Mining the E-commerce Data to Analyze the Target Customer Behavior

Yuantao Jiang Siqin Yu
School of Economics & Management, Shanghai Maritime University, P.R.China, 200135
jiangytao@yahoo.com.cn ysq@shmtu.edu.cn

Abstract

In the advent of the information era, e-commerce has developed rapidly and has become significant for every business. With the advanced information technologies, firms are now able to collect and store mountains of data describing their myriad offerings and diverse customer profiles, from which they seek to derive information about their customers' needs and wants. Traditional forecasting methods are no longer suitable for these business situations. This research used the principles of data mining to cluster customer segments by using K-Means algorithm and data from web log of various e-commerce websites. Consequently, the results showed that there was a clear distinction between the segments in terms of customer behavior.

1. Introduction

It may be observed that customers drive the revenues of any organization. Satisfying or even exceeding customers' expectations while simultaneously reducing costs are frequently quoted buzzwords in various business media. In addition to acquiring new customers, predicting potential buyers, firms are devoting a great deal of resources to delighting and retaining existing target customers to cultivate a long-term, close relationship with them. At present, Internet technologies have seamlessly automated interface processes between customers and retailers, retailers and distributors, distributors and factories, and factories and their myriad suppliers. E-commerce is changing the face of most business functions in competitive enterprises.

In the context of e-commerce, generating large-scale real-time data has never been easier, and there are numerous opportunities for gathering customer information in electronic form. With data pertaining to various views of business transactions being readily available, it is only apposite to seek the services of data mining to make (business) sense out of these data. Data mining (DM) has as its dominant goal, the generation of non-obvious yet useful information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations, and characterizations of data[1]. These forms, in turn, are the result of applying sophisticated modeling techniques from the diverse fields of statistics, artificial intelligence, database management and computer graphics.

Collecting consumer information seemed have been available, but still how to analyze these data effectively is of interest to marketers and researchers. The traditional methods for predicting and analyzing customer demands have found a wide range of applications. They are mainly used for predicting the total quantity of products that belong to the same family rather than the relationship between the different customer groups and associated product groups. This paper clusters customer segments by using K-Means algorithm and data from web log of various e-commerce websites. Consequently, the results showed that there was a clear distinction between the segments in terms of customer behavior. It is seen that this data mining model can serve as an efficient vehicle for firms not only to predict the products or services that should be provided or improved for their target customer groups, but also to identify the right customers for a specific product family or service.

2. Literature review

E-commerce is changing the face of business. It allows better customer management, new strategies for marketing, an expanded range of products, and more efficient operations. A key enabler of this change is the widespread use of increasingly sophisticated data mining tools. The term 'data mining' is used to describe the process of analyzing a company's internal data for customer profiling and targeting. In e-commerce application, the end goal of data mining is to improve processes that contribute to delivering value to the end customer.

0-7695-3090-7/08 $25.00 © 2008 IEEE
DOI 10.1109/WKDD.2008.90

At the most basic level, the information available in web log files can illuminate what prospective customers are seeking from a site. Are they purposefully shopping or just browsing? Buying something they're familiar with or something they know little about? Are they shopping from home, from work, or from a hotel dial-up? The information available in log files is often used to determine what profiling can be dynamically processed in the background and indexed into the dynamic generation of HTML, and what performance can be expected from the servers and network to support customer service and make e-business interaction productive.

E-commerce data are classified as usage data, content data, structure data, and user data. Usage data contain details of user sessions and pageviews[2]. The content data in a site are the collection of objects and relationships that are conveyed to the user. Structure data represent the designer's view of the content organization within the site. Structure data also include the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page. The user data may include demographic or other identifying information on registered users, user ratings on various objects such as pages, products, or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of a users' interests. Once the data types are clear, data preparation is easily achieved. The author then proposes association rules, sequential and navigational patterns, and clustering approaches for personalization of transactions as well as web pages[3].

Liuying Shen and Jana Hawley describe an approach to predict user behavior in e-commerce sites[4]. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users (obtainable from web server logs) to predict the purchase and traversal behavior of future users. Strader, T.J. and Shaw, M. J propose a methodology to improve the success of web sites, based on the exploitation of navigation-pattern discovery. In particular, the authors present a theory, in which success is modelled on the basis of the navigation behavior of the site's users[5]. They then exploit web usage miner (WUM), a navigation pattern discovery miner, to study how the success of a site is reflected in the users' behavior. With WUM the authors measure the success of a site's components and obtain concrete indications of how the site should be improved.

In the context of web mining, clustering could be used to cluster similar click-streams to determine learning behaviors in the case of e-learning, or general site access behaviors in e-commerce. Most of the algorithms presented in the literature to deal with clustering web sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the click-stream visitation. This has a significant consequence when comparing similarities between web sessions. Quinlan, J.R propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses[6]. Nonetheless, reviews and research in this area are handicapped by the proprietary nature of the data and algorithms. A great deal of effort is being expended in this area, but most of it is secret. Certainly Amazon, Google, and Microsoft are deeply engaged in statistical research, and in time the broader research community may learn more about their findings. But for now, all this paper can really be done is to lay out the main strategies in the relevant areas.

3. Date mining model

A systematic method was used to collect e-commerce transactions. The target data were customer transactions from web log file of the e-commerce site. There were 2518 transactions collected from 1st to 31st of December 2006. Each session gave details of web usage including user accounts of those who accessed the web sites, requested web pages and their order, and the period of time pages were viewed. This data were used as the basis for analysis in this study.

3.1 Preprocessing of data

Data preprocessing techniques can improve the quality of the collected data, thereby helping to increase the accuracy and efficiency of the subsequent mining processes. It is clear to see that data preprocessing is an important step in the knowledge discovery process, as high-quality decisions must be based on high-quality data[7].

Detecting data anomalies, rectifying them early, and reducing the data amount to be analyzed can result in substantial benefits and advantages for the decision makers. For usage profiles, a session file from data preparation stage was used as input in data mining. While K-means algorithm was applied for the purpose of clustering some incomplete data were deleted. The usable data were 2363 transactions.

3.2 Data analysis

Data mining software, Mineset was used in data analysis[7]. The usable customer data of 4263 e-commerce transactions were divided into two groups. Group 1 was about 70% of the total transactions and

was used as training data. Group 2 was about 30% of the total transactions and was used as testing data.

Five factors used in data segmentation included: age, gender, online in time, address, language, and target customer behavior type.

(1) Age was divided in 6 kinds:
Age 1 - 11 years old to 15 years
Age 2 - 16 years old to 20 years
Age 3 - 21 years old to 25 years
Age 4 - 26 years old to 30 years
Age 5 - 31 years old to 35 years
Age 6 - 36 years old to 40 years

(2) Gender was divided in 2 kinds:
Gender 1 - man.
Gender 2 - woman.

(3) How long online was divided in 4 groups:
Group 1 – 0.00 hours to 05.59 hours
Group 2 – 06.00 hours to 11.59 hours
Group 3 – 12.00 hours to 17.59 hours
Group 4 – 18.00 hours to 23.59 hours

(4) Online Adress:
Adress 1 - At work
Adress 2 - At home

(5) Language:
Language1 - China
Language2 - English

(6) Target customer behavior:
Behavior 1 – Buying computer products
Behavior 2 – Buying cloth
Behavior 3 – Buying gifts and flowers
Behavior 4 – Buying books
Behavior 5 – Buying CDs and DVDs
Behavior 6 – Buying toys and children products
Behavior 7 – Buying airline tickets and other tickets
Behavior 8 – Online trading
Behavior 9 - Online banking

3.3 Result analysis

By using K-Means algorithm to cluster, results from Table 1 show that data were segmented in five clusters. Based on the statistical results of customer usage, e-banking transactions can be classified into 5 clusters.

Cluster 1: The male customers who are 21 to 30 years old used personal computers to purchase computer products and books. They accessed e-commerce site via personal computers between 6.00 hours to 17.59 hours at home. The language was Chinese.

Cluster 2: This was the second smallest cluster. The male Customers were online between 6.00 hours to 17.59 hours. They accessed e-commerce site to buy books, airline tickets and other tickets at work. The language used was Chinese and their age was from 26 to 35.

Table 1 Segmentation of customer behavior

Type		Cluster					
		1	2	3	4	5	6
(1)	1	0	0	13	101	5	0
	2	82	26	28	97	12	9
	3	114	21	226	8	148	13
	4	103	79	176	5	153	27
	5	21	138	74	2	56	147
	6	5	23	31	2	18	169
(2)	1	283	262	387	197	62	318
	2	42	25	161	18	330	47
(3)	1	37	32	14	158	46	0
	2	132	117	80	23	148	0
	3	89	106	216	21	153	135
	4	42	32	238	5	45	230
(4).	1	297	237	327	12	26	327
	2	28	50	221	203	366	38
(5)	C	325	255	516	215	0	125
	E	0	32	32	0	392	240
(6)	1	98	9	17	0	8	8
	2	7	5	11	12	113	0
	3	3	12	8	92	21	2
	4	124	58	187	5	8	0
	5	9	23	172	2	68	0
	6	6	3	7	104	127	0
	7	39	117	15	0	23	61
	8	28	52	5	0	15	253
	9	11	8	126	0	9	41

Cluster 3: This cluster gained a majority of the e-commerce application since it had the largest population (26%). By using personal computers in Chinese language at home or at work, the female and male Customers accessed e-commerce site to purchase books, CDs and DVDs, airline tickets and other tickets. The online time was between 12.00 hours to 23.59 hours and their age was from 21 to 30.

Cluster 4: The male customers who are 11 to 20 years old used personal computers at home. They accessed e-commerce to purchase gifts (including Jewelry) and flowers, toys and children products between 0.00 hours to 05.59 hours. The language was Chinese.

Cluster 5: This was the second largest cluster, where female customers used personal computers for e-commerce between 6.00 hours to 17.59 hours. The language used was Chinese and transactions were bill payments. Their age was from 21 to 30 and they purchase cloth, toys and children products.

Cluster 6: Male customers who are 31 to 40 years old used personal computers for e-commerce site between 12.00 hours to 23.59 hours at work. This was

the only cluster where customers used English. The major behaviors were online trade.

4 Conclusions

E-commerce companies are shifting from the old world of mass production where standardized products, homogeneous markets, and long product life and development cycles were the rule to the new world where variety and customization supplant standardized products. Instead of tens of thousands of products in a superstore, consumers may choose among millions of ones in an online store to satisfy the personalization demands. It is clear that target customers marketing can be effective when a e-commerce company is able to collect rich information about buyers behavior on e-commerce site.

According to this study, the majority of the customers in e-commerce were male and online period was between 6 hours to 17.59 hours. The age is also an important factor that affects customer behavior. This can cause a market segment in the e-commerce. Personal computer at home are more popular for e-commerce than at work that can be attributed to the convenience of online purchasing at home. Chinese is more popular than English because Cluster 6 which used English was about 16%. It shows that some customers begin to access English e-commerce site to engage in international trade.

5. Acknowledgements

This work is supported by the Important Program of Management Science & Engineering of Shanghai Maritime University(No. XR0101).

6. References

[1] Randall S. Sexton, Richard A. and Michael A. “Predicting Internet/e-commerce use”, Internet Research, vol.5,2002, pp. 402-410.

[2] N R Srinivasa Raghavan. “Data Mining in E-commerce: A Servey”. Sadhana, vol,30, no.2, 2005, pp.275-289.

[3] Mobasher. B. Web Usage Mining and Personalization. Practical Handbook of Internet Computing (ed.) M P Singh (CRC Press), 2004.

[4] Liuying Shen, Jana Hawley, etc. “E-commerce Adoption for Aupply Chain Management in U.S. Apparel manufacturers”. Journal of Textile and apparel, technology and management, vol.4, no.1, 2004, pp.1-10.

[5] Strader, T.J., Shaw, M. J. Electronic Markets: Impact and implications, in: Shaw, M., Blanning, R., Strader, T., Whinston, A. Handbook on Electronic Commerce, Springer, Berlin, 2000, pp.77-98.

[6] Quinlan, J.R., Induction of Decision Trees, Machine Learning, Morgan Kaufmann Publishers, Inc, CA, 1990.

[7] John A. Rodgers, David C. Yen and David C. Chou. Developing E-business: a Strategic Approach, Information Management & Computer Security, no.4, 2002, pp.184-192.