CONVERTER MAGAZINE
Volume 2021, No. 7

Collaborative Filtering Method of Web Service Recommendation Based on Content Awareness and Personalization

Yajuan Sun

Henan University of Animal Husbandry and Economy, Zhengzhou City, Henan Province, China

Abstract

The rapid development of web has brought the status quo of information explosion. Now the research of personalized web service recommendation information system has become a hot research direction in the field of service computing. The research of web service recommendation system mainly solves two problems: prediction and completion of sparse QoS data, user personalized recommendation. Firstly, this paper proposes an improved collaborative filtering web service recommendation algorithm based on user preferences. Secondly, based on the improved collaborative filtering web service recommendation algorithm based on user preference (UPCF), this paper proposes an improved collaborative filtering web service recommendation algorithm based on joint user preference (CUPCF). The algorithm extracts user preference data from QoS data and uses it to select similar neighbors. After the Top-k algorithm is used to determine the set of similar neighbors of target users and services, the QoS data is used to calculate the similarity of neighbors. Finally, based on CUPCF algorithm, this paper proposes an improved collaborative filtering web service recommendation algorithm based on user location and preference. Experimental data show that the algorithm can improve the efficiency of recommendation and filtering.

Keywords: *Content awareness, web service recommendation, collaborative filtering, UPCF*

I. Introduction

Web service is a service-oriented architecture of network interaction technology, also known as WWW (World Wide Web), composed of many interlinked hypertext documents (HTML). These hypertext documents can be accessed through the Internet [1-2]. The interaction between different webs is completed by extensible markup language (XML). Users on different platforms in the network can find their target information through web browsers and jump to other pages according to hypertext links. The advantages of web service, such as easy browsing, easy searching and cross operating system, make it develop rapidly in the information age.

With the development of the web, it is difficult for users to get the information they want quickly in the network [3]. They usually need to choose between a large number of redundant and useless web pages, which is the problem of information overload. Full text index is the most familiar keyword index. The most representative search engines are Google, Baidu, Sogou and haoso [4-5]. As the entrance of the Internet, search engine has a huge market space. The job of the search engine is to save a large amount of Web information on the local server, and then sort the web links in the server cache through its own sorting algorithm. When the user sends out the query demand, the priority list of web pages will be fed back to the user according to the pre agreed keywords. The results obtained by users are sorted according to the degree of relevance, and the web links with the most relevant keywords are displayed in priority.

This kind of retrieval method requires users to know their query needs and type keywords in advance. If they can't get satisfactory results, they need to modify their query points and start the query again. When users can not query in the way prescribed by the system, this kind of search engine can not provide users with satisfactory search results. Furthermore, when users' needs are not clear, this kind of search engine has nothing to do [6-7].

The emergence of recommendation system makes up for the deficiency of search engine. Different from the targeted search of the search engine, the recommendation system is based on the analysis of the user's access history data. Through the analysis of these data, the system establishes a user access model, finds the user's interest in the model, and then recommends the candidate service to the target user. The recommendation system has been used in practice [8]. For example, Amazon's Kindle Bookstore uses recommendation technology to recommend books and CDs, and netflix.com, which has more than 17000 movies, also uses recommendation technology to recommend movies. They have achieved good market results.

II. Improved collaborative filtering web service recommendation based on user preference

2.1 Similarity calculation based on user preference

For the collaborative filtering algorithm based on access history, finding the preferred similar neighbors of the target user or service is the most important part of the whole algorithm, because the prediction accuracy of the recommendation algorithm depends on the accurate similar neighbors.

The existing similarity calculation methods include co rotation similarity, Pearson correlation coefficient and so on. Pearson correlation coefficient has better similarity calculation results. In this paper, Pearson correlation coefficient is used to calculate the similarity between users or services, and to find the preferred similar neighbors of the target users or services. The specific similarity calculation is shown in formula (1) and formula (2) [9-10]:

$$\mathrm{si}\,m_{sat}(u,v)=\frac{\sum_{i\in Isat}\left(sat_{u,i}-\overline{sat_u}\right)\left(sat_{v,i}-\overline{sat_v}\right)}{\sqrt{\sum_{i\in Isat}\left(sat_{u,i}-\overline{sat_u}\right)^2}\sqrt{\sum_{i\in Isat}\left(sat_{v,i}-\overline{sat_v}\right)^2}} \quad (1)$$

$$\mathrm{si}\,m_{sat}(i,j)=\frac{\sum_{u\in Usat}\left(sat_{u,i}-\overline{sat_i}\right)\left(sat_{u,j}-\overline{sat_j}\right)}{\sqrt{\sum_{u\in Usat}\left(sat_{u,i}-\overline{sat_i}\right)^2}\sqrt{\sum_{u\in Usat}\left(sat_{u,j}-\overline{sat_j}\right)^2}} \quad (2)$$

In formula (1), u and v represent user u and user v respectively; Simsat(u,v) represents the preference similarity between users u and v; Isat represents a Web service set accessed by user u and user v, and I belongs to Isat, which is a service in the set; Satu,i and Satv,i respectively represent user u and user v's preference for service I. satu and Satv respectively table the average value of user u and user v's preference for Web services they visit. the purpose of formula (1) is to calculate the similarity between users based on user preference.

In formula (2), I and j represent service I and service j respectively; $\mathrm{Sim}_{\mathrm{sat}}$(i,j) represents the preference similarity between services I and j; Usat represents the set of users that service I and service v are accessed together, and u belongs to Usat and is a user in the set; Satu,i and Satu,j represent the preferences of service I and service j for user u respectively, and Sati and Satj table the average values of preferences of service I and service j respectively. the purpose of formula (2) is to calculate the similarity between services based on user preferences.

2.2 QoS prediction of target users or services

After similar users are found, we get the neighbor list of users or services based on user preference, and the similarity matrix between users or services. The second stage is to predict the missing QoS of target users or services on this basis.

We define Pu(u,i) as the QoS value of user u to service I based on user prediction, which is calculated as shown in formula (3), and define Pi(u,i) as the QoS value of user u to service I based on service prediction, which is shown in formula (4):

$$Pu\left(u,i\right)=\overline{q}_u+\frac{\sum_{v\in N(u)}\mathrm{si}\,m_{sat}\left(u,v\right)\times\left(q_{v,i}-\overline{q}_v\right)}{\sum_{v\in N(u)}\mathrm{si}\,m_{sat}\left(u,v\right)} \quad (3)$$

$$Pi\left(u,i\right)=\overline{q}_i+\frac{\sum_{j\in N(i)}\mathrm{si}\,m_{sat}\left(i,j\right)\times\left(q_{u,j}-\overline{q}_j\right)}{\sum_{j\in N(i)}\mathrm{si}\,m_{sat}\left(i,j\right)} \quad (4)$$

In formula (3), N (u) is the set of preferred similar neighbors of user u selected according to the top-k algorithm, simsat (u, v) represents the preference similarity between user u and v, and $q_{v,\ i}$ represents the QoS value of user v accessing service i. qv represents the average value of user v's access to all known services, and qu represents the average value of user u's access to all known services except i. The formula is a user based collaborative filtering missing QoS prediction method.

In formula (4), N (i) is the set of preferred similar neighbors of service i selected according to the top-k algorithm, simsat (i, j) represents the preference similarity between service i and j, and qu, j represents the QoS value of user u accessing service j. qj represents the average value of service j accessed by all known users, and qi represents the average value of service i accessed by all known users except u. The formula is a service-based collaborative filtering method for missing QoS prediction.

In the WSREC method proposed by zheng et al., which combines UPCC and IPCC, his experiments prove that combining user-based collaborative filtering algorithm and service-based collaborative filtering algorithm can effectively solve the data sparsity problem of collaborative filtering and improve the accuracy of predicting QoS. We define a harmonic parameter to adjust the proportion between Pu(u,i) and Pi(u,i). The value of this harmonic parameter belongs to the closed interval [0,1]. We define the QoS value P(u,i) of user U to service I which is comprehensively predicted. A parameter of 0 means that the value of P(u,i) is completely determined by Pu(u,i). Parameter 1 means that the value of P(u,i) is completely determined by Pi(u,i). The variation of this parameter indicates the influence of Pu(u,i) and Pi(u,i) on the final prediction result. P(u,i) is calculated as shown in formula (5):

$$P\left(u,i\right)=\lambda Pu\left(u,i\right)+\left(1-\lambda\right)P_i\left(u,i\right) \quad (5)$$

2.3 Improved collaborative filtering algorithm flow based on user preference

To sum up, the specific steps of upcf algorithm are as follows:

Enter: TrainData set of QoS data, test set of QoS data, approximate neighbor number Topk of Users, approximate neighbor number Topk of Services, harmonic parameter λ.

Output: QoS prediction value P(u,i).

1. compare QoS data in training set TrainData.

2. For each user in the SAT user preference matrix, use formula (1) to calculate the preference similarity between users, which is stored in the user similarity matrix.

3. Select the user's priority similar neighbor set N(u) according to the neighbor number top-k of similar users.

4. For the missing QoS values in the test set TestData, use formula (3) to calculate the missing QoS prediction

value Pu(u,i).

5. For each service in the SAT user preference matrix, use formula (2) to calculate the user preference similarity between services, which is stored in the service similarity matrix.
6. Select the priority similar neighbor set N(i) of the service according to the number of similar service neighbors top-k.

7. For the missing QoS values in the test set TestData, use formula (4) to calculate the missing QoS prediction value Pi(u,i).

8. Adjust the ratio of Pu(u,i) and Pi(u,i) with harmonic parameter λ, and finally get the best prediction value P(u,i) of UPCF algorithm.

III. Improved collaborative filtering Web service recommendation based on joint user preference

3.1 Combine user preference with QoS data

How to combine user preference data with QoS data needs to study the characteristics of the data itself. User preference data projects the original QoS data from the actual space to the feature space. In the feature space, all users have the same access data range. This solves the problem of imbalance among users, and the lower limit and upper limit of data are consistent from the perspective of user preference. The QoS value of users' access to services belongs to the sample in the real space, which truly reflects the calling state of users to services and contains rich network state attributes. Therefore, the focus of this chapter is to combine user preference data in feature space with QoS data in real space to calculate the similarity between users or services.

3.2 Improved collaborative filtering Web service recommendation algorithm based on joint user preference

In this section, Pearson correlation coefficient is used to calculate the QoS data similarity between users or neighbors between services. The specific similarity calculation is shown in Formula (6) and Formula (7):

$$\mathrm{si}\,m(u,v)=\frac{\sum_{i\in I}\left(q_{u,i}-\overline{q_u}\right)\left(q_{v,i}-\overline{q_v}\right)}{\sqrt{\sum_{i\in I}\left(q_{u,i}-\overline{q_u}\right)^2}\sqrt{\sum_{i\in I}\left(q_{v,i}-\overline{q_v}\right)^2}} \quad (6)$$

$$\mathrm{si}\,m(i,j)=\frac{\sum_{u\in U}\left(q_{u,i}-\overline{q_i}\right)\left(q_{u,j}-\overline{q_j}\right)}{\sqrt{\sum_{u\in U}\left(q_{u,i}-\overline{q_i}\right)^2}\sqrt{\sum_{u\in U}\left(q_{u,j}-\overline{q_j}\right)^2}} \quad (7)$$

In formula (6), u and v represent user u and user v respectively; v is a member of the similar neighbor set N(u) of user u; Sim(u,v) represents the similarity between users u and v; i represents a Web service set accessed by user u and user v, and i belongs to a service in the set; $q_{u,i}$ and $q_{v,i}$ respectively represent QoS values of services i accessed by users u and v, and qu and qv respectively represent average QoS values of Web services accessed by users u and v. The purpose of formula (6) is to calculate the similarity of QoS data of similar neighbors based on user preference. The structure of formula (7) is the same as that of formula (6), which will not be repeated here.

We define Pu(u,i) as the QoS value of user U to service I based on user prediction, which is calculated as shown in formula (8), and define Pi(u,i) as the QoS value of user U to service I based on service prediction, which is shown in formula (9):

$$Pu\left(u,i\right)=\overline{q}_u+\frac{\sum_{v\in N(u)}\mathrm{s}im\left(u,v\right)\times\left(q_{v,i}-\overline{q}_v\right)}{\sum_{v\in N(u)}sim\left(u,v\right)} \quad (8)$$

$$Pi\left(u,i\right)=\overline{q}_i+\frac{\sum_{j\in N(i)}\mathrm{s}im\left(i,j\right)\times\left(q_{u,j}-\overline{q}_j\right)}{\sum_{j\in N(i)}sim\left(i,j\right)} \quad (9)$$

In formula (8), N(u) is the user preference similar neighbor set of user u selected according to top-k algorithm, sim(u,v) represents the QoS similarity between users u and v, and qv,i represents the QoS value of user v accessing service i. qv represents the average access value of user v to all known services, and qu represents the average access value of user u to all known services except i. This formula is a user-based collaborative filtering missing QoS prediction method. Formula (9) is a service-based collaborative filtering missing QoS prediction method, and its structure is consistent with formula (8), so it will not be described here.

We continue to use the joint collaborative filtering algorithm to calculate the final predicted value. We use harmonic parameters to adjust the proportion between Pu(u,i) and Pi(u,i). The value of harmonic parameter λ belongs to the closed interval [0,1]. We define the QoS value P(u,i) of user U to service I which is comprehensively predicted. The parameter λ is 0, which means that the value of P(u,i) is completely determined by Pu(u,i). The value of P(u,i) is completely determined by Pi(u,i). The variation of parameter λ indicates the influence of Pu(u,i) and Pi(u,i) on the final prediction result. P(u,i) is calculated as shown in formula (10):

$$P\left(u,i\right)=\lambda P_u\left(u,i\right)+\left(1-\lambda\right)P_i\left(u,i\right) \quad (10)$$

IV. Improved collaborative filtering of Web service recommendation based on user location and preference

4.1 The impact of user location on QoS

Web services are services provided on the Internet, so the QoS of network services depends on the network environment. If the performance of the network link between the target user and the target service is excellent, there is a great probability that the user will enjoy high-quality network service. The network link between users and services is good or bad, there are several factors that have a great impact. The most important two points are the network distance between users and services and the network bandwidth they enjoy. These two points have a great relationship with the location of users. Because the website providing services is usually operated by the company, when the company builds the website and provides services, the first thing to consider is the concurrency and pressure, so the service provider usually has a good network environment to cope with a large number of daily access needs.

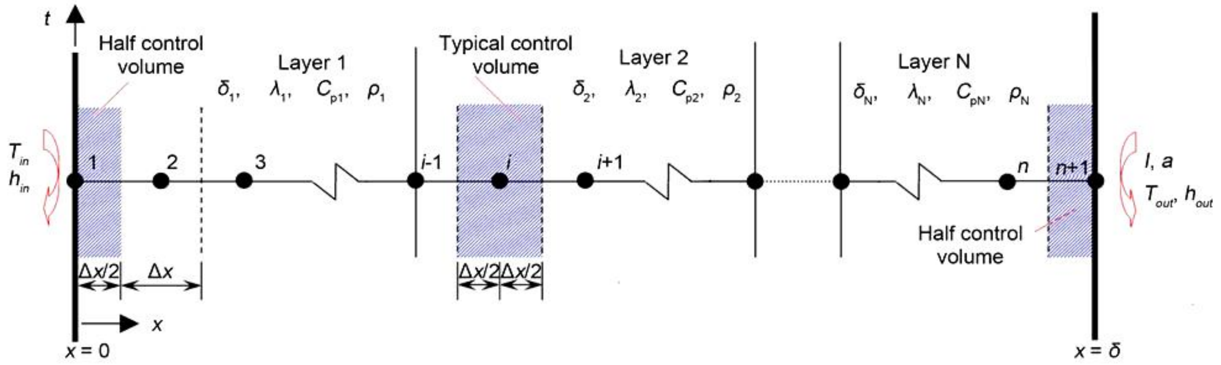

Fig 1: The influence of user location on QoS

When users and services log on to different networks, the two networks are far away from each other on the Internet, and the user's access experience is generally poor. This is due to the bandwidth limitation of different networks in communication and the delay caused by the storage and forwarding of information between different networks. On the contrary, when users and services are in the same network provided by the same network provider, users are likely to enjoy high-quality network services. Therefore, user location has a personalized impact on the QoS of Web services. Figure 5.1 vividly shows that user 1 and user 2 in the same network 1 can enjoy the high-quality access requirements of service 1 in network 1. However, user 3 in network 2 can only enjoy the service provided by service 1 in network 1 through Internet forwarding for many times, so the access of user 3 will be greatly affected.

4.2 Use of user location information

This section discusses how to use user location information, which is also the basis of our improved collaborative filtering recommendation algorithm based on user location and preference. We define the user's location generally bASed on three factors: the user's IP address, the user's access point as and the user's country. IP address specifies the user's address in the network, which is equivalent to the user's house number in the network world. The user's AS is the gateway for the user's IP address to intervene in the Internet, which is equivalent to the user's cell name in the network world. The user's country is not only the user's affiliation in the network but also the user's affiliation in the real world. IP address defines a user's unique network ID. existing Web service recommendation articles based on user's IP address use IP address as user's location information, and regard users with similar IP addresses as users of the same kind. based on this principle, users are clustered into several user clusters, and similar neighbors are calculated and selected in each user subset.

In the definition of IPv4, a user's IP address is represented by 32 bits in four fields, which are divided into a network field and a host field. Three types of addresses, A, B and C, are assigned to users. This is the standard set by the Internet at the beginning. Later, the development of the Internet can be described as explosive. With the access of a large number of hosts, the original address allocation method became unreasonable, and then the subnetting and classless inter-domain routing standards came into being. AS a result, two users who are very close in IP address do not strictly belong to the same AS or the same country. The IP address resources owned by AS863 in Canada are not continuous in network prefix, and the 4.67.64.0 network segment belongs to AS996 in Japan. This fact indicates that two users who are very close in IP address may not belong to one country, and it is unreasonable to use only IP address as user location information.

There are many AS in a country, one AS belongs to only one country, and thousands of AS are interconnected in the whole Internet. Generally speaking, the performance of network interconnection in AS is better than that between AS, because there are problems of transmission delay and link selection in communication between different AS. At the same time, the communication quality between AS with relatively close distance in Internet is better than AS with relatively long distance, which is due to the delay of routing relay in communication. Therefore, the distance at AS level has become one of the general criteria to measure the user's position. However, users in the same AS are not necessarily close in geographical location, and two users in the same city may access the Internet through different AS. This is because different network operators have different AS resources, and this phenomenon occurs when using different operators' services. Therefore, when we define the user location information, we need to comprehensively consider the user's IP address information, AS information and the user's geographical location information, that is, the user's country information.

V. Conclusion

The recommendation system based on collaborative filtering can predict the rating of current users according to the rating of similar users or similar services. There have been many research results in this research direction.

However, there are still many shortcomings in the accuracy of QoS data prediction and the reasonable line of personalized recommendation. On the basis of analyzing and summarizing the existing work, this paper focuses on the consideration of user preference in recommendation system, and proposes a series of effective personalized recommendation algorithms based on collaborative filtering and sparse QoS prediction.

Firstly, user preferences are extracted from QoS data and used as the selection criteria of similar users. Then, top-k algorithm is used to determine the target users and similar neighbors of services. Finally, the adjusted weighted sum method is used to predict the QoS value of target users. The adjusted weighted sum overcomes the problem of inconsistent QoS value range of different users or services, and improves the prediction accuracy.

Secondly, in order to solve the deficiency of using preference similarity to adjust weights and calculate the missing QoS prediction stage in collaborative filtering algorithm based on user preference. In this paper, an improved collaborative filtering Web service recommendation algorithm (CUPCF) based on joint user preference is proposed. Firstly, user preferences are extracted from QoS data and used as the selection criteria of similar users. Then, top-k algorithm is used to determine the similar neighbor sets of target users and services, and then QoS data is used to calculate the similarity of neighbors. Finally, the adjusted weighted sum method is used to predict the QoS value of target users. The adjusted weighted sum overcomes the problem of inconsistent QoS value range of different users or services, and improves the prediction accuracy.

Finally, considering that the preferences of the same service will vary greatly between two users whose positions are far apart. In order to overcome the disadvantage of low time efficiency of collaborative filtering algorithm based on historical data, and consider the influence of user location information on QoS prediction, we gather users who are close to each other. Considering the user's preference as well as the user's location attribute, an improved collaborative filtering Web service recommendation algorithm based on the user's location and preference is proposed.

Acknowledgements

This research was supported by key science and technology project of Henan Province (No.: 172102210298).

References

[1] Boulet, & Marie-Michele. Designing and developing an intelligent advisor system for transfer tasks in music. Computers & Education, vol.19, no. 4, pp. 341-357, 1992.

[2] Hodge, G. M., Jupp, J. J., & Taylor, A. J. . Work stress, distress and burnout in music and mathematics teachers. British Journal of Educational Psychology, vol.64, no. 1, pp. 65-76, 2011.

[3] Yang, L., Ketner, K., Luker, S., & Patterson, M. . A complete system for publishing music-related etds. Library Hi Tech, vol.34, no. 1, pp. 151-163, 2016.

[4] Hwong, N. C., Caswell, A., Johnson, D. W., & Johnson, R. T. . Effects of cooperative and individualistic learning on prospective elementary teachers' music achievement and attitudes. Journal of Social Psychology, vol.133, no. 1, pp. 53, 1993.

[5] Lyster, & Norman, C., The use of the music operating system to supplement the teaching of cobol. ACM Sigcse Bulletin, vol.18, no. 4, pp. 46-49, 1986.

[6] Oliveira, & A. Music teaching as culture: introducing the pontes approach. International Journal of Music Education, vol.23, no. 3, pp. 205-216, 2005.

[7] Ilhan Özgül. An analysis of the elementary school music teaching course in turkey. International Journal of Music Education, vol.27, no. 2, pp. 116-127, 2009.

[8] Maganioti, A.E., Chrissanthi, H.D. “Cointegration of Event-Related Potential (ERP) Signals in Experiments with Different Electromagnetic Field (EMF) Conditions”. Health, vol.12, no.2, pp.400-

406, 2016.

[9] Tianyi Qin, Drivers drowsiness detection in embedded system, IEEE International Conference on Vehicular Electronics and Safety, 2007. ICVES

[10] Yan Chen, Shunqing Zhang, Shugong Xu, G.Y. Li, ”Fundamental tradeoffs on green wireless networks,” Communications Magazine, IEEE , vol.49, no.6, pp.30,37, June 2011.