The Algorithm Research and System Design on User-Oriented Personalized Search Engine

Huifang Deng, and Qinwei Wang

***Abstract*—**The traditional search engine systems are usually not user-oriented and the search results may not be satisfactory for users. In this paper we proposed a novel search engine algorithm based on BP (Back Propagation) neural network. The algorithm uses the samples collected by the personalized search engine system for training. In order to obtain the input and output vectors of the leaning samples, we developed a user interest model and quantified user's behaviors which measure the degree of user's interest. Furthermore, the algorithm using the neural network's pattern recognition capability, combined with user's interest and the search engine system effectively, makes search results more in line with the user's search query request.

***Index Terms*—**BP Neural Network, Personalized Search Engine Algorithm, User Interest Model.

I. Introduction

Search engine technology has been one of the most popular Internet applications in recent years. It is an important way for Internet users to obtain concerned information resources and they are increasingly dependent on the search engine systems. However, the currently widely-in-use search engines cannot accurately understand the user's intent. Queries from different users often return to the same search results, rather than the user-oriented or personalized results. Users have to spend so much time and energy to select satisfactory information from huge number of results. In order to improve the user experience in using search engine systems, the personalized search engine has come out. And it has become the spotlight of the search engine technologies.

There are diverse forms to realize personalization for search engine systems. Regardless of which form they used, the user interest model is needed to describe the interests of users [1]. The system usually collects user's characteristics, preferences and interests information actively or passively to establish a user interest model for each user. Then, combined with the information of the user interest model and some suitable algorithms, the search engine system can show the most satisfactory results to users.

Meanwhile, the search engine algorithm is playing a very vital role and it is the key to realizing the personalization. It is directly related to the final experience of users and the quality of search engine system. Generally, the traditional algorithms evaluate the importance level of a web page by analyzing its content or the linking relationships. A typical example is the PageRank algorithm, proposed by Google's founders Larry Page and Sergey Brin [2], which is based on the theory that the link structure within web pages could give some approximation on evaluating the quality of a web page. It is similar to what is happening in academic citation where the importance of paper is proportional to the citation a paper received from other papers. We can conclude that analyzing object by the traditional search engine algorithm is web page, and the users' interest is not considered so that the search results may not meet the user's demand. Therefore, the user-oriented search engine algorithm is worth to explore.

In this paper we mainly study the personalized algorithm in search engine system. We first introduce the concepts of "user interest model" and "user interest degree" and then propose a new user modeling method in Section II. The algorithm based on BP (Back Propagation) neural network is presented in Section III. In Section IV, we conduct some verification tests and give a description of system design based on the algorithm. The last Section is the conclusions.

II. User Interest Model And User Interest Degree

Before describing the personalized algorithm of search engine system, we must introduce the concepts of "user interest model" and "user interest degree". They are the premise of the algorithm proposed in this paper.

A. User Interest Model

At present, the user interest model has become the key technology of the personalized information service. It is used for describing and recording users' preference information, and catching potential interest spot of users. Specifically, the user interest model is one kind of formalized user description which has fixed data structure. In general, it is the set of some key words which can express the user's interest, and each word has the weight information which can be defined as a number. The higher the weight, the more the user is interested in.

There are many kinds of methods used to represent the user interest model, such as the theme representation, the key words list representation, the VSM (Vector Space Model) representation [3], etc. In this paper, we choose the VSM representation.

VSM representation is the most common text representation method which uses the vector in key words to represent the user interest model. The basic idea is that the text can be represented by a vector. For example, document D can expressed as : $D = \{(k_1, \omega_1),...,(k_i, \omega_i),...,(k_n, \omega_n)\}$, where k_i (i=1, 2, …,

Manuscript received on 5 February, 2016.

Huifang Deng is with South China University of Technology

Qinwei Wang was with South China University of Technology, Guangzhou, China. He is now with the No. 722 Institute of CSIC, Hubei, China

n) stands for the *i*-th key word of document *D* and ω_i is the corresponding weight of the word. Correspondingly, the user interest model can be represented as follows: $P = \{(pk_1, pw_1),...,(pk_i, pw_i),...,(pk_n, pw_n)\}$ .Where pk_i is the *i*-th interest word of the user interest model, and pw_i denotes the interest degree of the word, ranged from 0.0 to 1.0. A bigger pw_i means that users have more interest in the word pk_i .

Furthermore, the method used to establish the user interest model is called user modeling. It refers to the process of establishing the user interest model through mining the meaningful words from the user's background and behavior information. In terms of the degree of user's participation, the user modeling method is usually divided into three forms: the user manual made customized modeling, the demonstration modeling and the automatic modeling. The user manual made customized modeling method gets the interest word by hand inputting or choosing from the set of key words, which is easy to realize and needs user's participation greatly. The demonstration modeling method provides user with demonstrations and interest categories. According to user's browsed content and behaviors, the automatic modeling method automatically constructs the user interest model which does not need the user providing the information initially.

Combined with the advantages of the manual custom-made modeling and the automatic user modeling, in this paper we propose a new way to establish the user interest model. In detail, the user interest model can be divided into two parts: one is composed of the words the user inputs; the other contains the words by machine learning. Then, using the VSM representation method, the two parts can be expressed as follow:

$$S = \{(sk_1, sw),(sk_2, sw),...,(sk_n, sw)\} \quad (1)$$

$$L = \{(lk_1, lw_1),(lk_2, lw_2),...,(lk_m, lw_m)\} \quad (2)$$

The *S* stands for the first part; *sw* refers to the weight of the interest word which is set as a constant. And the *L* represents the second part, lw_i (*i*=1, 2, ... , *m*) is the word weight of this part which is calculated by TFIDF (Term Frequency and Inverse Document Frequency) algorithm [4]. The TFIDF algorithm, which is a classic method, can evaluate the importance of the word in documents. Nevertheless, it's just the initial state of the weighting factors. Generally, the interest word in the *S* part would be more representative than the one in the *L* part so that *sw* should be greater than lw_i universally. But the weighting factors of the two parts may rather differ from each other. In order to avoid this situation, in this paper we take a new way and the concrete steps to establish the user interest model as follow:

1) Calculate the average value of the weighting factors in *L* which is represented as μ_l .
2) Multiply μ_l by α , here α is set manually ($\alpha \in [1.0,3.0]$), the result is the *sw*.
3) Combine the *S* and the *L* together as the initial user interest model P'
4) Normalize the weighting factors of P' through (3).

$$pw_i = \frac{w_i - \mu_p + 3\sigma_p}{6\sigma_p} \quad (3)$$

5) Where μ_p is the average value of the weighting factors in P', σ_p is the standard deviation, w_i is the initial weight and pw_i is the final weight.

According to above steps, a useful user interest model *P* is established.

$$P = \{(pk_1, pw_1),(pk_2, pw_2),...,(pk_t, pw_t)\} \quad (4)$$

B. User Interest Degree

From the psychological perspective, the behaviors generally reflect the interest and purpose of human beings. In real life, we can conjecture someone's potential interest and intention through reading his face or analyzing what he said. Correspondingly, the information system could get the user's interest by analyzing his behaviors when surfing the internet. Some behaviors (such as collecting, frequently accessing and long staying in a web page) could imply that the user is of great interest in the web page.

In some researches, the user browsing behaviors can be divided into the following categories [5]:

1) Marking behaviors: appending bookmark, storing and printing the web page, etc.
2) Operation behaviors: copying, cutting, pasting the content of web pages and clicking hyperlinks, etc.
3) Repetition behaviors: repeatedly accessing a web page, etc.

The concept "user interest degree" refers to the degree of the interest in a web page or document for users. By quantitative analysis of user behavior, we can get numerical user interest degree which can be applied in related algorithm to realize the personalized information service.

Some behaviors, such as marking behaviors, are so certain that we can assign a constant to the user interest degree. However, the behavior that the user browses a web page is uncertain and the corresponding user interest degree needs to be calculated. It is in direct proportion to two parts: the browsing duration and the click counts. We can get it by following equations:

$$t(P) = \sum_{i=1}^{freq(P)} t(P,i) \quad (5)$$

$$ID(P) = \frac{t(P) - \mu_t + 3\sigma_t}{6\sigma_t} \quad (6)$$

Where *t(P)* is the total time when browsing web page *P*, *freq(P)* is the click counts and *t(P, i)* is the time of each (*i*-th) browsing; μ_t is the average value of *t(P)*, σ_t is the standard deviation and *ID(P)* is the user interest degree of web page *P* after normalization.

III. SORTING ALGORITHM BASED ON BP

Traditional sorting algorithms in search engine system are not user-oriented which just took the factors of web pages into consideration. In order to make search engine systems became

personalized, artificial intelligence technology has been used in these system increasingly. In this paper we proposed a novel sorting algorithm based on BP neural network, we will explain it in detail in the following content.

A. BP Neural Network Algorithm

BP Neural Network algorithm also called Error Back-Propagation algorithm [6], it's the most widely used Artificial Neural Network algorithm. BP Neural Network algorithm can learn and store large amounts of input - output mode mapping relationship without knowing the mathematical equations that describe the mapping relationship. As long as there is a sufficient number of learning samples for BP neural network to learn and train, the n-dimensional input space to m-dimensional output space nonlinear mapping can be completed. It has great advantages for solving the questions whose internal rules are hard to describe.

BP algorithm can get knowledge from samples by training the neural network, and it will be stored in the connection weights in the form of numerical values. The main idea is transferring information layer by layer forwardly, then the output-layer error will propagate reversely, and the error of hidden layer can be calculated indirectly. The entire process is shown in Fig. 1 [7]. In first phase (forward process), the information will be input into input-layer and then calculate the output values of the respective units. In second phase (back-propagation process), the error of the hidden-layer units will be calculated layer by layer, and the values of the front layer will be corrected. And in BP algorithm, Gradient-Method is used to correct weights.

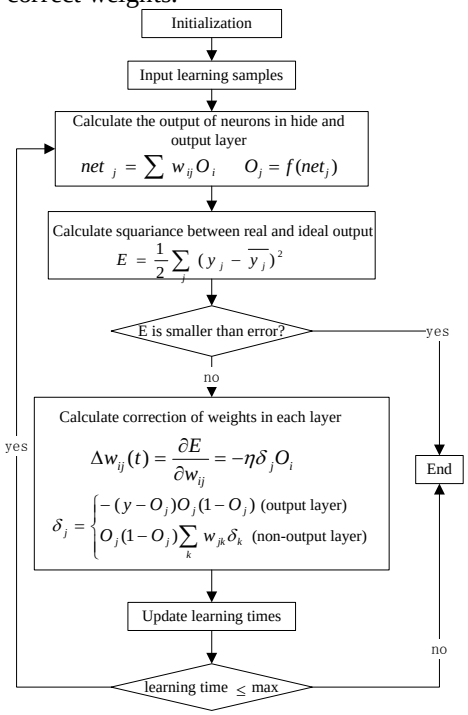

Fig. 1 BP Neural Network Algorithm Flow Chart

B. Learning Samples

Generally, the BP neural network needs a plenty of learning samples for training. The learning samples are usually representative web pages or documents and they definitely determine the quality of the neural network. Besides, how to collect the learning samples is an important work for personalized search engine systems.

Specifically, a learning sample is composed of input vector and ideal output vector. The input vector of a document has the same structure of the user interest model and it can be obtained after following processes:

1) Get the set of words by preprocessing which contains word fragmenting process and abandoning independent words.
2) Compare the set with the user interest model *P*: if the word exists in *P*, the corresponding weight is reserved; otherwise the weight is set as 0.

The ideal output vector in this paper is just the user interest degree of a document. We can obtain it by (5) and (6).

C. Sorting Algorithm Based on BP

After getting the respective learning samples, we can use them to train the neural network. Then, the neural network just likes another brain which knows the exact interest and preference information of users. By using the trained neural network, we can judge whether a document satisfies the requirement of users or not. And the suitable documents will be listed in the front to make users more convenient.

The specific steps of the algorithm can be described as follow:

Step 1) Get the stable BP neural network by training using the representative learning samples.

Step 2) Preprocess the disordered document to get its words set $D = \{d_1, d_2, ..., d_n\}$.

Step 3) Obtain the input vector of the document using the same way as for getting the input vector of learning samples.

Step 4) Acquire the output vector of the document through the trained neural network.

Step 5) Search the untreated documents. If there are untreated documents, go to step 2); otherwise, go to step 6) below.

Step 6) Sort the documents according their user interest degree.

Fig. 2 shows the specific flow of the algorithm.

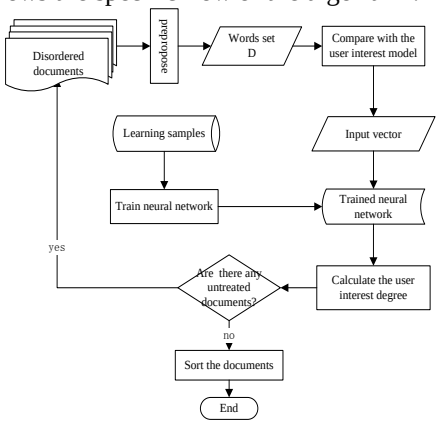

Fig. 2 Algorithm Based On BP Flow Chart

IV. Verification Tests And System Design

In order to verify the efficiency of the algorithm proposed in this paper, we conducted some experiments. Furthermore, a personalized search engine system is developed based on the algorithm.

A. Verification Tests

JOONE (Java Object Oriented Neural Network) is an open source project to establish the neural network in the "sourceforge" web site [8]. And the "joone-editor" is a graphical development environment of JOONE, which can construct a neural network rapidly. In this paper, we use it to carry out the verification tests.

In the first place, we list some representative learning samples. In order to improve the recognition capability of network, the learning samples are composed of two parts: relevant samples and irrelevant samples. In addition, the user interest model is extracted from the relevant samples. We select 10 words by the ways mentioned in this paper to establish the user interest model.

TABLE I
USER INTEREST MODEL

Interest	South china	Technology	University	Library	Job
Weight	0.72018	0.72018	0.77394	0.35126	0.37693
Interest	Postgraduate	Computer	Basketball	Sport	Soccer
Weight	0.45371	0.39778	0.37693	0.52891	0.30016

According to the method mentioned above, we can get the input vectors and ideal output vectors of learing samples. In order to reduce the problem complexity, we just consider one factor of ideal output vector which is click counts of documents.

Then, as shown in Fig. 3, we have established a neural network for test in the joone-editor.

Through the learning samples, we can train the neural network. Fig. 4 shows the control panel of joone-editor. After 10000 times of trianing, a stable network is established that has pattern recognition capability. RMSE represents the error of this network, its numerical value is 0.07207 which reaches the requirement of test (the required value should be less than 0.1).

Finally, we conducted two tests to verify the algorithm. The documents of test 1 contain the interest words of user interest model increasingly. And the doucments of test 2 contian the interest words separately, in which doucment 1 contians the most important interest word "job". Then, by using the trained neural network, we can calculate the user interest degree of each document.

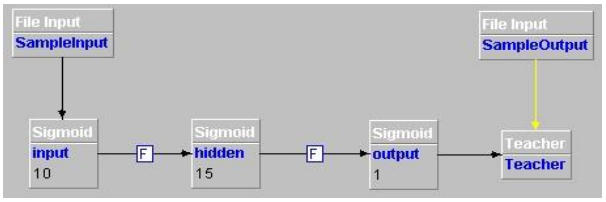

Fig. 3 Neural Network In Joone-Editor

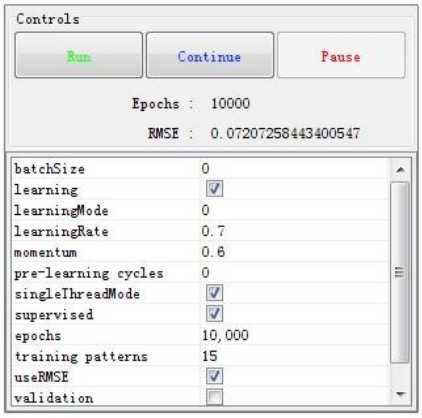

Fig. 4 Contorl Panel Of Joone-Editor

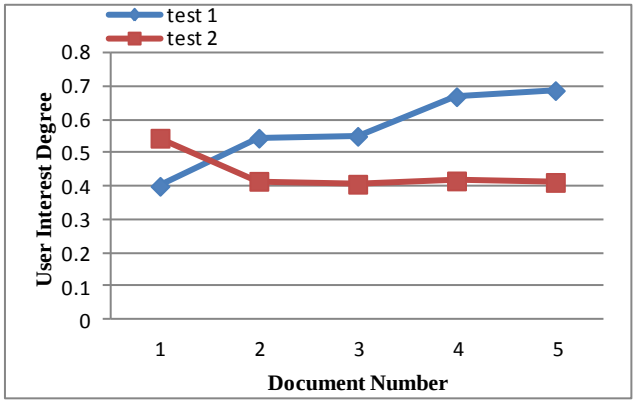

Fig. 5 Verification Tests

As shown in Fig. 5, the line of test 1 has the same trend as the situation that the documents contain the interest word increasingly. And the line of test 2 shows that the user interest degree of the document which includes word "job" is much greater than others. Therefore, we can draw two conclusions:

1) The more words a document contains, the greater its user interest degree.
2) The more important word a document contains, the greater its user interest degree.

Obviously, the test results verify that the proposed algorithm has taken into account the interest of users upon calculating the importance of documents. And it can meet the personalized requirement efficiently.

B. System Design

Based on the contents mentioned above, we design a user-oriented personalized search engine system. The structure of system functions is shown in Fig. 6.

Specifically, we use Lucene to realize the basic fuctions of search engine system [9].And the neural network is realized by the toolkit of JOONE, the word fragment model uses the JE toolkit [10]. In addition, the personalized sort model reorder the initial search results using the algorithm proposed in this paper, it is the key to achieve personalized search for the system.

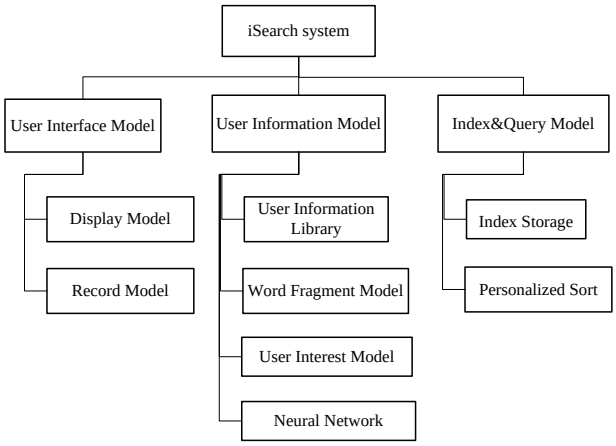

Fig. 6 System Functional Structure Diagram

V. CONCLUSIONS

In this paper, we proposed a user-oriented personalized search engine algorithm and realized a system based on it. The algorithm is based on BP neural network which uses its pattern recognition capability. For personalized service system, the user interest model is the indispensable part. So we give a novel user modeling method which contains user's interest comprehensively. And the user interest degree is treated as the measurable indicator to judge the importance of documents. The results of verification tests show that the algorithm has a good performance in reorder the documents according to user's interest and preference. At last, we described a system design of personalized search engine based on the algorithm.

However, the algorithm inevitably exposes some shortcomings of BP neural network, such as slow convergence and bad stability. In addition, we should find a better way to construct a representative user interest model. Therefore, we will pay more attention to solving these problems in our further work.

REFERENCES

[1] D.F. Liu, J.G. Duan. "The Design and Research of User Interest Model in Personalized Search Engine". Aisa-Pacific Conference on Information Processing, 2009.

[2] L. Page, S. Brin, R. Motwani, T. Winograd. "The PageRank citation ranking: Bringing order to the web". Technical report, Stanford University, Stanford, CA, 1998.

[3] G. Salton, A wong, CS Yang. "A vector space model for automatic indexing". Communications of the ACM, vol. 18,no. 11, pp. 613-620, 1975. http://dx.doi.org/10.1145/361219.361220

[4] Salton G, Clement T Y. "On the construction of effective vocalularies for information retrieval". Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval, New York, ACM, pp. 11, 1973.

[5] Y. Seo, B Zhang. "Learning user's preferences by analyzing web-browsing behaviors". Artificial Intellegence, vol. 15, no. 6, pp. 381-387, 2001.

[6] Hechi Nielsen R. "Theory of the back propagation neural network". Int. J. Conf. On Neural Network, no. 1, pp. 593-605.

[7] B. Yang, X.H. Su, Y.D. Wang. "BP Neural Network OPtimization Based on an Improved Genetic Algorithm". Proeeedings of the First International Conference on Machine Learning and Cybemeties, pp. 64-68, 2002.

[8] JOONE. JOONE project. http://sourceforge.net/projects/joone.

[9] The Apache Software Foundation. "Apache Lucene –Overview". http://lucene.apache.org/. 2008-10.

[10] JE-analysis. http://www.jesoft.cn.

Huifang Deng, born in Hunan, China, in 1957. He got his BSc degree in physics from Hunan Normal University, Changsha, China in 1981, his MSc degree in theoretical physics at Wuhan University, China in 1984, and his PhD degree in computer modeling (computational physics) from University College London (UCL).

He stayed in the UK for more than 15 years and worked in succession in Manchester University, Bristol University, Liverpool University, London University Queen Marry College, as well University College London before he came to China to take the position of the Dean of Software School at South China University of Technology (SCUT) in 2004. From 2001-2004, he served Sunrise Systems Limited at Cambridge as a Chief Scientist and Chief Technical Officer. He is now working at SCUT as a full Professor and undertaking researches in Internet of things, cloud computing as well as big data.

So far he has got over 110 papers published and chaired state-level and provincial-level key research projects on RFID technology and applications in logistics and automatic customs clearance. In addition, he holds several patents of invention.

Qinwei Wang, born in Hunan, China in 1988. He got his Bachelor degree in mathematics and computer science from Hunan Normal University, Changsha, China in 2010, and his Master degree in computer technology at South China University of Technology, Guangzhou, China in 2013.

He is now serving No. 722 Institute of CSIC, Hubei, China, as an Assistant Engineer and engaged with the computer networks and communication systems