PREDICTIVE DATA MINING ON WEB-BASED E-COMMERCE STORE

Jidi Zhao, Tianjin University of Commerce, zhaojidi@263.net
Huizhang Shen, Tianjin University of Commerce, hzshen@public.tpt.edu.cn
Duo Liu, Tianjin University of Commerce
Lei Ding, Tianjin University of Commerce

ABSTRACT

E-Commerce has had a significant impact on merchandise transaction which allows people to transcend the barriers of time and distance and take advantage of global markets and business opportunities, opens up a new world of economic possibility and progress. However, e-Commerce brings opportunities as well as drastic competition. The desire to prevail in the market competition has generated an urgent need for new techniques and automated tools that can intelligently assist corporations in transforming the vast amounts of data collected from web-based e-Commerce sites into useful information and knowledge. In this paper, the authors describe the data mining process on web-based e-Commerce store and present a new predictive optimistic algorithm used to predict customers' behaviors by data mining the information of customers. In this way, the competitive capacity and maximize the overall profits of a corporation can be effectively improved. In the end, the application of this algorithm on e-Commerce store is illustrated by a section of an example analysis.

Keywords: Optimistic Algorithm, data mining, e-commerce store, Web-based, competitive capacity

INTRODUCTION

Although e-Commerce is no longer an unfamiliar topic to people, it continues to be a popular one. Few innovations in human history bring as many potential benefits as e-Commerce does and some people maintain that the e-Commerce revolution is "as profound as the change that came with the industrial revolution"(Clinton and Gore 1997). The benefits to organizations include expanding the marketplace to home market and world market, reducing production cost, distributive cost and stock-holding cost, etc. At the same time, customers enjoy quicker delivery, more purchase choices, less expensive products and services, etc.
E-Commerce brings opportunities as well as drastic competitions. In order to gain existence and development under the market economy environment, most enterprises are crying for effective approaches to improve the attraction to more customers and transform the tremendous amount of data collected from web-based e-Commerce sites into useful information and knowledge.
Data mining technology has been received considerable research attention for more than one decade and several fast algorithms have been developed. But predictive data mining technology arises just during the recent years.

Data Mining on Retail E-Commerce Store

In recent years, popular use of the World Wide Web and web-based e-Commerce sites has flooded us with a tremendous amount of data and information. Comprehensive databases that integrate operational data with customer, supplier, and market information have resulted in an

explosion of information. Competition has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. Using Structured Query Language (SQL) and ODBC can access relational data in massive database and using On-line analytic processing (OLAP) can support making decisions (1), but these are not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems.

Data mining is a powerful technology for recognizing and tracking patterns within data. It helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, customer needs. Data mining tools can answer business questions that traditionally were too time consuming to resolve and predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The most commonly used techniques in data mining are (2)(3): Artificial neural networks, Decision trees, Genetic algorithms, Nearest neighbor method, Rule induction. 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.

In one aspect, data mining is the process that data mining algorithms analyze the historical data, creating mathematical functions that can be used to predict customer behaviors in the future. For example, now you are responsible for a sale campaign using advertisements, coupons and various kinds of discounts to promote products and attract customers. You should carefully analyze the effectiveness of the sale campaign in order to improve company profits. You can use some data mining algorithms for this purpose by comparing the amount of sales and the number of transactions containing the sales items during the sales period versus those containing the same items before or after the sales campaign.

The way in which e-Commerce stores interact with their customers has changed dramatically over the past few years. A customer's continuing business is no longer guaranteed. As a result, companies have found that they need to understand their customers better, and to quickly respond to their wants and needs. In addition, the time frame in which these responses need to be made has been shrinking. It is no longer possible to wait until the signs of customer dissatisfaction are obvious before action must be taken. To succeed, companies must be proactive and anticipate what a customer desires.

After deploying successful applications of data mining, a wide range of companies have leverage the knowledge about customers implicit in a database or data warehouse to reduce costs and improve the value of customer relationships. These organizations can now focus their efforts on the most important (profitable) customers and prospects, and design targeted marketing strategies to best reach them.

Now we use an example to explain the process of data mining in e-Commerce.

Assume that you are the marketing manager for the retail e-Commerce store. Now you are responsible for marketing three merchandises to your customers which are three kinds of shampoo, here we simply number them shampoo I, shampoo II, shampoo III. You are going to deliver three kinds of shampoo gifted products (which are free to customers to introduce products) and their explanations to certain customers. Your goal is to find out customers who

might be interested in these merchandises. Beyond this goal, you are also interested in maximizing the profitability of this marketing campaign.
Besides, there are some constraints about this marketing campaign that should be taken into consideration, which are listed as:
(1) There is a budget limitation for the campaign, which limits the number of offers that could be made.
(2) Only one offer of the three is mailed to each customer.
(3) The number of each offer is the same as others.
How to achieve the aims without violating the specified constraints?
You may have already done some thinking about your customers and their motivations in this area and came up with three kinds of people that might be interested in these merchandises:
(1) Customers who purchase most daily necessities for the whole household.
(2) Customers who have been using or ever used one of these shampoos.
(3) Customers who complain about a kind of shampoo other than the three mentioned above.
The next step is to collect data to support the analysis. This is a relatively easier step because the database in the e-Commerce system has collected huge amounts of data on sales, customer shopping history, goods transportation, consumption and service records, and so on. The historical data contains all information that you have about your customers and their purchases, including demographic and account level information (age, income, marital status, and zip code). In this case, we simply use some SQL operations such as join and selection to select appropriate customers who are one or more of the three kinds people mentioned above. In the end you will end up with a collection of several thousand pieces of information about each customer.
The next step is to choose appropriate model. Many companies such as Pilot, Lockheed, IBM, SGI and numerous startups now can afford various kinds of data mining tools, some of which include a predictive model. The data mining process (which can be called scoring) of a predictive model is described in **Figure 1** with details about the customer as inputs and a number between 0 and 1 as the output.

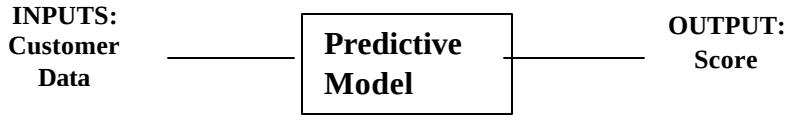

Figure 1 The Data Mining Process

Assume IBM DB2 Intelligent Miner Scoring Service is chosen. The predictive model in the software will predict the probability that a customer will go on making purchases on-line. The score, a number between 0 and 1, represent the probability that a customer will purchase a specific shampoo two months in the future. Since we have three different offers there will be three different scores for each customer. This ends up producing a table of scores like **TABLE 1**, with one row for each customer and one column for each offer score. For each customer, you will have three different scores. If the customer/score entry is NULL, it means that the customer was not eligible to receive an offer. By ranking the customers by their predicted probability, you will be able to identify the best prospects for your merchandises.

TABLE 1 The Probability Scores

Customer Name	Shampoo I	Shampoo II	Shampoo III
Jiawei Han	0.2422	0.4926	0.0872
Michel Kamber	0.8600	0.4465	0.0982
Karen Forcht	NULL	0.9700	0.4453
Robert Marcus	0.7854	NULL	NULL
Susan Hayen	0.5063	NULL	NULL
Stacey Duff	0.8210	0.5014	0.6386
Bill Haugen	NULL	0.5057	0.9177
Terry Jones	0.2226	0.1352	0.0888
Jeretta Horn	0.2928	0.1732	0.5244

Once the scores are available, the next step is to convert each of the scores into profitability values. Each offer has an economic value in return associated with it (and each offer value can be different). Assume that the value (to the company) for Shampoo I is $6 per customer, Shampoo II is $5, and Shampoo III is $2. Multiplying the economic values for each offer by each probability that a customer will respond to a particular offer will generate the expected average economic value for each customer / offer combination. **TABLE 2** shows the value matrix.

TABLE 2 The Value Matrix

Customer Name	Shampoo I	Shampoo II	Shampoo III
Jiawei Han	$1.4532	$2.463	$0.1744
Michel Kamber	$5.16	$2.2325	$0.1964
Karen Forcht	NULL	$4.85	$0.8906
Robert Marcus	$4.7124	NULL	NULL
Susan Hayen	$3.0378	NULL	NULL
Stacey Duff	$4.926	$2.507	$1.2772
Bill Haugen	NULL	$2.5285	$1.8354
Terry Jones	$1.3356	$0.676	$0.1776
Jeretta Horn	$1.7568	$0.866	$1.0488

The optimization process operates on the constraints mentioned above and the value table. The primary objective of the optimization is to maximize the overall profits of the marketing campaign without violating those specified constraints. This can be described as follows:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ a_{91} & a_{92} & a_{93} \end{pmatrix}$$

For the specified value matrix (on the right) in our example, we hope to find the combination expressed in the following formula:

Max V(i,j,k,l,m,n,o,p,q)=ai1+aj1+ak1+al2+am2+an2+ao3+ap3+aq3

(i,j,k,l,m,n,o,p,q=1,2,3…9 and i≠ j≠ k≠ l≠ m≠ n≠ o≠ p≠ q)

There are many algorithms that can be used in the process of optimization. One of them is the method of exhaustion. The process enumerates all the possible combinations, calculates the sum of each combination and sorts them in descending order and then keep on matching each combination with the constraints in turn until it finds the optimal solution or finishes checking the last combination which means there is no optimal solution. When dealing with little amount of data, it does well. But when it comes to massive amount of data, it will be time-consuming and impossible because the complexity of this algorithm is Xn (X is the column number of the matrix while n is the row number).

Here we found out a simpler and much more convenient method to get a good solution of the matrix above, the result is shown in **TABLE 3**. First, marking on the cell in rows that each has two null cells. Second, marking on the bigger value cell on the rows with one null cell. Third,

TABLE 3 The Solution of the Value Matrix

Customer Name	Shampoo I	Shampoo II	Shampoo III
Jiawei Han	$1.4532	$2.463	$0.1744
Michel Kamber	$5.16	$2.2325	$0.1964
Karen Forcht	NULL	$4.85	$0.8906
Robert Marcus	$4.7124	NULL	NULL
Susan Hayen	$3.0378	NULL	NULL
Stacey Duff	$4.926	$2.507	$1.2772
Bill Haugen	NULL	$2.5285	$1.8354
Terry Jones	$1.3356	$0.676	$0.1776
Jeretta Horn	$1.7568	$0.866	$1.0488

browse the whole table from the beginning to the end to find out the accurate cells. For each row, if it is not marked on, selecting out the max of the three numbers. If in the column where the max was in has marked on three numbers, then go on with the hypo-max. If this data matches the constraint, then mark on it, otherwise mark on the third data. The output of the method is the

table below which shows the values of each offer for each customer as well as the offer that the method selected for that customer which is highlighted in the table. Although the good solution may be not the optimal solution, it is convenient and can be easily implemented. The summation is 25.2553, which means that the store can get $25.2553 from the nine customers at most.

CONCLUSION

In this paper, we briefly gave an overview of data mining, applied it to Web-based e-Commerce store and put forward a data mining application on predicting customers' behaviors. This study is designed to offer a few commonplace remarks by way of introduction on the fields of data mining, we strongly hope this study will throw out a brick to attract a jade and spark much more interest in the fresh, yet evolving field.

REFERENCES

l. Turban, E. Electronic Commerce: A Managerial Perspective. Beijing: Higher Education Press and Prentice Hall, 2001, pp. 475-477
2. Kurt, K. An Introduction to Data Mining, http://www.shore.net/~kht/text/dmwhite/dmwhite.htm
3. Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Beijing: Higher Education Press and Morgan Kaufmann Publishers, 2001, pp. 10-12