A user interest model based on the analysis of user behaviors

Zhu Jinghua

College of Network Communication, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, China

Email address: uuem@163.com

To cite this article:

Abstract: Understanding the users' interest is the base for the industrialization of website. In order to provide individualized service better for the users, on the basis of analyzing the users' browse behavioral characteristics and according to the users' retention time in the page, and users' click frequency to the hyperlink and page, a model of computer user interest degree is established, and a neutral network is proposed to describe their correlation, and the reasonableness and effectiveness of this model are verified through experiment. The experiment result shows athat this model can accurately find out the page that the users are interested in.

Keywords: Individualization, User Browse Behavior, User Interest Degree, RBF Network

1. Introduction

With the application and development of network technology globally, network is influencing people's work and lifestyle in various aspects. However, the existing information system has significant defect, such as scattered resources, concentrated retrieval, the information provided to all users is the same and there is response only in case of demand. For the ordinary users, the "information disorientation" and "information overload" on the internet have become increasingly serious problems. The key to solve these problems is to transform the Internet from passively accepting the browsers' request to actively perceiving the information demand of browsers' [1], so as to achieve the individualized active information service of Internet system to the browser.

In order to achieve the individualized service, first it is needed to trace and learn the users' interest and behavior, depict the relation between users' characteristics and the users. It is an important direction in the current individualized service research to analyze and capture the user interest according to the browse behavior or browse contents[2]. Radial basis function (RBF) neutral network, with its profound physiological basis, simple network structure, rapid learning ability and excellent approximation performance, is also well applied in the recommendation of webpage individualization[3]. A method of to calculate the users' interest in the webpage is given in this paper through the analysis on the users' browse behavior, and several important characteristics are grasped to describe the users' browse behavior, and RBF is used to describe their correlation.

2. Characteristic Extraction of User Access Behavior

A lot of researches show that the users' interest in the webpage is closely related with their browse behavior on such webpage. Users visit a webpage usually with a certain hobby, and different users have different interests and hobbies. The users' access path contains the users' interest in a website and the shifting of user interest. Literatures [4] point out that many actions of uses can imply their hobby, for example inquiry, browsing webpage and article, labeling bookmark, feeding back information, clicking mouse, dragging scroll bar, advance and retreat. [5] point out that the retention time, access frequency, saving, edition, modification and other actions of the users during access can reveal their interest. However, these articles do not quantitatively estimate how these behaviors reflect the users' interest.

Superficially, there are many browse behaviors that can decide the users' interest to the webpage, but through analysis, we find that there are three behaviors playing the key role: browse time in webpage (the browse behavior is classified according to the specific time interval, and an individualized
2.2. Judge the Users' Interest Degree According to the Users' Link to the Hyperlink in the Webpage

Let that users obtain the URL of a site through the retrieval to a key word, and the browse to each page in this site is Page A→B→C→⋯→N, and then the link is from this site to other sites \( S_1, S_2, \cdots, S_s \), and the calculation formula of the weight is:

\[
R(P) = (1 - J_1) \sum m_j / c_j + (1 - J_2) \left[ \sum \sum m_{ji} / c_{ji} \right] 
\]

Where \( J \) represents the information disorientation rate when the users enter a new site each time. Because hyperlink is different from the traditional information carrier, there are different hyperlinks in the hypertext, which indicate to different contents, users are easy to follow the hyperlink when browsing the webpage and will be disoriented in the complex network information space of Internet, do not know their position in the information space, cannot return to a node and forget the original retrieval objective.

As shown in Table 1, the weight calculation formula of site 1 is:

\[
R(S_1) = (1 - J_1) \times (4/8 + 3/6 + 2/6) + (1 - J_2) \times (2/6) = (1 - J_1) \times (4/3) + (1 - J_2) \times (1/3) 
\]

2.3. The Method to Determine the Webpage Weight According to the Users' Click Rate to the Webpage

If a user is interested in a webpage, then he will spend more time in browsing the webpage and will frequently access this webpage, and this is a method to measure the user interest quantitatively. However, the user's click frequency cannot accurately reflect the users' interest, because with the accumulation of time, the users must have more clicks to a webpage.

As shown in Table 2, the total click of Page A is 273, more than that of Page B 247, but in the 4th week, it is obvious that users are more interested in Page B, so the users' click rate can more reflect the change and strength of change.

The click rate can be described with the following formula:

\[
\text{Click rate} = \frac{\text{Number of clicks}}{\text{Time}}
\]
\[ C(P) = \frac{m}{M} \quad (5) \]

Where \( m \) is the access time of the node, and \( M \) is the total access time in all nodes.

### 3. Modeling and Verification

According to the characteristics of the user access behavior extracted above, the calculation method of determining the users’ interest degree according to the users’ browse behavior is given. The following formula is constructed:

\[ W(CRT) = f[C(P), R(P), T(P)] \quad (6) \]

Where \( W \) is the users’ interest in webpage \( P \) obtained through weight calculation.

Webpage interest degree means the degree of interest of users in a webpage, which is expressed with the real number between 0 and 1, where 0 and 1 respectively reflect no interest and maximum interest.

Literature [3] uses the characteristics of strong adaptability and learning ability of neural network to train the extraction of users’ different demands, the users are classified into different clusters, and then are applied in the individualized recommendation of webpage to improve the problem of information overload on the electronic commerce website, in which the most common neutral network is BP network, also called multi-layer feed-forward network. When BP network is used for the function approximation, the weight value is adjusted with the gradient descent method. This method weight adjustment has its limitation and has the disadvantages of slow convergence method and being extremely small locally, while radial neutral network is superior to BP network in the aspects of approximation ability, classification ability and learning speed. In this paper, with the characteristics of self-adaptive determination of radial based network (RBF network), no relation between output and initial weight value and high efficiency, a model based on RBF is designed.

The basic idea of RBF-based network model is: first, \((G_{S_1}, \ldots, G_{S_n}, \ldots, G_{S_m})^T\) is taken as the input vector of the network and \(GV=(g_{v_1}, \ldots, g_{v_k}, \ldots,g_{v_p})\) as the target vector to train RBF network and get a well trained RBF network, and according to the actual condition of the network, \(GV=(g_{v_1}, \ldots, g_{v_k}, \ldots,g_{v_p})\) is output, and the calculation accuracy of the target vector is compared, and then the parameters are adjusted, so that \(GV\) and \(GV\) approach to each other as far as possible.

The specific steps to establish the determination method of RBF network-based model are as follows:

Step 1: Initialization. Let the input vector of RBF network is \((G_{S_1}, \ldots, G_{S_n}, \ldots, G_{S_m})^T\), \(GV\) is the target vector, and set the parameters of RBF such as the number of neuron in the hidden layer.

Step 2: Train the neural network \(N\);

Step 3: Adjust the parameters with the result of test data, so that the actual output of network approaches to \(GV\) as far as possible, let the actual output is \(GV=(g_{v_1}^1, \ldots, g_{v_k}^1, \ldots,g_{v_p}^1)\).

We use the newrb command in MATLAB statistical tool to train the RBF network model, and then call sim and test whether the network model is reasonable with the test set. The calling format of newrb is:

\[ \text{net} = \text{newrb}(P, T, \text{spread}) \]

Where, \(P\) is the distribution density of the radial base function, the larger spread is, the smoother the network forecast value performance will be. \(P\) and \(T\) respectively represent the input vector and target vector in the training sample, newrb can create an accurate RBF network, that is to say, the network creation process is also a training process, and the error of the network created is 0.

<table>
<thead>
<tr>
<th>Interest degree</th>
<th>Numerical expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very interested</td>
<td>(0.8,1.0]</td>
</tr>
<tr>
<td>Relatively interested</td>
<td>(0.6,0.8]</td>
</tr>
<tr>
<td>Ordinary</td>
<td>(0.4,0.6]</td>
</tr>
<tr>
<td>Not very interested</td>
<td>(0.2,0.4]</td>
</tr>
<tr>
<td>Very uninterested</td>
<td>[0,0.2]</td>
</tr>
</tbody>
</table>

In this paper, 80% of the samples are taken as the training sample, and the remaining 20% samples are taken as the test sample. The test code is:

\[ Y = \text{sim}(\text{net}, P_{\text{test}}) \]

In this paper, spread=9 is taken, and the forecast error at this moment is 0.

In order to verify the effect of RBF network model forecast user to the interest in webpage, we ask the users to give the pre-forecast interest degree when browsing the webpage, as shown in Table 3:

In Table 4, SPSSS statistical analysis software is adopted for relevant analysis, and the method of distance analysis is adopted to determine the similarity, and a similarity matrix is obtained (aa represents the estimated value).

<table>
<thead>
<tr>
<th>Correlation between Vectors of Values</th>
<th>1: 10</th>
<th>2: 30</th>
<th>3: 60</th>
<th>4: aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 10</td>
<td>1.000</td>
<td>.960</td>
<td>.955</td>
<td>.943</td>
</tr>
<tr>
<td>2: 30</td>
<td>.960</td>
<td>1.000</td>
<td>.977</td>
<td>.969</td>
</tr>
<tr>
<td>3: 60</td>
<td>.955</td>
<td>.977</td>
<td>1.000</td>
<td>.995</td>
</tr>
<tr>
<td>4: aa</td>
<td>.943</td>
<td>.969</td>
<td>.995</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The difference between the estimated value and calculated value is compared in the experiment. In the test, the number of webpages browsed by users increases from 10 to 60, the distribution density of radial base function spread is 9, and RBF network model is used to respectively calculate the corresponding result, which is compared with the pre-forested interest degree. The result is as shown in...
Figure 1, in which the transverse coordinate represents the number of the webpage browsed by users (extract 10 pages at random, and sort them according to the value of page interest degree from small to large), and the longitudinal represents the interest degree. It can be seen from the figure that first, with the increasing of number of webpages browsed by the users, the estimated more and more approaches to the calculated value, and the change trend tends to be consistent (this phenomenon can be found in the similarity matrix in Table 4), indicating that the capturing of users' interest in webpages is more and more accurate with the increasing of webpage browsing; second, when the interest degree is relatively low and high, the estimated value approaches to the calculated value relatively; when the interest degree is between 0.3 and 0.7, there is a great error, and the reasons for the above phenomena might be the following: users are very sensitive to the webpage that they are very interested and uninterested in, the subjective scoring is relatively accurate, while for the scoring to the webpages that they are not so interested in, there might be a great deviation, and in this way, compared with the forecasted value, there is a significant error. When there is a large data volume, the neutral network model trained can more reflect the real state, so the calculated value and pre-forecasted interest value approach to each other very much.

4. Conclusion

It is very significant to find out the interesting user access mode, interesting webpages and user interest migration etc. from the log data record for the auxiliary design of website and strategic decision making of electronic commerce. In this method, how the users' browse behavior reflects the users' interest is quantitatively analyzed and estimated. Several important characteristics of user browse behavior are extracted, it is proposed to use RBF network model to describe the correlation between these characteristics and the users' interest degree, and the reasonableness of this model is verified through experiment. However, as the users' browse behaviors are different and because of the randomness of user browse, it is difficult to extract all characteristics of user behavior. The method to expect the user interest in this paper is just a relatively reasonable calculation method, if it is required to more accurately judge the users' interest degree, it is required to consider the correlation between the user browse behavior and page more, and consider how to mine the high-quality log data.

Acknowledgements

This work is supported by Zhejiang Province Education Department projects (No. Y201330252).

References