IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 2, February 2022

DOI: 10.17148/IJIREEICE.2022.10219

Research on the Influence of Sampling Methods for the Accuracy of Web Services QoS Prediction

Prof. Vishal V. Mehtre[1], Mr. Aditya Kisan Chavan[2]

Department of Electrical Engineering , Bharati Vidyapeeth (Deemed to be university) College of Engineering ,Pune[1,2]

Abstract: In recent years, as the number of Web services, increases dramatically, the personalized Web service recommendation has become a hot topic in both academia and industry. The quality-of-service (QoS) prediction plays a key role in Web service recommendation systems. However, how to further improve the accuracy of QoS prediction is still a problem. Traditional QoS predicting models do not consider the impact of sampling methods on the accuracy of QoS prediction. However, the outstanding sampling method can train the predicting model more effectively and obtain higher accuracy. Therefore, it is necessary to study sampling methods based on the QoS dataset in order to obtain sample distribution closer to the original distribution, so as to improve the accuracy of the predicting models. In this paper, we first discuss how to apply several existing sampling methods to QoS datasets and then analyze their advantages and disadvantages. Finally, a novel sampling method, enhanced importance resampling (EIRS), is proposed and applied. The experiments on the real-world datasets show that our method can not only sample efficiently and accurately but also can greatly improve the accuracy of Web service QoS prediction.

I. INTRODUCTION

With the rapid growth of the number of Web services, personalized Web service recommendation has become a hot topic in both academia and industry. Web service QoS (Quality of Service) prediction plays an important role in the process of personalized Web service recommendation. However, how to further improve the accuracy of Web services QoS prediction is still a problem. Traditional researches [1]–[3] mainly focus on increasing the complexity of the predicting models for fixing the problem and simply assume that the probability distribution of QoS datasets is uniform. However, in the real world, QoS datasets tend to follow a complex distribution, that the sampled data (training data of the predicting models) based on such assumption is biased and leads inaccurate prediction. Therefore, it is necessary to study sampling methods based on QoS dataset which can obtain sampling distribution closer to the original distribution, so as to improve the accuracy of predicting models. In this paper, we firstly discuss how to apply several existing methods [4]–[7] to QoS datasets and then analyze The associate editor coordinating the review of this manuscript and approving it for publication was Anton Kos. their advantages and disadvantages. Finally, a novel sampling method (Enhanced Importance ReSampling, EIRS) is proposed and applied. Experiments on real-world datasets show that our method can not only sampling efficiently and accurately, but also can greatly improve the accuracy of Web service QoS prediction. The remainder of the paper is organized as follows. Section IV discusses related works. Section III provides the background and motivations of our work. In Section IV, we firstly discuss how to apply several existing sampling methods to QoS datasets and analyze their advantages and disadvantages. Then a novel sampling method is proposed and applied based on QoS dataset. In section V, we discuss our experimental results in detail. Finally, we conclude our work in Section VI.

II. RELATED WORK

Collaborative filtering (CF)-based approach has been widely used in Web services QoS prediction. There are two main types of CF methods, memory-based CF method and model-based CF method. Memory-based CF methods can be further divided into three categories: User based CF methods, Item based CF methods, and hybrid based CF methods. The main steps of memory-based CF methods firstly obtain preferences of users, then calculate similarities between users or services and finally predict QoS values. Memory-based CF method is simple to be implemented and is a computational model of early commercial recommendation system. However, problems such as cold start and inability to handle large-scale and time-aware datasets hinder the popularity of memory-based CF methods. The model-based prediction methods, utilize statistical learning and machine learning techniques to mine and extract the learning model from the historical records of web service invocations, and achieve QoS prediction by matrix decomposition technique. Model-based CF methods can deal with sparse and large-scale datasets better than memory-based CF methods while predicting web

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 2, February 2022

DOI: 10.17148/IJIREEICE.2022.10219

services QoS. However, they are more complex and time-consuming. Furthermore, most of the recent model-based CF methods , focus on adding additional domain information including context, time and location to improve the accuracy of QoS prediction. Although such additional information can improve the predicting accuracy, those models all use simple random sampling method to obtain training data from the original datasets, which makes the training data biased and leads to poor prediction accuracy. In the field of statistics, many representative sampling methods such as Rejection Sampling method (RJS) , Metropolis-hastings sampling method (MHS) and Importance ReSampling(IRS) have been proposed. RJS is an advanced random sampling method for complex problems with high complexity. MHS is a sampling method based on Markov chain Monte Carlo (MCMC) stochastic process , random number sequences with specific probability are sampled to make the sample distribution approximately to target distribution and IRS is an effective sampling method for estimating the target distribution of original datasets. However, to the best of our knowledge, those sampling methods have not yet been used on the QoS dataset. Therefore, in section 4 of this paper, we will discuss them in detail and apply those sampling methods to the QoS datasets, and analyze their advantages and disadvantages.

III. MOTIVATION

In this section, we firstly observe the distribution of a real world QoS data (WSDream1), then we propose a framework of on-line Web service recommendation system and emphasize the importance of sampling in the process of QoS prediction. A. OBSERVATIONS OF REAL WORLD QoS DATASET WSDream is a real world QoS dataset which has been widely utilized by many mainstream predicting models. There are two sub-data sets in the dataset, Response Time (RT) and Throughput (TP) respectively. In Figure 1, the upper and lower parts show the distribution of the QoS value according to five randomly selected users based on RT and TP respectively. We can obviously see that the data present a long tail distribution rather than uniform distribution and most of the values are concentrated in a very small ranges. Traditional Web Services QoS prediction models use Simple Random Sampling method to obtain the samples, which mean they simply assume the distribution of the original data is uniform, resulting in inaccuracy prediction results.

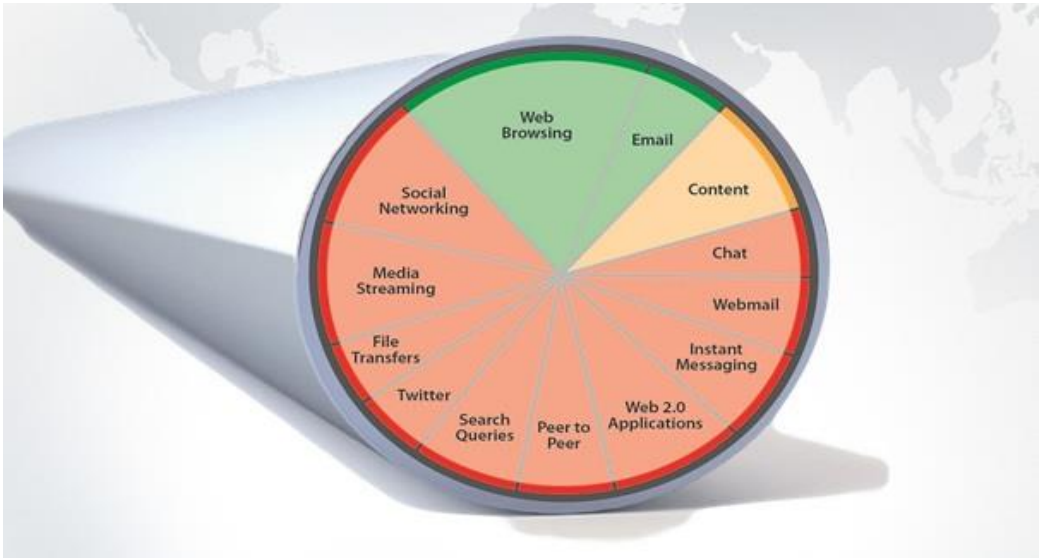

Fig-2.The framework of on-line web service recommendation system.

B. THE FRAMEWORK OF ON-LINE WEB SERVICES RECOMMENDATION SYSTEMS

Figure 2 shows the framework of QoS prediction based online Web service recommendation system. We can see that the framework contains five steps. First, the system collects the original QoS values. Second, the system uses sampling method to obtain training data. Third, prediction algorithms are used to train the model. Fourth, the system predicts the QoS value based on the trained model and personal user requirements. Fifth, the system recommends the personalized web services. Finally, once the user selected one of the recommended services, the scheduling system will schedule the service to the user. QoS Prediction is the key step in the on-line Web Services Recommendation System. It requires not only accuracy but also efficiency. The mainstream works focus on designing prediction models to improve the accuracy of QoS Prediction. However, such behavior often brings unnecessary system overhead and longer response time due to the complexity of the models. In order to improve the accuracy of recommendation without reducing user experience,

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 7.047 ⁜ Vol. 10, Issue 2, February 2022

DOI: 10.17148/IJIREEICE.2022.10219

we take the sampling step of on-line web service recommendation system into account, not only because the sampling step is off-line and has no affects of the user experience, but also a good sampling method effectively reduces the bias between training data and original data which helps improving the accuracy of predicting models.

IV. OUR WORK

As far as we know, existing works only use simple random sampling without considering the influence of different sampling methods when predicting Web services QoS. In this section, we will discuss how to apply different sampling methods to the QoS datasets and analyze their advantages and disadvantages, then propose a novel sampling method named Enhanced Importance ReSampling method (EIRS).

A. USER-BASED AND SERVICE-BASED RANDOM SAMPLING BASED ON QoS DATASET

Traditional simple sampling method (RS) assumes that the dataset is uniform distribution and samples globally according to a certain sampling density. However, we observed that some users or services data will never be sampled by using RS, resulting recommendation system unable to recommend services for such users. There are two variants of RS can fix such problem, one is user-based random sampling (URS) method which samples the data randomly according to each user for all users in the dataset and the other is service-based random sampling method (SRS) which samples the data randomly according to each service for all services in the dataset. RS, URS and SRS are all easy to be conducted on the QoS dataset.

B. DOMAIN BASED RANDOM SAMPLING BASED ON QoS DATASET

The domain information such as location and time is closely related to the QoS of Web services. By considering those domain information, domain based random sampling (DRS) method firstly divides the services into different domains and then samples the data randomly in each domain. When conducting a Domain based random sampling method (DRS) on WSDream, we firstly divide the dataset into different parts according to the 'AS' attribute which describe the location of services, then use RS on each part to obtain the samples. However, the sample distribution will be unbalanced because some parts have more data while others have less or even no data.

C. REJECTION SAMPLING BASED ON QoS DATASET

Rejection sampling (RJS) is an advanced random sampling method which can generate complex sample distribution. Figure 3 shows an example of RJS, where q(x) represents a presumed sample distribution (reference distribution) which can be adjusted after the process of sampling, p(x) represents the distribution of the original dataset (target distribution) and k represents a parameter for scaling all x subject to $kq(x) \geq \tilde{p}(x)$, where $\tilde{p}(x)$ represents the distribution of the sampled data in the process of sampling (observation distribution). RJS firstly samples the data x0 randomly according to q(x), then samples the value u0 randomly in the interval [0, kq(x0)] and compares $\tilde{p}(x0)$ to u0. If $u0 < \tilde{p}(x0)$, then accepts the sample with a certain probability, otherwise, rejects. The acceptance probability of the sample can be calculated according to equation (1) p(accept) = Z $\tilde{p}(x)$ kq(x) q(x)dx = 1 k Z $\tilde{p}(x)$dx (1) The RJS can be conducted on the QoS dataset according to Algorithm 1. The inputs of Algorithm 1 include the reference distribution q(x), the scale parameter k and the observation distribution $\tilde{p}(x0)$. We choose the normal distribution for q(x) according to the distributions of WSDream and specify a large number k in order to cover the range of the target distribution p(x). However, in real applications, it is difficult to find a suitable q(x) because of that when the target distribution is a distribution with spikes, a large number of unwanted samples will be sampled. The algorithm terminates until a certain number of samples are obtained. However, it converges slowly because lots of data are probably be rejected in the iteration step.

D. METROPOLIS-HASTINGS SAMPLING

Metropolis-hastings sampling (MHS) is based on Markov chain Monte Carlo (MCMC) stochastic process [20] . The basic idea of MHS is firstly constructing Markov Chain from the reference distribution q(x), then randomly selects an initial state of Markov Chain and begin to transfer until the state to be stable. Finally, the obtained state sequence can be used to estimate the target distribution. Considering the complexity of distribution of the QoS dataset and in order to satisfy the fine stationary condition of Markov chain, we calculate the acceptance probability of samples according to equation (2): $A(j|i) = \min\{1, \tilde{p}(j)q(i|j)\ \tilde{p}(i)q(j|i)\}$ (2) where A(j|i) represents the acceptance probability of sample j condition on the sampled sample j, $\tilde{p}(j)$ and $\tilde{p}(i)$ can be calculate by utilizing observation distribution $\tilde{p}(x)$, q(i|j) and q(j|i) can be calculate by reference distribution q(x). The pseudo code of MHS algorithm based on QoS dataset is described in Algorithm 2. We can see that in the iteration step of Algorithm 2, a candidate state s 0 is generated and then calculate the conditional probability A(s 0 |st), where st is a sample already be sampled. If A(s 0 |st), is larger than u

DOI: 10.17148/IJIREEICE.2022.10219

which is a random number between [0, 1], accepts s 0 , otherwise, rejects. Similar to RJS, MHS converges slowly because lots of data are probably be rejected in the iteration step.

REFERENCES

1) A Agarwal , S N Negahban , M J Wainwright A. Agarwal, S. N. Negahban, and M. J. Wainwright, "Stochastic optimization and sparse statistical recovery: An optimal algorithm for high dimensions," in 2014 48th Annual Conference on Information Sciences and Systems (CISS), mar 2014, pp. 1-2.

2)R Salakhutdinov, Mnih R Sala, khutdinov and A Mnih, "Probability Matrix Factorization," International Conference on Neural Information Processing Systems, pp. 1257-1264, 2007.

3) Ary, D., Jacobs, L. S., & Razavieh, A. (1985). Introduction to research in education(3rd ed.) New York: Holt Rinehart and Winton.

4) Bartlatte, J. E. Kotrlik, J. W., & Higgins, C. (2001). Organizational research: determining appropriate sample size for survey research. Information technology, learning, and performance. Journal 19(1) 43-50.

5) Denga, D. I., & Ali, A. (1998). An introduction to research methods and statistics in education and social science (3rd ed.). Calabar: Rapid Educational Publishers. Denga, D. I. (2003). Educational measurement, continuous assessment and psychological testing (3rd ed.). Calabar: Rapid Educational Publishers.

6) Frohle, P. (2001). Influence of sample data on the statistical analysis of wave measurements. American society of Civil Engineer Journal, 273 (44), 424-433.

7) Gay, L.R. & Airasian, P. (2000). Educational Research: Competencies for Analysis and Application. New Jersey: Prentice-Hall.

8) Gy, P. (1992). Sampling of heterogeneous and dynamic material systems: theories of heterogeneity, sampling and homogenizing. Wikipedia-the free encyclopedia.

9) Isangedighi, J. A., Joshua, M. T., Asim A. E., & Ekuri, E. E. (2004). Fundamentals of research and statistics in education and social sciences. Calabar: University of Calabar press.

10) Kerlinger, F. N. (1986). Foundations of behavioural research (3rd ed.). New York: Holt, Rinehart and Winston