Webpage Recommendation Model in Personalized Service based on the Group Clustering

Rong Fu [a], Yi He [b], Yingqian Zhang [c]

City Institute, Dalian University of Technology, Dalian Liaoning 116600, China

[a]furong880@126.com, [b]heyi517@126.com, [c]zhangyq@dlut.edu.cn

Keywords: Group Clustering; Webpage Recommendation; Personalized Service

Abstract. This paper proposes an effective model to realize the Webpage recommendation in personalized service based on the data mining technology. Applying user interest feature vector and user session feature vector to represents user's interest and generate user groups' interests by clustering, and extends the individual user interest. The experiment proves that the model is effective and accuracy.

Introduction

With the development of Internet, computer users are constantly submerged by huge information. Now people are looking for a service mode which can actively provide user interested information to the user and provide different service strategy and service contents to different users, that is, the personalized information service mode. Many researchers have proposed a variety of solutions to meet the individual needs of users. Information retrieval is the recommended technique based on content, it finds the qualified content in users' given retrieval commands and conditions. Information filtering technique using the user model to describe user interest, it calculate the similarity of new Webpage and user model, and recommend the high similarity Webpage to the user.

Ref. [1-3] proposes method to provide personalized service based on collaborative filtering, but they need user input additional information to determine the users personalized behavior. Personalized service based on the categories depends on the category setting and Webpage content classification. The rough classification leads to recall and precision of information is not high, and it is difficult to recommend accurately and interested Webpage to the user. Now people use data mining technology to mine the users which have similar browsing interest, the organizational structure of the web designers accordingly to adjust the page, so as to provide high quality service for the user. Ref. [4] provides three methods of real-time personalized recommendation, including two methods for clustering method and association rules. Ref. [5-7] takes different strategies to get the user's access pattern by mining users' log, but they do not consider personalized recommendation. Personalized service based on previous browsing content is hard to capture the fresh Webpage and change of users' interest.

This paper combine the information retrieval, information filtering and data mining technology to construct an effective model to realize the Webpage recommendation in personalized service and to meet the user needs of checking the website.

The Recommendation Model Structure

This paper proposes user interest model which based on mining user's browsing behavior (recorded in the server logs), information retrieval and filtering technology, so the user's interest is represented by two features. The structure of model is shown in Fig.1. After getting the interest of individual user, we could found group of users' interest by clustering, and the individual user interest can learn and extended from the group of user interest. In the process of user browsing, we will record, analysis and extract feature of the user access behavior and browsing Webpage, and transfer the user feedback to the individual user interests' database, thus further supplement and

improve the individual user interests' database. At the same time through the group clustering feedback to group user interests database, and then update the group user interests' database.

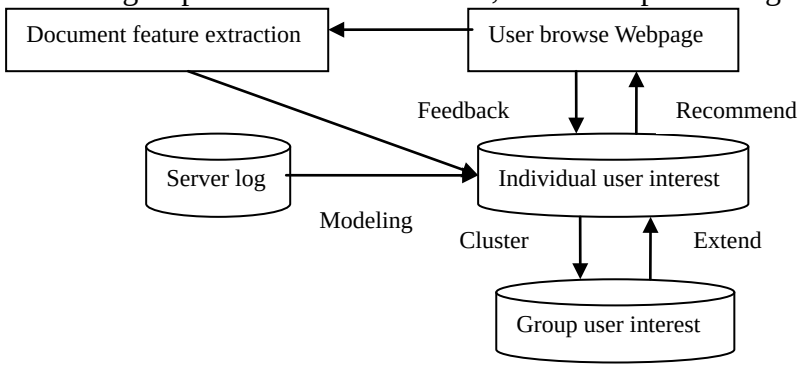

Fig.1. Recommendation Model Structure

The Key Technology of User Interest Modeling

The way of user interest modeling: (1) Building and updating user interest database according to the feedback information of search results through the user browsing Webpage. (2) In the case of without explicit engaging in of user, establish and update user interest through the observation of the user's browsing behavior.

At present, mainly uses the Vector Space Model (VSM) to represent the documents. According to the attribute, document frequency and anti frequency of the term in the document, the document is represented as a feature vector of terms. If the term T_i appears in the document title and author, then its weight is 1. If the term appears in the main body of a document, we calculate the weight according to the term frequency and anti frequency. The weight W_i is as follows:

$$W_i = \frac{(0.5 + 0.5^{tf(i)}_{tf\ \max})(\log^{n}_{\mathrm{tf(i)}})}{\sqrt{\sum_{T_i \in F}((0.5 + 0.5^{tf(j)}_{tf\ \max})^2 (\log^{n}_{tf(j)})^2)}} \quad (1)$$

In the Eq. 1, *tf(i)* is the frequency of the term T_i appears in documents, *df(i)* is the number of documents containing the term T_i. tf_{max} is the number of most frequently occurring terms appear in the document. In user interest modeling method(1), the user interest is considered as a document. First select a set of term $(T_1, T_2, T_3, \ldots, T_n)$ which is suitable to represent the user interest. Second, calculate the weight W_i $(i=1, 2, \ldots, n)$ according to the important degree of the term T_i in the user's browsing Webpage file. Then the user interest file is represented by the term feature vector $(W_1, W_2, W_3, \ldots, W_n)$, W_i is the weight of the feature item *i*. User interest feature vector is got by calculation of the user' behavior based on content viewed.

Definition 1 Vector of Any m terms representation for the user interest feature vector $P\ (W_1, W_2, W_3, \ldots, W_n)$, W_i is the weight in the user interest in the feature vector.

In order to maintain the interest of a user, from the user's feedback: If the user likes this page, the weight of term extracted form this page will be added to the corresponding weight of user interest feature vectors, this process is called relevance feedback. Generally produced by iteration as follows:

$$P_{k+1} = P_k + \beta \sum_{k=1}^{n_1} \frac{R}{n_1} - \gamma \sum_{k=2}^{n_2} \frac{S_k}{n_2} \quad (2)$$

In the Eq. 2, P_{k+1} is a new user interest feature vector, P_k is a new user interest feature vector, R_k and S_k respectively represent interest in and not interested in the content of user feedback. n_1 is the number of documents which interested in, n_2 is the number of documents which not interested in, β and γ determines the relative role of positive and negative feedback.

The Key Technology of User Browsing Behavior Description

Web server logs include access log, reference log and the agent log. After preprocessed the

diaries, classified the log file information by user, then got the each user access. *L=(ip, uid, url, time, hits)* is used to represent the Web server log. The *ip, uid, url, time, hits* respectively represents the user IP address, user ID, user request URL and the corresponding browsing time and the number of clicks. After further processing, a user in a certain period of time browsing behavior is reflected.

Definition 2 The user browsing behavior has the following form: $B=(ip_B, uid_B, \{(l_B.url, hits, time)\}^n)$, $l_B \in L$, $l_B.ip = ip_B$, $l_B.uid = uid_B$, $n \geq 1$, *hits* and *time* respectively represents the number and time of customer uid_B browse the page of $l_B.url$.

This paper uses the two most basic features hits and access time which contained in the browse behavior to represent the user session feature vector. Reference the abstract tree structure of website. For each page of the characteristic sequence of user browsing, we use the Eq. 3 to calculate the weight of the page in a user session feature vector:

$$w_i = (\frac{l}{L} + \frac{t}{T}) \cdot \frac{n}{N} \qquad i = 0,1,2\ldots \tag{3}$$

If the page in this sequence without adjacent nodes:

$$w_i = (\frac{l}{L} + \frac{t}{T}) \cdot \frac{n}{N} + \sum_{(w_i)sub} w_i \qquad i = 0,1,2\ldots \tag{4}$$

In the Eq. 3, *l* is the number of times a web page be visited, *L* is the total number of visit, *t* is the time that consumed in visit node, *T* is the total time of visit, *n* is the total length of the visit, *N* is the total length of the site. Considering the changes in the structure of website, In the Eq. 4, we introduced the web length parameter compared with the user interest. We think that the user's browsing behavior is contained in the site topology, considering the Inheritance of the user's browse interest, referring to the website structure we calculate the parent node and child node interest separately in the equation. The two most basic characteristic in user sessions which we chose make the user feature vector expression more accurate, objective, and reasonable. Although the sequence information of user access is lost, the calculation is simplified and the cost is reduced.

Definition 3 User session feature vector is described by $S(W_1, W_2, W_3 \cdots W_n)$, W_i is the weight of page in the session which is calculated by Eq. 3 and Eq. 4.

The Key Technology of Group Clustering

This paper introduces two kinds of features to represent the user interest vector. In previous research work, user clustering usually uses similar browsing path to clustering, and can not accurately capture user interest. Because people often browse Webpage with jumping, but real interest is the Webpage content, we integrate the user's two feature vector *P* and *S*, the similarity of user α and β are described in Eq. 5:

$$\sin(\alpha, \beta) = \sin_P(\alpha, \beta) + \sin_S(\alpha, \beta) \tag{5}$$

This article introduces the weight parameter λ to balance the different characteristics in similarity calculation:

$$\sin(\alpha, \beta) = \sum_1^q \lambda_i \sin_{F_i}(\alpha, \beta) \quad \sum \lambda = 1 \tag{6}$$

In the Eq. 6, F_i represents different characteristics, *q* is the number of features which extracted from system. This article extracts two features, so *q*=2. λ_i represents the different characteristics impact on calculating the similarity. By calculating the Euclidean distance between the vectors to compare the similarity between feature vectors *S*, the value is smaller, the more similar. Expression is as follows:

$$D(A,B) = \sqrt{(\sum_{i=1}^{n} (a_i - b_i)^2)} \tag{7}$$

Similarity is calculated by inner product between the vectors, which shown as follows:

$$\sin(V_i, D_j) = \cos\theta = \sum_{k=1}^{m} v_{ik} \cdot d_{jk} / (\sqrt{\sum_{k=1}^{m} v_{ik}^{\ 2}} \cdot \sqrt{\sum_{k=1}^{m} d_{jk}^{\ 2}}) \tag{8}$$

Clustering is the process of data set classification which according to the similarity of data sets. This paper adopts Leader level algorithm.

Algorithm 1 Leader algorithm of group clustering

Input: user feature vector set U (composed of P and S)

Output: a group of cluster $C=\{C_1,C_2,\ldots\}$ $C_i \in U$

Process:

(1) Initial value of C is empty.

(2) Looking for the clustering c for each user feature vector u of U, the shortest center of mass distance (similarity) between U and C is recorded as d_{min}. If d_{min} is less than the distance threshold Distance, then u will join the c, otherwise the $\{u\}$ will join the C. We can use the average value to calculate the center of mass, suppose there are clustering $C=\{u_1,u_2,\ldots\}$, the property set of the center of mass m of c is:

$$F(m) = \frac{F(u_1) + F(u_2) + \cdots + F(u_k)}{k} \tag{9}$$

$F(u_i)$ is a feature set of user vector. In this paper, $F(u_i)= F_P(u_i) \cup F_S(u_i)$. Combining with the characteristics calculation method, this clustering algorithm is simple and can realize the dynamic clustering. For example, when a new user feature vector u is added to the cluster, feature attribute sets of clustering can be conveniently got by following formula:

$$F(m) \leftarrow \frac{k \cdot F(m) + F(u)}{k+1} \tag{10}$$

Compared with the common clustering algorithm, the calculation cluster method can save a lot of computation time, it also achieve the purpose of dynamic clustering and improve the accuracy of clustering similar users.

The Establishment of User Interest Entity

Through the application of vector space model, we construct a group of user interest, individual user interest learning and extended (added dimension) group of user interest. The establishment of user interest files by following the idea of ontology, we use the tree type hierarchical classification method to dynamically show the user interest, makes the concept gradually thinning [8]. Each node in the tree has a user interest classification and the value of the extent of the classification of user preferences. Each user has the same hierarchical classification tree in the system, although this part lost personal information, but it simplifies the computation complexity and reduces the cost of memory system. The extent of the classification of user preferences in the tree is defined as follows:

$$P(N_{child} \mid N_{parent}) = \frac{W_{child}}{\sum_{i \in \{childofNparent\}} W_i} \tag{11}$$

In the Eq. 11, N represents the nodes of tree, W_i is the value which association with the node.

When get the user interest feature vector P, we treat it as document to generate the user interest file. Any user interest feature vector and the classification similarity in tree type structure are defined as follows:

$$\Delta W_i = \lambda^{days} \cdot similarity(P, c_i) \tag{12}$$

Taking into account the user interest changes, we introduce the time factor λ which less than or equal to 1. The user interest constantly be captured, but the user's interest may change as time goes on. It can reflect the interest change, especially change in user’s short-term interest.

Webpage Recommendation and User Feedback

Webpage recommendation includes separate and joint recommendation. Separate recommendation is that recommend similar to our own interests of user browsing document, or define a threshold concentration of I_h, if the *similarity(V,U)*> I_h of the document, it is recommended to the user automatically by the system. Joint recommendation first obtains similar browsing interest users browse the next page set, and calculate the similarity of the document and the user, to sort the similarity, the higher the document is recommended to the user. The user can mark for the filtered documents in document sets according to the degree of correlation of own demand, and scores can be set into several levels. According to the user's feedback, it need adjust user interest timely, and adjust group user interest when necessary. The structure of the model can easily realize the feedback process, so as to improve the precision.

Simulation Results Analysis

In this paper, the experiment uses precision which commonly used in information retrieval systems to evaluate the effect of the system. The recommendation model is recommend personalized and interest information to the user, so the results of the recommendation system can only be judge through the collection of user feedback. The precision can accurately reflect the ratio of useful information to users which system brings. This paper chooses four users to participate in the test, after a week of user information collection, the system collect the results are shown in Tab.1:

Tab.1. Experimental Results

user	The number of page views	The number of recommended information	The number of interest	precision
A	84	840	471	0.561
B	61	610	331	0.543
C	78	780	356	0.456
D	112	1120	595	0.531

Tab.1 shows that four users browsing are not same in a week, but the rate of precision is basically stable. The system has a better recommendation effect and precision slightly higher than the traditional recommendation systems.

Conclusion

This paper proposes personalized Webpage recommendation model, which can generate two different feature vectors for users. The user groups' interests are generated by clustering, and extend the individual user interest. The experiment proves that the model is effective and accuracy. In order to provide better personalized service to the users, the following research work is that the model how to combined with user browsing path prediction technology.

References

[1]Konstan J, et al. Apply Collaborative Filtering to Usenet News [J].Communications of the ACM, 1997 40(3) 103-110.

[2] Herlocker J, Konstant J, et al. An Algorithmic Framework for Performing Collaborative Filtering[C]. New York: Proc. Conference on Research and Development in Information Retrieval, 1999 57-63.

[3] Shardanand U Maes P. Social Information Filtering Algorithms for Automating Word of Mouth[C]. Los Angeles: Proc. ACM CHI Conference, 1995 127-131.

[4] MobasherB, et al. Creating Adaptive Web Sites Through Usage-based Clustering of URLs[C]. London: Proc. the 1999 IEEE Knowledge and Data Engineering Exchange Workshop, 1999 112-123.

[5]Spiliopoulou M. The Laborious Way From Data Mining to Web Mining [J]. International Journal of Computer System Science & Engineer Special Issue on Semantics of the Web, 1999 3(1) 105-113.

[6] Cooley R, Mobasher B, et al. Data Preparation for Mining World Wide Web Browsing Patterns [J]. Knowledge and Information Systems, 1999 1(1) 17-24.

[7] Buchner A G, Mulvenna M D. Discovering Internet Marketing Intelligence through On-line Analytical Web Usage Mining [J]. SIGMOD Record, 1998 27(4) 54-61.

[8] Xing Xie, Huajun Zeng, Weiying Ma. Enabling Personalization Services on the Edge Microsoft Research Asia [EB/OL]. http: //research.microsoft. com/asia/download/disquisition/0208. asp.

[9] Guiyang Su, Yongcheng Wang, Yinghua Ma. WebPage Recommendation Structure Model in Personal Service [J]. Computer Science, 2003 30(4) 99-101.