Research Journal of Applied Sciences, Engineering and Technology 5(4): 1257-1262, 2013
DOI:10.19026/rjaset.5.4859
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 28, 2012 Accepted: August 08, 2012 Published: February 01, 2013

Research Article

Design and Implementation of the Personalized Search Engine Based on the Improved Behavior of User Browsing

[1]Wei-Chao Li and [2]Jin-Guang Liu
[1]Computer Center, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, 450015, China
[2]Department of Institute of Information Engineering, Huanghuai University, Zhumadian, 463000, China

Abstract: An improved user profile based on the user browsing behavior is proposed in this study. In the user profile, the user browsing web pages behaviors, the level of interest to keywords, the user's short-term interest and long-term interest are overall taken into account. The improved user profile based on the user browsing behavior is embedded in the personalized search engine system. The basic framework and the basic functional modules of the system are described detailed in this study. A demonstration system of IUBPSES is developed in the .NET platform. The results of the simulation experiments indicate that the retrieval effects which use the IUBPSES based on the improved user profile for information search surpass the current mainstream search engines. The direction of improvement and further research is proposed in the finally.

Keywords: Behavior of user browsing, personalized service, search engine, user profile

INTRODUCTION

The user profile is the feature set which is used to store the user's interest preferences, to store and manage user’s behavior history, to store and learn the knowledge of user’s behavior and the knowledge of relevant derivation (Martin-Bautista *et al.*, 2003). The user profile is the starting point of realizing personalized service to the search engine and also is the foundation and core of the personalized service to the search engine. The quality of the user profile is directly related to the quality of personalized service, the combination of the use profile and the user's retrieval demand can be more approximation to the user’s real information needs. We can improve the precision of search engine through filtering and screening to the retrieval results.

The user profile based on the user browsing behavior mainly considers the various operations of the user on the web pages, such as the saving of the page, printing the page, favorite the pages, the time of browsing the web pages and so on. All of these browsing behaviors to the web pages reflect the user to the level of interest on the web pages. As the user profile based on the user browsing behavior overall takes into account the user to the level of interest to the web pages and key words in the weight of each web page, fully reflects the user's individualized characteristic, it can meet the user’s actual needs (Li and Fu, 2011).

IMPROVED USER PROFILE BASED ON THE USER BROWSING BEHAVIOR

The user profile based on the user browsing behavior overall takes into account the user to the level of interest to the web pages and key words in the weight of each web page; however, any user’s interest preference is not unchangeable. In different periods of time, the user’s point of interest is different. Therefore, by tracking the operating history of the user’s access to the web pages, mining the visited web pages, which can be implicitly getting the feedback information. And using the feedback information to update the user profile, thus we can realize the personalized service to the search engine. In this study, overall taking the level of user’s interest to the web pages and the level of key words into account and considering the user’s short-term interest and long-term interest, we establishes the user profile.

In general, if someone accesses a web page, he is interested in the web page. But there is such the situation, the user may be interested in certain areas in the short-term while neglecting the areas of long-term interest, which is bound to affect the user’s retrieval quality and efficiency within a period of time (Shan, 2010; Li *et al.*, 2008). So, the user’s short-term interest of the user and long-term interest must be taken into account in the user profile. We take the user's interest preference of that very day temporarily as the short-term interest.

Corresponding Author: Wei-Chao Li, Computer Center, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou, 450015, China

The user's short-term interest: Assume that the user visits L web pages altogether that very day, the vector today can state the user interest preferences that very day. For the user's short-term interest, the user browsing behavior and the level of interest to the key words also should be taken into account. Therefore, the user profile of user's short-term interest can be expressed as follows:

$$P^{today} = \{P_{11}^{today}, P_{12}^{today}, P_{21}^{today}, ..., P_{ij}^{today}\} \quad (1)$$

where, the degree to the keyword j of the user's that very day can be expressed as:

$$P_{ij}^{today} = \frac{1}{L}\sum_{i=1}^{L} \omega_{ij} \quad (2)$$

The user's long-term interest: As for the user's long-term interest, we can construct the user profile using the user's visited history to the web pages in N days. Here we introduce the concept of window scale, we define L_t as the number of web pages viewed by the user at the t day, t = 0 states that very day, So we can construct the long-term interest P^{per} through setting the size of window scale N(N = 1, 2,… 30). We can use vectors state the long-term interest as follows at the same construction method of the short-term interest P^{today}.

$$P^{per} = \{P_{11}^{per}, P_{12}^{per}, P_{21}^{per}, ...P_{ij}^{per}\} \quad (3)$$

where,

$$P_{ij}^{per} = \frac{1}{L_t}\sum_{i=1}^{L_t} \omega_{ij} \times e^{-\frac{\log 2}{h}(t-t_j)} \quad (4)$$

L_t states the total number of web pages which the user viewed at the t day in formula (4), where t<N. Here we simulate the human brain by introducing a forgetting factor. The weights of all key words in the user's long-term interest model decline as time elapses away. E -log^{2}/h (t - t_j) States the time reliability of the user profile in the formula (4), where h is the life cycle parameter (That is the cycle of decay) (Shan, 2010). As human's memory begins to decay after exposure to the new knowledge one week later, the life cycle parameter value is generally equal to 7, where t_j states that day which the key word T^{per}_{ij} appears on the web page i. t-t_j states the number of days which the key word T^{per}_{ij} is stored in the knowledge base if t states that very day. That is to say, t-t_j states the number of days which the user is interested in.

The finally user interest model: As time goes on, the user uses the personalized search engine system for information retrieval constantly and the quantity of the user's visited web pages also is increasing constantly. Now the user's long-term interest is need to update dynamically. Comprehensive consideration to the user's short-term interest and long-term interest, the finally user interest model is expressed as follows:

$$P_{ij} = c \times x \times P_{ij}^{today} + y \times P_{ij}^{per} \quad (5)$$

where, {x, y| x>0, y>0, x + y = 1}, x and y indicate the influence weight individually which the user's short-term interest and the user's long-term interest to the finally user interest model. In general, the value of x and y is given according to the experience, It is artificially given. In this study, through the observation of a large number of retrieval results and the comparison to the value x and y, we finally determine that x = 0.6 and y = 0.4 and c is a constant, the c is defined as follows:

$$c = \begin{cases} 1 & dt \geq Thresh \\ 0 & dt < Thresh \end{cases} \quad (6)$$

dt states the average time which the user spends on the each key word browsing the web pages. Thresh is the threshold value, In the view of a statistical point, Threshold is usually the sum of the mean and standard variance, That is Thresh = μ + σ. We can determine the relationship between the user's short-term interest and the long-term interest through the threshold.

PERSONALIZED SEARCH ENGINE BASED ON THE IMPROVED USER PROFILE

The user profile and the display of personalized retrieval results are the key technologies of the personalized search engine. The user profile gets the user's interest preference continuously through the user interaction process from the user interface and then constructs the user profile, updates optimization constantly. Transferring the information which the user is concern about to the search engine system, the search engine system can retrieve the information which the user is interest in. At the same time, comparing the retrieved information with the user profile, the search engine system gets the ranking of the web page which the user is interest in. And then displays the retrieved results according to the sort of the user's interest degree

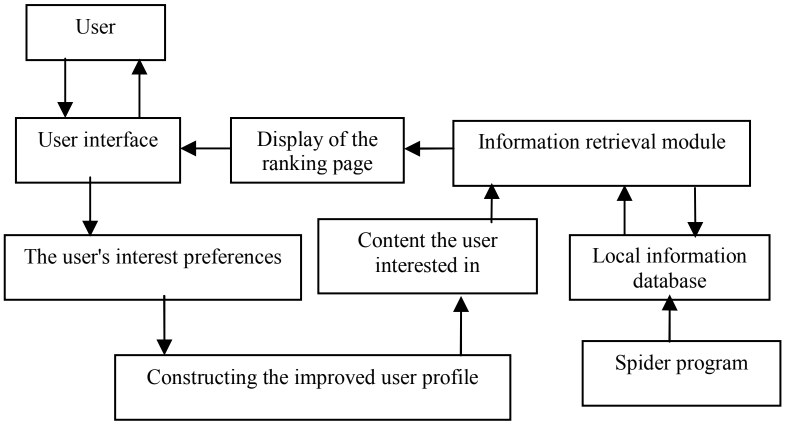
User
User interface
Display of the ranking page
Information retrieval module
The user's interest preferences
Content the user interested in
Local information database
Constructing the improved user profile
Spider program

Fig. 1: The working principle of the personalized search engine

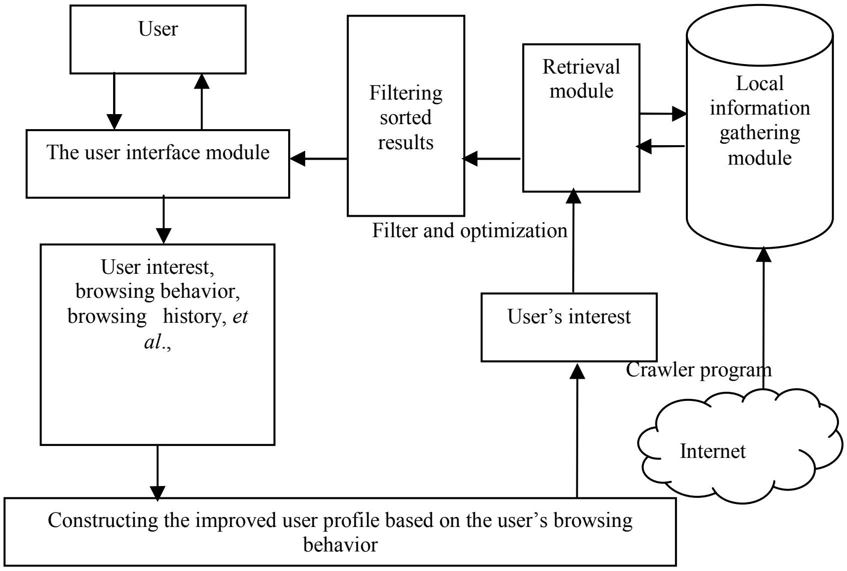
User
Filtering sorted results
Retrieval module
Local information gathering module
The user interface module
Filter and optimization
User interest, browsing behavior, browsing history, et al.,
User's interest
Crawler program
Internet
Constructing the improved user profile based on the user's browsing behavior

Fig. 2: The framework of the personalized search engine system based on the user profile

to the web pages. The working principle of the personalized search engine is shown in Fig. 1 (Zhou and Wang, 2006).

The basic framework: The framework of the personalized search engine system based on the user profile is shown in Fig. 2. The workflow of the system is as follows, when the user retrieves information, he will access to the user interface module at first. He will enter into the retrieval module, when he input the search keywords. The personalized user profile gathers the user's personalized information from the user interface module. In the retrieval module, the user can directly input the keywords to retrieve information and the retrieval module will record the user's search history at the same time. The personalized user profile extracts the user's search history from the retrieval module. Then, it processes the user's personalized information through sorting the retrieval results in order of relevance according to the user's interest preference and then feedbacks to the retrieval module. The retrieval module searches out the web pages which the user is interested in and at the same time filters out the web pages which the user is interested in and is not relevant to the user's behavior. As to serve the user's retrieval behavior better, the retrieval module sorts the retrieval results in order of relevance. Finally, the personalized system submits the eventually disposal retrieval results to the user.

From Fig. 2 we can see that the information of the user profile is from 2 aspects, the user's browsing behavior and search history, which are used to construct the user's short-term interest and long-term interest. By establishing the dynamic user profile, the information which the user is interested in can be searched out quickly and at the same time, the useless information to the user is filtered out. The retrieval module sorts the retrieval results in the order of relevance and feedbacks the most appropriate information to the user, so as to realize the intelligent and personalized information retrieval.

The basic functional modules:

The user interface module: The user interface module is the interface which provides the registration or login interface for user. In the user interface module, the user may input some personal information, such as nickname, gender, age, education, specialty, the research direction, hobbies and so on. Such information is called the user's explicit personalized information, which can compose the default information of the user profile. That is the initial user interest vector. The default interest vector of unregistered user may think as empty. The user's browsing behavior and search history in a period of time can help construct the user profile.

The retrieval module: The user can directly input the query keywords to retrieval information in the retrieval module and at the same time the user interface module will record the user's search history, such as keywords, visited URLs, the user's operation (collection, copy, save, etc.) and so on. The retrieval module searches the user's retrieval keywords from the local information database first. If there is no relevant information in the local information database, the crawler program is called for gathering information from Internet. Then, the retrieval module will dispose the collected web pages and adds the disposed web pages to the local information database.

In addition, the retrieval module filters the retrieval results according to the user profile. The web pages which the user is uninterested in are not shown to the user. For the web pages which the user is interested in, calculate the relevance of the web pages and the user profile. Finally, the retrieval module sorts the web pages in the order of relevance and shows them to the user.

The user profile module: The user profile module includes three aspects, collecting and recording the needs of the user's interest, constructing the personalized user profile, filtering and sorting the retrieval results and getting feedback on the results to the user. The user profile is mainly used to collect, record, manage the user's interest preferences. And it is also used to describe the user's potential interest demand. In this study, the user profile module is better to facilitate the user's personalized retrieval through disposing the personalized information such as the user's browsing behavior, the user's interest preference to the keywords, the user's short-term interest, the user's long-term interest, feedback on the disposed results to the retrieval module and so on.

The local information gathering module: The establishment of the local information database is of great practical significance for increasing the retrieval speed of personalized search engine. When the user use the traditional search engine for retrieval information, after the user submits the keywords, in general, the crawler program of the search engine collects the web information on the Internet immediately. Or the meta-search engine calls a number of special search engines at the same time collecting the web information on the Internet. After searched the retrieval records for the conditions, the retrieval module disposes the results and then gets feedback on the disposed results to the user.

Established the local information database, when the user uses the personalized information retrieval system, the retrieval module searches the information from the local information database first of all, so that the query scope is greatly reduced and the retrieval speed has been greatly increased in the local area network. When there is no keywords information in the local information database, the crawler program is called immediately for collecting information on the Internet. And put the retrieval results in the local information database. Then, the retrieval module disposes the results and gets feedback on the disposed results to the user. That is to say, keep the local information database update in real-time. If there is no retrieval keywords information in the local information database, update it in real-time, to ensure the integrity of the local information database. In order to ensure that the user searches the stylish information, there is a regularly update on the corresponding information of existed keywords library in the user profile. In general, we set the frequency of update for 7 days (current news excluded).

Table 1: The comparative table to keyword precision

Key words	Google	Reference Zhou and Wang (2006)	Reference Liu (2005)	IUBPSES
User profile	0.200	0.467	0.533	0.587
Word segmentation	0.467	0.800	0.800	0.830
Human-computer interaction	0.467	0.600	0.667	0.738
Information retrieval	0.267	0.533	0.600	0.566
Feature extraction	0.467	0.733	0.733	0.718
Text classification	0.600	0.667	0.733	0.812
Information retrieval	0.267	0.600	0.800	0.880
Search engine	0.667	0.800	1.000	0.858
Ontology	0.333	0.533	0.600	0.788

In order to simplify the complexity of the algorithm, the crawler program directly calls to the Google's crawler program in this study. That is to say, the information in the local information database is a direct call to the Google's search engine's retrieval results. The personalization of the user's retrieval information is a personalized service utilizing the user profile and the mining algorithms of the user's interest based on the retrieval results of Google search engine.

EXPERIMENT OF PERSONALIZED SEARCH ENGINE

To illustrate the effectiveness of the improved user profile in this study, an Improved Personalized Search Engine System based on the Users' Browsing behavior (IUBPSES) is developed on the Microsoft .NET platform at May, 2012. The improved user profile is integrated into the IUBPSES, which the retrieval module directly calls the Google's retrieval module.

At the same time, we also put the IUBPSES system based on the improved user's browsing behavior have a contrast through simulation experiments to Google search engine, the intelligently adjust algorithm based on vector space model reference to Zhou and Wang (2006), the adjust algorithm of the interest mode reference to Liu (2005). We have simulation experiments to the 8 retrieve keywords respectively in the above four model algorithm. For each keyword, take the top 20 retrieve results respectively in the above four model algorithm and calculate the corresponding keyword precision respectively. The comparative table to keyword precision is shown in Table 1. The Fig. 3 is the comparative figure of keyword precision.

From the above results of simulation experiments, we can observe that the retrieval efficiency which uses the improved user profile based on the user's browsing behavior is obviously superior to the general search engines. As using the IUBPSES system usually, the user's interest preference is gradually learned, the superiority of the IUBPSES system is obviously more apparent. From Table 1 we know that the retrieval efficiency which uses the improved user profile based on the user's browsing behavior is higher than without using the IUBPSES system in case of existing the interest preference. And as the user uses the IUBPSES system for a long time, the user's interest preference is stronger, the superiority of the IUBPSES system is more obvious significant.

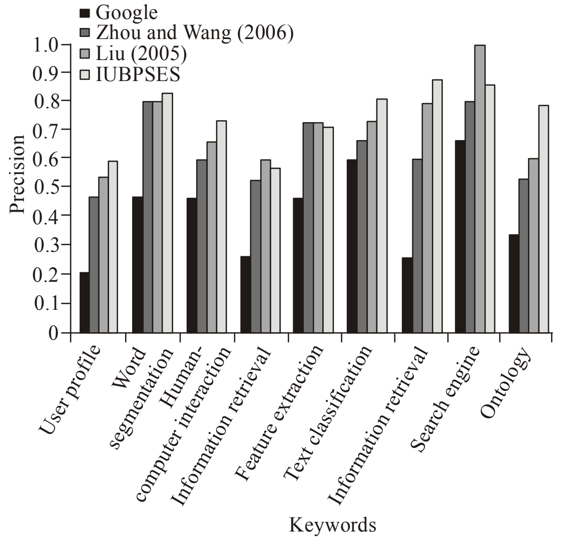

Fig. 3: The comparative figure of keyword precision

CONCLUSION

The improved user profile based on the user's browsing behavior proposed in the study can accurately describe the user's interest preference, the simulation experiments of the IUBPSES demonstration system developed on the Microsoft .NET platform, also show that the retrieval efficiency which uses the improved user profile based on the user's browsing behavior is obviously superior to the general search engines. As using the IUBPSES system usually, the user's interest preference is gradually learned, the superiority of the IUBPSES system is obviously more apparent. In case of existing the interest preference, the user uses the IUBPSES system for a long time, the user's interest preference is stronger, the superiority of the IUBPSES system is more obvious significant.

However, there is a lot of improvement and further research in this study:

- We could use the more efficient mining algorithms to the user interest on the basis of the improved user profile, to be faster and more accurate in mining the user's interest, so that establish the user profile which meets the user's interest preference and user's features more than before.
- In the aspect of the personalized display to the retrieval results, we could introduce more parameters comprehensive considering the order to the retrieval results, except only considering the similarity of the web pages and the user profile, so as the retrieval results meet the user's needs more.
- In the study, there is no clustering and classification to the user, clustering and classification to the user will divide the users whose needs are the same or similar into the same category. For the retrieval of the same or similar keywords, other users' retrieve results may recommend to the user. So as to help the users find their own interested information more conducive and this is also conducive to the improvement of recall ratio and precision.

ACKNOWLEDGMENT

This study is supported by the foundation and frontier technology research project in Henna province, China (No. 112300410304).

REFERENCES

Li, F., J. Pei and Z. You, 2008. Adaptive user interest model based on the implicit feedback. Comp. Eng. Appl., 44(9): 76-79.

Li, W. and Y. Fu, 2011. Improved user profile based on the user browsing behavior. Telecommun. Sci., No. 5: 77-81.

Liu, J., 2005. The studying of agent based user interest model. J. Syst. Software, 3: 7-49.

Martin-Bautista, M.J., D.H. Kraft and M.A. Vila, 2003. User profiles and fuzzy logic for web retrieval issues. Soft Comp., No. 6: 365-372.

Shan, R., 2010. New user's interest model updated based on browsing behaviors. Elec. Design Eng., 18(4): 61-62.

Zhou, X. and S. Wang, 2006. Construction and update to the user profile in Web text mining. J. Xiangtan Normal University Nat. Sci. Edn., 28(3): 33-36.