based on this information the mailing company could target their mailing to a more select group of customers, thereby reducing their costs and increasing their profits. One important thing to note about data mining is that it does not replace the need for a good understanding of the data and business, and sound judgment in making evaluations between the spurious and significant knowledge discoveries from data.

Increasing Productivity and Profitability Using Data Mining With the intentions of increasing productivity and profitability through the use of data mining, there are two types of data that can be gathered when it is applied to a given situation. First are ”descriptive tasks.” This is data which describes general, overarching properties. Second are ”predictive tasks. This is data which infers on data that is available. Below these two levels of classification, there are also varying levels of data mining functionality. These are as follows:

1. **Characterization**
2. **Discrimination**
3. **Association**
4. **Prediction**
5. **Classification**
6. **Clustering**
7. **Characterization**
8. **Outlier Analysis**
9. **Time Series Analysis**

Current Fortune 500 Companies and Data Mining Figure 1 (above) illustrates the recent willingness of largescale companies to adapt to changing times and begin to integrate data mining applications into their businesses.

Current Research in Data Mining Currently, several programs exist that perform similar data mining tasks to my program. It is possible to implement a

```
C#project.php.NET Object
```

Oriented Application with the .NET Framework, XML.NET and ADO.NET for analyzing customer purchases, orders and transactions for buying patterns that reflect items frequently purchased together. The

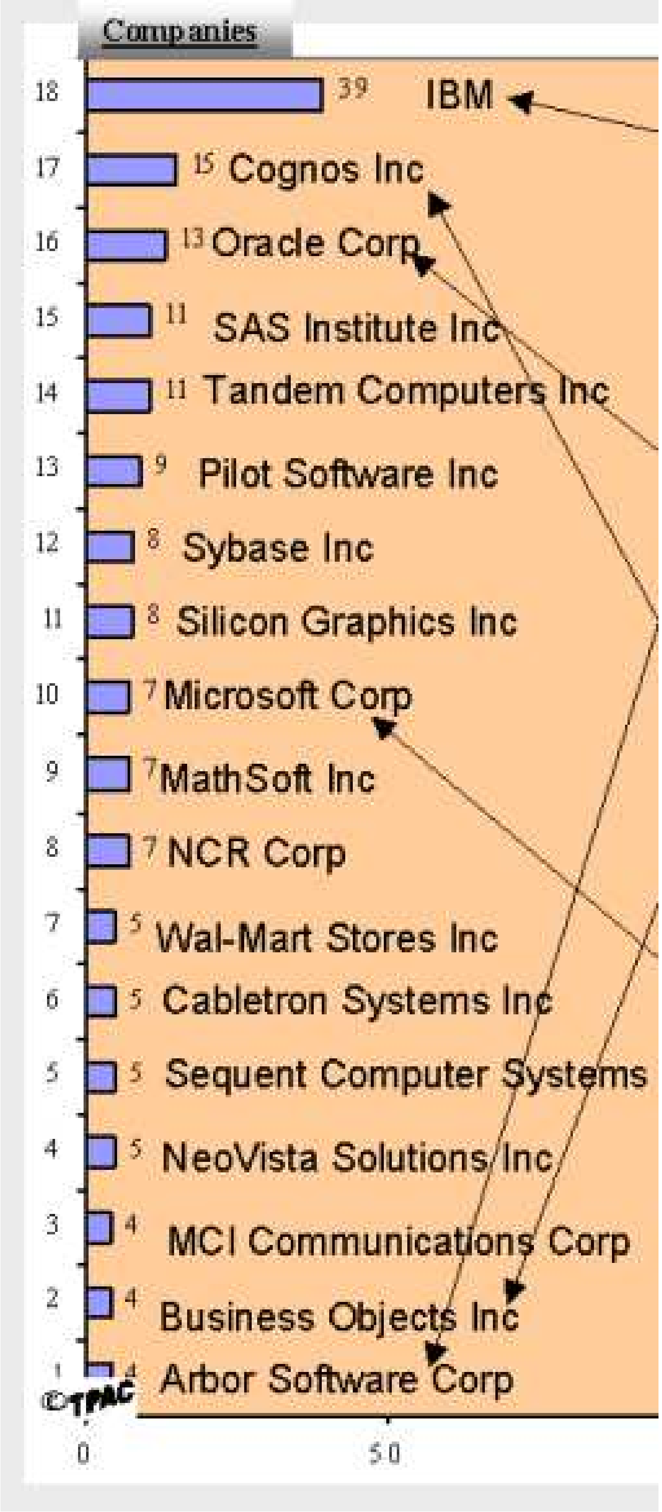

Figure 1: Company Utilization of Data Mining

`C# .NET Sales`

Data Mining Application mines customer transactions (in XML Files and in-memory ADO.NET DataSets) from databases, shopping carts and retail stores and the results are used for marketing, advertising and sales management.

Decision-Tree Data Mining Decision Trees are one of the most popular implementations of data mining. A decision tree data mining algorithm is a visual tool, as it presents the predictions or analysis of data in a tree format. There are three basic components of a decision tree: the decision node, branches and leaves. The first component is the top decision node, or root node, which specifies a test to be carried out. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree you can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Data Mining on the Internet Data mining has its most applicability on the World Wide Web. Data mining can be used to:

1. Extract navigational behavior of web site visitors.

2. Personalize web sites

3. Make product offers customized for the visitor (check Amazon.com)

4. Clickstream mining is the analysis of data created by a user while browsing the web.

5. Clickstream includes items like where someone has gone on the Web, how long they stayed, what they buy, what their interests are, and what sites they return to.

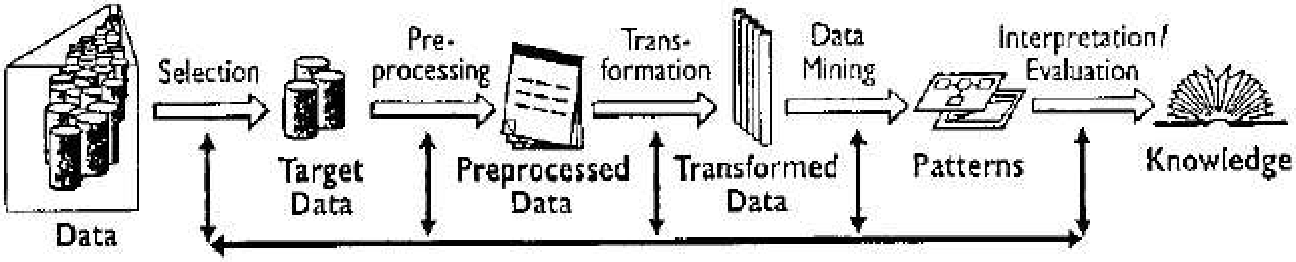

Figure 2: The Data-Mining Process

6. Through data mining, the compiling of detailed demographic information and individual Web-user traits from clickstream data, retailers will be able to create ever more targeted ads with effectiveness rates that are off the charts.

2.1 Data Mining for the Stock Market

As pointed out in above sections, one key aspect of this simulation is the acquiring of real data about stocks, ranging from their current prices to volume to even growth potential. This is done through the use of data mining, and is accomplished with the help of XML. Eventually, however, the data mining will also be used to help advise users as what the best options are for their current portfolio.

2.2 Data Mining Using XML

In order to get data regarding current stock information, a data parser is used that connects to a special Nasdaq website. Thanks to Nasdaq's awareness of the need for open-source real-time data about stocks, they have created an XML-based page (http://quotes.nasdaq.com/quote.dll), where the stock symbol can be inputed in the appropriate place. What results is a page detailing the stock's current data. A PHP script (http://www.tjhsst.edu/ abalding/techlab/parsexml.phpp) then takes this data and reads it into a format the application can understand.

2.2.1 Selecting Data to Be Mined

It is quite easy in this application to define what data needs to be mined, since it is simply the data based on a given stock ticker symbol.

2.2.2 Preprocessing the Data

The way that my application is setup, there are only certain variables that are extracted from the Nasdaq XML webpage. Therefore, preprocessing, or getting rid of superfluous data, is not needed in this application.

2.2.3 Transforming the Data

This step is not needed because XML is provided by Nasdaq. If it was not, the application would need to turn the data into a format that contained certain patterns of data.

2.2.4 Mining the Data

The application tranverses the XML page, pulling certain chunks of data from within their respective bracketed areas. For example, when the application sees the start element, {PE-ratio}, it knows that what it is about to read is the PE ratio of the stock. When it reaches the end element, {/PE-ratio}, it knows to terminate the current case. The data is stored in global variables in the PHP script, thus, the user's portfolio has access to this retrieved data.

2.3 Other Applications of Data Mining

At the current time, data mining methods such as neural networks and decision trees cannot fully replace a financial advisor, but they can certainly aid in the process. Current data mining techniques usually require large amounts of historical data on individual stocks, and their output can be difficult for users to understand. However, through careful planning, they can be morphed to be a useful tool for potential investors.

A data mining tool that would be able to predict stock market returns, for example, would need historical data for the stock dating back upwards of ten years. By extrapolating this data and taking into account exogenous factors, such as growth rate, a model can be outputed. A good model would be characterized by two things:

1. Predictors which are highly statistical and repetitive

2. Predictors which show replication of patterns occurring over history

Stock market returns and other financial data that is mined in this application can be classified in four different areas:

1. Time Series (ie. Opening/Closing Prices, High/Low Prices)

   A time series is a sequence of observations which are ordered in time (or space).

2. Fundamental Factors (ie. Exchange Rates, Interest Rates)

   Fundamental factors consist of assessments of where a currency should be trading based on virtually any criteria but the price action itself. These criteria often include the economic condition of the country that the currency represents, monetary policy, etc.

3. Lagged Returns from the Time Series

4. Technical Factors (ie. Moving Averages)

   Moving averages are an average of data for a certain number of time periods. They ”move” because for each calculation, we use the latest x number of time periods’ data. By definition, a moving average lags the market. An exponentially smoothed moving average (EMA) gives greater weight to the more recent data, in an attempt to reduce the lag.

Based on these classifications, the data can be used in different ways. Time series data, for example, used in conjunction with moving averages, can produce extrapolations proving to be very accurate. My job will be to find the best combination of these classifications and apply it to each individual user's portfolio.

3 Application 1: StockSim

This application applies the idea of web data-mining, or gathering data from other websites. It has two components-the overall portfolio management and teaching simulation, which is web-based, and the "number cruncher" applications, which work behind the scenes. These applications apply data mining to produce real-time results, and eventually, to help guide the user in making smart investment choices.

3.1 The Simulation Design

A user's portfolio in most simple terms is their personal trading area - they are able to make all the decisions on where their money goes. They can buy stocks, sell stocks they own, track certain stocks they are interested in, and research stocks they know nothing about.

3.2 User Portfolio

This portfolio is for the student, and consists of four main pages - current holdings, transaction history, trades, and research.

3.2.1 Buying

A user can buy stock in the simulation by entering the number of shares they wish to purchase and by entering the stock symbol. Upon entry of this information, the application first finds the current price of the stock using data mining techniques described in Section

3. After determining whether the variables that were entered were valid, and whether the user has enough cash on hand to make the purchase, the order is sent to the ”processing” database. This database is used every ten minutes to make trades, thus simulating the real-life time lapse between an order being placed and being processed. After the order has been processed, the money is withdrawn from the user’s account and they are credited with shares in the company. Commission costs can be set by the user to be based on a number of factors, but the default is a charge of ten dollars per trade.

3.2.2 Selling

Upon viewing their current portfolio holdings, or what they currently own, a user is given the option to sell a portion or all of their holdings in a certain company. If the user wishes to do this, they enter the amount of holdings they want to sell. The current price is determined, and the order is sent to the ”processing” database (the same as is used for buying). When the order has been processed, money is credited to the user’s account for the value of the stock and their ownership in the company is terminated. Commission works the same way as for buying.

3.2.3 Stock Tracking

When users wish to investigate further a stock they are interested in buying, they may add it to their personal stock tracker. The stock tracker keeps track of all types of information, including the tracking start day, high and low prices for several different time periods, and earnings ratios.

3.2.4 Stock Research

When users are interested in a stock, but have limited knowledge of it, they may do research through the integrated research feature in their portfolio. This research tool pulls all type of current information about the stock, ranging from current prices to volume to earnings ratios. This information is acquired through data mining applications.

3.2.5 Current Holdings

This window is utilized by users in order to view which stocks they currently own. The window provides information ranging from the date the stock was purchased and the buying price to the current price and current value of the stock. From this window, the user also has the option to sell individual stocks.

3.2.6 Transaction History

This window is utilized by users in order to view previous trades that they have made, including buying and selling. The window provides the user with data such as the date the trade was made, the name of the stock, the amount bought/sold, and the type of transaction that was made.

3.3 Teacher Portfolio

This portfolio is for the teacher, and consists of three main pages - class rankings, simulation data, and class holdings.

3.3.1 Class Rankings

This window is utilized by teachers in order to view the positions of their students in the current simulation. Teachers have the option of ranking their students according to name (alphabetically), total value of portfolio, or total transactions. This page allows the teacher to grade students easily and efficiently. In addition, this page provides links to each user's individual portfolio.

3.3.2 Simulation Data Window

This window is utilized by teachers to control the overall flow and design of their simulation. From this page, statistics such as the date the simulation begins and ends and the commission price that their students must pay can be set and changed. Teachers also have the option of adding 'market pitfalls,' which would include events such as a war, natural disaster, or terrorist attack that would have an effect on the stock market. The teacher can then track student response to these events and monitor students' actions.

3.3.3 Overall Class Holdings

This window is utilized by teachers to view each user's personal trading history and current holdings. This is most important when it comes time to evaluate students, and a history of activity must be generated. The teacher can simply go to this page and generate a student report, which provides all vital information about a student's performance during a given time period.

3.4 The Simulation Hierarchy

The importance of the hierarchy in this simulation is not to be understated, as it deals exclusively with the relationships between the individual user (usually the student) and the teacher portfolios.

3.4.1 Relational Databases

The teacher and student portfolios are linked through the use of relational databases. Each teacher is given a unique ID number, and upon registering, each user provides this ID. Two separate databases are maintained for user data, one for teachers and one for students. The student database holds each user's respective teacher ID. The teacher database holds information about a teacher's simulation; for example, the commission price for their students. When a student logs into the system, the teacher simulation data is retrieved, and used for their trades.

3.4.2 Other Databases

Besides the teacher and student databases which contain pertinent data about the users and their simulations, there are three other databases which drive this application.

Transactions Database This database holds a record of every transaction that is made in the system. It includes the following fields: username, stock, amount, type, date, and time. When individual users or a teacher wishes to view a particular user's transaction history, the username field is invoked and all corresponding trades are noted.

Holdings Database This database holds a record of every current holding of users. It includes the following fields: username, stock, amount, lastdatetraded, buyingprice, and value. As with the transactions database, when individual users or a teacher wishes to view a particular user's current holdings, the username field is invoked and all corresponding holdings are noted.

Stock Tracking Database This database holds a record of every stock that users have currently designated as wishing to track. It includes the following fields: username, stock, dateadded, and startprice. When individual users or a teacher wishes to view the stocks a particular user is tracking, the username field is invoked and all corresponding stocks are noted.

4 Application 2: JTrader

4.1 Application Summary

JTrader, which stands for "Jefferson Trader," is a student fundraiser similar in some respects to eBay, but quite different, as well. This application applies the idea of predictive data mining through data gathered from online users. The overall idea is to provide an efficient way for students and teachers in the Thomas Jefferson community to exchange property that they no longer want, but for which they would like to receive money. At the same time that it serves as a fundraiser for FBLA (Future Business Leaders of America), JTrader also functions as an educational tool, allowing students interested in business to obtain real life experience developing a business plan and carrying it through. JTrader uses data mining techniques to build a user "profile," which is then used to target items which they would be most likely to buy.

4.2 How JTrader Works

When a student or teacher wishes to sell an item, they bring it to FBLA on a designated drop off day. Once the item has been approved for sale (see What Can Be Sold?), the item is retained by FBLA and posted on the JTrader website, located on the TJHSST server. The ad remains online for a period of one month, at the end of which time the student may request that the item remain online or take their item back if it has not sold. In exchange for the publicizing of the item, FBLA takes a one dollar commission on the sale (only if the item sells) . Once the item is online, other interested students and teachers may reserve the item by entering their school network username. A confirmation email is sent to their school email address (user@lan.tjhsst.edu) to confirm that they are a Thomas Jefferson student or staff. Once identity is confirmed, the item is automatically removed from the site. This reservation is in no way contractually binding, but simply allows FBLA to pull the item so that others do not see items that are actively being pursued. FBLA then contacts both the seller and the buyer and arranges a meeting between the two during designated eighth periods. The exchange of the item and money takes place at this time, as well as the commission that FBLA takes. If the person who reserved the item does not claim it within a period of five days, the item goes back onto the website.

4.2.1 Security

Because this service is intended for only the TJHSST community, the online JTrader website requests confirmation of anyone who wishes to reserve an item. The user enters their TJHSST domain username (first initial + seven letters of last name) . An automatic email is then sent to username@ lan. tjhsst. edu, from which they must respond.

4.2.2 Item Status

When a seller brings an item to one of the designated drop off days, the seller' s item is assigned a unique item code. The seller may use this code to check their item' s status online at any time. The item status feature on the JTrader website simply provides a quick way for the seller to check in on their item' s status. They are still contacted, however, when their item sells, or their ad time has expired.

4.3 Data Mining for JTrader

Although still in the planning stages, there are several ways I plan to implement data mining techniques in JTrader. Data mining is frequently used to assign a score to a particular customer or prospect indicating the likelihood that the individual will behave in the way you want. For example, a score could measure the propensity to respond to a particular offer. It is also frequently used to identify a set of characteristics (called a profile) that segments customers into groups with similar behaviors, such as buying a particular product.

4.3.1 Data Mining Used to Increase Selling Ability

Catalogs frequently group products by type to simplify the users task of selecting products. In an on-line store, however, the product groups may be quite different, often based on complementing the item under consideration. In particular, the site can take into account not only the item youre looking at, but what is in your ”shopping cart” (in JTrader’s case, what you have reserved) as well, thus leading to even more customized recommendations.

4.4 Data Mining Implementation

The first step to using data mining techniques with JTrader is clustering to discover which products group together naturally. Some of the clusters are obvious, such as shirts and pants. Others are a little more surprising. These groupings are then used to make recommendations whenever someone looks at an item for sale. A customer profile is then built to help identify those customers who would be interested in the new items. Another option down the road would be for customers to elect to receive e-mail about new products that the data mining models predicted would interest them.

5 References

1. http://www.postech.ac.kr/ bkyi/research.htm

2. http://www.richard.peterson.net/buyontherumor10.html

3. http://mscmga.ms.ic.ac.uk/jeb/track.html

4. http://www.fool.com/ddow/1999/ddow990513.htm

5. http://www.nasdaq.com/reference/glossary.st

6. http://quotes.nasdaq.com/quote.dll

7. http://www.cs.waikato.ac.nz/ ml/weka/book.html

8. http://www.tutorgig.com/encyclopedia/getdefn.jsp

9. http://www.businessweek.com/1999/9930/b3639027.htm

10. http://ftp.cwi.nl/CWIreports/INS/INSR9908.pdf

11. http://www.spss.dk/wp/webmining.pdf

12. Levin, N. and Zahavi, J. (1998), Continuous Predictive Modeling - A Comparative Analysis, The Journal of Interactive Marketing, Vol. 12, pp. 5-22.

13. Levin, N. and Zahavi, J. (1997), Applying Neural Computing to Target Marketing, Journal of Direct Marketing, Vol. 11, pp. 76-93.

14. Levin, N. and Zahavi, J. (1996), Calculating the Regression to-the-Mean Effect: A Comparative Analysis, Journal of Direct Marketing, Vol. 10, pp. 29-40.

15. Liu, H. and Motoda, H. (1998), Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, Boston, MA.

16. Long, S.J. (1997), Regression Models for Categorical and Limited dependent Vari-ables, Sage Publications