Personalized and Accurate QoS Prediction Approach Based on Online Learning Matrix Factorization for Web Services

Jian-Long XU[1,2*], Chang-Sheng ZHU[1]

[1]*Shantou University, Shantou, P.R.China*
[2]*Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R. China*
cszhu@ stu.edu.cn
** Corresponding author: xujianlong@stu.edu.cn*

Abstract: Quality of Service (QoS) prediction has played an important role in service computing. However, in the real-world scenario of Web service, many user-observed QoS values are unknown and vary over time. In order to provide high accurate and efficient QoS prediction performance for Web services, we propose a personalized and accurate QoS prediction approach namely PAOMF. Our prediction model is built by employing matrix factorization and online stochastic gradient descent algorithm. Extensive experiments are conducted on real world public datasets, which demonstrate the effectiveness and efficiency of our proposed approach.

1 Introduction

Quality of Service (QoS) of Web services has been widely concerned and researched [1-2] in recent years. QoS is widely used evaluation to select suitable service from many services with similar or equivalent functionality [3]. Users (applications that invoke the services) can select optimal service by ranking the QoS of services. Many QoS-based approaches have been proposed for Web service composition [4], Web service selection [5], fault-tolerant Web services [6], etc. Accurate QoS values of Web services are desired to work well for these approaches. However, in many cases, QoS properties (e.g., response time, invocation failure rate) observed by different user (e.g., located in different geographical location) are usually different. Additionally, it is expensive and too time-consuming for users to directly invoke all of Web services. QoS values of most services are unknown and unfixed. There are not sufficient personalized user-observed QoS for users to select optimal Web services. Moreover, the QoS properties may vary when users invoke the same Web service over time. The users need new and optimal Web services to replace low quality Web services with better ones. Therefore, it is an urgent task to predict the unknown QoS based on known QoS, which can be able to guarantee the accuracy performance.

To address the problems above, this paper presents a personalized real-time QoS prediction approach based on online learning matrix factorization for Web services, named PAOMF. In this approach, we build a QoS prediction model by employing matrix factorization and online stochastic gradient descent algorithm. We conduct extensive experiments in real world public datasets and compare with other well-known methods.

The paper is organized as follows. Section 2 presents related work. Section 3 describes our approach. We demonstrate the experimental results in Section 4. Finally, we conclude the whole paper in Section 5.

2 Related Work

Personalized QoS prediction methods for Web service have caused much attention recent years. In most of the existing reports, many researchers explore to obtain high accurate prediction result, and the popular method is collaborative filtering (CF). CF can be divided into neighborhood-based CF and model-based CF. Typical neighborhood-based CF are UPCC (User-based Pearson Correlation Coefficient), IPCC (Item-based Pearson Correlation Coefficient) and UIPCC (IPCC+UPCC) [7]. Considerable research has been conducted based on these methods. Zheng et al. [8] proposed a hybrid collaborative filtering algorithm which combines UPCC and IPCC. Ma et al. [9] presented a highly accurate prediction algorithm (HAPA) to predict unknown QoS values by keeping the original linear relationship. However, neighborhood-based CF method predicts QoS by employing the values of similar users or similar items. When the QoS data are very sparse, the prediction accuracy is not good.

Matrix factorization (MF) is a typical model-based CF approach. MF-based QoS value prediction method is to train a model according to the available QoS data in the user-service matrix. MF-based method has been

widely applied in Web service QoS value prediction. Tang et al. [10] proposed a network-aware web service QoS prediction approach by integrating MF with the network map. Su et al. [11] proposed neighbor information combined non-negative matrix factorization algorithm by utilizing the information of the observed data. He et al. [12] proposed a location-based hierarchical matrix factorization (HMF) method to perform personalized QoS prediction by using the global QoS matrix and local QoS matrices. In our previous work [13], we presented reputation-based Matrix factorization (RMF) method which integrated MF with reputation to achieve accurate unknown QoS values prediction results. Memory-based methods are easy to implement and understand. Relative to neighborhood-based CF, MF can achieve better performance. In this paper, we focus on MF to construct the QoS prediction model.

In recent years, online learning has received emerging attention. Existing method [10, 11, 12, 13] above are based on batch learning techniques which generate the predictor by learning on the entire training data set at once. When the data come sequentially, batch learning method cannot update the prediction model in time. The online learning method is an effective way to handle large scale data, especially streaming data. It can quickly adjust the model to reflect the change of the data timely and improve the online prediction accuracy [14]. Many researchers focus on integrating online learning with collaborative filtering. Abernethy et al. [15] present an algorithm for learning a rank-k matrix factorization online for collaborative filtering tasks. Qiao et al. [16] present an online nonparametric max-margin matrix factorization for flexible recommendation. Lin et al. [17] present First Order Sparse Collaborative Filtering (SOCFI) and Second Order Sparse Online Collaborative Filtering (SOCFII) to deal with the user-item ratings for online collaborative filtering. In this paper, we study online learning algorithms to solve the issues facing batch-trained MF algorithms and integrate online learning with MF for QoS values prediction in Web services.

3 Personalized and Accurate QoS Prediction Approach Based on Online Learning Matrix Factorization

To provide high performance prediction service, we design a personalized and accurate QoS prediction framework which is based on online learning matrix factorization (PAOMF). Figure 1 shows the system architecture of our framework.

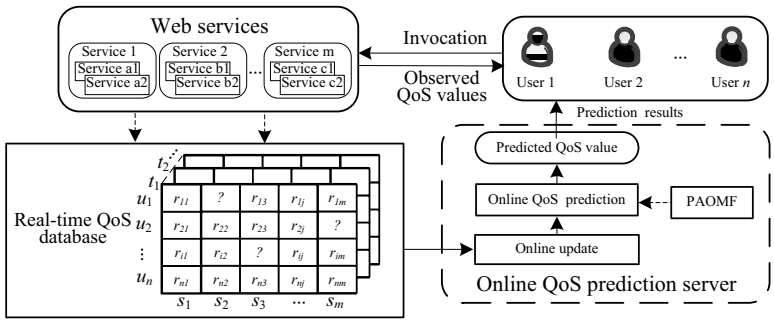


Figure 1. Framework of QoS prediction

The framework works as follows: 1) The online QoS prediction server collects the user-observed QoS data in real-time and save them to database. These data are transformed to normalized QoS data. 2) The PAOMF model performs update if new data come. 3) Online QoS prediction server prediction the unknown QoS value and returns prediction results to the target users who can use these QoS values to invoke the optimal Web services. Our goal of QoS prediction is to employ the observed QoS data to estimate the unknown values at each time slice. Because a user may invoke a few servers (not all the servers) and the quantity of QoS value obtained is limited, so many entries in the user-service-time invocation matrix are unknown. Thus, our main task is to fulfill unknown values in the matrix. However, since the QoS values may vary over time, the prediction model must adapt this condition and work effectively.

Let $U=\{u_1, u_2,\ldots, u_m\}$ be the set of m users, $S=\{s_1, s_2,\ldots, s_n\}$ be the set of n services, R be a $m\times n$ user-service sparse matrix $R \in \mathbb{R}$, r_{ij} ($i\le m, j\le n$) represent the value of a certain QoS property (e.g., response time, throughput) at a certain time slice. We can decompose QoS value matrix to multiplication of two low dimensional matrices with the following equation:

$$R \approx \tilde{R} = U^T S \tag{1}$$

where $U \in \mathbb{R}$ denotes the feature matrice of user U, and $S \in \mathbb{R}$ represents service latent feature matrices, the factor l is called dimensionality [7]. $\tilde{R} = \left\{U_i^T S_j\right\}_{n\times m} = \left\{\tilde{r}_{ij}\right\}_{n\times m}$ $(1 \le i \le n,\ 1 \le j \le m)$ is the approximate matrix of R. U_i and S_i denote the $i^{\text{th}}$and j^{th} column of U and S, respectively. In the real-time condition, we suppose the new coming QoS value is $(i, j, r_{ij}, t_{ij}) \in R_t$, where t denotes each time slice, R_t is the QoS value matrix at t slice. The objective function of MF for personalized QoS prediction can be represented as:

$$\min_{U,S} \mathcal{L} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} E_{ij}\left(r_{ij} - U_i^T S_j\right)^2 + \frac{\lambda_u}{2}\|U\|_F^2 + \frac{\lambda_s}{2}\|S\|_F^2 \quad (2)$$

where E_{ij} is a indictor function, whose value is 1 if r_{ij} is known or otherwise. $\|\cdot\|_F$ represents Frobenius norm which is employed to avoid the over-fitting issue during the learning process. λ_u, λ_s are both small positive decimals. In Eq.2, the first term presents the squared error between the observation and predicted value, and the last two terms are the corresponding regularizations. In order to get a local minimum of Eq.(2), we employ online algorithm named stochastic gradient descent algorithm and obtain the following update equations:

$$U_i \leftarrow U_i - \alpha(I_{ij}(U_i^T S_j - r_{ij})(U_i^T S_j)^{'} S_j + \lambda_u U_i), \quad (3)$$

$$S_j \leftarrow S_j - \alpha(I_{ij}(U_i^T S_j - r_{ij})(U_i^T S_j)^{'} U_i + \lambda_s S_j), \quad (4)$$

where α is the learning rate. The main idea of personalized and accurate QoS prediction based on online learning matrix factorization (PAOMF) algorithm can be simply described as follow: at each time slice, when a new data sample (i, j, r_{ij}, t_{ij}) comes, PAOMF performs online updating on its corresponding factors U_i and S_j using Eq. (3) and Eq. (4).

4 Experiments

As illustration in section 3, our main task is to employ the observed QoS data to estimate the unknown values at each time slice. In this section, we conduct the experiments and compare the prediction accuracy of our approach with other methods. We also discuss the key parameters which impact the prediction model.

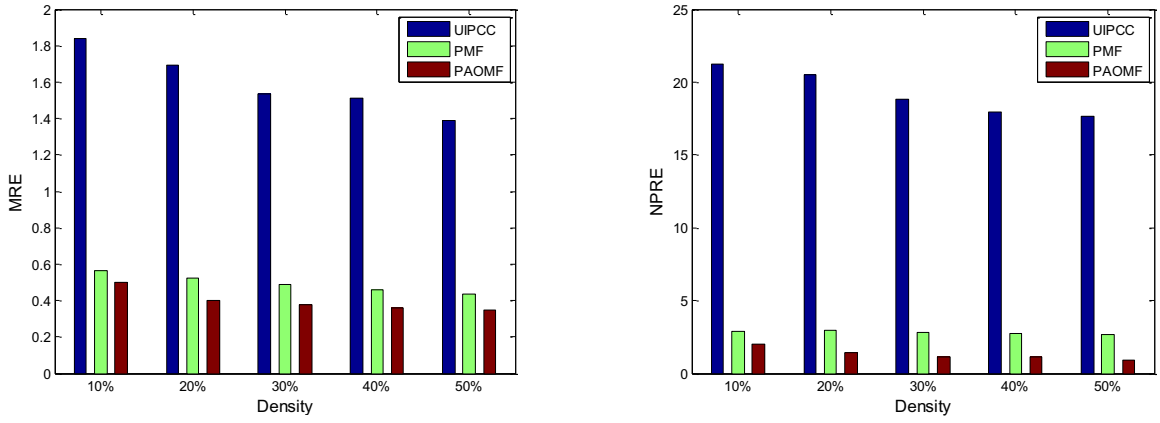
In this paper, the real world Web service QoS datasets released by Zheng et al. [18] are used to conduct all experiments. These released datasets are obtained based on PlanLab [18]. The datasets contain 142 users and 4,500 Web services for 64 consecutive time slices, at an interval of 15 minutes, and their corresponding QoS values are response time and throughput. In our experiments, we use the throughput datasets to verify our approach. This datasets can be expressed as a $142 \times 4{,}500 \times 64$ matrix.

In the experiments, the evaluation metrics of prediction accuracy are MRE (Median Relative Error) and 90% NPRE (Ninety Percentile Relative Error) which are defined as follows:

$$MRE = \underset{I_{ij}=0}{median}(\left|\tilde{r}_{ij} - r_{ij}\right| / r_{ij}) \quad (5)$$

$$NPRE = 90\% \times \left|\tilde{r}_{ij} - r_{ij}\right| / r_{ij} \quad (6)$$

We conduct compare our method with UIPCC [7] and probabilistic matrix factorization PMF [19] which employ batch learning to update the prediction model. We use different matrix density whose densities are 10% to 50% at a step increase of 10%. The dimensionality is set to 10. λ_u and λ_s are set to 30 and 0.001with PMF and PAOMF, respectively. The learning rate is set to 0.01. Figure 2 shows the MRE and NPRE results of different methods with different density.



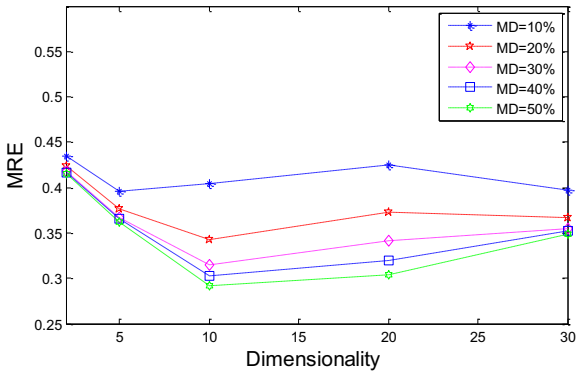
(a) MRE for throughput

(b) NPRE for throughput

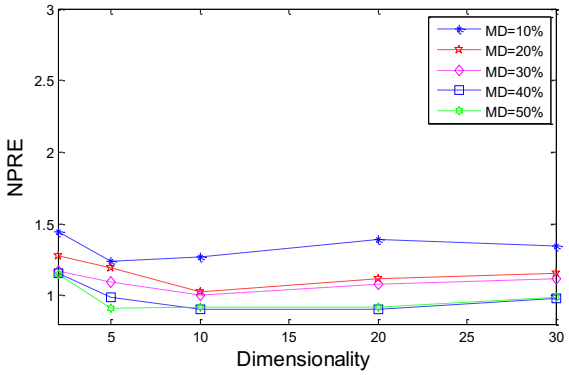
Figure 2. Accuracy comparison

The experimental results show that that no matter what the matrix density is, our PAOMF approach has smaller MRE and NPRE values relative to PMF for throughput with different matrix densities. In average sense, PAOMF can achieve 53.2% improvement in MRE and 63.5% improvement in NPRE than PMF model, which indicates further higher accuracy and effectiveness of our approach. Due to the sparsity of data, UIPCC has the lowest accuracy. As opposed to PAOMF with dynamically adapting to new patterns, PMF has a declining accuracy of the predictions after each factorization since its static nature.

To study the impact of dimensionality, we assess how many potential dimensionalities in the model learning is enough to character user and service latent features. We conduct experiments using different number of latent feature in the model by varying the value of dimensionality from 2 to 30.



(a) MRE for throughput

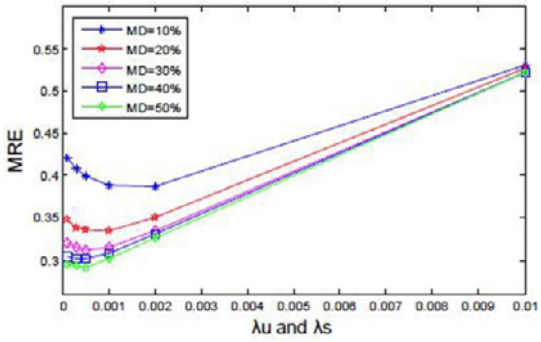


(b) NPRE for throughput

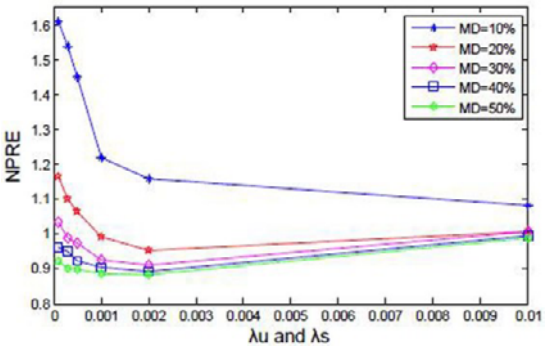
Figure 3. Impact of dimensionality

Figure 3 (a) and (b) shows the impact of dimensionality on MRE and NPRE, respectively. Generally, a higher dimensionality means the more latent features are used to characterize users and services for training the prediction model, which may enhance the prediction performance. However, we notice that: 1) MRE and NPRE drop quickly when the dimensionality increases from 2 to 10. 2) When the dimensionality is larger than 10, MRE and NPRE increase slowly with the increasing of dimensionality. This is due to the fact that too many latent features might cause the over-fitting problem which will do harm to the performance. Furthermore, the higher value of dimensionality means the more time of learning these features. Therefore, too small or large dimensionality value will affect the prediction accuracy and efficiency. It seems that the best value of dimensionality is about 10 in this experiment. In other experiments, we set dimensionality =10.

To study the impact of λ_u and λ_s, we also conduct some experiments. Similar to dimensionality, λ_u and λ_s are used to avoid over-fitting through controlling the proportion of the two regularization terms which are used to in Eq. (2). In this experiment, we assume $\lambda_u = \lambda_s$, and set λ_u and λ_s from 0.0001 to 0.01, vary the density from 10% to 50% with a step value of 10% for each matrix corresponding each time slices. Figure 4 shows that the experimental results.



(a) MRE for throughput



(b) NPRE for throughput

Figure 4. Impact of λ_u and λ_s

From Figure 4, we can observe that: 1) When the matrix density increases, performances of MRE and NPRE all improve. 2) With the increasing of λ_u and λs, the overall accuracy of prediction result first increases, then drops after reaching an optimal value. 3) If λ_u and λs are large (e.g., $\lambda_u = \lambda_s$=0.01) or too small (e.g., $\lambda_u = \lambda_s$=0.0001), the prediction accuracy is unsatisfactory. 4) The optimal value seems to be about 0.0005 for MRE and 0.002 for NPRE. Therefore, the optimal value of λ_u and λ_s can be set in accordance with the matrix density and the evaluation metrics.

5 Conclusion and Future Work

In order to provide high performance QoS prediction result for Web services in real-time condition, we design a personalized and accurate QoS prediction approach based on online learning matrix factorization (PAOMF). In this approach, we build a prediction model based on matrix factorization and online stochastic gradient descent algorithm. Sufficient experiments based on real-world datasets show that our model can achieve 53.2% improvement in MRE and 63.5% improvement in

NPRE than PMF model, which indicates the outstanding performance our approach. In future, we plan to employ some techniques to further improve the prediction performance, such as clustering techniques, taking cold start into consideration, and so on.

Acknowledgment

This research was financially supported by the Guangdong High-Level University Project "Green Technologies for Marine Industries", Guangdong Common Colleges Young Innovative Talents Project (No. 2016KQNCX056), and Shantou University National Fund breeding project (No.NFC16001).

References

1. Mousa A, Bentahar J, An Efficient QoS-aware Web Services Selection Using Social Spider Algorithm, Procedia Computer Science. 94 (2016) 176-182.
2. Christi J C R, Premkumar K, Survey on recommendation and visualization techniques for QOS-aware web services, IEEE International Conference on Information Communication and Embedded Systems, 2015, pp.1-6.
3. Yu T, Zhang Y, Lin K J, Efficient algorithms for Web services selection with end-to-end QoS constraints, Acm Transactions on the Web, 1 (2007), Article 6.
4. Karimi, Mohammad Bagher, A. Isazadeh, and A. M. Rahmani, QoS-aware service composition in cloud computing using data mining techniques and genetic algorithm, Journal of Supercomputing (2016) 1-29.
5. Wang S, Hsu C H, Liang Z, et al. Multi-user web service selection based on multi-QoS prediction[J]. Information Systems Frontiers, 16 (2014) 143-152.
6. Zhu X, Qin X, Qiu M. QoS-Aware Fault-Tolerant Scheduling for Real-Time Tasks on Heterogeneous Clusters, IEEE Transactions on Computers, 60 (2011) 800-812.
7. Zibin Zheng, and M.R. Lyu, Personalized Reliability Prediction of Web Services, ACM Transactions on Software Engineering and Methodology (TOSEM), 22 (2013) 12:1-12:25
8. Z. Zheng, M. Lyu, and I. King, "Wsrec: A collaborative filtering based web service recommender system," in Proceedings of the 2009 IEEE international Conference on Web Services, 2009, pp. 437–444.
9. Ma, Y., Wang, S., Hung, P. C. K., Hsu, C. H., Sun, Q., & Yang, F. A highly accurate prediction algorithm for unknown web service qos values. IEEE Transactions on Services Computing, 9 (2015) 511-523.
10. Tang M, Zheng Z, Kang G, et al. Collaborative Web Service Quality Prediction via Exploiting Matrix Factorization and Network Map. IEEE Transactions on Network & Service Management, 13 (2016) 126-137.
11. Su K, Ma L L, Sun Y F, et al. Non-negative matrix factorization model for Web service QoS prediction, Journal of Zhejiang University (Engineering Science Edition, 49 (2015) 1358-1366.
12. Pinjia He, Jieming Zhu, Zibin Zheng, Jianlong Xu, and Michael R. Lyu, Location-based Hierarchical Matrix Factorization for Web Service Recommendation, International Conference on Web Services, Alaska, 2014, pp.297-304.
13. Jianlong Xu, Zibin Zheng, and Michael R. Lyu, Web Service Personalized QoS Prediction via Reputation-based Matrix Factorization, IEEE Transactions on Reliability, 65 (2016) 28-37
14. Shalev-Shwartz S: Online Learning and Online Convex Optimization. Foundations and Trends in Machine Learning, l4 (2011) 107-194.
15. Abernethy J, Canini K, Langford J, and Simma A, Online Collaborative Filtering Di.ens.fr, 2011, pp. 271-280.
16. Zhi Qiao, Peng Zhang, Wenjia Niu, Chuan Zhou, Peng Wang, Li Guo, Online Nonparametric Max-Margin Matrix Factorization for Collaborative Prediction, 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, 2014, pp.520-529
17. Lin F, Zhou X, Zeng W. Sparse Online Learning for Collaborative Filtering. International Journal of Computers Communications & Control, 2016, 11(2):248.
18. Yilei Zhang, Zibin Zheng, Michael R. Lyu: WSPred: A Time-Aware Personalized QoS Prediction Framework for Web Services. In: the 22th IEEE Symposium on Software Reliability Engineering (ISSRE), Los Alamitos, California, 2011, pp. 210-219.
19. R. Salakhutdinov and A. Mnih. Probabilistic Matrix Factorization. Advances in Neural Information Processing Systems, 2007, pp. 1257-1264.