International Journal of Intelligent Information Systems
2015; 4(2-2): 5-8
Published online February 11, 2015 (http://www.sciencepublishinggroup.com/j/ijiis)
doi: 10.11648/j.ijiis.s.2015040202.12
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

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:
Zhu Jinghua. A User Interest Model Based on the Analysis of User Behaviors. *International Journal of Intelligent Information Systems.* Special Issue: Content-Based Image Retrieval and Machine Learning. Vol. 4, No. 2-2, 2015, pp. 5-8. doi: 10.11648/j.ijiis.s.2015040202.12

Abstract: Understanding the users' interest is the base for the industralization 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 experiemtn result shows aathat 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 activley 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 analuze 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 grapsed 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, fo example inquiry, browsing webpage and articlee, 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: borwse time in webpage (the browse behaivor is classified according to the specific time interval, and an individualized

recommendation model is established in the method of clustering [6]), the link clicked under a webpage (forecast the users' next behavior, calculate the webpage weight and recommend the webpage by collecting the information of users' browse behavior and through a certain mechanism), and the click frequency of a webpage (how users use mark and collection to further analyze the user's browse and search behaviors when using the browser[7]). There are three reasons: a. inquiry, edition, modification and other behaviors must increase the webpage browse time and page turning times (after page turning, what is positioned is still the same URL, so the manifestation is still to increase the webpage browse time), so it can be indirectly reflected through the webpage browse time. b. The page implementing the actions like saving and labeling bookmark, if really concerned by users, will be called for many times in the future for rebrowse, so it can reflect the access times. c. If the users retrieve based on the site, and there are more links making the users interested in the site, there will be more user click links, and the webpage is more important.

Table 1. *Link click condition*

Site		**Site 1**		**Site 2**
Page	**Page 1**	**Page 2**	**Page 3**	
Existing links	8	6	6	6
Clicked links	4	3	2	2

2.1. Method of Determining Page Weight According to the Users' Retention in the Page

Assume that there are n hyperlinks in a page accessed by the users, in which the accessed hyperlinks are respectively $\{H_1, H_2 \cdots, H_i\}$, the retention time is $\{T_1, T_2 \cdots, T_i\}$, and T(P) represents the weight of page P determined based on the users' retention time. Let a is the effective time of user retention in the current page, then the recurrence formula to calculate the page weight is:

$$T(P) = a + \sum (T_i / n) \tag{1}$$

This formula considers that the child nodes have some characteristics of the parent node, and some child nodes are even the embodiment of the parent node, so the interest in child nodes increases, which can be seen as the contribution to the interest strength of the parent node.

Actually, the length to browse the webpage is always closely associated with the total information in the webpage. In consideration of this situation, the above formula can be changed as:

$$T(P) = \frac{[a + \sum (T_i / n)] * B_i}{\sum B_i * \sum T_i} \tag{2}$$

Where B_i represents the information amount in page i, i.e. the multiple of the ratio between the webpage length and total browse length by the ratio between the number of byte of the webpage and the totla information.

2.2. Judge the Users' Interest Degree According to the Users' Link to the Hyerlink 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 $S_1, S_2, \cdots, S_i$, and the calculation formula of the weight is:

$$R(P) = (1 - J_1) \sum m_{i1} / c_{i1} + (1 - J_2)[\sum \sum m_{ji} / c_{ji}] \tag{3}$$

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:

$$R(S_1)=(1\text{-}J_1)\times(4/8+3/6+2/6)+(1\text{-}J_2)\times(2/6)$$

$$=(1\text{-}J_1)\times(4/3)+(1\text{-}J_2)\times(1/3)$$

Where J_1 and J_2 are the information disorientation rate when the users enter a new site each time that should be given according to the statistics. According to the formula above, the weight formula to calculate each specific webpage 1 in the website i is:

$$R(I) = m_{i1} / c_{i1} \tag{4}$$

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 weill 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.

Table 2. *Link click condition*

Time	**Click condition (time)**	
	Page A	**Page B**
1st week	192	18
2nd week	58	42
3rd week	15	85
4th week	8	102

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:

$$C(P) = m / M \quad (5)$$

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 unders 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.

Literautre [3] uses the characteristics of strong daptability and learning ability of neutral 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 informaiton overlod 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 negative-gradient descent method. This method of weight value adjustment has its limitation and has the disadvantages of low 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, $(GS_1, \ldots, GS_j, \ldots, GS_h)^T$ is taken as the input vector of the network and GV =$(gv_1, \ldots, gv_k, \ldots, gv_p)$ as the target vector to train RBF network and get a well trained RBF network, and according to the actual condition of the network, $GV'=(gv'_1,\ldots, gv'_k,\ldots,gv'_p)$ 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 faras 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 $(GS_1, \ldots, GS_j, \ldots, GS_h)^T$, 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 neutral 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 $GV'=(gv'_1, \ldots,gv'_k,\ldots,gv'_p)$.

We the newrbe command int eh 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 newrbe is:

net=newrbe(P, T, spread)

Where, spread 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, newrbe 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 3. *Evaluation of pre-forecast interest degree*

Interest degreE	Numerial expression
Very interested	(0.8,1.0]
Relatively interested	(0.6.0.8]
Ordinary	(0.4,0.6]
Not very interested	(0.2,0.4]
Very uninterested	[0.0,0.2]

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

Y=sim(net, P_test)

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

In order o 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 obtianed (aa represnets the estimated value)

Table 4. *Similarity matrix result*

Correlation between Vectors of Values				
	1: 10	2: 30	3: 60	4: aa
1: 10	1.000	.960	.955	.943
2: 30	.960	1.000	.977	.969
3: 60	.955	.977	1.000	.995
4: aa	.943	.969	.995	1.000

The difference between the estimated value and calculated value is compared in the experiemnt. 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-forecasted interest degree. The result is as shown in

Figure 1, in which the transver coordinte 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 webpges browsed by the users, the extimated more and more approaches to the calculated value, and the change trend tends to beconsistent (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 repatively low and high, the estimated value approaches to the calcualted 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 webpge 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 dta record for the auxilairy design of website and strategic decision making of elecronic 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 reasonablenes 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 calcultion 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

[1] Enrique Frias-Martinez, Sherry Y. Chen, Xiaohui Liu. Investigation of Behavior and Perception of Digital Library Users: A Cognitive Style Perspective[J]. International Journal of Information Management , 2008(28): 355-365.

[2] Zhang Haitao, Jing Jipeng, Method of Determining Webpage Level According to User Browse Behavior [J]. Intelligence Journal, 2004, 23(3): 303-306.

[3] Cheng Chih Chang, Pei-Ling Chen, Fei-Rung Chiu, et al. Application of Neural Networks and Kano's Method to Content Recommendation in Web Personalization[J]. Expert Systems with Applications, 2008.

[4] A. Georgakis, H. Li. User Behavior Modeling and Content Based Speculative Web Page Prefetching[J]. Data & Knowledge Engineering, 2006(59): 770-788.

[5] Wang Jimin, Peng Bo, Analysis on Click Behavior of Search Engine Users [J]. Intelligence Journal, 2006(2): 154-162.

[6] Feng-Hsu Wang, Hsiu-Mei Shao. Effective Personalized Recommendation Based on Time-Framed Navigation Clustering and Association Mining [J]. Expert Systems with Applications, 2004(27): 365-377.

[7] Mrugank V Thakor, Wendy Borsuk, Maria Kalamas. Hotlists and Web Browsing Behavior - an Empirical Investigation [J]. Journal of Business Research, 2004(57): 776-786.

[8] Zeng Chun, Xing Chunxiao, Zhou Lizhu, Technical Overview of Individualized Service [J]. Software Journal , 2002(10): 1952-1961.

[9] Shuchih Emest Changa, S Wesley Changchiena. Assessing Users' Product-Specific Knowledge for Personalization[J]. Expert Systems with Applications, 2006(30): 682-693.

[10] Shu-Hsien Liao, Chih-Hao Wen, Artificial Neural Networks Classification and Clustering of Methodologies and Applications Literature Analysis From 1995 to 2005[J]. Expert Systems with Applications, 2007(32): 1-11.

[11] Huang Xiaoyuan, Tian Peng, Securities Selection Decision-making Tools based on Neutral Network[J]. Application of Systematic Engineering Theory Method, 1995(2): 60-65.

[12] Tan Qiong, Li Xiaoli, Shi Zongzhi, A Method to Realize the Individualzied Service of Search Engine[J]. Computer Science, 2002, 29(1): 23-25.