RESEARCH ON PERSONALIZED INFORMATION FILTERING OF SEARCH ENGINE

Shu Zhang [1] , Xinrong Chen [1,*] , Changshou Luo [2,*]

[1] *Department of Computer Science, College of Information and Electrical Engineering, China Agricultural University, Beijing, P. R. China 100083，China*

[2] *Beijing Academy of Agriculture and Forestry Sciences, Beijing 100089, China*

[*] *Corresponding author, Address: Xinrong Cheng, Department of Department of Computer Science, College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, P.R.China, Tel: +86-13521369058, Email: hh0188@sina.com*

[*] *Corresponding author, Address: Changshou Luo, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100089, P.R. China,Tel:+86-13683248103, Email: luochangshou@163.com*

Abstract: Since network has been created and developing rapidly in recent years, the age of information exploding is coming. The Search Engine becomes more and more important for people, but the traditional search engine retrieves and provides information just according to the keywords that users input. How to recommend the right information to users has become the hot point. The technology of personalized information filtering brings people hope .The paper I present analyzed the achievements of those filtering technologies ,and adopted user-system complex-operating modeling to build User-activity-collecting module, User-interest-updating module and User-searching module, in order to meet theme-oriented searching's needs. Experiments showed that the user-interest model can provide personalized service and enhance search engine's precision.

Keywords: information filtering, user interest-model, personalized searching, vector space model

1. INTRODUCTION

With the rapid development of Internet information, billions of webpages have been piling up, it makes the most valuable information searching more and more difficult .Search engine (Li Xiaoming et al., 2004) becomes an important retrieval tool for people. Traditional search engines can retrieve and provide information according to the keywords users input, then the system will recommend the same results for the same keywords input by different users. In fact, it depends on what users need. Subject search engine (Liu Huilin et al., 2007) and personalized service (Zeng Chun et al., 2002) emerge as the times require. Subject search engine can filter the unrelated field information during the crawling to reduce the searching resources. The search engine based on technology of personalized service will provide users information which they require, and prevent users from confusing in the information sea.

2. INFORMATION FILTERING AND PERSONALIZATION

Information filtering (Liu Baisong et al., 2003) (ab. IF), contains two significations: the first one which is applicable to rubbish E-mail filtering is how to delete the unrelated from the massive inordinate information; the other used in information recommending in favor of user interests. is to extract the related information which is demanded by users from dynamic information streams. Compared with huge dynamic information, user interests will not change in general. Information filtering on search engine means it can recommend users how to get the required information from information repository.

Search engine's personalized service will select the different information for different users from huge information repository according to their different interesting. Technology of information filtering is the implemental technology of search engine personalization, which can recognize the users' need from dynamic web resources.

Information filtering (IF) technologies (He Jun et al., 2001) can be classified into three, such as:

1) IF technology based on rules, by means of designed rules library to filter information.

2) IF technology based on collaboration, using the comparability of the different users to filter information.

3) IF technology based on content, taking advantage of the comparability between resource and user interest to filter information.

In terms of the research on three personalization technologies' application

in theme-oriented search engine, we find that: 1) based on rules. It is difficult to design the rules. This technology has high uncertainty, which should not be adopted. 2) based on collaboration. The key point of this technology is how to cluster the users, but theme-oriented search engine has its own field-orientability. That is said, theme-oriented search engine has locked the users from the certain field. It is unnecessary to cluster users, which is time-consuming and will gain little significant effect. 3) based on content. This technology will lead us to the result information based on the user-interest model, which makes the tiny change on the results which can be gotten from general search engines.

It can also provide the users more related information, and be applicable to the personalized service of theme-oriented search engines. This research is based on theme-oriented search engines which are applied to a certain field and have certain users, equaling to achieving user clustering in general search engines. Therefore, this research adopts the information filtering technology based on content to implement the personalization of theme-oriented search engines.

Information filtering system contains 4 foundation models (Pang Yali et al., 2007):

1) Web resource analyzing model, analyzing and describing the webpages crawled by spider.

2) User-interest model, obtaining and describing user information by obvious or hidden meanings.

3) User-interest updating model, tracking and analyzing the users' behaviors to get the users' current interest.

4) Filtering model, matching the web information description with user interest information in terms of given rules, the filtering model will afford users the web information required in descending order.

From above 4 foundation models, we can see how to obtain the user interests and create user interest models, which will influence the effect of information filtering. The creating of user interest model is one of the key technologies to implement personalized information filtering. Furthermore, filtering model's design also is the key point during the creating process of search engine personalization.

3. USER-INTEREST MODEL

3.1 User-interest Description

User interest description is the key technology in information filtering, which is related to the filtering effect directly. At present, there are 3

descriptions (Zhang Meixiang et al., 2005):

1) Keyword Description. According to certain rules, each interest's weight has not only been valued, but the user interest vector has been set up, so that each interest keyword can possess the branch vector and each keyword's weight can get the value of the branch vector.

2) Fixed Document Set Description (ab. FDS). This description selects the most representative FDS which can reflect all kinds of user information of a certain field sufficiently. The FDS description is used to solve some problems that are hard to be described with exact keywords.

3) Paragraph Description. In lengthy web text information, what an user is interested maybe just has several paragraphs. Paragraph is the minimum unit of articles, so the meaning of paragraph description analyzes is to find out some which the users are interesting in.

The theme-oriented search engine in this paper is the search engine facing the certain field. It possesses its own thematic words library, which can express the user interest exactly. It is hard to use FDS and Paragraph description which are applicable to general search engine. Therefore, the study adopts Keyword Description to express user interest.

The Keyword Description is implemented in Vector Space Model (Zeng Chun et al., 2002).

Given a user-interest vector I, I=（I_0, I_1, I_2 …I_i, I_{i+1} …I_n）, i=0, 1, 2…n， Where, I_i is the branch vectors of user-interest，which are interest keyword； the value of I_i is the weight W_i of each keyword .

3.2 Selection of User-interest Modeling

At present, in the field of personalized information service, there are 3 main methods (Ji Meijun et al., 2006) to build user-interest models.

1) Manual customization modeling. It is a modeling method through users' self-input or selection. But it counts on user interests totally and can not track the changes of an user's interests timely;

2) Demonstration modeling. It is up to users to provide the demonstrations of relevant interests. However, the method requires users to mark webpages in order to obtain corresponding demonstrations. Therefore, users' normal browsing behaviors will be disturbed;

3) Automatic modeling. It is constructed automatically according to users' browsing behavior, which would not interfere with users.

Considering the disadvantage of the front two modeling methods above, the study adopts the complex user-interest modeling methods, as illustrated in Fig.1, which is the cooperation of users' inputting and collecting of users' behaviors by the Log-collector. Adopting complex modeling method can avoid two things, which are negative information out of date and dynamic IP

false work in manual customization modeling of users.

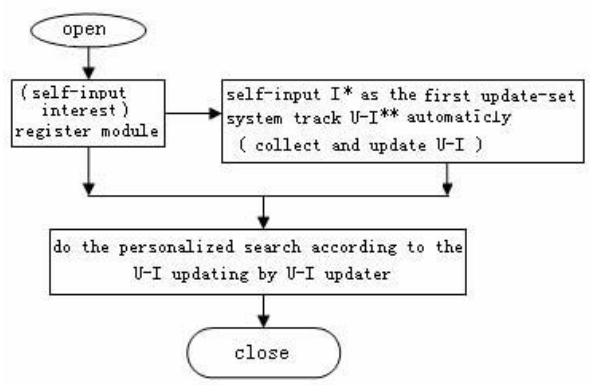


Fig.1: Illustration of complex U-I modeling
* I is the abbreviation of interest; ** U-I is the abbreviation of user-interest

4. INFORMATION FILTERING

During the implementation process of IF, the system adopts user-behavior collector to obtain user behaviors, then update user-interests and analyzes them, finally uses search-recommender to achieve information's filter and recommendation, as illustrated in Fig.2.

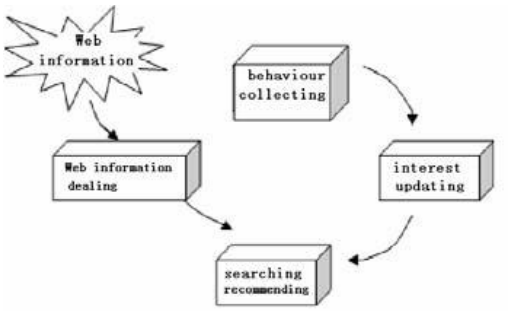


Fig.2: Framework of personalized information filtering

4.1 User-behavior Collecting Module

Users' historic searching behavior can reflect the trend of users' interests, which is collected by the user-behavior collector. The behavior contains search time, search content, search number and so on.

4.2 User-interest Updating Module

4.2.1 Module Function

User-interest Updating Module is the important part of the personalized search system, which is consisted of log obtaining part, log splitting part, log stating part and weight computing part, as illustrated in Fig.3. When users have no search behavior, the module adopts users' registered interests as first updating set to avoid updating error. When users have already some search behaviors, the module will collect user logs and extract the logs considering the factor of logs' preserved time. Finally, a vector of five branch vectors is figured out.

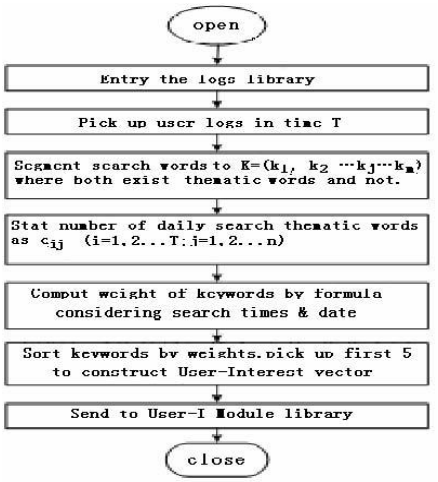


Fig.3: Flow chart of user-interest updating module

4.2.2 Weight Computing of User Interest

Weight computing is the primary part of user-interest updating. During the weight computing, the search number of keywords determines how much the user is interested in . So we regard it as the basic factor of computing weight. Otherwise, we can also consider the search date as the important factor to hope this module will do its best to obtain users' closest interests.

User interests may be changed, so the date as the keyword is searched determines whether the user's interest about this keyword is reduced or enhanced. What would be happened, when some keywords were paid more attention previously but little attention now. The time function can solve this problem. It is a degressive function, which would depress the former keywords' weights and increase the more lately searched keywords' weights.

Below is the formula to compute weight by this system.

$$W = \sum f_i t_i ,$$

Where, i =1,2,3….T (T is the period of user interest updating);

f_i is the number of the keyword's searched times at day i in period T;

$$t_i = 1 - (T+1-i)/T$$

this function is time-factor function, which is aiming at decreasing the former keywords' weights.

4.2.3 About Keywords long time unvisited

The capacity of hardware is so limited that it can not preserve all the users' search logs and information processed by log processor endlessly. Thus, we regard the keywords long time unvisited as the users' uninterested and discard them.

The formula below is to process the keywords,

$$C_{i,j} = \begin{cases} C_{i,j}, & D_{now} - D_{visited} \le 5T \\ 0, & else \end{cases}$$

Where, T is the period of User-Interest Updating;

$C_{i,j}$ is the search frequency of keyword. When becomes zero, it will be deleted from the log library;

D_{now} is the current date of system;

$D_{visited}$ is the final visited date of $C_{i,j}$;

D_{now} -$D_{visited}$ is the days between current and final search of $C_{i,j}$.

If a keyword can not be searched by user more than 5 times, its search frequency will become zero and it would be canceled from user logs library.

4.3 Implementation of Information filtering

The implementation of information filtering includes four steps. Firstly, the search module with User-interest Model can get the user interest vector from User-interest Updating Module; secondly, extract the keywords and their weights from webpages model; then compute the relevance degree between user interest vector and webpage model vector; the last step is to sort the webpages obtained by general keyword searching according to interest-webpage relevance degree. The flow chart is illustrated below, Fig.4.

During the computing of relevance degrees, the study adopts typical Vector Space Model relevance degree method (Zeng Chun et al., 2002),

$$r = \cos < \alpha, \beta >= (\alpha, \beta) / |\alpha||\beta| = \sum_{i=1}^{n} (F(i) x_i w_i^2) / \sum_{i=1}^{n} (F(i) x_i w_i)^2$$

Where, α equals the user interest vector I; β equals the vector F extracted by the webpage model;

the value of their branch vector is the weight of the branch one.

This module adjusts the webpages tinily according to the magnitude of the relevance degree, and then attains the search result.

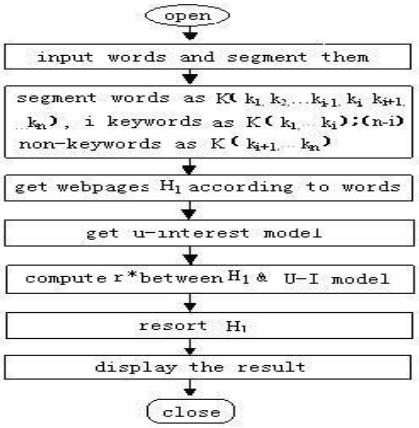


Fig.4. Implementation of information filtering

5. EXPERIMENTS AND ANALYSIS

The experimental system adopts Windows platform, PHP and Visual c++ programming language, and Apache server, as well as SQL Server database. In the experiments, the sample set came from the flower-relevant webpages crawled by thematic spider; the user behavior was preserved by logs; the user interest updating period was 5 days; the quantity of webpages was 8000.

5.1 Micro-comparison Experiment

According to above works, we designed 3 micro-comparison experiments:

1) Same user at two states, both personalized search and general search, inputting the same keyword.

2) Same user at two states, both personalized search and general search, inputting non-flower keyword, which is experimental by non-keyword "technology".

3) Different user at the state of personalized search and by inputting the same keywords.

From above experiments, it can be seen that:

1) The results show that:

Non-personalized search returns the same results for different people;

Search with registered information, which is static personalized search, turned out that the results were not changed as the user interest changed. The logs show that the user was interested in "yulan" then.

Search with user-interest updater, that is dynamic personalized search, can return the result which could mostly reflect the information needed by users.

2) The results gained by inputting general word “technology” in two states show that:

Search without the personalized technology obtains mess result;

Search with the personalized technology acquires user-interest relevant information.

3) The results we gain by inputting the same keyword from different user shows that:

The system leads to different results by inputting the same keyword for different users according to each user’s different interest updated by themselves.

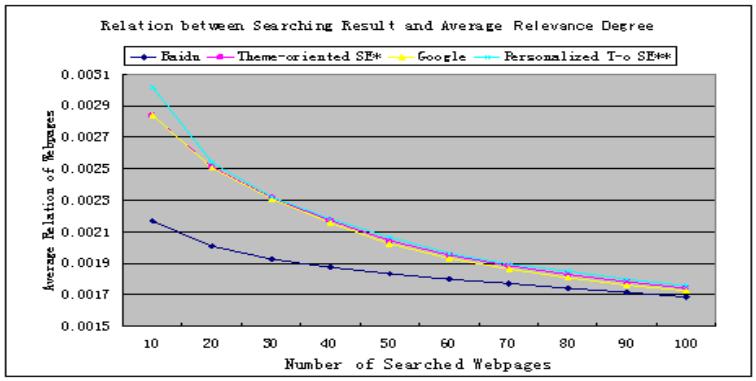
All above analyses prove that the system could be used to implement personalization.

5.2 Macro-comparison Experiment

This experiment involved user Xia Yan, who has used the system for 7 days. His search behavior was analyzed by user-interest updating module to: Plum blossom,1.1666669999999999;Peony,1.;Clove,1.; Rose,0.833332999 99999999; Jasmine,0.33333299999999999 (sorted by each flower’s weight and the data was the lately updated). Keywords were plum, blossom and peony.

The comparison experiment was performed among Baidu, Google, Thematic search engine and Personalized thematic search engine. We searched keywords relevant website both in baidu and google, then got 31 seed websites by picking out the same, finally crawled 61358 webpages as the experimental webpage sample.

Meanwhile, the keywords “peony” and ”plum blossom” were searched at above 4 systems. We got the relation between webpages quantity and average relevance degree, as illustrated in Fig.5.



* SE is the abbreviation of search engine.

** T-o SE is the abbreviation of theme-oriented search engine.

Fig.5: Relation between Webpages Quantity and Average Relevance Degree

6. CONCLUSION

The study took advantage of manual customization modeling and automatic modeling to construct a complex user-interest model, and proposed the algorithm of building and updating user personal interests based on the model. Experimental results prove that the model and algorithm can enhance the searching precision. The development of search engine can easily integrate the user profile, semantic and syntax technology to service the users. The next step is to join ontology technology effectively in the preprocessing and service part of the search engine, consequently achieve the true semantic search.

ACKNOWLEDGEMENTS

Fund for this research was provided by Beijing Natural Science Foundation Committee (P. R. China). The first author very appreciates what the College of Information and Electrical Engineering has contributed at the moment she pursues a master degree at the China Agricultural University.

REFERENCES

He Jun, Zhou Mingtian. Information Filtration Technology in Information Network. Journal of Systems Engineering and Electronic Technology, 2001, 23(11): 76-79.

Ji Meijun. Research on Related Question of Personalized User Modelling. Journal of Information, 2006, 25(3): 77-79.

Li Xiaoming, Yan Hongfei, Wang Jimin. Search Engine— Theory, Technology and System, Science and Technology publishing house, 2004, 29-54.

Liu Baisong. Research on Information Filtration. Journal of Modern Books and Information Technology, 2003(6): 23-26.

Liu Huilin, Guo Laigang, Liu Lanzhe, Wang Xingguang. Design and Implementation of Chinese Agricultural Subject Search Engine. Zhengzhou college newspaper (Neo-Confucianism version), 2007, 39(2): 74-77.

Pang Yali, Wang Caifen. Personalized Information Filtration Technology. Journal of Gansu Science and Technology, 2007, 23(3): 124-126, 171.

Zeng Chun, Xing Chunxiao, Zhou Lizhu. Summary on Personalized Service Technology . Journal of Software, 2002, 13(10): 1952-1961.

Zhang Meixiang, Chen JunjieZhao Shuanzhu. User Interest Model Expression of Information Filtration. Journal of Computer Development and Application, 2005, 18(5): 2-3, 14.