Journal of Theoretical and Applied Information Technology
30th November 2012. Vol. 45 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTEREST DEGREE ANALYSIS BASED ON BROWSING BEHAVIOURS

XIAO MA
School of Information, Xi’an University of Finance and Economics, Xian 710100, Shaanxi, China

ABSTRACT

This paper analyzes the relationship between user's browsing behavior and interest. User interest is reflected by typical user’s browsing behavior that can be categorized as: Saving page, Printing page, Adding page to Favorites, copying page content, the same page browsing times and page browsing time. Considering the five factors of user’s browsing behaviors as well as the size of a page, an algorithm based on the speed of a page browsing was derived for the user interest degree. Meanwhile, a BP neural network was utilized for the fusion of the user interest degree. The training samples of the BP neural network were feed by collected user's browsing behaviors. The rationality and feasibility of the algorithm for the user interest degree was verified in the research.

Keywords: *User’s Browsing Behaviors, Degree of the user interest, Browsing Speed, Web Interest Degree*

1. INTRODUCTION

In the Internet information retrieval, the personalized result based on user interest is one of the main approaches to improve the ratios of its completion and accuracy. It is also one of the important researches for personalized search engines.

Generally, there are two approaches to collect user interest. One of them is asking users to label the level of their interest before their information retrieval. In theory, this approach is able to obtain accurate user interest model relatively. However, users are often not willing to have such extra procedure for their Internet search in the real world. It is not only annoyed to be asked to fill the labeling pages while users are browsing, but also has privacy issues. Furthermore, even users mark the level of their interest as requested, it may not be able to accurately reflect the degree of their interest. The other collecting approach of user interest is based on user’s browsing behaviors. The degree of the user interest is extracted from user’s actual browsing behaviors without user’s knowledge. User’s actual browsing behaviors are presenting the user's current practical interest. This method has become the mainstream model for collecting user interest [1, 2].

There are all kinds of browsing behaviors. To estimate the degree of the user interest by their browsing behaviors, one of the key factors are determine which browsing behaviors are able to represent user interest. Another key factor is how to quantify collected user behaviors so that they can represent the user interest correctly. In this paper, we analyzed user browsing behaviors as well as the characters of user browsing. We focused on the characters of the page residence time, the time of page visiting, browsing speed, etc. to calculate the degree of the user interest.

2. CLASSIFICATION OF USER BROWSING BEHAVIORS

Existing researches indicate that user’s browsing behaviors relate to user interest closely. The behaviors are include querying, browsing pages and articles, tagging bookmarks, feedback information, clicking mouse, scrolling bar, forwarding, backward rolling and so on. The performance and movements of a page residence time, the time of page visiting, save, edit, modification also present the user interest.

According to the browsing behaviors related to user interest, the user's browsing behaviors can be categorized into three types of behaviors, physical activity, significant behavior and indirect action. The physical activity is those browsing behaviors that reflect to user’s thinking and mood swings. Physiological psychology research shows that when people find something interest, a serial of physiological reactions, such as the eye movement, the changes of heart beat and skin temperature, the user's facial expressions, etc., occurs [3, 4].

The significant behavior directly reflects the browsing behaviors of user interest. It includes page saving, page printing, adding a page to favorites, frequently visiting the same page, and so on. It is certain that a user has the high degree of interest on a page whenever a significant behavior happens.

The indirect action indirectly reflects the browsing behavior of user interest. It includes the time presence on the page, scrolling bar, moving or clicking mouse, using the UP/Down arrow keys to scroll pages, pressing Page Up/Page Down keys to flip, etc. during browsing pages. The indirect action itself does not determine whether a user is really interested in a page. However, the number and duration of indirect actions could reflect the degree of the user interest.

In the user's browsing behaviors, the physical activity cannot be used for the calculation of the user interest because of technical limitations at the present. The significant behavior and the indirect action can be used for the calculation of the user interest degree. Some research indicates that the smallest browsing behavior combinations of the user interest calculation can be the following five kinds: save page, print page, add to Favorites page, copy page content, the same page browsing times and page browsing time.

Analyzing the five kinds of smallest browsing behaviors above, it can be found that a page has the high degree of the user interest if the save page, print page and add to Favorites page happen. Generally, printing a page is less happening in the normal browsing behavior, and saving a page and adding a page to Favorites can be treated as the same type of behaviors. Therefore, the estimate of the user interest should be based on the comprehensive consideration of save page, add page to favorites, print page, the times of a certain page visiting and the dwell time on a page.

Between the numerous user's browsing behaviors, they could be related to each other, or independent. User’s browsing habits are also various. Therefore, the accuracy can be an issue if there are not enough browsing behaviors selected for the analysis. However, the complexity of computing interest, and the quantification of various browsing behaviors could be issues as well if there are too many browsing behaviors selected [5, 6].

Considering the user's browsing behavior comprehensively, we determined to estimate the level of the user interest by the following three factors: the user's actual browse action, the page of visits, the page browsing time.

3. BROWSING BEHAVIOR BASED CALCULATION OF INTEREST

Assuming a user browses multiple pages in a certain period of time, and the user could also revisit the same page several times. The degree of the interest to different pages is the key factor for the modeling. Provided a user has visited the pages w_1, w_2, …, w_n.

Analysis of user browsing behaviors, save pages, add pages to Favorites, print pages, the same page browsing times and page browsing time, these five categorized behaviors can be approximately represent all kind of the typical browsing behaviors. Thus, the degree of the user interest can be indicated as a function of the above five variables [7, 8, 9].

Provided $Interst(w)$ be the degree of the user interest to the page w, it can be expressed as:

$$Interst(w) \\ = f(Save(w), Keep(w), Print(w), Freq(w), Time(w)) \qquad (1)$$

Among them, $Save(w)$, $Keep(w)$, $Print(w)$, $Freq(w)$ and $Time(w)$ were five behavior functions for save pages, add pages to Favorites, print pages, the same page browsing times and page browsing time respectively. Also define the range of user interest degree value from 0 to 1, that is:

$$Interst(w) \in [0,1]$$

For the three actions, save pages, add pages to Favorites and print pages, there is no issue on the magnitude of the three actions. They can only have two states, occurring or not occurring. It indicates a relatively high degree of the user interest when the three actions occur. There is no need to further analyze the times of a certain page visiting and the dwell time on a page in such situation. Therefore, the function is given the greatest interest degree 1 when the three actions occur, and stop to continue analyzing the other two behavior functions.

There is the rate of changes on the times of a certain page visiting and the dwell time on a page. The magnitude of the difference means that the different degree of the user interest. For example, two pages w_1 and w_2, if in the same time period, the number of page w_1 visiting is 1, and the number of page w_2 visiting is 10, then the degree of the user interest for the page w_2 is significantly higher than the degree of the user interest for the page w_1. The dwell time on a page has the same characteristics.

According to the analysis above, to calculate the degree of the user interest, it is necessary to consider the times of a certain page visiting and the dwell time on a page only when none of the behaviors of the save pages, add page to Favorites and print pages occur.

Thus, the degree of the user interest for browsing behavior is expressed as:

$$Interest(w) = \begin{cases} \varphi(w) & Save(w), Keep(w), Print(w) \ \ No\ occurrence \\ 1 & Save(w), Keep(w), Print(w) \ There\ happened \end{cases} \quad (2)$$

Of which $\varphi(w)$ is the function of the times of a certain page visiting and the dwell time on a page.

In the real world, the browsing behaviors, saving, add page to Favorites and printing pages happen not very often. The estimate user interest for a page mostly comes from the times of a certain page visiting and the dwell time on a page. In other words, $\varphi(w)$ is the key for calculating the degree of the user interest that based on the browsing behaviors.

3.1 The Calculation Of User Interest Degree Based On The Times Of A Certain Page Visiting

The more visits on a page, the more interested in the page a user is. That is, the larger $Freq(w)$ in a period of time, the greater the user interest $Interst(w)$. The degree of the user interest based on the times of a certain page visiting can be described as:

$$Interest_{Freq}(w) = \frac{Freq(w)}{\max\limits_{v \in W}(Freq(v))} \quad (3)$$

Where, W is the set of pages visited in the period of time.

This method is a kind of quantitative measure method of the user interest. With the accumulation of time, a user clicking on a web page will also accumulate gradually to a lot, these historical accumulation does not necessarily be accurate to the user's current interest. Therefore, it is more important to set up the reasonable statistical browsing number of periods and corresponding update mechanism. To utilizing it, set up week for a statistical cycle. Every other week, reset the browse number of updates.

Set $Freq_{old}(w)$ for the former statistical cycle browsing times, $Freq_{new}(w)$ for the current statistical cycle browsing times, p is update proportion for this page browsing number, it can be expressed as:

$$p = (|\,Freq_{new}(w) - Freq_{old}(w)\,|) / Freq_{old}(w) \quad (4)$$

If $p < 0.5$, it shows that the change of small degree on the former statistical cycle and the current statistical cycle for a page to browse. Explain to the user page interest with the passage of time did not have larger transfer, that is:

$$Freq(w) = (Freq_{new}(w) \ + \ Freq_{old}(w)) / 2 \quad (5)$$

On the contrary, if $p<0.5$, it shows that the change of larger degree on the statistical cycle. Explain to the user page interest with the passage of time have larger transfer, that is:

$$Freq(w) = Freq_{new}(w) \quad (6)$$

3.2 The Calculation Of The User Interest Degree Based On The Browsing Speed

The more the browsing time on pages, the more interested in the pages a user is. On the other hand, the browsing time on a page is closely related to user's operation habit, speed and the page size. For the better consideration of all factors, we made the browsing time, or dwell time corresponds to the current browsing speed. The calculation of the user interest degree based on the browsing speed relates not only to the dwell time on the page, but also to the page size.

Define the browsing speed of page w:

$$Speed(w) = \ Size(w) / Time(w) \quad (7)$$

Among them, $Size(w)$ is the page size and $Time(w)$ is the browsing time visited currently.

$Time(w)$ calculation are influenced by many factors, for example the user's operation speed, users browsing speed, the current network transmission delay, server corresponding delay, etc..

Here, $Time(w)$ is defined as the time interval to visit two pages, that is:

$$Time(w_i) = T(w_{i+1}) \ - \ T(w_i) \quad (8)$$

Where, $T(w_i)$ is the request time of page w_i.

Browsing speed is the amount of bytes browsed by a user in a unit of time. For each page w, the faster a user browses, the larger $Speed(w)$ is. A large $Speed(w)$ presents that the user is not interested in the page w. Therefore, the interest level of a user to the web pages $Interst(w)$ and the user's browsing speed $Speed(w)$ is an inverse relationship. That is proportional to the user's dwell time, and is inversely proportional to the page size.

Then the degree of the user interest to a page can be calculated by the following formula:

$$Interest_{Time}(w) = \frac{1/Speed(w)}{\max_{v \in W}(1/Speed(v))} \quad (9)$$

As the user’s browsing speed could be various widely, the linear normalization process makes it hard to distinguish the differences for the most of the degrees of page interest. In order to improve the situation, the nonlinear normalization of browsing speed is introduced [7, 8, 9]. The degree of the user interest in the page w can be presented as the following formula:

$$Interest_{Time}(w) = \frac{\log_{10}(1/Speed(w)) - \min_{v\in W}(\log(1/Speed(v)))}{\max_{v\in W}(\log(1/Speed(v))) - \min_{v\in W}(\log(1/Speed(v)))} \quad (10)$$

Here, $0 \le Interst_{Time}(w) \le 1$.

In the reality, the user browsing status and environment could be pretty complex. Unexpected dwell time happens quite often. For example, a user leaves for something after he has opened a page. Because of the absence with the open page, it makes the dwell time of the page is far greater than the normal. Therefore, the degree of the user interest is close to 0, and the calculated degree of the user interest is meaningless.

To rule out the negative impact of user's browsing behavior in the calculation of user interest, an exception handling strategy is introduced for the abnormal dwell time. When a user’s dwell time on a page exceeds a predetermined threshold, the value of the dwell time will be set to the maximum. The dwell time value maintains its actual value when it is under the threshold. It is very important to select the threshold. If the value is too high, the most of the degrees of page interest will be close to 0. The result will be meaningless. On the other hand, if the threshold is too low, the most of the degrees of page interest will be close to 1. It is not conducive to distinguish the degree of page interest. Existing research indicates that for normal page’s visiting, 90% of the dwell time is ranged in 3 to 5 minutes. Therefore, it is reasonable to select 5 minutes as the threshold.

3.3 The Combination Of The Two Kinds Of The Degrees Of Interest

Considering both the times of page visiting and the browsing speed for the degree of the user interest, the BP neural networks is utilized for fusion of the two. BP neural network consists of input layer, output layer and several hidden layers. The number of hidden layer can be single or multiples. The more the hidden layers it has, the more complex the neural network structure is, and the longer the training time it needs. Typically, the hidden layer of a three-tier structure should meet the most of application requirements. Such structure is selected in this paper.

First, make the two browsing behaviors, the times of page visiting and the browsing speed for the degree of the user interest as a neural network input data. Its output is the fused degree of the user interest. According to the input and output data of neural network, the neural network consists of the two-node input layer, the single-node output layer, and six-node hidden layer. The structure of BP neural network is shown in Figure 1. The neural network toolbox in Matlab software is utilized to implement the neural network. Training sample data is selected for the neural network training. The trained neural network data is eventually used for the fusion of the degree of the user interest.

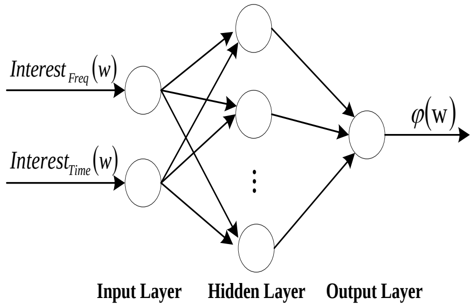

Figure.1 BP Neural Network Structure

4. TRAINING SAMPLE SELECTION AND EXPERIMENTAL ANALYSIS

In order to implement BP neural networks for the fusion of degrees of interest, a large volume of training sample data is required. A large volume of experimental data is needed as well to validate the rationality for the algorithm of interest degree. In order to obtain the raw experimental data, the Windows hook function in the function of the browser extension was used for embedding Java script into the browser, that recorded user's valid browsing behaviors effectively[10, 11, 12]. During one month period, we collected the data from the specific user groups who browsed the 280-pages website http://www.sina.com.cn with totally 3186 browsing behaviors. The collected raw data of the browsing behaviors is utilized as the training sample data that fed to the BP neural network for training.

150 typical pages with different themes were selected from the original recorded data. The web browsing behaviors of the 150 pages were collected from the web logs. The degree of the user interest was calculated based on the times of page visiting and the browsing speeds. The degree of page interest was obtained by the fusion of degree of interest in the trained neural network. Meanwhile,

we conducted a survey among the particular user groups. They provided their own evaluation of the degree of interest on the pages. Finally, the user evaluation result was compared with the calculated degree of page interest. Some typical data of the comparison are shown in Table I.

Through the error statistics in the result comparison, over 80% of the difference between the self-assessment and the computed degree of the user interest are less than 0.1. The average difference is 0.086. This indicates that web census interest measure and projected web interestingness is very close. Therefore, the rationality and the accuracy of the calculated degree of the user interest based on user's web browsing behaviors are verified.

Table I: The Comparison Of The Calculated Degree Of The User Interest And The User's Self-Assessment

URL	The Calculated Degree of User Interest	The User's Self-assessment	Difference
http://slide.news.sina.com.cn/......26485.htm	0.9821	0.9	0.0821
http://finance.sina.com.cn/......109002.shtml	0.6744	0.8	0.1256
http://tech.sina.com.cn/....../637077.shtml	1	1	0
http://blog.sina.com.cn/......9rr.html?tj=1	0.7521	0.6	0.1521
……	……	……	……

5. CONCLUSION

The behavior of user's browsing comprehensively reflects user interest. According to the relationship between the user interest and the browsing behaviors, the typical behaviors of user browsing can be reduced as saving pages, printing pages, adding pages to favorites, the times of a certain page visiting and the dwell time on a page. We discussed the calculation of the user interest degree based on the five categorized browsing behavior.

We also discussed the algorithms of the user interest degree based on the times of a certain page visiting and the dwell time on a page respectively, considered the influence factors from the page size, calculated the degree of the user interest by the interestingness algorithm based on web browsing speed. Meanwhile, BP neural network was utilized for the fusion of interestingness. A scientifically calculation method was obtained. Finally, the rationality and accuracy of the algorithm was verified through the calculation of the experimental sample data.

ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Shaanxi Province of China (Program No.2012JM8034 and Program No.2010JM8029).

REFRENCES:

[1] Songjie Gong, "Learning User Interest Model for Content-based Filtering in Personalized Recommendation System", International Journal of Digital Content Technology and its Applications, Vol.6, No.11, 2012, pp. 155-162.

[2] Yingxu Lai, Xin Xu, Zhen Yang, Zenghui Liu, "User Interest Prediction Based on Behaviors Analysis", International Journal of Digital Content Technology and its Applications, Vol.6, No.13, 2012, pp.192-204.

[3] A. Georgakis, H. Li, "User Behavior Modeling and Content Based Speculative Web Page Perfecting", Data & Knowledge Engineering, Vol. 59, 2006, pp.770-788.

[4] Zhu Zhen, Wang Jing-Yan, Chen Mei-Lan, "User interest modelling based on access behavior and its application in personalized information retrieval", Proceedings of 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, IEEE Conference Publishing Services, November 26-28, 2010, pp.266-270.

[5] Zheng Ling, Cui Shuo, Yue Dong, Zhao Xinyu, "User interest modeling based on browsing behaviour", Proceedings of 3rd International Conference on Advanced Computer Theory and Engineering, IEEE Conference Publishing Services, August 20-22, 2010, pp.V5455-V5458.

[6] Xing Kun, Zhang Bofeng, Zhou Bo, Liu Yucong, "Behavior based user interests extraction algorithm", Proceedings of IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, IEEE Conference Publishing Services, October 19-22, 2011, pp.448-452.

[7] SHAN Rong, "New user's interest model updated based on browsing behaviors", Electronic Design Engineering, Vol.18, No.4, 2010, pp.61-62.

[8] YIN Chunhui, DENG Wei, "Extracting User Interests Based on Analysis of User Behaviors",

Computer Technology and Development, Vol.18, No.5, 2008, pp.37-39.

[9] ZHU Zhengyu, ZHOU Zhi, LUO Ying, LI Lipei, "Extraction of User-Interested Web Page Based on the Browsing Action Quantitative Analysis", Journal of Chongqing Institute of Technology, Vol.23, No.7, 2009, pp.79-84.

[10] Pan Shouhui, Wang Li, Xia Guoping, "Personalized User Profile Mining Based on User's Browsing Behaviors", Journal of Computational Information Systems, Vol.7, No.14, 2011, pp.5041-5049.

[11] Zhang Yuchen, Chen Weizhu, Wang Dong, Yang Qiang, "User-click Modeling for Understanding and Predicting Search Behavior", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 21-24, 2011, pp.1388-1396.

[12] Velayathan Ganesan, Yamada Seiji, "Investigating User Browsing Behavior", Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Conference Publishing Services, November 2-5, 2007, pp.195-198.