

Electronic Commerce Customer Churn Prediction Model Based on Web Data Mining

Weihua Zhang, Li Zhu
Jiangxi Science and Technology Normal University
Nanchang, China
WHZeuh951@163.com



ABSTRACT: Web data mining is a new subject which has been applied in many fields. With the rapid development of electronic commerce, the consumers have started to accept this kind of online shopping purchasing way. E-commerce websites have increasingly brought economic income. But on the other hand, this also brings several problems. Some of them are the instability of customers and high customer churn rate. So in this paper, we will analyze the electronic commerce customer churn prediction model based on Web data mining. Network and information security issues, have become a bottleneck for further development of the network economy. Web data mining technology is the key to improve the performance of network information security technology. And we will use the Pareto / NBD model to figure out the customers' activity to predict the trend of the customers, which can be useful for the enterprise to formulate retaining strategies and resistance strategies to control website customers churn.

Keywords: Electronic commerce, Customer churn prediction model, Web data mining

Received: 18 April 2017, Revised 28 May 2017, Accepted 6 June 2017

© 2017 DLINE. All Rights Reserved

1. Introduction

Since 1990s, the rapid development of electronic commerce has aroused the attention of the business and theory fields. Electronic commerce is a new business model based on advanced computer network technology and communication technology platform on the Internet. Its appearance caused great changes in the traditional market concept in the amount of the scope, including the expansion of the time dimension and space dimensions [1]. So that the E-commerce makes the traditional geographical, political concepts, awareness, and other differences in the formation of the market boundaries blurred. But now, with the increasingly fierce competition in the market, the enterprises pay more attention to customer resources under such electronic commercial circumstances [2].

In the current information analysis technology, web data mining technology is the most powerful and efficient technology. In order to promote the electronic commerce, web data mining technology is applied to predict the customers churn online. The enterprises will analyze the behavior of the customers through a variety of web data mining analysis, which can help to reduce the operating costs and enhance the competitive power of the enterprises. Web data mining technology is the combination of data mining and web research. The web data includes the web page content, the structure of the page, the user, access information, e-commerce information and so on.

The left part of the article is organized as follows. Chapter two will introduce the state of the art, including the Web data mining and its application in E-commerce customer churn forecasting model. Chapter three will show the readers the method to study the relation between customer activity and the customer churn. Chapter four will analyze the result. Chapter five will be the conclusion.

2. State of the Art

2.1 Web Data Mining

In today's e-commerce environment, enterprises can use the powerful data processing ability of Web data mining technology easily, excavating the potential and valuable customer behavior information through the huge data which can assist the enterprise marketing activities and decisions. Customer behavior analysis based on a large amount of customer data is a kind of Web data mining technology which is widely used. Through Web data mining, the customer behavior data is conducive for the company to develop marketing programs and plans. Last but not least, the most important is to establish the electronic commerce customer loss prediction model. In the face of the fierce market competition, how to reduce the loss of online customers is becoming more and more important. The most effective method is to build up a suitable forecasting model to predict the loss of the customers online accurately. Once found out their trend of loss, take measures and make plans to retain the actively. In recent years, with the rapid spread of the Internet, the network becomes user-related business which is very important and indispensable tool. Born in network economic base, e-commerce, both in foreign and domestic, has been developed. On the other hand, network security in facing the current fiscal situation is not optimistic. Network and the information system's own defects, vulnerabilities and threats are surfaced. Network security has not only become an important part of the State and national security, but also an important bottleneck restricting the development of the network economy. One can say that the Web data mining technology is upgraded network information security prevention performance of the key technology. Various network servers, and Agent servers are retained to have access log. The records have network system occurred during various important events, through it we can understand the system run status and can monitor user behavior, audit security event, and diagnosis errors exception. Although these data providers have security prevention application possibilities, but often need more high abstract level of analysis. Relations in the period of time between multiple events and event identification, pattern recognition are required in order to make better use of these data. This was the identification usually depends on intuition and experience, or based on attack patterns, security experts and the analysis of the weaknesses of the system. This is very time consuming and inefficient and imprecise, and are only effective against known security threats and new security threats could not be found. And Web data mining through access path analysis, and associated rules found, and sequence mode analysis, and classification rules found, and poly class analysis, technology, get resources data in the extraction and security related of system features property. According to system features property, the automatically generated security event of detection model, for the security event of automatically identification avoid artificial analysis and coding detection model. It greatly improves the mode recognition, and rules structure of efficiency makes the network information security prevention performance. (as shown in Figure 1).

Analyzing the customer behavior through web data mining and monitoring the actual purchasing behavior of the customers (such as the website they browse, the keywords and the products they purchase online), will help the enterprises to enhance the judgment of the prospect. Here is a status about the expression of the E-commerce. Data mining refers to the large, incomplete, noise, fuzzy and random data extraction implicit in it and do not know in advance, but it is also potentially useful information and knowledge. But most of the classical data mining methods are only homogeneous, where the homogeneous data analysis for text on the Internet to a large number has heterogeneous information, log information, hyperlinks, and so on where it does not apply. In order to solve this problem, people combine traditional data mining and web, and create a new mining technology—Web data mining. Web data mining from Web documents and Web activities in the discovery and extraction are potentially useful model of interest and hidden information. It targets the mining useful information from the Web, based on the data mining, document mining, multimedia mining, and comprehensive use of computer networks, databases and data warehouses, artificial intelligence, information retrieval, visualization and natural language understanding technology. It will combine



Figure 1. Web data collection

traditional data mining and the web technology. The basic web data mining process is as shown in the Figure 2. The resources found refer to are from web which get and returned to the text resources process. It is by processing of the object including static page, web database, web structure, and user records and the related information. The data pretreatment refers to according to mining of purpose, is found in the original resources data for extraction, and decomposition. This is merged, last into data mining in data format, and save to relationship type database table or data warehouse will wait for further processing. The Pattern recognition is the use of various algorithms for data mining, and generation mode, where the algorithm includes an access path analysis, association rules, sequence analysis, classification rules and clustering analysis. The user access pattern analysis is a mode of analysis, which will have a value of pattern extraction process.



Figure 2. Express state of the E-commerce

3. Methodology

Pareto/NBD model is proposed by Schmittlein, which is mainly used to describe the model of customer repeat purchasing behavior [3]. The results of the model can predict the customers churn. First of all, the customer churn is the random events. We do not know the state of customer churn actually in advance. The data all we known are from the web, such as the purchase behavior of the customers. So here we should estimate the customer activity[4]. Pareto/NBD model from Schmittlein, Morrison or Colombo's description of customer repurchase behavior model, also known as SMC model. Pareto/NBD model is applicable to the following customer churn is a random event and cannot be directly observed when a client loss. The unique performance customers stop buying in a long time; during the active period, customers buy a random interval of time and customers buy event could occur at any time. The wastage rates and purchase rates for different customers have a big difference[5]. Before that, scholars have put forward a lot of client repurchase behavior prediction models, such as NBD (Negative Binomial Distribution) model, LSD (Logarithmic Series Distribution) model and NBD-Dirichlet model[6]. Pareto/NBD model is the first ever to customer churn prediction model of customer repurchase phenomenon, where it cannot only predict the repeat purchase, and can provide customers active, customer base, and other important information. In recent years, customer relationship management and customer asset management become the focus of marketing management, where Pareto/NBD is a basis for prediction of consumer

behavior models caught the attention of researchers. Schmittlein, and Peterson improve the estimation of the model, join the customer purchase amount prediction model and it is an empirical study for the industrial market[7]. Reinartz and Kumar research customer profitability and customer relations in time directly use the Pareto/NBD model to estimate the length of time for the customer relationship. Through the digital Corporation of data, it is found that the Pareto/NBD model was empirical research. Fader and Hardie by CDNOW online CD vendors to verify the data to the Pareto/NBD model[8]. But generally speaking, Pareto/NBD models also lack a wide range of empirical research, empirical methods need to be improved and Pareto/NBD model has yet to be developed[9]. Pareto/NBD model has very strict assumptions limit, but under certain conditions, it can be more effective to capture hidden information in the sales data[10]. Customers' repeat purchases and loss of effectiveness have early warning forecast. Application section of this article is based on the Pareto/NBD model of customer lifetime value calculation method and customer churns alarm point setting method and these methods have real meaning, which can provide guidance for enterprise marketing decision[11] (as shown in Figure 3).

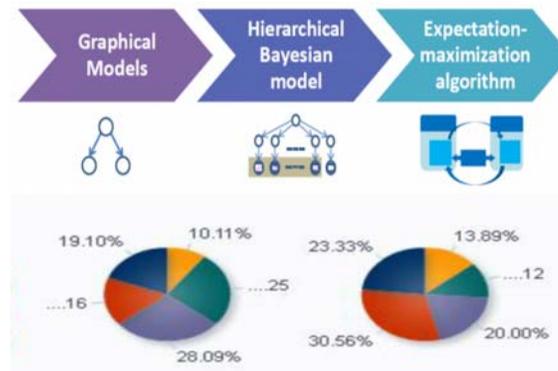


Figure 3. Pareto/NBD model

Enterprises involved in this research belong to the professional retail industry, which is engaged in a variety of products. The product prices vary for small number of customers, mainly for television stations. The schools and other institutions, operational units and customers closely associated with business, have long customer relationship in a typical market. Poisson's purchase rate is one when the customer is active and the customer repeating purchase rate is a Poisson distribution. When the individual customer is active, the transaction rate is λ , and the number of transactions for X observed period length for $(0, T)$, where 0 is the time of the first transaction and the future purchase probability is Poisson distribution. The customer survival time length is τ :

$$P[X = x | \lambda, \tau > T] = e^{-\lambda T} \frac{(\lambda T)^x}{x!}, x = 0, 1, 2 \quad (1)$$

$$E[X | \lambda, \tau > T] = \lambda T \quad (2)$$

$$Var[X | \lambda, \tau > T] = \lambda T^2 \quad (3)$$

According to the calculation of customer purchase data, the enterprises can predict the electronic commerce customer churn which can be the reference and specific implementation of the follow-up to maintain the strategy work. Customer churn management is an important component of the customer relationship management. In the customer relationships without contract conditions, the customers can change providers at any time, without having to inform the vendor, resulting in customer loss concealment of non-contractual customer relationships. If one can detect abnormal changes in customer buying behavior, early detection of brain drain, can promptly take measures to investigate the cause of customer churn, where customer and the new supplier contacts early to win back customer trust. In the empirical part, one can see the active reflect of the survival probability of customers, as growth is not buying time, the activeness will decline, so it can be used as indicators of customer churn alarm. But action is a continuous variable, in practical applications it also need to be converted to binary variables, that is, if active value is reduced to a certain threshold alerts. It leads to take the appropriate measures. In the existing literature, generally set the threshold to 0.5, that is, active 0.5 alert, where the advantage of this method is relatively simple and the disadvantage is that treats all customers equally. 20/80 principle shows that 20% customers bring a profit of 80%, so the loss of high-value customers are more which has severe impact on the enterprise. The high-value customers get more early warning and the low value customers can get later warning. For customers with negative value or zero value, the warning is not required.

4. Result Analysis and Discussion

The model samples are selected from electronic commerce operating conditions from 2005 to 2010. tx , T , P_2 indicates the purchasing amount during the time, the latest customer bought the commodity, time interval and the flag that whether people have bought things. And the four variables descriptive statistics are as follows:

		Statistic			
		X	tx	T	P_2
N	Effective	2357	2357	2357	2357
	Loss	0	0	0	0
	mean value	1.04	6.8458	32.7158	.29
	standard value	2.190	10.7314	3.32691	.454
	variance	4.796	115.163	11.068	.206
	maximum	0	00	27.00	0
	minimum	29	38.43	38.86	1

Table 1. Four variables descriptive statistics of the sample

From the table data we can draw the conclusion that the customers online are still active, and the web data dining can analyze more relevant statistic to help predict the E-commerce customer churn.

P	Customer number	Ratio of customer
0.9-1	299	9.7%
0.8-0.9	113	4.8%
0.7-0.8	112	4.8%
0.6-0.7	64	2.7%
0.5-0.6	56	2.4%
0.4-0.5	64	2.7%
0.3-0.4	1423	60.4%
0.2-0.3	197	8.4%
0.1-0.2	78	3.3%
0-0.1	21	0.9%

Table 2. The ratio of customer activity in observation period

If the customer activity value is 0.5, and as the customer is active boundaries, from table above the current activity of the site is 24.4%, which indicates the E-commerce customers are still in active state.

5. Conclusion

With the development of the data mining technology, Web data mining begin to be applied in many technological fields. and especially in the model of electronic commerce customer churn forecasting. At the same time, adopting scientific and effective methods to predict the loss of customers in the next period of time and analyze the reasons for the loss of customers will help the

leadership of the enterprises to make decisions and development plan. This article is based on web data mining, using the Pareto/NBD model, and put forward the strategy to control the loss of E - commerce customers. According to the experimental data, we can know that the enterprise should know the activity of their customers and monitor their purchasing behavior. With such data, people can build up the E-commerce customer churn model more efficiently.

References

- [1] Ngai, E W T., Xiu, L., Chau, D C K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*. 36 (2) 2592-2602
- [2] Cho, Y H., Kim, J K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications*, 26 (2) 233-246
- [3] Verbeke, W., Martens, D., Mues, C. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*. 38 (3) 2354-2364.
- [4] Schmittlein, D C., Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next?. *Management Science*. 33 (1) 1-24.
- [5] Blattberg, R C., Deighton, J. (1996). Manage Marketing by the Customer Equity Test. *Harvard Business Review*. 74 (4) 136-44.
- [6] Peterson, R A. (1994). Customer Base Analysis: An Industrial Purchase Process Application Marketing. 34 (2) 354-367.
- [7] Reinartz, W J., Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*. 67 (1) 77-99.
- [8] Reinartz, W J., Kumar, V. (2013). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. *Ukraïnsky- Biokhimichny- Zhurnal*. 40 (5) 532-547.
- [9] Fader, P S., Hardie, B G S. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*. 8 (4) 415-430.
- [10] Fader, P S., Hardie, B G S., Lee, K L. (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Social Science Electronic Publishing*. 24 (2) 275-284.
- [11] Berger, P D., Nasr, N I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*. 12 (1) 17-30.