Predication of Quality of Service (QoS) in Cloud Services: A Survey

Muayad N. Abdullah¹,a, Wesam S. Bhaya¹,b
¹College of IT, University of Babylon, Iraq.

*muayad.masudi@uobabylon.edu.iq; bwesambhaya@uobabylon.edu.iq

Abstract. Quality of Service (QoS) in cloud computing and web service has been commonly used as an essential measure for evaluating services of nonfunctional attributes. However, QoS prediction has become one of the key factors in services recommendation and selection to build high-quality software systems, so it becomes an urgent and crucial research area in both academia and industry. Therefore, several methods and techniques of service recommendation based on collaborative filtering (CF) have been exist for optimizing service selection. We investigate the different exist methods to solve challenging problems. Thence, we discuss and compare these approaches to provide an overview of the latest researches in the field.

1. Introduction

Recently, cloud computing has become more popular and scalable services [1]. However, which is considered as a new type of Internet-based computing that provides resources such as software, platform, and infrastructure as services [2]. Service-Oriented Architecture (SOA) is one the most technical foundation in cloud computing and web services, whereby a very large number of web services offered by various service providers are discovered and to allow for integration with other systems over the Internet [3]. Therefore, more and more competitive services with the same functionality emerge in recent years. Thus, one of the crucial issues in SOA is the selection of the best web services from a large number of the same functionality Web services [4].

Quality of service (QoS) is a quality measure that refers to a set of features for evaluating the performance of candidate services. However, QoS becomes an important research topic in cloud computing for service selection and optimal service composition [5] [6].

In a cloud environment, there are two different types of attributes which utilized to evaluate Quality of Services (QoS) for cloud computing services, functional and non-functional attributes. Functional attributes refer to the performance of QoS services on the service provider side such as availability, price, and reliability that is usually unchangeable for different users, while non-functional attributes refer to QoS observation for invoicing services at the user side such as response time, throughput, which is not identical for different users due to the unpredictable links at network environment and diverse in contexts [1] [4].

Due to the unpredictable QoS values observed by different users at the user side, several methods and approaches have been proposed to improve the whole quality of candidate services based on non-functional features.

There are many issues that need to be addressed when building high-quality cloud services. First, without sufficient QoS information from the users’ perspective, it is quite difficult for conducting a real-
world evaluation on the candidate services. Therefore, in practice, it is time-consuming or even impossible to invoke a large number of services. Second, it is too expensive, since cloud providers may charge for their services invocations. Third, unpredictable network links and the distribution of different user locations are highly influenced in user-side QoS observation of services [1] [7].

Typically, predicting process is required to build recommendation systems. Actually, there are two primary techniques in recommendation system field: Content based approaches and Collaborative Filtering approaches. 1- Content based approaches concentrate on the properties of items, where the similarity of items is evaluated by the equivalence of their own properties. 2- Collaborative filtering approaches concentrate on user-to-items relationship where the similarity is determined by measuring user/service similarity [8]. However, Collaborative filtering (CF) approaches are commonly used in a recommender system, where services are recommended to users on the basis of similarity measures between users and/or services [4] [9].

2. Collaborative Filtering
Collaborative filtering is a widely known method of predicting unknown ratings based on user/item similarity. Items that are chosen by similar users are recommended to an active user. However, this methods are typically employed utility matrix as data represented by two entities refer to users and items [8] [10]. Collaborative filtering was used for the first time by Shao et al. to make QoS predictions, using user-based CF algorithm to predict missing QoS values [11].

Generally, QoS prediction methods for candidate services can be classified into three main categories: Time series, context-based and Collaborative Filtering CF-based methods [1] [12]. There are also two classes for CF methods: memory-based (neighborhood-based) and model-based. [2].

2.1. Memory-based CF methods
Memory-based CF algorithms are one of the most common methods in CF-based systems. However, these methods have two essential steps for predicting QoS values: The first step is the measurement of similarities and the second step is the prediction of unknown QoS values by the significance of similar users or services. However, unknown QoS values are predicted by using similarity measurements in historical usage experience, based on observing QoS for similar users and/or similar services. Therefore, data-sparse in QoS utility matrix is the critical challenge in existing approaches, which It leads to inaccurate predicting unknown QoS values when calculating the similarity of users and/or services [13][14]. Typically, Pearson correlation coefficient (PCC) algorithm has been utilized in memory-based methods [15].

Zheng et al. presented a hybrid collaborative filtering algorithm (UIPCC) to predict unknown QoS values, which actively incorporates user-based and item-based methods to fully exploit user-item matrix relevant data. If the missing value does not have similar users, they employed information from similar items to predict the missing value, and vice versa [13].

Overall, memory-based algorithms are relatively simple to implement and have reasonable results when the user-item matrix is almost complete, but there are several challenges that have had an impact on accuracy: data sparsity problem, cold-start problem, scalability and trust of users [4] [16] [17].

2.2. Model-based CF methods
Model-based CF methods, on the other hands, are highly efficient for prediction of unknown QoS values, since they assume that a limited number of features have an effect on the efficiency of the service. However, these techniques are introduced to solve memory-based methods problems. A model predicts missing entries by creating a pattern and then learning from training dataset [18].

There are many techniques including in these approaches, such as latent factor models (Matrix Factorization), clustering model, and time series [4].

2.2.1. Latent factor (Matrix Factorization) models. Matrix Factorization algorithm is the dominant mode-based technique used for predict unknown QoS values. However, this technique is most suitable
for low-dimensional feature vectors [8]. The primary benefit of this approach is that data sparsity and scalability limitations can be somewhat overcome [1]. The key idea of MF is to eliminate the less important factors in the user-service matrices. Therefore, the prediction missing QoS influencing by only a small number of highly important factors [8] [19]. Zhang et al. have proposed a framework called WSPred to predict missing QoS values, which employed latent features of users and service and time by conducting tensor factorization [20]. Xie et al. proposed an asymmetric correlation MF framework to address data sparsity to achieve acceptable degree of prediction accuracy in service recommendation, which assume there is exist a hidden correlation between users and services when analysis a historical real word QoS dataset [21]. Kai et al. presented a hybrid approach which is a combination between model-based and memory-based methods. This hybrid approach incorporated the MF with a neighborhood-based model to alleviate matrix sparsity and improve predictive accuracy [22].

2.2.2. Time series. Due to the dynamic environment of the network and its variable traffic, some QoS attributes, for example response times and throughput, have different values at different times. In such conditions, When the value of a property depends on time, time series analysis can be a good predictive tool. In these approaches, the values of a QoS attribute are predicted by the previous attribute values of that attributes. Zadeh et al. have presented a method, called Time Series Forecasting (TSF) to predict response time based on previous service behavior, using a neural network of three layers [23]. Hu et al. proposed a method that took into consideration both the time variance of QoS attributes and the individual user factors. In this method, Where ARIMA model and kalman filtering have been combined to provide a time-based prediction model and overcome ARIMA’s weaknesses. This modified method indicate that sufficient past usage experiences are necessary to ensure efficient and accurate QoS prediction [24].

2.2.3. Clustering algorithms. Clustering is the process by which a set of objects are examined and grouping it into clusters according to a certain distance scale, where similar objects are placed in a cluster and objects are not same to the other clusters [8].These approaches minimize data volumes by clustering data, increasing scalability and reducing calculation time, the problem of the cold start is also somewhat resolved. Thus, clustering is combined with a collaborative filtering method for predicting missing QoS values. Pinjia et al. have proposed a method which employed a K-means clustering algorithm for cluster users and services according to their geographical location, and MF was then used to estimate missing QoS values for some user/service groups in order to make a local prediction. However, this method combines global and local predicting missing QoS values. Therefore, by using geographic location, this technique could achieve good accuracy [25]. Chen et al. have proposed an approach called Credibility-aware prediction (CAP) that used two-phase K-means to identify unreliable users, and predict unknown QoS value according to credible clustering information [9].

3. Context-based method
There have been a number of a hybrid approach for prediction missing QoS values that combine collaborative filtering with context-based method. In these hybrid methods, Context information can help to improve prediction accuracy [26]. Geographical location for user or service and invocation time are the most important contextual information used in these approaches. Therefore, due to the dynamic network environments as well as geographical location for users and services, for the same service, different users will have different QoS values. Additionally, unreliable data which contributed by untrusted users is a critical problem in collaborative filtering methods. Thus, geographical location, invocation time and trust are three important contextual factors in predicting missing QoS values [18].

3.1. Location-aware methods
Studies and experiments have shown that the services' QoS values are strongly influenced by network conditions. Since users use various network links for invoking services, their observations of services are different in several QoS attributes such as throughput, response time and availability [1]. Therefore,
users who are located in the same geographic location have a similar QoS value for the same service, since they have a similar distance for their service provider and similar infrastructure and users located at different locations have observed different QoS values for the same service [27].

Kuang et al. have presented a context-aware method CASR that combined context information such as location, network speed and invocation time with Bayesian inference to predict missing QoS values-based CF memory-based approach. However, the algorithm proposed consist of two steps. First, clustering the past QoS experiences by their context and then use the closest cluster to the active user. Second, Missing QoS values are estimated on the basis of filtered records by Bayesian inference [26]. In another study, the user’s geographic location and service invocation context have been combined with a model-based CF method using a matrix factorization approach [28]. Xu et al. have presented a context-based approach to predict QoS values where user context and service context were combined to make a more accurate prediction. In this method, the prediction of missing QoS values is based on the geographical location as context information on the user side as well as the country and company affiliation as context information on the service side. This context information from the user and the service is then combined with the MF approach [27].

In general, the above studies have shown that the use of context information can reduce the problems of data sparsity of the matrix, scalability and cold start and improving prediction accuracy.

3.2. Trust-aware methods
As we have seen in the most previous research studies, it is assumed that every user is equal in credit, and submitted values are also valid. However, in fact, this is not correct, users may submit unreliable QoS values. Therefore, unreliable data is one of the challenges with collaborative filtering approaches. As a result, user reputation is usually considered as a factor [29]. Some approaches have been introduced to improve the prediction of unknown QoS values, and to obtain more accurate results.

Qiu et al. presented a reputation-aware (RAP) approach, which first calculates the reputation of each user and then rank a reputation values to remove the users with low reputation. Therefore, Users with low reputation values are considered unreliable users and their QoS data is not used for the missing QoS prediction. Hence, excluding QoS values provided by untrustworthy users increases utility matrix sparsity and less predictive accuracy [30]. Xu et al. have proposed a method called RAM based on user reputation and matrix factorization. Once the reputation of users is calculated, users are divided into different grades and then integrated into a matrix factorization (MF) prediction. However, this method may not produce efficient results if inappropriate parameters are selected [29]. In another method, TAP is a trust-aware that first cluster users with k-means algorithm and then each user’s reputation is calculated based on the number of positive and negative feedback provided in the past by updating the Beta probability density features statistically. However, this method used a PCC algorithm for calculating user similarity, resulting in an inaccurate prediction if the original utility matrix is extremely sparse [31].

4. Discussion
As a summary of the research work reviewed in the previous sections, non-functional QoS features on the user side is significantly affected by unpredictable network links, and their values are different for each user. Collaborative filtering is one of the most important approaches to predict unknown QoS values. Many researchers combine CF techniques with other approaches and provide a hybrid approaches for accurate QoS prediction. However, all of these approaches used historical QoS experiences of services invocation of other users. CF-based methods can be divided into three main categories, memory-based methods, model-based methods and context-based methods. As a result of this review research, we can say that:

- The first methods for predicting unknown QoS values are memory-based approaches. They are relatively simple, easy to implement and their accuracy prediction is acceptable when there are sufficient data in the utility matrix, but they suffer from problems of data sparsity, cold start, and scalability.
• Model-based methods are presented to address memory-based methods problems. However, it provides better prediction results when the utility matrix is sparse, and the numbers of users and services is very high. Thus, the missing QoS values will be predicted by a model based on the experiences of other users or services as well as data sparsity and new users or services have a lower effect on these approaches.

• Latent factor in model-based methods employ matrix factorization for prediction. This method widely used in prediction QoS values and main benefit of this approach is that data sparsity and scalability can be overcome.

• Analyzing time series can be a good predictive tool in dynamic environment of the network where QoS attributes will be change over time.

• Clustering algorithms can minimize data volumes by clustering data, increasing scalability and reducing time for calculations, and resolving the problem of cold starts. Thus, it is helpful to use this technique together with other prediction approaches.

• Model-based methods can be enhanced with contextual information including user location, service location, service invocation time and user’s trust. Thus, this hybrid approaches can improve prediction accuracy.

• Identifying unreliable users and deleting their invalid data leads to more accurate prediction. So, determining users' trustworthiness is an important issue.

5. Conclusion

In this survey, we have presented a review for some research in the field of cloud computing and web services to predict unknown QoS values. We have summarized and classified different prediction methods. We also summarized the challenges for QoS prediction and determined the strength and weakness for some of these works. Finally, an analysis of the reviewed methods is presented.

References

[1] Y. Zhang and M. R. Lyu, QoS Prediction in Cloud and Service Computing: Approaches and Applications. Springer, 2017.

[2] Tang, T. Zhang, J. Liu, and J. Chen, “Cloud service QoS prediction via exploiting collaborative filtering and location-based data smoothing,” Concurr. Comput. Pract. Exp., vol. 27, no. 18, pp. 5826–5839, 2015.

[3] L.-J. Zhang, H. Cai, and J. Zhang, Services computing. Springer, 2007.

[4] L. Kuang et al., “A personalized qos prediction approach for cps service recommendation based on reputation and location-aware collaborative filtering,” Sensors, vol. 18, no. 5, p. 1556, 2018.

[5] X. Luo, Y. Ly, R. Li, and Y. Chen, “Web service QoS prediction based on adaptive dynamic programming using fuzzy neural networks for cloud services,” IEEE Access, vol. 3, pp. 2260–2269, 2015.

[6] T. Chen, R. Bahsoon, and X. Yao, “Online QoS Modeling in the Cloud: A Hybrid and Adaptive Multi-Learners Approach,” 2014 IEEE / ACM 7th Int. Conf. Util. Cloud Comput., 2014.

[7] Z. Chen, L. Shen, and F. Li, “Exploiting Web service geographical neighborhood for collaborative QoS prediction,” Futur. Gener. Comput. Syst., vol. 68, pp. 248–259, 2017.

[8] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of massive data sets. Cambridge university press, 2020.

[9] C. Wu, W. Qiu, Z. Zheng, X. Wang, and X. Yang, “Qos prediction of web services based on two-phase k-means clustering,” in 2015 ieee international conference on web services, 2015, pp. 161–168.

[10] Z. Chen, L. Shen, F. Li, D. You, and J. P. B. Mapetu, “Web service QoS prediction: when collaborative filtering meets data fluctuating in big-range,” World Wide Web, pp. 1–26, 2020.

[11] L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, and H. Mei, “Personalized qos prediction forweb
services via collaborative filtering,” in IEEE international conference on web services (ICWS 2007), 2007, pp. 439–446.

[12] H. Ma, H. Zhu, Z. Hu, W. Tang, and P. Dong, “Multi-valued collaborative QoS prediction for cloud service via time series analysis,” Futur. Gener. Comput. Syst., vol. 68, pp. 275–288, 2017.

[13] Z. Zheng, H. Ma, M. R. Lyu, and I. King, “Wsrec: A collaborative filtering based web service recommender system,” in 2009 IEEE International Conference on Web Services, 2009, pp. 437–444.

[14] Z. Zheng, H. Ma, M. R. Lyu, and I. King, “Collaborative web service qos prediction via neighborhood integrated matrix factorization,” IEEE Trans. Serv. Comput., vol. 6, no. 3, pp. 289–299, 2012.

[15] Z. Zheng, H. Ma, M. R. Lyu, and I. King, “Qos-aware web service recommendation by collaborative filtering,” IEEE Trans. Serv. Comput., vol. 4, no. 2, pp. 140–152, 2010.

[16] Z. Zheng, L. Xiaoli, M. Tang, F. Xie, and M. R. Lyu, “Web Service QoS Prediction via Collaborative Filtering: A Survey,” IEEE Trans. Serv. Comput., 2020.

[17] A. Puri and M. Bhonsle, “A Survey of Web Service Recommendation Techniques based on QoS values,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 12, 2015.

[18] S. H. Ghafouri, S. M. Hashemi, and P. C. K. Hung, “A Survey on Web Service QoS Prediction Methods,” IEEE Trans. Serv. Comput., 2020.

[19] D. Yu, Y. Liu, Y. Xu, and Y. Yin, “Personalized QoS prediction for web services using latent factor models,” in 2014 IEEE international conference on services computing, 2014, pp. 107–114.

[20] Y. Zhang, Z. Zheng, and M. R. Lyu, “WSPred: A time-aware personalized QoS prediction framework for Web services,” in 2011 IEEE 22nd International Symposium on Software Reliability Engineering, 2011, pp. 210–219.

[21] Q. Xie, S. Zhao, Z. Zheng, J. Zhu, and M. R. Lyu, “Asymmetric correlation regularized matrix factorization for web service recommendation,” in 2016 IEEE International Conference on Web Services (ICWS), 2016, pp. 204–211.

[22] K. Su, L. Ma, B. Xiao, and H. Zhang, “Web service QoS prediction by neighbor information combined non-negative matrix factorization,” J. Intell. Fuzzy Syst., vol. 30, no. 6, pp. 3593–3604, 2016.

[23] M. H. Zadeh and M. A. Seyyedi, “Qos monitoring for web services by time series forecasting,” in 2010 3rd International Conference on Computer Science and Information Technology, 2010, vol. 5, pp. 659–663.

[24] Y. Hu, Q. Peng, X. Hu, and R. Yang, “Web service recommendation based on time series forecasting and collaborative filtering,” in 2015 ieee international conference on web services, 2015, pp. 233–240.

[25] P. He, J. Zhu, J. Xu, and M. R. Lyu, “A hierarchical matrix factorization approach for location-based web service QoS prediction,” in 2014 IEEE 8th international symposium on service oriented system engineering, 2014, pp. 290–295.

[26] L. Kuang, Y. Xia, and Y. Mao, “Personalized services recommendation based on context-aware QoS prediction,” in 2012 IEEE 19th International Conference on Web Services, 2012, pp. 400–406.

[27] Y. Xu, J. Yin, S. Deng, N. N. Xiong, and J. Huang, “Context-aware QoS prediction for web service recommendation and selection,” Expert Syst. Appl., vol. 53, pp. 75–86, 2016.

[28] L. Kuang, Z. Liao, W. Feng, H. He, and B. Zhang, “Multimedia services quality prediction based on the association mining between context and QoS properties,” Signal Processing, vol. 120, pp. 767–776, 2016.

[29] J. Xu, Z. Zheng, and M. R. Lyu, “Web service personalized quality of service prediction via reputation-based matrix factorization,” IEEE Trans. Reliab., vol. 65, no. 1, pp. 28–37, 2015.

[30] W. Qiu, Z. Zheng, X. Wang, X. Yang, and M. R. Lyu, “Reputation-aware QoS value prediction
of web services,” in 2013 IEEE International Conference on Services Computing, 2013, pp. 41–48.

[31] K. Su, B. Xiao, B. Liu, H. Zhang, and Z. Zhang, “TAP: A personalized trust-aware QoS prediction approach for web service recommendation,” Knowledge-Based Syst., vol. 115, pp. 55–65, 2017.