Analysis and Prediction of Mobile Internet Users’ Behavior Preferences

Ming-sheng ZHAO¹,* and Wei WU²

¹Department of Information Technology, Nanjing Forest Police College, 28 Wenlan Road, Nanjing 210023, China
²School of Economic and Management, Nanjing University of Science and Technology, 200 Xiao Ling Wei Street, Nanjing 210094, China

*Corresponding author

Keywords: Mobile internet, Behavior preferences, Data mining, Support vector machine.

Abstract. The number of mobile Internet users is mounting, and the behavior preferences of its users also show some new characteristics and a tendency of diversification. In this paper, we combining theoretical research and empirical research, the key factors of mobile Internet users’ behavior preferences would be explored from multiple perspectives, in order to understand the underlying reasons of differences in behavior preferences between different users. Based on those factors and SVM algorithm’ adoption, the mobile Internet users’ behavior preferences would be classified. Finally, some effective suggestions would be put forward in network continued development and personalized service.

Introduction

In recent years, with the gradual promotion of smart phones and 4G networks, mobile Internet users and related applications present an unprecedented growth in the scale. This status will not only promote the innovation wave of mobile Internet, but also positively influence the concern degree which the researchers in related fields to mobile Internet. Although the scope of the mobile Internet’s influence is growing, the number of mobile Internet users are constantly refreshed, the mobile Internet is still in the development stage, there is still much growth room for mobile Internet in the telecommunications industry, and the analysis researches of the mobile Internet users behavior is still very little. These limitations greatly negatively affect the effectiveness of the mobile Internet operations strategy, so understanding mobile Internet users behavior become a pressing need.

Analyzing the impact factors of mobile Internet user’s behavior preferences and predicting user behavior is an important part which promoting mobile Internet to develop further, thus exploring mobile Internet user’s behavior preferences is extremely urgent and meaningful. In this paper, our main work is to analyze the key factors of mobile Internet users’ behavior preferences, to select the appropriate dimensions and to classify mobile Internet users’ behavior preferences based on SVM algorithm through the combination of literature analysis and practice investigation. We expect the research results could give the operators of mobile Internet some inspiration in strategy formulation and strategy optimization.

Related Works

Despite the rapid development of mobile Internet, there are still relatively few researches on mobile Internet users, and the researches primarily focus on three aspects, user attitudes, user characteristics and user behavior. In both user attitudes and user’s individual characteristics aspects, Okazaki (2006) [1] used a two-step clustering method to divide Japanese mobile Internet users. He found that the groups who have an optimistic and positive attitude towards the use of mobile are company's managers and housewives generally, and the users who less use mobile Internet mainly are highly educated experts and freelancers. Seongil (2009) [2] explored the attitude of using mobile Internet in South Korea from the perspective of user expectation, and the result indicated...
that mobile Internet users' age, the types of mobile Internet services and service delivery styles will affect the attitude of using mobile Internet. Moreover, Okazaki and Romero (2010) [3] also explore the effects of gender on the use of mobile Internet, the results explained that women account for the majority among the users who use both mobile Internet and PC Internet user groups, while men primarily use the PC Internet. As for the studies on user’s behavior preference, researchers pay more attention to the influence factors and model of user’s behavior preference. However, in the factors research, experts concentrate on investigating the effects of external environment, stimulus, and product information on user’s behavior preference and rarely consider user’s characteristics. Catherine & Anthony, et al. (2015) [4] found feeding style of parents plays a very important role on children’s future food preferences. For the impacts of product information on user’s behavior preference, Zenebe (2010) [5] thought the reason why users prefer certain product is this kind of product’s related attributes, such as brand, price and so on. In addition, in order to verify this conclusion, Diane (2015) [6] set college students as experimental subjects and conducted some experiments to explore the impact factors of academic reading preferences in a view of product information. He found that users prefer to choose printing rather than electronic version when they intend to learn, and the factors affecting user’s actual behaviors included the reading’s accessibility, cost, complexity and importance, etc. In terms of user behavior preference model, Shang, et al. (2002) [7] developed a model of internal preference and provided conceptual support to deeply explore the changing process of user's internal preferences.

Methodology

First, the considered dimensions of the natural characteristic are physiological characteristics, geographical attributions and social attributions. Physiological characteristics commonly includes age, gender, height, etc.; geographical attribution mainly refers to the province or city that users are live in; social attributes generally deal with some problems of marital status, occupation, level of education, income, etc. In our study, considering the research object is a group of students in the same region and their natural attributes’ similarity is relatively high.

The problems of this part mainly are designed from three aspects, namely WIFI, smart phones and 4g networks. For the perception characteristics, the problems in the design of questionnaire mainly included whether the communication service and mobile traffic in the phone plans are appropriate. Are you willing to add the cost of accessing Internet by mobile phone? Do you think the Internet speed is very important? The specific items designed in all aspects are shown in Table 1.

Table 1. Questionnaire items list.

<table>
<thead>
<tr>
<th>Attributions</th>
<th>ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior characteristics</td>
<td>B1</td>
<td>How much mobile traffic do you use per month?</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>How long do you use WLAN per month (minutes)?</td>
</tr>
<tr>
<td></td>
<td>B3</td>
<td>Which business do you usually use when you are on mobile</td>
</tr>
<tr>
<td></td>
<td>B4</td>
<td>Do you use WIFI to surf the mobile Internet?</td>
</tr>
<tr>
<td></td>
<td>B5</td>
<td>Which way do you use to get WIFI?</td>
</tr>
<tr>
<td></td>
<td>B6</td>
<td>Are you using a smart phone now?</td>
</tr>
<tr>
<td></td>
<td>B7</td>
<td>Are you using a 4g phone now?</td>
</tr>
<tr>
<td>Perception characteristics</td>
<td>P1</td>
<td>Do you think the communication service and mobile traffic contained in the current mobile phone plan are appropriate?</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>Do you think the Internet speed is more important?</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>Do you think the calling quality is more important?</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>Are you willing to add the cost of accessing the Internet by</td>
</tr>
<tr>
<td>Natural characteristics</td>
<td>N1</td>
<td>Your gender?</td>
</tr>
<tr>
<td></td>
<td>N2</td>
<td>Education background?</td>
</tr>
<tr>
<td></td>
<td>N3</td>
<td>Your average consumption per month?</td>
</tr>
</tbody>
</table>
The Logit model is a kind of nonlinear classified statistical method, it can effectively solve the problem that dependent variable only has two options and these options are discontinuous. In order to adapt to the condition of the Logit model, there designed specific options for some open questions in the questionnaire items, as shown in Table 2, and modified these problem options to yes-or-no questions. For example, “you usually browse microblog when you are on mobile Internet”, would be modified to ask, “Would you like to browse microblog when you are on mobile Internet?”, if the answer is true, then marked it as 1, otherwise marked it as 0.

Table 2. Open questions option design.

<table>
<thead>
<tr>
<th>ID</th>
<th>Items</th>
<th>Options</th>
</tr>
</thead>
</table>
| B3 | Which business do you usually use when you are on mobile Internet? | B31 Browsing the microblog.  
B32 Using instant messaging tools such as WeiXin  
B33 Surfing social-networking site and campus BBS.  
B34 Downloading or streaming music and video  
B35 Playing online games.  
B36 Doing E-reading.  
B37 Onlineshopping (Mobile payment, group purchase, download coupons, buy train ticket, etc.)  
B38 Information retrieval and browsing.  
B39 Sending and receiving e-mail. |
| B4 | Do you use WIFI to surf the mobile Internet? | B41 Frequently.  
B42 Occasionally  
B43 Don’t use, mobile traffic is enough.  
B44 Don’t use, WLAN’s price is expensive.  
B45 Don’t use, using WLAN is troubled. |
| B5 | Which way do you use to get WIFI? | B51 WLAN in the mobile phone plan and the free WLAN operators provide.  
B52 WLAN paid individually.  
B53 Free WIFI in the schools and public places.  
B54 WIFI of dormitory routers.  
B55 else. |

These actual results would be normalized to avoid the deviations that result from different dimensions. The specific normalized formula is as follows:

$$x' = \frac{x - minValue}{maxValue - minValue} \quad (1)$$

$x'$ is the processed data, $X$ is the original data, $minValue$ is the minimum data in the data set, $maxValue$ is the maximum data in the data set.

Assume the probability that users are willing to add the fee for mobile Internet is $p_i$, and then $(1 - p_i)$ is the probability that the users mean not to add the fee for mobile Internet. So, the dualistic Logit model of describing whether the users are willing to add the cost of access the Internet by mobile phone can be represented as:

$$log \frac{p_i}{1 - p_i} = \alpha + \beta X = \alpha + \beta x_{B1} + \beta x_{B2} + \beta x_{B31} ... + \beta x_{N3} + \varepsilon_i \quad (2)$$

Among them, $\alpha$ is constant. When $\beta$ is a positive number, it means that the correlation between dependent variable and $x_i$ is positive; When $\beta$ is a negative number, it means that $x_i$ is negatively related to the dependent variable, and $\varepsilon_i$ is random variable. Finally, the questionnaire data would be brought into the formula to compute the probability $p_i$ of every participant.

The training set $T$ is known.

$$T = \{(x_1 \cdot y_1), (x_2 \cdot y_2) ... (x_n \cdot y_n) \in (X \times Y)^m\} \quad (3)$$

$X_i \in X=R^n, \ y_i \in Y=\{1,-1\}, \ i=1,2,3, ... ,m; X_i$ is feature vector.
Select the appropriate kernel function \( K \) and parameter \( C \), then construct the optimization problem and solve it.

\[
\begin{align*}
\min & \quad \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j y_i y_j K(x_i \cdot x_j) - \sum_{j=1}^{m} a_j \quad s.t. \sum_{i=1}^{m} a_i y_i = 0 \\
0 & \leq a_{i} \leq C, \quad i = 1,2,3 \ldots, m, \text{get optimum solution: } a^{*} = (a_{1}^{*}, \cdots, a_{m}^{*})^{T}.
\end{align*}
\]  

Select a positive classification \((a_{j}^{*})\) that less than \( C \) in \( a^{*} \) set, and calculate \( b^{*} \).

\[
\begin{align*}
b^{*} &= y_j - \sum_{i=1}^{m} a_{i}^{*} y_i K (x_i \cdot x_j) \quad (5) \\
\text{Construct the decision function.}
\end{align*}
\]

\[
f (x) = sgn (\sum_{i=1}^{m} a_{i}^{*} y_i K (x \cdot x_i) + b^{*}) \quad (6)
\]

\( K(x \cdot x_{i}) \) is a kernel function. Among the above decision functions, we need to determine the kernel function and the model parameter \( C \). In this study, the RBF (radial basis function) would be used as a kernel function.

\[
K(x \cdot x_{i}) = \text{Exp} \left(-\gamma ||x - x_{i}||^2 \right), \gamma > 0 
\]  

At present, the kernel functions be used most widely mainly has four types, namely the linear kernel function, polynomial kernel function, the RBF and sigmoid kernel function. There are three reasons why not choose other kernel functions.

**Results and Analysis**

**Descriptive Statistical Analysis**

This research chooses 200 samples and collects 168 questionnaires, then 10 useless questionnaires were refused and finally 168 available questionnaires were attained, the effective rate of questionnaire is 79%. In the sample, there are 76 participants are female and they cover 53.16% of the whole sample, on the contrary, the rest 82 participants are male and they occupy the 51.90%. It is in accordance with the ratio of schools that be selected. The concrete features are in the Table 3.

### Table 3. Demographic statistics of the sample.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Options</th>
<th>Number of participants</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>82</td>
<td>51.90%</td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>76</td>
<td>48.10%</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Junior college/bachelor</td>
<td>84</td>
<td>53.16%</td>
</tr>
<tr>
<td></td>
<td>Master</td>
<td>48</td>
<td>30.38%</td>
</tr>
<tr>
<td></td>
<td>Doctor</td>
<td>26</td>
<td>16.46%</td>
</tr>
<tr>
<td>Average consumption per month</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Below 500</td>
<td>1</td>
<td>0.63%</td>
</tr>
<tr>
<td></td>
<td>500-1000</td>
<td>122</td>
<td>77.22%</td>
</tr>
<tr>
<td></td>
<td>More than 1000</td>
<td>35</td>
<td>22.15%</td>
</tr>
</tbody>
</table>

**The Critical Factor Analysis of Mobile Internet Users’ Behavioral Preference**

This study used SAS to analyze the result data. In SAS, dualistic Logit model was chosen to gain the correlation between independent variable and dependent variable, the optimization method is Fisher scoring method, the amount of observations is 158. It’s turned out that only 12 users are willing to add the cost of accessing Internet by mobile phone among 158 mobile Internet users. In addition, SAS also generated a maximum likelihood estimated value table. The absolute value of their estimated value is all less than 0.001, so the six variables which have best relevancy in negative and positive separately was selected as the classification features in later stage, taking consideration into p value (it shows significant influence when p value is less than 0.1), so the final result is in Table 4.

**694**
Table 4. The results of maximum likelihood estimation.

<table>
<thead>
<tr>
<th>Number</th>
<th>ID</th>
<th>Content</th>
<th>Estimated value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P2</td>
<td>Speed of Internet</td>
<td>2.4982</td>
</tr>
<tr>
<td>2</td>
<td>B37</td>
<td>online shopping</td>
<td>2.1321</td>
</tr>
<tr>
<td>3</td>
<td>B44</td>
<td>Overcharge of WLAN</td>
<td>1.9972</td>
</tr>
<tr>
<td>4</td>
<td>B43</td>
<td>Enough mobile phone plan to use</td>
<td>-12.3247</td>
</tr>
<tr>
<td>5</td>
<td>B33</td>
<td>Browsing Social network site, school forum</td>
<td>-3.2820</td>
</tr>
<tr>
<td>6</td>
<td>B32</td>
<td>Usage of instant messaging applications as WeiXin,</td>
<td>-10.412</td>
</tr>
</tbody>
</table>

From table 4, some conclusions could be gained. Speed of Internet through mobile phone, shopping online and price sensitivity of WLAN positively influent the intention that mobile Internet users add the cost of accessing the Internet by mobile phone. And the accessing Internet by mobile phone is to scan social network and school forum, and to use instant messaging applications like WeiXin significantly and negatively influent the intention of mobile Internet users adding the cost of accessing the Internet by mobile phone.

**Classification of Mobile Internet Users’ Behavioral Preference Classification**

Model parameters C is the parameter c, the parameter $\gamma$ in kernel function is the parameter g in matlab. SVMcgForClass function was used to get the optimal parameter values, the specific code is:

$\{\text{bestacc, bestc, bestg}\} = \text{SVMcgForClass(train_labels, train, -10, 10, -10, 10)}$;

After running program, some data was attained: Best Cross Validation Accuracy = 91.4651%, Best c = 2, Best g = 0.1, so when the parameter C is 2, g is 0.1, the accuracy of the classification result is 91.4651%. after calculating the parameters value, the class of training set and test set could be predicted by using the ‘svmtrain’ and ‘svmpredict’ functions, the of classification accuracy of training set and test set is 100% and 91.4651% respectively. It shows that SVM method works well in classifying mobile Internet users’ behavioral preferences.

**Conclusion**

With the rapid development of mobile Internet and the growing popularity of smart phones, the influence scope of mobile Internet is expanding, there are more and more people tend to use mobile Internet, which brings great changes to people’s lives, study and works. However, the mobile Internet is still in early stage and still has development space. On the one hand, there are a great number of potential mobile Internet users, they may not contact or accept this emerging information technology for some reasons, such as charges, lack of knowledge, etc. On the other hand, it is a hard issue to address that how to make existing users continuous use the mobile Internet. In order to explain the mobile Internet user behavior, we explored the factors which influent mobile Internet users’ behavioral preference in three aspects, namely natural characteristics, behavioral characteristics and perceptual characteristics, from view of data mining based on Logit model, then six most important factors were extracted as SVM model’s features to classify mobile Internet users’ behavior preferences. This design of research methods breaks the traditional methods and ideas of user behavior analysis, it not only can help us solve the nonlinear issue in empirical study effectively, but also broaden the application scope of support vector machine model.

**Acknowledgment**

This research was supported by Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (No. PPZY2015A058).

**References**


