Summer 5-26-2017

Predicting Purchase Proneness of Anonymous User in Mobile Commerce

Wenming Huang
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China,
1196027536@qq.com

Li Li
School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China,
lily691111@126.com

Predicting Purchase Proneness of Anonymous User in Mobile Commerce

Wenming Huang[*1], *Li Li*[1]

1.School of Economics and Management, Nanjing University of Science and Technology, Nanjing, 210094, China

Abstract: In recent years, mobile commerce is developing rapidly because of the popularity of mobile devices. However, for the difficulty of the mobile device input, the users of the e-commerce websites usually don't log on the website when they are browsing, which resulting in a situation that a large number of website visitors are anonymous users. In order to increase sales revenue and expand market share, an effective prediction of anonymous users' purchases proneness is very helpful in providing targeted marketing strategy for website to induce anonymous users to purchase. In the past, customer segmentation was mainly analyzed and modeled by customers' historical data. But the history data of anonymous users can't be obtained on mobile commerce sites. This method is difficult to put into management practice. In order to solve this problem, this paper proposes a method based on random forest of using user clickstream data to forecast purchase proneness in real time. This method includes two stages: the model training part and the user purchasing proneness prediction part. In the model training part, a classifier based on random forest algorithm is trained. In the users' predicting part, the classifier is used to predict the user's purchase proneness in real time. The method proposed can be effectively applied in the real-time prediction of anonymous users' purchasing proneness, and the results of prediction will help enterprises implement the marketing measures in real time.

Keywords: mobile commerce, data mining, purchase proneness

1. INTRODUCTION

The rapid development of mobile commerce enables users to search goods and pay bills anywhere and anytime, more and more users are using mobile devices for shopping, making enterprises accumulate a large number of user data, especially the user clickstream data. On the one hand, mobile commerce allows enterprises to understand the user behavior in detail, all user requests are saved to the server's log database, helping the enterprise possess all data generated when users visit, search and browse in the site. Particularly, the enterprise can know those pages browsed by the user who purchased in the site [1]. On the other hand, even though highly detailed user behavior data can help us understand user behavior better, the volume of user's click stream data presents a challenge to analyze user behavior, leading enterprises to introduce more and more data mining methods of user behavior research. In addition, users are becoming more sensitive, a large number of users didn't log on before payment. For these anonymous users, it's difficult for companies to get their transaction data in the past and registration information [2].The previous research usually focused on the analysis of purchasing information and demographic information left by users in the past [3]. Marketing strategy tends to lag if produced in this way, and only when the user revisits the site those marketing strategies can be effective, for some site which users re-visit ratio is low, this marketing model is usually very difficult to achieve. In addition, B2C site visitors are usually anonymous users, the previous information of visitors isn't available, companies can't analyze users by historical data. This paper shows how to use the user's clickstream data to predict the

* Corresponding author. 1196027536@qq.com(Wenming Huang); lily691111@126.com(Li Li)

purchase proneness of the anonymous users, and the enterprises can select the target customers according to the forecast results.

Market segmentation is an important topic in marketing, traditional market segmentation is often based on demographic characteristics such as gender, age and education level. This kind of market segmentation methods often ignore the differences in user behavior, and with the growth of people's awareness of privacy in Internet era, the user's demographic characteristics are becoming increasingly difficult to obtain, while user behavior records are more accessible in recent years, more and more scholars have begun to pay attention to market segmentation based on user behavior[4].

This paper presents a customer segmentation method based on users' purchase proneness. According to the massive click stream data provided by the website, including the login user and anonymous user, the user can be divided into two groups according to the random forest algorithm, which is inclined to purchase or not. Based on the customer segmentation of purchasing proneness, we can provide decision support for enterprise to do the real-time analysis and marketing strategy loading for anonymous users.

In the first section, the background and meanings of this research are briefly described. In the second section, we will review the related work of the scholars in the market segmentation. The third section will introduce the research methods of this paper. The fourth section will practice our method in an assurance sales platform. Final section will make a conclusion of this paper.

2. RELATED WORKS

As the kinds of goods continue to increase and needs of user increasingly become personalized, market competition is imperfect[5]. Based on the theory of imperfect competitive markets, in 1956, marketing master R. Smith thought the successful marketing plan not only includes product differentiation, but also should include market segmentation. Now product categories and product differentiation are quite considerable compare to 1956, with the improvement of living standards, people's individual needs more and more intense, obvious differences in the user, the market segmentation has become basis of the enterprise to adapt to customer personalization demand. Demographics based segmentation is the earliest customer segmentation method, but with the development of the times, such as globalization, informationization weakened the relationship between users and demographics features [6]; and because of user privacy awareness demographic information are difficult to obtain and most of the demographic information is acquired pre-post, demographic based methods are more often used to understand the market structure. Because of the difficulty in obtaining demographic information, more and more scholars are devoting themselves to other market segmentation methods. Here are four main market segmentation methods.

1) RFM analysis

The RFM analysis was performed by Hughes using "recency", "frequency", "momentary" to analysis customers. RFM analysis can assess the value of the user, for enterprises to develop marketing decisions, and it's an easy way for enterprises with a database. On the basis of Hughes, Marcus proposed a market segmentation based on the user value matrix, which provides a way for enterprise to segment customers by different customer values[7]. Colombo uses a stochastic approach to analyze customer segments for a company by RFM data to establish user models to help companies find the target user[8].

However, in RFM, there are serious multicollinearity between the purchase frequency and momentary, it's difficult to separate customer by these two indicators. Although the RFM model is a classic model of marketing, the number of variables analyzed by the RFM model is too small to fully represent the user characteristics. In order to solve this problem for database marketing, data stored in the enterprise database is utilized to represent the user characteristics.

2) Factor analysis

Factor analysis is a method that extracts the factors or principal components less than the original variables and interpret the factors or principal components by comparing the original multiple observation variables. Kahle used the factor analysis method to study the value of the list (LOV) to extract three factors by LISREL model to predict whether the user will buy [9]; Hassan use principal component analysis for the global market division[10]. However, Hassan's study lacks an assessment of customer values in different market segments, it's hard for enterprise to decide the ratio of input in each segment markets.

Factor analysis can help marketers to find the latent variables from a large number of observation variables to segment market, so as to formulate data-supported market segmentation rules according to latent variables. In the market segmentation, factors are orthogonal to each other, which avoids the multicollinearity, and the number of variables is more than RFM model. But the data of this research come from the user questionnaire. On the one hand, there are high data collection costs when collecting data through questionnaires. The results of questionnaire and data quality are influenced by the structure of the questionnaire and the proficiency of the researchers. On the other hand, the interpretation of the principal components or factors after rotation is often subjective, and the interpretation of variables also requires expert experience. As the result of factor analysis is a standard normal distribution of variables, for this variable is difficult for enterprises to make quantitative decision.

3) Cluster analysis

Clustering analysis based on a large number of user data, according to similarity between users, the user clustered into different clusters, and explain the various clusters . Sandra used clustering algorithm to study demographic characteristics of inpatient patients[11]; Saunders clustered 200 students which data came from questionnaires about the importance of car attributes[12]; Walsh reanalyzed EU audiences for smoking bans campaign in 2006-2007[13]; Wang used a biclustering-based method to analysis user pain points[14].

Clustering is a kind of unsupervised learning. The result of this mining method may be difficult to explain, which makes some clustering results difficult to serve the management decision.

4) Classify

Classification algorithm according the user data and user tags, apply such as decision tree-based or neural network based machine learning algorithm find a series of rules to divide user into different classes. A new evolutionary learning algorithm has been proposed by Au for telecom customer churn prediction[15]; Lessmann et al. Used SVM to classify multiple public datasets come from real databases[16];Kim et al. Used neural networks and genetic algorithms to predict the users to find the users who were most likely to produce daily consumer behavior[17].

Many scholars have analysis the users of large data set, and many of them predicted user purchase proneness. However, most of their data come from the enterprise business database or from the public data set. Those data records belong to log on users, so this study did not address the anonymous user.

3. PROPOSED METHODOLOGY

This paper suggests a methodology for real time anonymous user purchase proneness predict. A real time predicts requires predictive process speed. Because the nonparametric learning method does not train the model in advance, it can only be classified according to the algorithm and historical data when obtaining records to be classified. Although this method does not need training, the prediction speed is slow; the parameter learning method uses the data to train a classifier in advance, when there is a record need to be classified, the model is used to classify the record, such learning algorithms need to complete the training procedure, but the advantage is that the prediction speed is relatively high. In addition, because most of the users in the website don't generate

purchase behaviors, the data onto forecasting the user's purchasing tendencies are imbalanced and the algorithm needs to deal with the problem of imbalanced data very well. For these reasons, in this paper, random forest algorithm is adopted as learning algorithm. The research process in this paper mainly includes the following steps: data collection, data cleaning, user identification, session identification, feature extraction, model training and model adjustment. It is divided into two stages: model training process and predicting user purchase proneness. Research process as shown in Fig.1. Firstly, users' click stream data is obtained from the log file. Then, data cleansing, user identification and session identification are carried out. Model are trained after feature extraction. If the model is not ideal, the features are re-extracted and re-trained. And then use the model to predict real time user data, and enterprise carry out marketing strategy according to the result of predict.

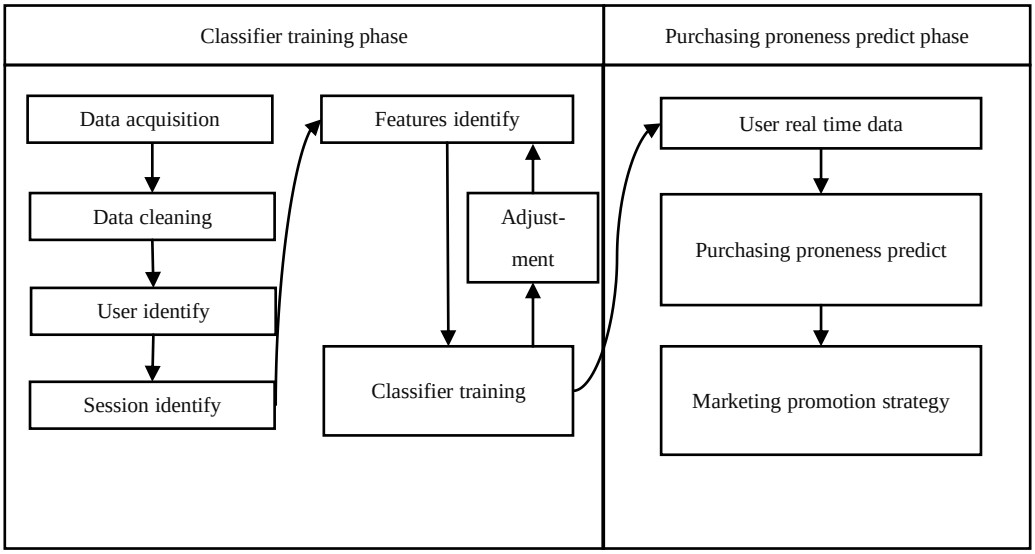

Fig.1.Procedure of anonymous user purchasing proneness predict

3.1 Data preprocessing

1)User identify

In the earlier literature, scholars often use the IP address to identify the user [18], but because of the lack of IPv4 address resources, many users have access to the Internet by sharing IP. Therefore, we can't only use IP to identify user. Later, scholars used Cookie which placed by the sever while user asked for web site to identify user. Nowadays, people clean their browser time to time, Cookie in the browser is often cleared by the user, so Cookie unique identification users can't be very accurately.

We append an attribute ID to each record to identify user. ID is a piecewise function of IP address, user's agent and Cookie. It's shown as Formula 1. ID is a function of Cookie when user Cookie exists in the log record. Otherwise, ID is a function of user's agent combine with IP address.

$$\mathrm{ID} = \begin{cases} f(Cookie) & Cookie! = "Null" \\ f(ip, Agent) & Cookie == "Null" \end{cases}$$

Formula 1. User identify function

2) Session identify

Session identification is an important step after user identification, users often visit a site more than once, or when a user visits a website will complete multiple transactions. So that we need separate user records into sessions. There are three classical methods in practice:

(1) Set a time limit for every session, from the user's first request to start time, after the time limit as the next session.

(2) Create a time threshold for the user session, if the time between every nearly two requests is shorter than time threshold, treat those requests for the same session. If the time between two requests is longer than time threshold, separate those requests into different sessions.

(3) According to whether the reference page in the log is consistent with the most recent access request, the user's reference page should point to the previous visited page of the user. If the reference page of the user does not match the last visited page, treat them in different session.

Set up the time limit is relatively simple way, but the disadvantage of this session identification method is when the user is accustomed to a long time browse through a site, this session identification method may lead to the wrong separations.

Session segmentation based on the reference page is often useful when accessing WEB pages, since on WAP pages, the browser will read the mobile device cache, this time will not post a request to the server, the reference page will not be able to connect to the user's latest request site. At this point will lead to the user session segmentation error.

To sum up, in this study, we set the time threshold for session identify. In order to determine the time threshold more precisely, the time threshold is not selected by the classical 10 minutes. We count the time of two adjacent requests of each user first, and the last request in a certain session is not counted. Because the duration of the last request for a user's session can't be determined from the user's clickstream data. Then, according to the results of the distribution graph, find the peak in the distribution as a time threshold.

3) Transaction identify

Transaction identifies will extract user behavioral attributes and user tag. The result of transaction identifies is the attributes of each session, those attributes will be input into learning algorithm to train a model. Learning algorithm base on those attributes we extracted, to find hidden rules between behavioral attributes and the user tag. Server log transaction identification is often based on the specific keywords URL contained, the user's request is used to identify the transaction. For example, if the "Pay / success" is included in the URLs requested by a user's, the user is considered to have completed the payment process. He's user tag set to "1".

3.2 Model training

Random forest is a very effective way to deal with unbalanced data. Random forest is used to train multiple classifiers for a classifier problem. Multiple classifiers are jointly predicted and finally results were given by voting.

Since the selection of attributes in training is randomly selected, there is no need for manual feature selection. And because the number of selected attributes is far less than all attributes, the depth of each CART tree is very shallow, so the learning and forecast speed of random forests are very fast, and very suitable for concurrency.

Random forest uses the Bootstrap sampling method to sample training from the records, and finally, about 1/3 of the records have not been sampled, to these records called OOB, the random forest using the OOB assess the model accuracy. In addition, the value of an attribute's record is changed at random, and the importance of the attribute is measured by the degree of decline in accuracy.

3.3 User purchase proneness predict

Since the random forest constitute with a large number of CART trees, random forest prediction will integrate the classification results of all the decision trees, predict the unknown records. When a new record needs to be classified, all of the established CART trees classify the record. The classification result of the unknown record R is finally obtained from the trees vote result. The classification process is shown in Fig.2.

This part of the goal is to classify R into a class of C, C_j is a certain value of C. This paper aim to predict user purchases intension, so the purchase proneness is the attribute C, the user is divided into two types of purchase and not to purchase, Purchase as C_1. R is first classified by a tree t in the random forest, and then the decision tree t gives the classification result C_j, which is incremented by 1 on the C_j. Determine whether all the trees have been voted, if not, then use the tree without voting to continue to classify R until all the trees have been voted. The maximum number of votes Cj are taken as the prediction result of R at random forest.

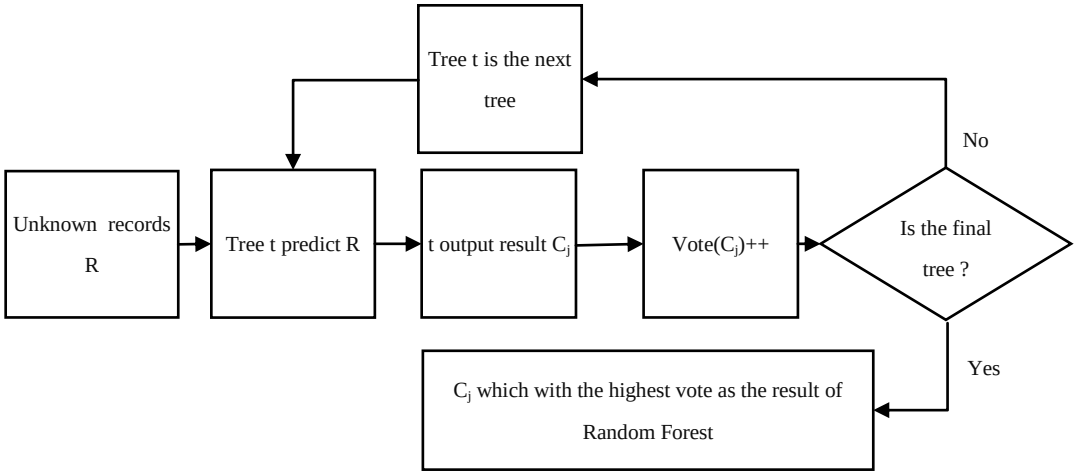

Fig.2.Procedure of Random Forest classification

4. CASE STUDY

We use the real server log data to predict the purchase proneness of anonymous users. The framework of purchase proneness forecast of anonymous users is shown in Fig.3. When an anonymous user lands on the web site, he or she will have some browsing behavior, which will send a request to the server to be saved as a server log. In order to get the training data, we need modify the log data to extract attribute, using the procedure we have already mentioned user identification, session identification and transaction identification to obtain structured data. The structured data is input into the trained model, and the model will output the purchasing proneness of the user, choose the different marketing plan according to the different purchasing proneness of the user, and then load the web pages with marketing information into the page that the user is browsing. Users in the process of browsing may purchase before leaving, may also leave directly, enterprises need to load the marketing strategy, as far as possible to induce the user to purchase behavior.

4.1 Experimental process

The experimental data from an Internet insurance trading platform two weeks server log, sum to 1.5 million records. Data structure is shown as Table 1.

Table 1.Attribute of server log data

IP	CALNUMBER	URL	REQUEST	VINFO	REFERE	VISITTIME	AGENT

After the user identification we need to determine the time threshold of the session segmentation, after drawing the request interval, we get a few peaks, in order to determine which high point is the best time threshold, we counted some special pages' interval.

1) The interval between the page near product details pages.

2) The interval between the page near product lists pages.

3) The interval between the page near payment pages.

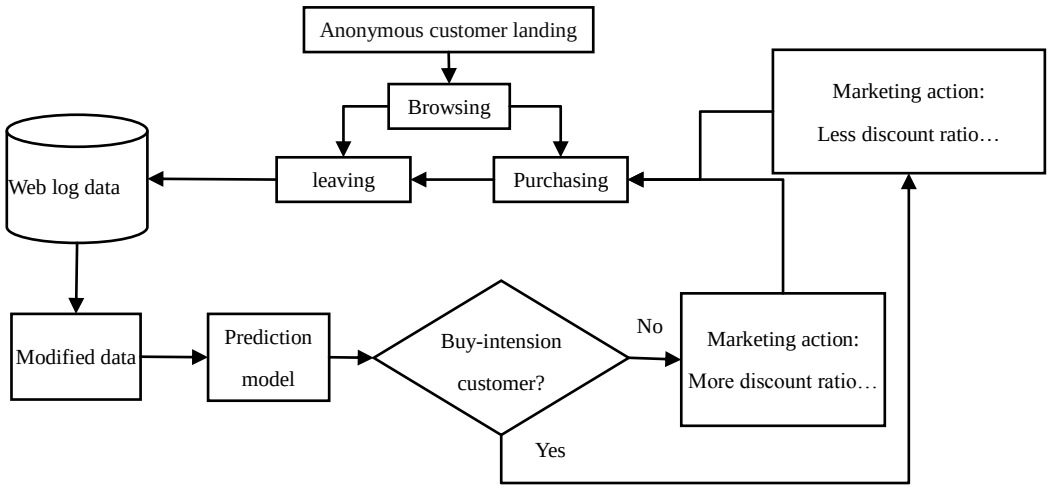

Fig.3.Procedure of real time user buy-intension predict

Table 2. Log data missing value

Records with “agent”:	128483	Records missing “agent”:	132
Records with “Cookie”:	80421	Records missing “Cookie”:	48194
Records with “IP”:	128615	Records missing “IP”:	0

The interval between the pages near payment page usually more than ten minutes. So the time threshold should more than ten minutes. We draw an interval distribution graph shown as Fig.4. So we set 800 seconds as time threshold.

After the user identification and session segmentation, feature extraction is performed in the manner described above, kind of features extracted and the data type of features as shown in Table 4. Finally, this paper selects the number of pages visited, the average page stay time, log flag, the number of searches and other 11 features. And features of each extracted from user session are input into the random forest algorithm to train. And adjust features according to the OOB test.

4.2 Result

As the number of trees increasing, the accuracy of the random forest will increase, but when the number of trees reaches a certain level, the correlation between the trees’ result is too high, the new tree will almost no longer improve the accuracy of classifier. It’s can be seen from Fig. 5, in this experiment, when the number of trees reaches about 20, the error converges.

There are some differences in the user behavior between WAP and WEB sites. The number of user searching is a very important user behavior on the WEB side prediction. However, due to the structure of the web page, the weight of searching in user purchasing proneness prediction is decrease.

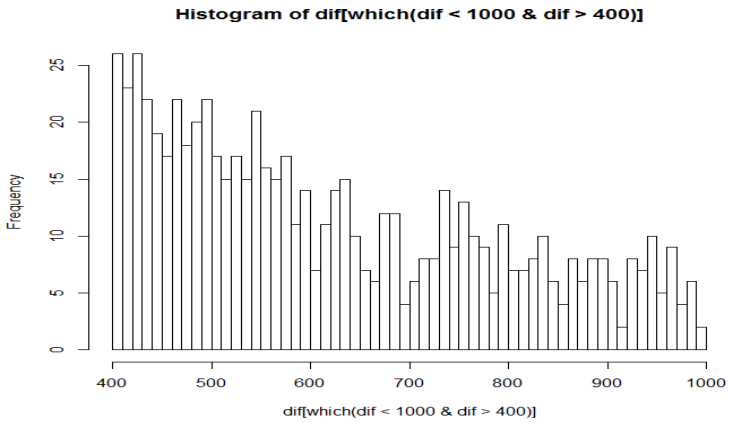

Fig.4.Distribution diagram of time between two request

Due to the convenience of the mobile platform, time of user visiting is flexible, the majority of purchase and access behavior occurred in the evening. As the user's reading habits change, mobile users tend to read faster and faster, combined with smaller mobile terminal screen size, the information presented is more refined, a large number of access staying is only tens of seconds.

Table 3 shows the results of random forest predictions. When most of the decision trees classify a record as "1", the final result of the random forest is also "1", and most decision trees classify a record as not to buy, marked 0, then the final result of the forest is 0. The accuracy of the final model is 97.3%, the sensitivity is 56.6%, and the recall rate is 87.2%.

Table 3. Result of random forest

ID	Tree 1	Tree 2	Tree 3	…	Predict result	Actual result
1034845192	1	1	0	…	1	1
1709342069	0	0	0	…	0	0
1728274468	0	1	0	…	0	1
1033803805	0	0	0	…	0	0

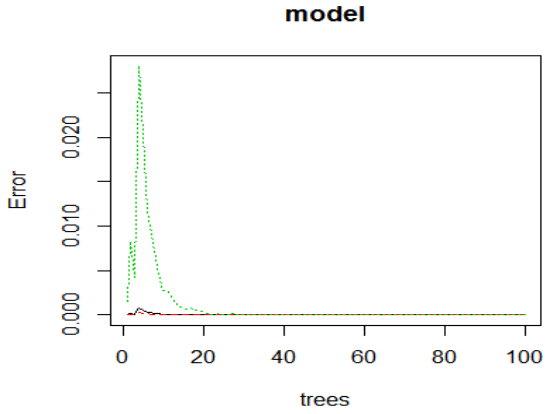

Fig.5.Relation between Number of trees and error ratio

5. CONCLUSION

Nowadays, e-commerce customers, especially mobile commerce, often don't log on when browsing sites with the difficulty of using mobile devices to input, so mobile commerce site visitors are often anonymous. The researches on the purchase intension of mobile commerce anonymous users can help enterprises identify high-value customers and adopt targeted marketing strategies for users with different purchasing intension. In the past, the customer segmentation and customer value usually focused on how to make use of users' history data and demographic information to segment customer. These data and information require the user to submit some specific information, such as the data about registration and login. These studies are not competent for online anonymous user segment, which may not help companies choose target customers from anonymous user.

In order to overcome the weaknesses of past research, we propose a method based on the random forest algorithm using user clickstream data to predict user purchase proneness. This method is consistent with two stages, the first stage is the classification model training, the second stage is using real-time users' clickstream data to predict. In the first stage, a classifier based on the random forest algorithm is obtained by data preprocessing and model training. In the second stage, the classifier obtained in the first stage is used to predict the online users of the website in real time, and the users can be targeted according to the predicted results in the marketing activities.

Our proposed approach has several advantages.

1)Model training and prediction of anonymous user data do not need additional transformations of pages, as we use the server logs data to analysis.

2) Because of the short time of the user browsing behavior in mobile commerce platform, if we want to load the marketing strategies in real time, we need do prediction as fast as possible. This paper adopts the method of parameter learning to predict, and this algorithm can predict the user's purchasing proneness quickly.

3) Our predicting process doesn't need the user's historical data nor the users' login, this work can effectively solve the problem of anonymous user historical data and demographic information and other data difficult to obtain.

4) Since we can obtain the user's clickstream data, we can use our method to predict the user's purchase proneness for all users online whether the user logs on or not.

5) The method proposed in this paper does not involve the user's privacy data, which can protect the user privacy.

We believe that the method we have proposed to predict the user purchasing proneness can effectively help enterprises to obtain new users and induce the user's buying behavior which help increase sales revenue. The main significance of this paper is to provide an effective way to predict the purchase proneness of anonymous users.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 71271115.

REFERENCE

[1] Nottorf F. (2014).Modeling the clickstream across multiple online advertising channels using a binary logit with Bayesian mixture of normal. Electronic Commerce Research and Applications, 13(1): 45-55.

[2] Suh E, Lim S, Hwang H, Kim S. (2004). marketing: A case study. Expert Systems with Applications, 27(2): 245-255.

[3] Shih T K, Chiu Chuan Feng, Hsu Hui Huang, Lin Fu Hua. (2002). An integrated framework for recommendation systems in e-commerce. Industrial Management & Data Systems, 102(8): 417-431.

[4] Kuo R, Ho L, Hu C. (2002). Cluster analysis in industrial market segmentation through artificial neural network. Computers & Industrial Engineering, 42(2-4): 391-399.

[5] Smith W R. (1995). Product differentiation and market segmentation as alternative marketing strategies. Marketing Management, 4(3): 63.

[6] Liu Ying Zi, Wu Hao. (2006). Market segmentation methods review. Journal of Industrial Engineering/Engineering Management, 20(1): 53-57. (in Chinese)

[7] Marcus C. (1998). A practical yet meaningful approach to customer segmentation. Journal of Consumer Marketing, 15(5): 494-504.

[8] Colombo R, Jiang W. (1999). A stochastic RFM model, Journal of Interactive Marketing, 13(3): 2-12.

[9] Kahle L R, Kennedy P. (1989). Using the list of values (LOV) to understand consumers. Journal of Consumer Marketing, 6(3): 5-12.

[10] Hassan S S, Craft S. (2012). Examining world market segmentation and brand positioning strategies. Journal of Consumer Marketing, 29(5): 344-356.

[11] Liu S S, Chen Jie. (2009). Using data mining to segment healthcare markets from patients' preference perspectives. International Journal of Health Care Quality Assurance, 22(2): 117-134.

[12] Saunders J A. (1980). Cluster Analysis for Market Segmentation. European Journal of Marketing, 14(7): 422-435.

[13] Walsh G, Hassan L M, Shiu E, Andrews J C, Hastings G. (2010). Segmentation in social marketing: Insights from the European Union's multi‐country, antismoking campaign. European Journal of Marketing, 44(7-8):1140-1164.

[14] Wang Bin Da, Miao Yun Wen, Zhao Hong Ya, Jin Jian, Chen Yi Zeng. (2016). A biclustering-based method for market segmentation using customer pain points. Engineering Applications of Artificial Intelligence, 47: 101-109.

[15] Au Wai Ho, Chan K, Yao Xin. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. Evolutionary Computation, 7(6): 532-545.

[16] Lessmann S, Voß S. (2009). A reference model for customer-centric data mining with support vector machines. European Journal of Operational Research, 199(2): 520-530.

[17] Kim Y, Street W, Russell G, Menczer F. (2005). Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms. Management Science, 51(2): 264-276.

[18] Zhou Bao Yao, Hui Siu Cheung, Fong A. (2005). A web usage lattice based mining approach for intelligent web personalization. International Journal of Web Information Systems, 1(3): 137-146.