2012 IEEE 15th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops

A Clustering-based QoS Prediction Approach for Web Service Recommendation

Jieming Zhu†, Yu Kang†, Zibin Zheng†, and Michael R. Lyu†‡

†*Dept. of Computer Science and Engineering, The Chinese Univ. of Hong Kong, Hong Kong, China*
‡*School of Computer Science, National Univ. of Defence Technology, Changsha, China*
{*jmzhu, ykang, zbzheng, lyu*}*@cse.cuhk.edu.hk*

***Abstract*—The rising popularity of service-oriented architecture to construct versatile distributed systems makes Web service recommendation and composition a hot research topic. It's a challenge to design accurate personalized QoS prediction approaches for Web service recommendation due to the unpredictable Internet environment and the sparsity of available historical QoS information. In this paper, we propose a novel landmark-based QoS prediction framework and then present two clustering-based prediction algorithms for Web services, named UBC and WSBC, aiming at enhancing the QoS prediction accuracy via clustering techniques. Hierarchical clustering is adopted based on the real-word Web service QoS dataset collected with PlanetLab[1], which contains response-time values of 200 distributed service users and 1,597 Web services. The comprehensive experimental comparison and analysis show that our clustering-based approaches outperform other existing methods.**

***Keywords*-Web service recommendation; QoS prediction; clustering; landmark**

I. Introduction

Web services are self-contained and self-describing computational Web components designed to support machine-to-machine interaction by remote invocations [1], which are becoming a major technique for building service-oriented distributed systems and applications, such as e-commerce, automotive systems, multimedia services, etc [2].

According to the counter at seekda.com[2], there are totally 7,739 service providers and 28,606 public Web services in the Internet. With the rapidly growing number of Web services, evaluation and recommendation for numerous function-equivalent alternative Web services is thus becoming a crucial task. In order to address this problem, Quality-of-Service (QoS) is widely employed for describing and evaluating the non-functional characteristics of Web services, comprising response time, throughput, failure probability, reputation, etc [2]. Web service composition is to select and aggregate one or more Web services to construct a service-oriented system with good QoS performance and to meet different requirements of different service users and applications. However, the performance of service-oriented systems is greatly influenced by the unpredictable Internet environment and the locations of service users. Different service users may experience quite different QoS on the same Web service. When making service selection from a set of candidate Web services, QoS-aware Web service recommendation approaches provide useful information to assist service users to improve the performance of applications [2]. Personalized QoS-driven Web service recommendation is becoming a hot and challenging research problem in recent years.

The most straightforward approach of service selection is to monitor and evaluate all the candidate Web services at the user-side and select the Web service with the best QoS performance for Web service composition. However, this approach is impossible in practical, since some Web service invocations may be charged which heavily increase the cost of Web service users. In addition, it's time-consuming and resource-consuming to invoke all Web services each time, as there may exist a huge number of function-equivalent candidate Web services [3].

To address this challenge, several QoS prediction approaches [4], [5], [6], [7] have been proposed for Web service recommendation. These approaches mainly employ collaborative filtering (CF) based algorithms to model the user similarity and Web service similarity to predict the missing QoS values via sharing historical QoS information among users. However, these approaches heavily depend on the historical Web service invocation information. In reality, each user only invokes one or several Web services out of the numerous candidates at each time, resulting in the high sparsity of the available QoS data matrix. Besides, some historical QoS values are not updated in real-time or even out of date, which cannot provide accurate predictions as a result of the dynamic nature of the Internet environment.

To remit the problem of data sparsity, in this paper, we propose a novel clustering-based QoS prediction framework with a set of fixed landmarks (computers) distributed in the Internet. The landmarks can monitor available Web service periodically to enrich the QoS data for more accurate QoS prediction. Our approach take the real-time QoS information of the landmarks as a reference, and then apply clustering techniques to help service users to make QoS predictions. Extensive experiments are conducted using our real-world QoS dataset, which containing about 319,400 invocation result about response time from 200 service users on

[1]http://www.planet-lab.org
[2]http://seekda.com

978-0-7695-4669-8/12 $26.00 © 2012 IEEE
DOI 10.1109/ISORCW.2012.27

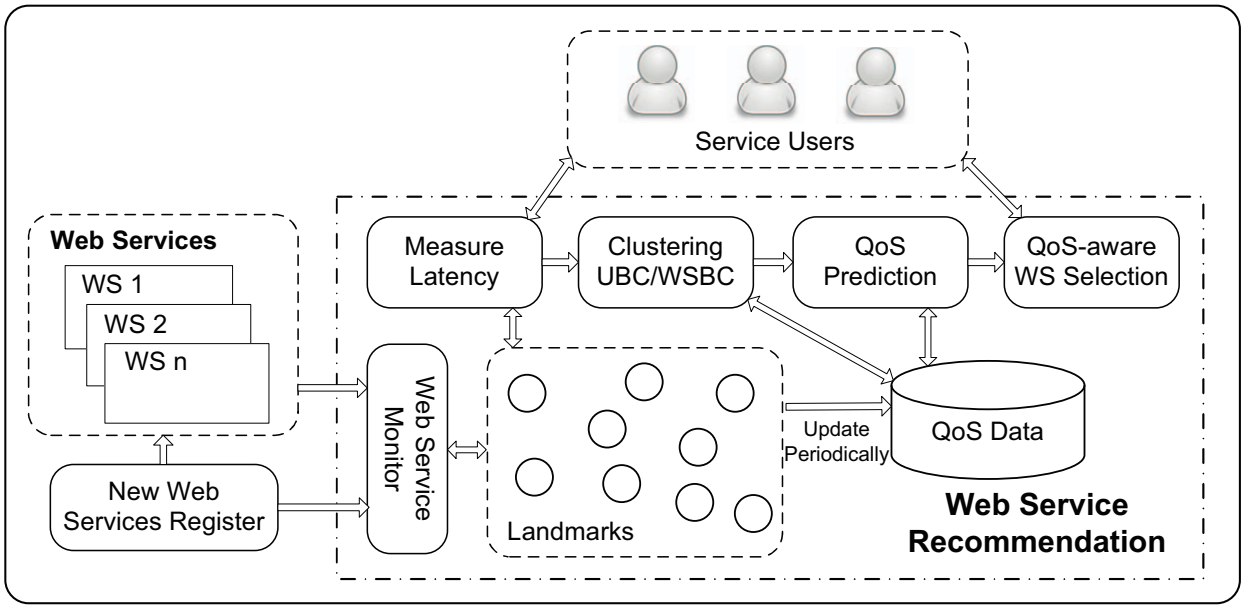

Figure 1. Web service Recommendation Framework

1,597 Web services. The experimental results show that our clustering-based approach outperform other collaborative filtering based algorithms.

The rest of this paper is organized as follows: Section II introduces the related work. Section III presents our landmark-based Web service recommendation framework. Section IV describes two clustering-based QoS prediction approaches. Section V discusses the experiments and results. Finally, Section VI concludes the paper.

II. Related Work

A number of QoS-aware approaches have been comprehensively investigated for Web service selection and recommendation in recent literature, e.g. [8], [9], [10]. It's time-consuming and resource-consuming for service users to monitor all the Web services in real-time, owing to the expensive Web service invocations and the enormous number of Web service candidates. Consequently, the Web service QoS prediction is proposed and gets much attention in research. QoS prediction for Web services targets at predicting the unknown QoS values between different service users and different Web services, with partially available information [2]. As a result, the optimal Web service with best QoS value can be recommended to the service user for composition.

Collaborative filtering, one of the most popular recommendation algorithms for commercial systems, is introduced to make QoS predictions for Web service recommendation, e.g. [4], [5], [6], [7], [11], [12]. Shao et al. [5] propose a user-based CF approach to predict the QoS values based on the similarity between service users. That is to employ the available QoS data of similar neighbors to predict the unknown QoS of each service user. WSRec [6] puts forward a hybrid collaborative filtering based QoS prediction approach by taking advantage of both the similarity information between service users and the similarity information between Web service items, which has been shown to achieve a good overall prediction accuracy. Here is a brief description of WSRec.

First, the similarities between service users and between Web service items are computed using the Pearson Correlation Coefficient (PCC) [6] of the historical QoS data. Then each user can find Top-K similar neighbors and each Web service item can find Top-K similar items. As a result, the user-based collaborative filtering methods use similar users to predict the unknown QoS values, while the item-based collaborative filtering methods employ similar Web service items to predict the unknown QoS values, as described in equation in the following equations, respectively,

$$p_u = \bar{u} + \frac{\sum_{u_a \in S_u} s(u_a, u)(q_{u_a i} - \bar{u}_a)}{\sum_{u_a \in S_u} s(u_a, u)}, \quad (1)$$

$$p_i = \bar{i} + \frac{\sum_{i_k \in S_i} s(i_k, i)(q_{u i_k} - \bar{i}_k)}{\sum_{i_k \in S_i} s(i_k, i)}, \quad (2)$$

where $\bar{u}$ and $\bar{i}$ are the average QoS values of service user u and Web service item i. $s(u_a, u)$ and $s(i_k, i)$ are the similarity between users and items. S_u and S_i are the Top-K similar set of users and items. $\bar{u}_a$ and $\bar{i}_k$ denote the average

QoS values of users and Web service items in the similar set. q_{ui} is the entry of user-item matrix of QoS data. p_u and p_i are the predicted QoS values with user similarity and Web service item similarity respectively.

In order to fully utilize the similarity information of the user-item matrix, WSRec propose to combine the user-based and item-based methods to improve the prediction accuracy. At last, the final prediction result is obtained by employing the following equation:

$$p = w_u P_u + w_i P_i. \tag{3}$$

The hybrid collaborative filtering method is the linear combination of P_u and P_i, where w_u and w_i are the corresponding coefficients, and $w_u + w_i = 1$. However, the collaborative filtering based approaches heavily depend on the historical QoS data. The historical Web service QoS data may be very sparse or not updated in real-time, which will lead to the inaccurate QoS predictions and Web service recommendations.

In this paper, we propose a clustering based QoS prediction approach to address this challenge. Our approach takes a set of fixed landmarks as references, which monitor QoS values of all the available Web services and update periodically. After clustering, we can use QoS information of the similar landmarks in one cluster to predict the QoS value of the users in the same cluster. There is no need for service users to take any historical QoS information to make predictions. Extensive experiments show that our approach outperforms the others.

III. Web Service Recommendation Framework

To address the limitations of collaborative filtering based approaches, we propose a novel landmark-based QoS prediction framework, as illustrated in Figure 1.

As we can see in the figure, the Web service recommendation framework mainly contains the following procedures: 1) The landmarks are deployed in the Internet, and monitor the QoS information of available Web services by periodical invocations. 2) Clustering the landmarks using the obtained QoS data. 3) Each service user measures the latencies to the landmarks and then be clustered into one exiting cluster. 4) QoS predictions are made by employing the QoS information of landmarks. 5) QoS-aware Web services selection and recommendation are made with the prediction results.

IV. QoS Prediction Algorithm

A. Landmarks Clustering

Given N_L landmarks are set up and distributed in the Internet, we denote $\{L_i, i = 1, 2, \cdots, N_L\}$ as the landmark set. We can use the landmarks to monitor the Web service and update the QoS values periodically. As the number of landmarks is small, the measurement overhead is acceptable. In this way, the QoS data (only response time is considered in this paper) between N_L landmarks and W Web services can be obtained, expressed as follows:

$$Q_L = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1W} \\ q_{21} & q_{22} & \cdots & q_{2W} \\ \vdots & \vdots & \cdots & \vdots \\ q_{N_L 1} & q_{N_L 2} & \cdots & q_{N_L W} \end{bmatrix} \tag{4}$$

Besides, in order to achieve more accurate clustering, we also measure the Round Trip Time (RTT) between the landmarks, which is a matrix as follows:

$$D_L = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N_L} \\ d_{21} & d_{22} & \cdots & d_{2N_L} \\ \vdots & \vdots & \cdots & \vdots \\ d_{N_L 1} & d_{N_L 2} & \cdots & d_{N_L N_L} \end{bmatrix} \tag{5}$$

1) User based Clustering (UBC): In this approach, we cluster the landmarks based on the latency matrix D_L. However, since the matrix provides the distance information (i.e. latency time) between the landmarks, which small distance represents high similarity, we then consider the hierarchical clustering algorithm [13] to cluster landmarks into N_C clusters. Hierarchical clustering method tries to build a hierarchy of clusters which could be organized as a tree structure. We summarize the clustering algorithm in Algorithm 1.

Algorithm 1: Hierarchical Clustering Algorithm: Hierarchical(D_L, N_C)

Input: Distance matrix D_L, cluster number N_C
Output: The specific N_C clusters and their members

```
proximity matrix D = DL;
assign each landmark to a single cluster;
N = N_L;
while N ≠ N_C do
    merge the two closest clusters;
    inter-cluster distance d_{c_i c_j} = min_{i∈c_i, j∈c_j} d_ij;
    update D with d_{c_i c_j};
    N=N-1;
end
```

2) Web Service based Clustering (WSBC): In this approach, however, we cluster the landmarks based on the N_L-by-W QoS matrix Q_L. Each row is considered a feature of a landmark. It's obvious that the dimension of the feature is W, with a set of homogenous attributes.

We take the Pearson Correlation Coefficient (PCC) as the similarity between two users. The PCC measure the linear relationship (i.e. correlation) between two users, which can be expressed as follows:

$$s_{ij} = \frac{\sum_{m=1}^{W} (q_{im} - \bar{q}_i)(q_{jm} - \bar{q}_j)}{[\sum_{m=1}^{W} (q_{im} - \bar{q}_i)^2 \sum_{m=1}^{W} (q_{jm} - \bar{q}_j)^2]^{1/2}}, \tag{6}$$

where $\bar{q}_i$ is the average of QoS value q_{im}, i.e. $\bar{q}_i = \sum_{m=1}^{W} q_{im} \Big/ W$. It's similar to $\bar{q}_i$. And s_{ij} denotes the similarity result.

However, the PCC values range from -1 to 1, where 1 denotes the most similarity and -1 denotes the most dissimilarity. Therefore, we transform the range via the following equation:

$$d_{ij}^{'} = 1 - s_{ij}, \tag{7}$$

which is usually referred as the distance between N_L landmarks. So we can obtain the distance matrix $D_L^{'}$ and cluster the landmarks using the hierarchical clustering algorithm, i.e. *Hierarchical($D_L^{'}$, N_C)*.

B. QoS Prediction

With the above cluster result, we then design to predict the QoS values between different service users and Web services to help service users to make Web service recommendation. In our QoS prediction framework, we don't employ the historical QoS data due to the dynamic Internet environment. However, each user measures the latency time to the landmarks so as to be clustered into one of the N_C clusters and make predictions based on the QoS information of the landmarks in the same cluster. Suppose the number of service users is N_U, the N_U-by-N_L latency matrix can be expressed as follows:

$$D_{UL} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N_L} \\ d_{21} & d_{22} & \cdots & d_{2N_L} \\ \vdots & \vdots & \cdots & \vdots \\ d_{N_U 1} & d_{N_U 2} & \cdots & d_{N_U N_L} \end{bmatrix} \tag{8}$$

In order to cluster each service user to a proper cluster, the distance between each user to each cluster can be obtained by taking the minimal distance to the landmarks in one cluster, according to the hierarchical algorithm. Then every service user can be merged into one cluster.

The QoS prediction of a service user depends on the landmarks in the same cluster, while the distance can be expressed as $\{d_{ul}, u = 1, 2, \cdots, N_U,\ l \in Cluster_u\}$, where $Cluster_u$ denotes the cluster of service u and l is the landmark in $Cluster_u$. Hence, the similarity between u and l can be denoted as $s_{ul} = 1/d_{ul}$, as small distance d_{ul} means high similarity. The unknown QoS values of service user u can, therefore, be predicted by the following equation:

$$p_{uw} = \frac{\sum\limits_{l \in Cluster_u} s_{ul} q_{lw}}{\sum\limits_{l \in Cluster_u} s_{ul}},\ w = 1, 2, \cdots, W \tag{9}$$

where q_{lw} denotes the QoS value when the landmark l invokes the Web service w and p_{uw} is the prediction result. This QoS prediction algorithm mainly enhances the prediction accuracy by deploying a set of landmarks in the Internet and monitoring the Web services periodically. The experiment results show that this new QoS prediction mechanism can significantly enhance the prediction performance for Web service recommendation and also update the QoS prediction results in real-time according to the new QoS information of landmarks.

V. Experiments

In this section, we evaluate the performance of the proposed clustering based QoS prediction approaches via a set of experiments and performance comparison with other existing prediction methods, using the measured QoS data between a large number of service users and Web services.

A. Data Collection

In this experiment, the data is a collection of the real-word Internet latency measurement, comprising the response time between 200 distributed nodes and 1597 Web services, and also the latency time between the 200 nodes.

To collect the data set, we have an access to the global research network, PlanetLab, which supports the development of distributed systems and networks, including distributed storage, network mapping, peer-to-peer systems, distributed hash tables, query processing and etc. Currently, PlanetLab consists of 1,093 nodes at 530 sites distributed all over the world, while most nodes are located in North America and Europe. We first get a list of the active PlanetLab nodes, about 588 nodes, after removing the inactive nodes which may shut down or get out of connection of the Internet. Considering that many nodes lying in the same lab must have similar performance, which may lead to the significant increase of QoS prediction accuracy but violate the practical condition, we remove these redundant nodes and get a new list of 288 nodes. Meanwhile, about 10,000 Web services' addresses are obtained from the release dataset at WSDream[3], which are mainly collected by crawling Web service information from the Internet. However, not all Web services are accessible or available to support the ICMP ping. Thus, we test and reserve 2,213 web services for our experiment.

To measure the distances between two hosts, we send 32 byte ICMP ping packets continually for 10 times and take the average round-trip time (RTT) from all replies as the response time. We use ICMP ping to measure the response time mainly based on the assumption that all the Web services are the same function-equivalent candidates and have the same service-running time. In this paper, we only focus on the latency time prediction under the dynamic Internet environment. After collecting the data collected on 288 PlanetLab nodes and omitting the failure ones, two data matrices are obtained. One is a 200-by-200 matrix of the RTTs between any two of 200 PlanetLab nodes, and the other is a 200-by-1597 matrix of the RTTs between 1,597 Web services and 200 PlanetLab nodes.

[3]http://www.wsdream.net

Table I
PERFORMANCE COMPARISON

Metrics	MAE	RMSE	MRE
UPCC	11.630	25.868	0.122
IPCC	9.155	25.103	0.087
WSRec	9.073	24.978	0.083
WSBC	8.182	24.468	0.080
UBC	**7.306**	**23.361**	**0.064**

B. Evaluation Metrics

To evaluate the QoS prediction performance, we employ several metrics, including *Mean Absolute Error (MAE)* [6], *Root Mean Square Error (RMSE)* [3], and *Median Relative Error (MRE)* for our experiments.

Mean Absolute Error (MAE) metric is widely employed to measure the prediction quality, which is defined as:

$$MAE = \frac{\sum_{i,j} |p_{ij} - q_{ij}|}{N}, \quad (10)$$

where p_{ij} denotes the predicted QoS value between service user i and Web service j, q_{ij} denotes the measured QoS value, and N is the number of predicted items.

Root Mean Square Error (RMSE) presents the standard deviation of the prediction error, which expressed as follows:

$$RMSE = \sqrt{\frac{\sum_{i,j} (p_{ij} - q_{ij})^2}{N}}. \quad (11)$$

Relative Error (RE) is also widely adopted for evaluating the prediction performance, in order to identify the error effect of different Internet latencies. Median Relative Error (MRE) is proposed to get the median recognition of all the RE values.

$$MRE = Median(RE_{ij}) = Median(\frac{|p_{ij} - q_{ij}|}{q_{ij}}), \quad (12)$$

which means 50% of the relative prediction errors are below *MRE*.

C. Performance Comparison

In order to study the prediction performance, we compare our approaches with the well-known collaborative filtering based prediction method (WSRec [6]), which is the hybrid recommendation approach combining both user-based CF algorithm and item-based CF algorithm.

In our experiment, we select 100 nodes as landmarks and the other 100 nodes as the service users, while the selection of landmarks will be discussed in Section V-D. And also the number of clusters, N_C, is set to 50. Table I shows the prediction accuracy of different approaches, where UPCC, IPCC and WSRec are collaborative filtering based prediction methods proposed by [6]. UPCC and IPCC denotes user-based CF method and item-based CF method, respectively. The results of UPCC, IPCC and WSRec are obtained under the 50% matrix density.

From the results in the table, we observe that both the UBC and WSBC methods have smaller MAE, RMSE and MRE values, outperforming the collaborative filtering based methods, and particularly UBC approach obtains the best prediction performance. The performance of the collaborative filtering based approaches highly depend on the Top-K similar neighbors, but the constant Top-K may lead to some inaccurate predictions influenced by the dissimilar neighbors in Top-K. In addition, we can't check whether the QoS predictions, using the the historical QoS data, are available in real-time. Hence the predictions are not confident enough due to the dynamic Internet environment from time to time. On the other hand, our approaches obtain higher accuracy because the number of similar neighbors in each cluster is different and adjustable for QoS prediction. UBC method achieves better performance than WSBC method, since UBC takes advantage of more latency information, i.e. the latency matrix D_L, where at the cost of more measurement overhead.

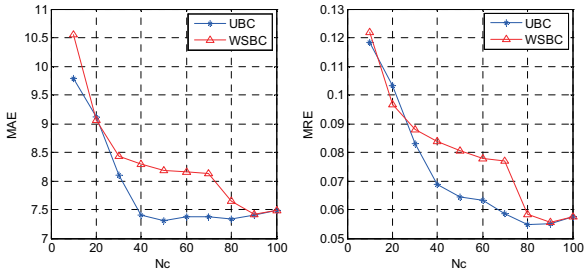

Figure 2. The Impact of N_C

D. Impact of N_C

The cluster number N_C plays an important role in the clustering-based QoS prediction algorithms. In order to study the impact of N_C, we vary the N_C from 10 to 100 and then present the prediction performance as shown in Figure 2.

We can observe that the prediction performance is highly influenced by N_C. For UBC method, the MAE attains its minimum when $N_C = 50$, and the MRE reaches the minimum when $N_C = 90$. On the other hand, for WSBC, both MAE and MRE obtain the optimum when $N_C = 90$. In general, the prediction errors first decrease when N_C increases, and then basically keep steady or slightly increase. This is because when N_C is small, the size of the cluster is too large. There may exist some landmarks in the same cluster of the service user that are not similar to this service user in fact. While N_C is too large, we can't find enough landmarks in one cluster to assist QoS predictions of service users. As a result, a optimal N_C will significantly enhance the prediction performance.

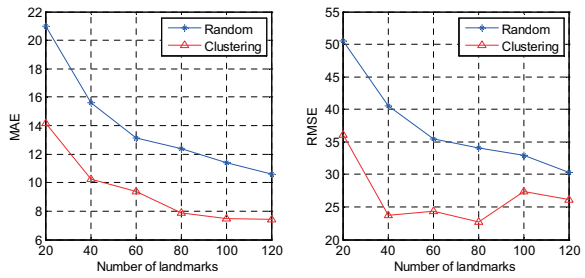

Figure 3. The Impact of Landmarks Selection

E. Impact of the landmarks selection

As the reference of service users, the deployment (i.e. the location and number) of landmarks in real world is very important. Ill-positioned landmarks can significantly degrade the performance of QoS prediction. In tuition, the landmark should be well separated to capture more information about the network topology. To study the impact of landmarks selection, we propose two selection methods, namely random selection and clustering based selection.

- **Random**: In this method, we randomly select landmarks by uniformly sampling from all candidate nodes, which means every node has the same probability to be selected as a landmark.
- **Clustering**: In order to well separate the landmarks, we cluster all candidate nodes into N_L clusters, and then select one landmark form each cluster. In this scheme, every pair of landmarks has a large average distance.

To compare the performance of different schemes, we conduct a set of experiments, and the results are illustrated in Figure 3. We can see that the clustering-based landmark selection scheme greatly outperforms the random method consistently for any number of landmarks, which verifies that well-separated landmarks can provide more information. Meanwhile, the MAE accuracy increases when the number of landmarks gets larger. This observation indicates that the prediction performance can be enhanced by deploying more landmarks as reference of service users.

VI. Conclusion

In this paper, we propose the landmark-based QoS prediction framework and clustering-based prediction approaches for Web service recommendation. By deploying a set of fixed landmarks in the Internet, the QoS data can be monitored and update periodically and then provides reference to service users to make more accurate QoS predictions. Experiment results show that our approaches outperform other existing collaborative filtering based prediction methods, and makes the QoS prediction more confident for Web service recommendation.

Acknowledgment

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 415410) and by a grant from the National Nature Science Foundation of China (No. 61100078). It was also sponsored in part by the National Basic Research Program of China (973) under Grant No. 2011CB302603.

References

[1] L. Zhang, J. Zhang, and H. Cai, *Services Computing: Core Enabling Technology of the Modern Services Industry*. Tsinghua University Press, 2007.

[2] Z. Zheng, "QoS management of web services," Ph.D. dissertation, The Chinese University of Hong Kong, 2011.

[3] Z. Zheng and M. R. Lyu, "Collaborative reliability prediction of service-oriented systems," in *Proc. of the ACM ICSE'10*, 2010, pp. 35–44.

[4] K. Karta, *An Investigation on Personalized Collaborative Filtering for Web Service Selection*, Honours Programme thesis, University of Western Australia, Brisbane, 2005.

[5] L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, and H. Mei, "Personalized QoS prediction for web services via collaborative filtering," in *Proc. of the IEEE ICWS'07*, 2007, pp. 439 –446.

[6] Z. Zheng, H. Ma, M. R. Lyu, and I. King, "WSRec: A collaborative filtering based web service recommender system," in *Proc. of the IEEE ICWS'09*, 2009, pp. 437–444.

[7] Q. Xie, K. Wu, J. Xu, P. He, and M. Chen, "Personalized context-aware QoS prediction for web services based on collaborative filtering," in *Proc. of the 6th Int'l Conference on Advanced Data Mining and Applications (ADMA'10)*, 2010, pp. 368–375.

[8] V. Cardellini, E. Casalicchio, V. Grassi, and F. L. Presti, "Flow-based service selection for web service composition supporting multiple QoS classes," in *Proc. of the IEEE ICWS'07*, 2007, pp. 743–750.

[9] J. E. Haddad, M. Manouvrier, G. Ramirez, and M. Rukoz, "QoS-driven selection of web services for transactional composition," in *Proc. of the IEEE ICWS'08*, 2008, pp. 653–660.

[10] Z. Zheng and M. R. Lyu, "A distributed replication strategy evaluation and selection framework for fault tolerant web services," in *Proc. of the IEEE ICWS'08*, 2008, pp. 145–152.

[11] X. Chen, X. Liu, Z. Huang, and H. Sun, "RegionKNN: A scalable hybrid collaborative filtering algorithm for personalized web service recommendation," in *Proc. of the IEEE ICWS'10*, 2010, pp. 9–16.

[12] Y. Jiang, J. Liu, M. Tang, and X. F. Liu, "An effective web service recommendation method based on personalized collaborative filtering," in *Proc. of the IEEE ICWS'11*, 2011.

[13] V. K. Pang-Ning Tan, Michael Steinbach, *Introduction to Data Mining*. First Edition. Addison Wesley Longman Publishing Co. Inc., 2005.