THE RESEARCH AND IMPLEMENTATION OF PERSONALIZED USER ATTRIBUTE MODEL FOR THE ACCURATE RETRIEVAL

[1]WU FEI, [2]ZHANG DONGSONG, [3]FU YU

[1]College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

[2]College of Computer, National University of Defense Technology, Changsha 410073, China

[3]Spears School of Business, Oklahoma State University, Stillwater, OK, 74075, United States

E-mail: [1]wufeimailbox@gmail.com , [2]dsongzhang@gmail.com , [3]yu.fu@okstate.edu

ABSTRACT

This paper presents a domain-oriented user attribute model for getting user personalized and real requirements on the field of the accurate retrieval. Resource Distribution Matrix (RDM) and User's Personalized Preference Vector (UPPV) are created by analyzing the user feature and resource characteristic on domain – specific condition. We then present a method to improve the efficiency of the personalized accurate retrieval with the user attribute model. Finally, the preliminary experiments with the prototype indicated that there was an increase in the precision of individualized resources retrieved as the model combines with Shanghai Education Resource Library.

Key words: *Accurate Retrieval, User Attribute Model, Personalized Character*

1. PROBLEM INTRODUCING

It is now of urgent and primary consideration to improve the individualized precision of resources retrieved on the domain-oriented accurate retrieval service. Owing to the diverse users' distinct characteristic from their different knowledge background, interest views and hobbies, the conventional retrieval mode, is hard to meet the personalized requirements of the users. Therefore, a large amount of research activities are done by many scholars. However, most researches focus only on the analysis of search engine and leave out important personalized retrieval factors, such as the inherent characteristic of users. For instance, American scholars Budzik and Watson [1] fetched and searched information by analyzing the local switch context, but no user profile was formed by machine learning. Another scholar L. Chen [2] only employed the user profile to extract the user's information and requirements without the experimental evidence provided. Chinese scholars, B. Song [3] et al, proposed the double information service model which is the user's interest template and grouping model based on the information of the conventional search engine. Meanwhile, they divided the user requirements as dominant and recessive requires, and presented relative algorithm, yet the specific experiment results are not presented. Scholars L. Li [4] et al proposed a tracking method based on the user behavior of Ajax. The method updates the strategy of user behavior information storage based on the interactive sessions and modifies the contents of vector space retrieval model of the user's document. Then, an experimental system of personalized search engine was designed and implemented based on this method but the personalized effect of this model needed further improving.

Many scholars attempt to implement the accurate retrieval by acquiring the user's real requirements. There mainly are three ways in the field. The first way is to acquire the user's requirement information through the interactive mode of natural language, and the well-known example is the Start System developed by the MIT (Massachusetts Institute of Technology) Artificial Intelligence Laboratory. The development of this aspect is mainly subjected to the difficulty of natural language comprehension. The second way is to acquire the requirement information through the specific retrieval language, such as Konopnicki's W3QL system [6] and Mendelzon's WebSQL system [7], and the limitation of this kind of research is that the user has to learn and master specific retrieval language. While the third type of researches focuses on the behavior analysis of the users, such as Liu [8、9] et al of Tsinghua University. They analyzed the macro-behavior and historical behavior of the users, and got deeper understanding of the users' requirement information to relief the bottlenecks for getting the real requirement information. The difficulty of this kind of research is the credibility problem of the random probability distribution on user's behavior.

Therefore, this paper presents a domain-oriented user attribute model that meets the user's real requirements. The model is derived from analyzing the user attribute and resource property in specific domain. Then, characterizing the user's personal character by employing the resource distribution matrix (RDM) and user's personalized preference vector (UPPV), and building a model of user attribute that user's identity is relatively fixed. The application of the model is to add to the various search engines, thus improve the efficiency of the personalized and accurate retrieval. Finally, a method is presented to improve the efficiency of the personalized accurate retrieval with the user attribute model.

The remainder of the paper is organized as follows. Section 2 describes the user attribute model. Section 3 presents a case study of network learner make use of Education Resource Library to measure the value of the user attribute model on personalized accurate retrieval. At last, Section 4 gives a brief summary.

2. THE CONCEPT OF USER ATTRIBUTE MODEL

The proposal of the model is mainly focus on the application of the domain-oriented Resource Library, and the specific details are as follows.

Definition 1: Suppose that in one resource library there are N items at the moment of t0 and the average modification period is T. The period T indicates the average time of one resource item is added or deleted in resource library. If the initial rank of all the resource is S0 = <s1,s2,…sN>, supposed the Relevance Rank(RR) corresponding the initial rank is R0 = <r1,r2,…rN>. So that the retrieval strategy Ss, which is carried out in the period [t0, t0+T], can be considered as the re-ranking of the initial rank S0 =< s1,s2,…sN >. The re-ranking result of all the resource items will be certainty and uniqueness with the alteration of the Relevance

Rank R0, and the alteration is marked as $R_0 \xrightarrow{Ss} R_S$.

Explanation: The definition is obvious, but the premise is that the average modification period T should be large enough. Otherwise, the internal initial ranking relation of all the resource items cannot exist stably. Therefore, the re-using of any identical search strategy Ss will present inconsistent result in a short time.

Because the average period T is $T \longrightarrow 0$, such as the retrieval service on the Internet, the addition or deletion of the resource item would occur very frequently. It is obvious that the re-using result of the identical retrieval strategy Ss will show significant difference even if any other conditions are unchanged in such circumstance. Moreover, the discrepancy increases as the time goes on.

On the contrary, the average modification period T of the domain-oriented Resource Library is long enough in general. Theoretically, the identical retrieval strategy Ss would get the same results while other conditions are unchanged. As for the kind of character, we claim that the Resource View of the Resource Library possesses global stability and predictability in a long enough time period.

Definition 2: Suppose that one user possess M items of personal character, and assign a revised retrieval strategy SS(Aj) for each character item.

Under the condition of definition 1, each retrieval operation will get a relevance ranking vector $R_j = < r_{1j}\ \ r_{2j}\ \ ...,\ r_{Nj} >$, and the mark is $R_0 \xrightarrow{Ss(A_j)} R_j$.

Now, a vector $P = < p_1\ \ p_2\ \ ...,\ p_M >$ is defined $p_j \longrightarrow Ss(A_j)$. Let:

$$R_0^T \times < Ss(A_1),\ \ Ss(A_2),\ \ ...,\ \ Ss(A_M) > \times P^T$$

$$= \ < R_1^T,\ \ R_2^T,\ \ ...,\ \ R_M^T > \times P^T$$

$$= \ < r_1',\ \ r_2',\ \ ...,\ \ r_N' >$$

$$= \ R', \qquad (1)$$

Here, R_i^T is the inversion of R_i , PT is the inversion of P.

As a result, the vector $R' = < r_1',\ \ r_2',\ \ ...,\ \ r_N' >$ is termed as the user personalized retrieval restriction vector (UPRRV). If re-arranging the initial rank of all the resource S0 according to the order of the vector R', we can get the new rank of the entire resource vector S'= <s1', s2',…sN'>. Then, the vector S' is called as the user personalize retrieval result vector (UPRRV).

Meanwhile the matrix $RS_{N \times M} = < R_1^T,\ \ R_2^T,\ \ ...,\ \ R_M^T >$ is called as the user personalize retrieval Resource distribution matrix (UPRRDM), and the vector $P = < p_1\ \ p_2\ \ ...,\ p_M >$ is called as the user personalized preference vector (UPPV).

Explanation: Based on definition 2, the ranking algorithm and results are significantly related to the user personalized preference vector (UPPV) for any retrieval strategy. For instance, assume that UPPV vector element pj=1/qj, and qj is the location number of the element rij in a non-increasing order vector <r1j,r2j,…rNj> | r1j≥r2j≥…≥rNj. Then, a frequently used measure score (qj), the Relevance Comprehension

Algorithm in the Meta Search Engine, is produced as follow:

Score(qj) = ri’ $= \sum_{j=1}^{M} r_{ij} / q_j$ (2)

Deduction 1: The necessary and sufficient conditions of the personalized retrieval operation in Resource Library are the two conditions. The first one is that the data view of Resource Library should be global stability and predictability in a certain time period (Definition 1), and the second one is that the profile of the user possesses M (M>0) items of personal attribute (Definition 2).

Proof: Based on the above definitions of the personalized retrieval, the sufficient condition is evident. Meanwhile, the proof of the necessary condition is shown by contraposition. Suppose that there is no global stability and predictability, obviously, no sufficient stable resource ranking relationship S0 in the system, then the retrieval strategy with personal parameter or no parameter won't work under the same condition, and the personal ranking relationship will be severely affected by time. On the other hand, if the personal attribute items M=0, it means that there is no difference between users. Hence, the modified retrieval strategy should be suitable for all the users; namely, the phenomenon that single retrieval ranking result is suitable for all kinds of users will be emerged. So, the personalized retrieval service loses its value.

Explanation: From Deduction 1, there obviously is a conclusion that the personal operation is quite limited in the Internet if without restrictions on user and resource. Thus, how to delimit the data resource chunk to lengthen the average modification period T in every small resource block would be an important approach for personal retrieval service. Namely, the important things are to increase the global stability of the resource view and to find the personal character of the user as much as possible. For example, many researches mainly focus on the domain-oriented retrieval service and the regional retrieval service.

Deduction 2: Suppose that some resource retrieval system meet the needs of Definition 1, and the user A and B adopt the same retrieval strategy. If both personalized preference vectors (UPPVs) are associated as P_A ≅ P_B, the personal retrieval ranking relationship would be regarded as Ss’_A ≅ Ss’_B, here Ss’_A and Ss’_B represent user retrieval resource distribution matrixes (UPRRDMs) of two users separately.

Explanation: Obviously, when there is the relationship P_A=P_B, the conclusion should be formed. However, when the relationship P_A≈P_B is set, Deduction 2 is lack of rigorous mathematic analysis. So, it is regarded as the empirical conclusion, and is measured by the way of collaborative recommendation algorithm in this paper.

2. MODEL ANALYSIS AND MEASUREMENT

3.1 Attribute Model Applied for Network Education

This study sets up two personal attributes for each learner's attribute structure feature based on the characteristic of the network education and training business, namely, basic information attribute of the learner and interest information attribute of the learner. The agreement lists as follows:

- Learner's Basic Information Model (*LBM*): The set is {*d_gender, d_age, d_speciality, d_grade*}. Element *d_gender* represents Gender attribute which is used for comparing the gender relationship between the learners. Element *d_age* represents Age attribute which is used for comparing the age relationship between the two learners. Element *d_specialty* represents Major attribute which used for comparing the professional similarity between the two learners. Element *d_grade* stands for Learning-Degree attribute which used for comparing the continuous learning status between two learners.

- Learner Interest Model (*LIM*): The set is {<*keyword*$_1$, *last_access*$_1$, *weigh*$_1$>, <*keyword*$_2$, *last_access*$_2$, *weigh*$_2$>, ……<*keyword*$_i$, *last_access*$_i$, *weigh*$_i$>}。 The vector element <*keyword*$_1$, *last_access*$_1$, *weigh*$_1$> is one interest point of learner interest attribute set, while the mark *'keyword'* records the keywords of the interest, and the mark *'last_access'* records learners recent visit to that interest point, the mark *'weigh'* labels the learner's interest degree of the interest, namely regarded as the learner's Interest Weight..

Based on the above definition, deduction and convention, a personal matching and similarity analysis would be operated with the suitable retrieval strategy. In the study, the basic retrieval and comparative strategy Ss is based on the Cosine Similarity Function Model, which is proposed by the information retrieval expert Gerrard Salton [10]. The formula is as follows:

$$Sim(D_1, D_2) = Cos\ \theta = \frac{\sum_{i=1}^{n} d_{1i} \times d_{2i}}{\sqrt{(\sum_{i=1}^{n} d_{1,i}^2) \times (\sum_{i=1}^{n} d_{2,i}^2)}} \qquad (3)$$

Therefore, the revised personal retrieval strategy and the some similarity evaluation algorithms are constructed as follows:

- Based on Learning Resource Base (LRB), the revised retrieval strategy S_s *(LIM)* is created with the learner interest model (LIM). According to the practical operation, that strategy can be divided into two sub-strategies.

Sub-strategy S_s *(LIM, LRB)*: Retrieval results are ranked in descending order with the similarity $Sim(LIM, LRB)$ between the learner's interest information and the resource base (LRB). The similarity formula is as follows:

$$Sim(LIM, LRB) = Cos\,\theta = \frac{\sum_{i=1}^{n} weigh_{keyword_i} \times fre_{keyword_i}}{\sqrt{(\sum_{i=1}^{n} weigh^2_{keyword_i}) \times (\sum_{i=1}^{n} fre^2_{keyword_i})}} \qquad (4)$$

here，$fre_{keyword}$ stands for the TF/IDF (Term Frequency/Inverse Document Frequency) of the key interest word 'word keyword' in LRB.

Sub-strategy Ss (LIM, LIM): Retrieval results are the most similar learner, whose interest attributes are the most similar to the interest attributes of the retrieval user. The similarity formula is as follows:

$$Sim(LIM) = Sim(LIM_1, LIM_2)$$

$$= Cos\theta = \frac{\sum_{i=1}^{n} weigh_{1,keyword_i} \times weigh_{2,keyword_i}}{\sqrt{(\sum_{i=1}^{n} weigh^2_{1,keyword_i}) \times (\sum_{i=1}^{n} weigh^2_{2,keyword_i})}} \qquad (5)$$

- Based on Learning Resource Base (LRB), the revised retrieval strategy S_s *(LBM)* is created with the learner basic information model (LBM). The same way, that strategy can be divided into two sub-categories.

Sub-strategy S_s *(LBM, LRB)*: Retrieval results are ranked in descending order with the similarity $Sim(LBM, LRB)$ between the learner's basic information attributes and the resource base (LRB). In the study, assume that $Sim(LBM, LRB) = 1$, namely, the system do not provide this kind of retrieval service. As well as requesting to the service, the original ranking sequence of the resource item will be gotten back.

Sub-strategy S_s *(LBM, LBM)*: Retrieval results are the most similar learner, whose basic attributes are the most similar to the basic attributes of the retrieval user. The similarity formula is as follows:

$$Sim(LBM) = Sim(LBM_1, LBM_2)$$

$$= Cos\theta = \frac{\sum_{i=1}^{n} d_{1,i} \times d_{2,i}}{\sqrt{(\sum_{i=1}^{n} d_{1,i}^2) \times (\sum_{i=1}^{n} d_{2,i}^2)}} \quad (6)$$

Here, mark d represents the corresponding quantized value of some element, and refer to the segmentation quantization method and the knowledge tree method [11]**.**

3.2 Experiment and Analysis

We collected some course resources from network learning resource library, which include various kinds of forms such as doc, html, ppt, pdf, and xls. These courses include operation system, human resource management, information retrieval, network security etc. These resources are indexed for preparing the record of the learners' evaluation with help of the Lucene tool. Meanwhile, learners of different specialized majors are registered in the system. For example, suppose that learner User1, User2 and User3 are specialized in the logistics management major, and User4 and User5 are specialized in the computer major. Moreover, User1 is first time to logon the system for User. The others have enrolled before, and given the evaluation on some learning resource, marked '√'. The specific details are as Table 1.

The experiments' conclusions are as follows: (1) When the learners are first time to enroll the system, for example, system will find the most similar learners for the newcomers and search the learning resources they preferred based on the attribute structure of the learner, which can help to solve the 'cold start problem' of the system with personalized push technology [12]. (2)When the learners go through a period of learning activities, system will record the interest information of the learners in LIM. System would recommend the similar learner and learning resource based on the learner's accumulated information in LIM. For example, system would recommend the information retrieval course resources for User2, and these course resources are preferred by User 3 and User. Similarly, system would recommend the resources of human resource management course that preferred by User 3 to User2. (3) For the personal combine service, system would adopt various personal revised strategies that system established, and figure out the most similar learning resources for the learner based on the current user's LIM and LBM information.

3. CONCLUSION

This paper presents a concept and method of user attribute model (UAM) that applied in the personal and accurate retrieval. Through the theory analysis and experimental test, its value mainly manifests as the following three aspects: First, it makes emphasis on the 'actively' acquiring the user attribute in human-computer interaction way. Meanwhile the auxiliary 'passive' perception by virtue of the analysis of user behavior, we will reduce the technological difficulty depending on the 'analysis of subjective intention with the external behavior ', and relieve the bottleneck of grasping the user's requirement. Second, the resource distribution matrix (RDM) and user's personalized preference vector (UPPV) are constructed to describe the global stable feature of the resource view in many network learning circumstances, Namely, differ from the resource's dynamic character under the open condition, the precise matching and retrieving service can be improved if worked together with the parameters that stand for the user's personal attributes. Third, the retrieval service of different personalized requirement are implemented based on the above two aspects in the rank of personal recommended result. Then, we will realize the hybrid recommendation that mixed the coordination and content filtering based on the meta search , and help to solve the cold start problem that exist in the personal recommendation and retrieval service.

ACKNOWLEDGMENTS

This work is supported in part by Innovation Program of Shanghai Municipal Education Commission No.06ZZ91 and No.12ZZ182. Meanwhile, the authors would like to thank Dr. Wu Bing and the reviewers for their constructive works.

REFERENCES:

[1] Budzik, J., Hammond K..“Watson: Anticipating and Contextualizing Information Needs[C]”. In Proceedings of the 62th Annual Meeting of the American Society of Information Science, 1999, 27-40.

[2] Chen L., Sycara K., “Webmate: A personal agent for browsing and searching[C]”. In Proceedings of the 2nd International Conference on Autonomous Agents, 1998, 132–139.

[3] SongBin, YuKai. “A Model of Agent-based Individuation Search [J]”. Journal of Nanjing University of Science and Technology,2002, 26(3)：295-298.

[4] Li Lei, Zhou Guomin. “Personalized search engine based on Ajax and VSM [J]”. Computer Engineering and Applications, 2007，43(19):89-91.

[5] START: National Language Question Answering System; http://start.csail.mit.edu/;

[6] Konopnicki W3QS: The WWW Query System http://www.cs.technion.ac.il/~W3QS /;

[7] Mendelzon WebSQL: Querying the WWW; http://www.cs.toronto.edu/~websql/;

[8] Yu Huijia, Liu Yiqun, Zhang Min, Ru Liyun, Ma Shaoping. “Research in Search Engine User Behavior Based on Log Analysis[J]”, Journal of Chinese Information Processing, 2007，01（21），pp:109-114.

[9] Liu Yiqun, Cen Rongwei, Zhang Min, Ma Shaoping. “Automatic Search Engine Performance Evaluation Based on User Behavior Analysis [J]”, Journal of Software, Vol.19, No.11, November 2008, pp.3023−3032.

[10] Salton G., Wong A., Yang C.S.,“A vector space model for automatic indexing[J]”. Communications of the ACM, 1975,11, pp:613-620.

[11] Wu B., Wu F., Ye C.M., “Personalized Recommendation System Based on Multi-Agent and Rough Set[C]”. Proceedings 2010 2nd International Conference on Education Technology. Shanghai: China. June, 2010, V4:303-307, ISBN: 978-1-4244-6368-8.

[12] Guo Yanhong, Deng Guishi. “Hybrid Recommendation Algorithm of Item Cold-start in Collaborative Filtering System[J]”, Computer Engineering, 2008,12，34(23):11-13.