WUJNS
Wuhan University Journal of Natural Sciences

Vol.12 No.5 2007 893-896

Article ID: 1007-1202(2007)05-0893-04
DOI 10.1007/s11859-007-0002-3

Research on the User Interest Modeling of Personalized Search Engine

□ LI Zhengwei, XIA Shixiong†, NIU Qiang, XIA Zhanguo

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, Jiangsu, China

Abstract: At present, how to enable Search Engine to construct user personal interest model initially, master user's personalized information timely and provide personalized services accurately have become the hotspot in the research of Search Engine area. Aiming at the problems of user model's construction and combining techniques of manual customization modeling and automatic analytical modeling, a User Interest Model (UIM) is proposed in the paper. On the basis of it, the corresponding establishment and update algorithms of User Interest Profile (UIP) are presented subsequently. Simulation tests proved that the UIM proposed and corresponding algorithms could enhance the retrieval precision effectively and have superior adaptability.

Key words: personalization; Search Engine; User Interest Model; intellectual agent
CLC number: TP 391.3

Received date: 2007-01-06
Foundation item: Supported by the National Natural Science Foundation of China (50674086), the Doctoral Foundation of Ministry of Education of China (20060290508) and the Youth Scientific Research Foundation of CUMT (0D060125)

Biography: LI Zhengwei(1977-), male, Ph.D. candidate, research direction: personalized service and information retrieval. E-mail: zwli@cumt.edu.cn
† To whom correspondence should be addressed. E-mail: xsx@cumt.edu.cn

0 Introduction

The explosion of Internet results in searching information as if finding a needle in sea and people have difficult to get the information fitting for their own intention rapidly. How to enable Search Engine to construct UIP initially, master user's personalized information timely and provide personalized services accurately have become the hotspot in the research of Search Engine area. Therefore, user Personal interest modeling has become the key point in the research of personalized information service.

At present, there are mainly three sorts of user interest modeling technology in the personalized information service. ① Manual customization modeling. It is a modeling method by user's self-input or selection, as in Refs.[1, 2], but it counts upon user totally and can not track user interest change timely. ② Demonstration modeling. It is up to user to provide the demonstrations of relevant interest. However, the method requires user to mark webpage in order to obtain corresponding demonstrations. Therefore, user's normal browsing behaviour is disturbed, as described by Ref.[3]. ③ Automatic modeling. It is constructed model automatically according to user's browsing behaviour, which belongs to the improved demonstration modeling technology and would not interfere with user, as described by Ref.[4]. Among Personalized search engines, user interest modeling is in the incipient stage and has not formed an integral technology system up to now[5].

Combining of manual custom modeling and automatic modeling techniques, a user Personal interest model is built in the paper. Based on it, the corresponding establishment and update algorithm of the UIP are presented accordingly. By mining user's browsing behaviour and abstracting user's interests, Simulation tests proved that such architecture and algorithms could recommend relevant information for user effectively and have superior adaptability.

1 UIM Adjustment

The UIM is a formalized representation of user's interest, which has specialized data structure[6]. In order to Personalize search result, the UIM needs to be constructed firstly.

Traditional search engine didn't make good use of user's interest or preference. So, a new type of user interest model is presented and a Sub-System of Personal Interest Search Intelligent Agent (SSPISIA) is constructed accordingly, which is employed to gather, mine and exert user personal interest information. SSPISIA is a Multi-Agent system, including user IA (Intellectual Agent), mining IA, resource IA and decision IA (Fig.1).

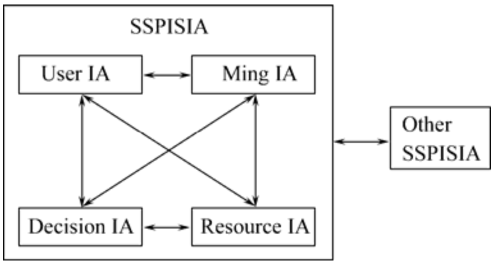

Fig.1 Organization of SSPISIA

Each IA in SSPISIA regards as a relevant independent intellectual unit and acts as distinct role. At the same time, they communicate and collaborate for each other and constitute an organic integrity. Aside from analogous query interface similar to ordinary search engine, the front end of the system also compromises several important components, such as query optimizer, dictionary, UIM and machine self-learning part, which constitutes the key points of personalized search engine system.

When user is employing search engine system, according to his own UIP, in the coordination of resource IA, user IA is responsible for revising query vector, sending it to retriever and returning the intermediate result to user IA. Then, user IA reassembles the intermediate result according to his UIP and recommends the information the user may be interested in. By means of this process, personalized search service is provided. Furthermore, the system revises his UIP in term of user's feedback and then provides higher quality service for next time.

Generally, according to user's registration, searcher engine system constructs user's UIP in the form of key words (such user ID, key word ID, key word, weight of keyword, creation time, last access time, and so on). During user's employing procedure, the system learns user's UIP automatically from his browsing behaviour and adjusts it dynamically, so it can provide better query quality subsequently. There are three kinds of adjustment for User interest model, which are as follows:

① If system learns new key words, it will calculate their corresponding weight and update user's UIP accordingly.

② If system learns old key words included user's UIP already, the only thing is to adjust its corresponding weight.

③ Because of limited UIP, each user has a maximum lexical capacity. When lexical quantity excesses its threshold, some low weight value lemmas will be deleted so as to make the lexical quantity maintain determinate extent. Of course, the capacity of UIP may change dynamically if there are enough space to be utilized.

2 UIM Adjustment Algorithms

2.1 Overview of UIM

In order to describe UIM, UIP is presented necessarily. At first, a hypothesis is given as follows:

Definition 1 UIP is defined as P_i, which is an aggregate that is composed of binary groups.

$$P_i = \{(k_{i,1}, \theta_{i,1}), (k_{i,2}, \theta_{i,2}), \cdots (k_{i,n}, \theta_{i,n})\} \qquad (1)$$

$k_{i,j} \in T, T = \{t_1, t_2, \cdots, t_m\}$, which is an aggregate of lexical items (dictionary). $\theta_{i,j} \in [0,1]$ is the weight of $k_{i,j}$ in the UIP[7].

As a rule, when a user opens a webpage and finds it does meet his need, he will close this webpage promptly. The duration time of it will not exceed 5 s generally. Furthermore, the weight of UIP reflects the relevant degree of user's interest. According to user's selection, search is performed from the root node to leaf node and corresponding UIP is built consequently.

2.2 Algorithm of User Interest Profile Building

Based on user's selection about interest tree, user's interest sub tree is obtained and then UIP, namely P_i, is formed accordingly.

① Set each leaf node's weight to $\mathrm{sub}\theta_1 = b + c$. In this expression, b is user's browsing time and its unit is minute, $0 \leqslant b \leqslant 30$. c is the count of user's press, $c \in [0,255]$. Furthermore, user may also set the weight of leaf node manually.

② Intermediate node's weight is the sum of its sub nodes' weight, namely $\mathrm{mid}\theta_j = \sum_{i=1}^{m} \mathrm{sub}\theta_i$, $\mathrm{mid}\theta_j \in [0,255]$, where m is the sub nodes count of the intermediate node.

③ Sort all nodes according to their weight descendingly and frontal n nodes is fetched, such as (k_1,v_1), (k_2,v_2) , $\cdots$, (k_n,v_n), where k_i is a lexical item and v_i is its corresponding weight.

④ Set $P_i = \{(k_{i,1},\theta_{i,1}),(k_{i,2},\theta_{i,2}),\cdots,(k_{i,n},\theta_{i,n})\}$, $(k_{i,j}, \theta_{i,j}) = (k_j, v_j / \sum_{k=1}^{n} v_k)$, $j \in [1,n]$.

2.3 Update Algorithm of UIP

Evidently, user's interests and preferences are not eternal and easy to change, so UIP should change itself according to the change of user's preference. Furthermore, the update of UIP is the key point to maintain the accuracy of UIM. As a rule, the retrieval information and browsing content are regard as the source to update UIP[8]. The update algorithm is described as follow:

Step 1 Check the query lexical item K inputted or captured and register its browsing time synchronously;

Step 2 If the lexical item K has existed in the set $\{k_{i,1},k_{i,2},\cdots,k_{i,n}\}$, then go to step 3, or else go to step 4;

Step 3 $(k_{i,p},\theta_{i,p} \pm r) \rightarrow (k_{i,p},\theta_{i,p})$, $\mathrm{mid}k_{i,p} := \mathrm{mid}k_{i,p} + c + b$. $r = \dfrac{c+b}{255n}$, b is browsing time span, $b \in [0,30]$. $c + b \in [0,255]$. Then go to step 6;

Step 4 $k_{i,q} \in \{k_{i,1},k_{i,2},\cdots,k_{i,n}\}$, $\omega_{i,q} = \min\{\theta_{i,j} \mid 1 \leqslant j \leqslant n\}$, compare the value of r and $\omega_{i,q}$, if $r > \omega_{i,q}$, the flow will jump to the step 5, or else go to step 6;

Step 5 $(K,r) \rightarrow (l_{i,s},\omega_{i,s})$;

Step 6 If there has any lexical item unchecked, the aforementioned procedure will repeated and go to step 1.

Step 7 Adjust $\theta_{i,j}$ by the formula $\theta_{i,j} := \theta_{i,j} / \sum_{t=1}^{n} \theta_{i,t}$.

After above process, the fresh UIP $\{(l_{i,1},\omega_{i,1}), (l_{i,2},\omega_{i,2}),\cdots,(l_{i,n},\omega_{i,n})\}$ is obtained, which will supervise the information search for next time.

2.4 Process of Keywords Long Time Unvisited in UIP

As note above, in the limit of UIP capacity, the weight of long time unvisited keyword $K_{i,j}$ will be recalculated by formula (2).

$$K_{i,j} = \begin{cases} K_{i,j} \ , & D_{\text{now}} - D_{\text{created}} \leqslant 15 \\ K_{i,j} \times \left(1 - \dfrac{D_{\text{now}} - D_{\text{visited}}}{D_{\text{now}} - D_{\text{created}}}\right), & \text{else} \end{cases} \tag{2}$$

D_{now} is the current system date; D_{visited} is the last accessed date of $K_{i,j}$; D_{created} is the original creation date of $K_{i,j}$; $D_{\text{now}} - D_{\text{created}}$ is the days between now and last access date of $K_{i,j}$.

Owing to the limit of lexical items in UIP, etyma abstraction is provided accordingly. As we know, verb in English has present tense, present perfect, past tense, pluperfect, and so on. For example, verb of do has the form of does, did, doing and done. Similarly, the noun, adverb and adjective also have several kinds of form. For instance, the noun wonder has the form of wonder, wonderful, wonderfully, etc. Obviously, each group lexical items above has same etyma, so they are identified as same lexical item. According to above-mentioned analysis, the system provides a primary lexicon, a thesaurus and a latent lexicon. When check the lexical item inputted, besides looking up the primary lexicon, the thesaurus and the latent lexicon are also examined synchronously[9].

3 Experiment Results and Analysis

According to above Algorithms, an experimental search engine prototype is built. Based on it, a simulation experiment is performed consequently. Firstly, key lexical items registered by user are added to his UIP, whose weight is set to 5 originally. For example, the lexical items selected by the user are computer network, Internet, software engineering, database, pattern recognition, news, harmonious society, tour, computer graphics, grid computing, data mining, and so on.

Comparison experiments are performed among the personalized search engine, Google and user automation algorithm based on TF-IDF[10]. We select 10 different kind search words (shown in Table 1) and carry through 15 times. Every time, the frontal 20 pages are fetched and the precision is calculated synchronously. The comparison results are shown in Table 1.

From the experimental results, it can be seen that the effect of foregoing algorithms excels Google in precision. In the case of user having preferences, its search effect is also better than search engine based on TF-IDF. Furthermore, with the increase of user's preferences degree, the difference between them becomes larger and

Table 1 Precision comparison among three systems

Sequence number	Lexical item	Precision of foregoing algorithms	Precision of based on TF-IDF	Precision of based on Google
1	Computer Net-work Internet	1.00	0.95	0.91
2	Computer Net-work	0.93	0.85	0.88
3	Internet	0.87	0.80	0.80
4	Software Engineering	0.59	0.51	0.46
5	Data Mining	0.81	0.68	0.72
6	Tour	0.72	0.51	0.67
7	Pattern Recognition	0.55	0.41	0.43
8	Harmonious Society	0.83	0.71	0.73
9	Grid Computing	0.58	0.49	0.46
10	Computer Graphics	0.66	0.49	0.52

larger. Namely, the stronger in user's preferences, the more obvious in superiority of search engine system adopted foregoing algorithm.

Moreover, the system permits user to maintain his or her profile. For instance, the user can update, add or delete his or her personalized information. The future information search service will be incorporated with other industry such as news, amusement, communication, and so on. The unique aim is to offer user a more comprehensive, initial and personalized information service space.

4 Conclusion

Aiming at the problems of UIM construction and combining of manual custom modeling and automatic analytical techniques, a UIM is provided in the paper. Based on it, the corresponding establishment and update algorithms of the user personal interest are presented accordingly. Simulation tests proved that such model and algorithms could enhance the precision of search results and have superior adaptability. The next step is to research the formation and application of user profile from the point of syntax and semantic.

References

[1] Venkat N Gudivada, Vijay V Raghavan. Information Retrieval on the World Wide Web [J]. *IEEE Internet Computing*, 1997, **1**(5): 58-68.

[2] Liebeman H. Letizia:An Agent that Assists Web Browsing [C/OL]//*Proceeding of the International Joint Conference on Artificial Intelligence*, Montreal, 1995: 924-929. *http://citeseer. ist.psu.edu/lieberman95letizia.html*.

[3] Chen L, Sycara K. WebMate: A Personal Agent for Browsing and Searching[C]//*Proceeding of the 2nd International Conference on Autonomous Agents and Multi Agents Systems*. New York:ACM Press, 1998:132-139.

[4] Zhao Zhongmeng, Yuan Wei, He Shili, *et al*. Research on the Intelligent Adjustive Algorithm for User Profile in Personalized Search Engine[J]. *Computer Engineering and Application*, 2005, (24): 184-187(Ch).

[5] Nahm U, Moony R. Text Mining with Information Extraction[C]//*Proceedings of the AAAI* 2002 *Sprint Symposium on Mining Answers from Texts and Knowledge Bases*. Standford: Springer, 2002:60-67.

[6] Xing Dongshan, Shen Junyi, Song Qinbao. Research on Mining Algorithm of User Browsing Preference Model[J]. *Xi'an Jiaotong University Journal*, 2002,(4): 369-372(Ch).

[7] Xu Baowen, Zhang Weifeng. *Technology of Search Engine and Information Retrieval*[M]. Beijing: Tsinghua University Press, 2003: 95-96(Ch).

[8] Xu Ke, Huang Guojing, Cui Zhiming, *et al*. Personalized Scheduling Algorithm Based on User Profile for Meta Search Engine[J].*Journal of Tsinghua University*(*Science and Technology*), 2005, **45**(s1): 1915-1919(Ch).

[9] Sun Tieli, Yang Fengqin. Construct and Maintenance User Interest Model according to User's Implicit Feedback [J]. *Northeast Normal Journal*, 2003, **35**(3): 99-104(Ch).

[10] Zhang Yu, Yuan Fang. A User Interest Model-Based Personalized Information Retrieval Method[J]. *Journal of Shandong University*, 2006, **41**(3):120-125(Ch).

□