Engineering Applications of Artificial Intelligence 38 (2015) 14–23

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Efficient web service QoS prediction using local neighborhood matrix factorization

Wei Lo, Jianwei Yin*, Ying Li, Zhaohui Wu

School of Computer Science and Technology, Zhejiang University, 310000 China

ARTICLE INFO

Article history:
Received 10 April 2014
Received in revised form
8 October 2014
Accepted 11 October 2014
Available online 12 November 2014

Keywords:
Web service application
Matrix factorization
Performance

ABSTRACT

In the era of Big Data, companies worldwide are actively deploying web services in both intranet and internet environments. Quality-of-Service (QoS), the fundamental aspect of web service has thus attracted numerous attention in industry and academia. The study on sufficient QoS data keeps advancing the state in Service-Oriented Computing (SOC) area. To collect a large amount of resource in practice, QoS prediction applications are designed and built. Nevertheless, how to generate accurate results in high productivity is still a main challenge to existing frameworks. In this paper, we propose LoNMF, a Local Neighborhood Matrix Factorization application that incorporates domain knowledge in modern Artificial Intelligence (AI) technique to tackle this challenge. LoNMF first proposes a two-level selection mechanism that can identify a set of highly relevant local neighbors for target user. And then, it integrates the geographical information to build up an extended Matrix Factorization (MF) approach for personalized QoS prediction. Finally, it iteratively generates results by utilizing hints from previous round computations, a gradient boosting strategy that directly accelerates solving process. Experimental evidence on large-scale real-world QoS data shows that LoNMF is scalable, and consistently outperforming other state-of-the-art applications in prediction accuracy and efficiency.

1. Introduction

Web service is a kind of self-describing programmable application, which is implemented in a standard language or a specific protocol to achieve inter-operability in network environments (McIlraith et al., 2001). In the era of Big Data, World Wide Web has witnessed a great boom in web services, thanks to the fact that they are ubiquitously and pervasively adopted everywhere in online networks. In addition, web services are broadly utilized in numerous cloud computing platforms, like Amazon Web Service (Huckman et al., 2013) and Microsoft Azure (Armbrust et al., 2010), which bridge the gap between companies and end users. The growth and development in industrial sector raises numerous practical problems to Service-Oriented Computing (SOC) academic community. In web service domain, a large number of problems need to be considered under the guidance of modern Artificial Intelligence (AI) techniques at this exciting age (Shenfield and Fleming, 2014).

In this context, the study on Quality-of-Service (QoS) (Zhang et al., 2007), which is a fundamental element describing the non-functional characteristics of web services, is boosting. In general concept, QoS includes price, response time, throughput and failure rate which deeply captures interaction behavior between users and services. With massive adoption in the dynamic Internet, sufficient QoS recourse benefits multi-disciplines in web service domain. For example during the process of selecting services, consumer tends to enjoy a mass of QoS data, among which it can select desirable subsets through comprehensive comparisons. When researcher wants to conduct complex service composition, abundant annotated resource is expected to describe a sophisticated process. And in building up a robust service recommendation system, architect hopes to train with plentiful QoS records, for the sake of accuracy improvement. In short, adequate QoS information is within the central component in many web services scenarios. And capturing the insight in this big QoS data helps to advance the state of SOC industry and academia.

However, this requirement is not so easy to satisfy in practice. The scarcity can be viewed from the following reasons: (1) Most web services are operated by commercial companies and in charge. Gathering this piece of QoS resource by web services execution can be too costly to organizations and individuals (Zheng et al., 2013b). (2) Nowadays there are vast novel services emerging on internet. It turns out impractical to acquire all this large volume of QoS data since time and cost is limited. To satisfy the urgent needs for SOC domains, a straight-forward manner is to

* Corresponding author. Tel.: +86 133 9681 1588.
E-mail addresses: spencer_w_lo@zju.edu.cn (W. Lo), zjuyjw@zju.edu.cn (J. Yin), cnliying@zju.edu.cn (Y. Li), wzh@zju.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.engappai.2014.10.010
0952-1976/

utilize modern AI techniques to build up novel applications that predict those missing QoS values.

There are a number of works that fulfill this task in recent years (Chang et al., 2012; Zheng et al., 2011). Generally, they achieve this goal under the inspiration of Collaborative Filtering (CF), the most widely used technique in Recommender Systems to predict the users' preference (Luo et al., 2012). The core idea behind is to identify a set of similar neighbors, and to collect the wisdom of crowds that can infer the next-step behavior of the target user. Therefore, the concept of neighborhood is vital and existing approaches mainly exploits users' historic information. However, we argue that users' geographical prior knowledge, an intrinsic element in web service domain, would contribute to the performance of QoS prediction system. To our common sense, if two users live in a local region and share similar basic IT infrastructure, they can be valuable reference to each other since QoS messages are non-functional and highly affected by subjective environment. In fact, there are a few QoS prediction applications emerging recently that confirm this intuition, and they continually outperform traditional frameworks in various metrics (Lo et al., 2012; Chen et al., 2010). Another advantage to leverage this compelling factor is that it helps to solve the "cold-start" problem, which means user does not leave any historic records before. Under this situation, many existing approaches fail to work. Despite the fact that the progress on integrating local information is accumulated, prediction accuracy still has plenty of room to grow in these works, and how to incorporate domain knowledge in applications is still an open question.

Besides prediction performance, efficiency is of the great concern to both end user and system designer. Intuitively, active user feels impatient and frustrated when it takes to wait for a slightly better prediction in a long time. Meanwhile, system designer is eager to use a faster model to keep track of the rule of massive QoS resource. As far as we can see, most existing approaches suffer from low efficiency problem and cannot provide detailed quantitative analysis in literatures. To build up a practical QoS prediction framework, it is crucial to strike a balance between accuracy and efficiency.

In this paper, to simultaneously achieve the high standard of accuracy and efficiency, we propose a novel Local Neighborhood Matrix Factorization QoS prediction application (LoNMF) that incorporates domain knowledge to Matrix Factorization (MF) technique. LoNMF first proposes a two-level neighborhood selection mechanism that can identify a set of highly relevant local neighbors for each target user. And then, it integrates the geographical domain knowledge to build up an extended Matrix Factorization (MF) system for personalized QoS prediction. Finally, it iteratively generates results by utilizing hints from previous round computations, a memo boosting strategy that directly accelerates solving process in response to practical requests. Experimental evidence on large-scale real-world QoS data shows that LoNMF is scalable, and consistently outperforming other state-of-the-art approaches in accuracy and efficiency.

In summary, the contributions of this paper are four-folds:

1. We implement LoNMF, a QoS prediction application that incorporates domain knowledge into MF technique to meet the high standard of accuracy and efficiency.
2. We design a two-level neighborhood selection mechanism to precisely capture the similarity relationship for target users.
3. We transfer geographical information to unified mathematical forms to extend MF technique in normal and cold-start cases.
4. We propose a gradient boosting strategy that directly accelerates the solving process in response to practical requests.

Experimental evidence on large-scale real-world QoS data shows that LoNMF is scalable, and generally achieving up to 15.10% in accuracy improvement compared with other state-of-the-art approaches. In addition, our proposed boosting strategy lowers the latency-accuracy ratios, with 57.10% better than the rest in general.

The rest of paper is organized as follows. We cover related work in Section 2, then formulate the scope of problem in Section 3. As the core part of our contribution, we elaborate the construction of LoNMF in Section 4 and report on our experimental results in Section 5. Finally we conclude this paper in Section 6.

2. Related works

The study on QoS plays a central role in Service-Oriented Computing (SOC) domain. Thanks to the development of industrial sector, many interesting QoS-based problems have been widely discussed in a number of literatures, covering the topic of service selection (Vahdani et al., 2013), service discovery (Meditskos and Bassiliades, 2010), service composition (Jimenez-Molina and Ko, 2011), service monitoring (Aiello et al., 2011), and so on. In existing works, sufficient QoS data is required for successful implementation and adoption. However in the current big data environment, this requirement is very difficult to fully satisfied. Hence a necessary preprocess is to predict the missing QoS values. In this paper, we build up a novel prediction application that predicts abundant QoS data to support the advancement of above topics.

Collaborative Filtering (CF) technique has been adopted to fulfill this task. Inspired by the idea of user-collaboration, the key step is to build up a proper neighborhood for each user. Existing works mostly focus on leveraging users' historic QoS records to find out similar users. Zheng et al. (2011) present a hybrid approach combining user-based and item-based approach to recommend the QoS-aware service. Zheng et al. (2009) propose WSRec, which employs an effective and novel hybrid CF method for web service recommendation. However these works suffer from the "cold-start" problem, and the performance is not so desirable. To advance the state of QoS prediction systems, we argue that incorporating the users' geographical feature matters. Recently, conclusions from a few emerging works have shown the promising future to this direction. Chen et al. (2010) support the idea that local knowledge contributes to the prediction performance. Lo et al. (2012) build geographical regularization terms to make prediction. Moreover, Yin et al. (2014) explain how to fuse different regularization parts into MF model. Even though some signals reveal geographical-based systems work in various situations, prediction accuracy still has plenty of room to grow in these frameworks, and how to incorporate geographical domain knowledge in systems is still an open question. In this paper, we incorporate local factor to extend the prediction power of Matrix Factorization (MF) model.

The driven force to harness MF model comes from the fact that its effectiveness has been widely verified in Recommender System. Koren (2008) smoothly merge the factor and neighborhood models to build an accurate framework. The ultimate goal of MF framework is to construct two latent space so that their inner product can be approximately to the original matrix. Koren (2010) propose a MF-based model to track the temporal behavior throughout the life span of the data. Besides, MF has also achieved great success in Machine Learning (Mitra et al., 2010), Computer Vision (Wang and Gao, 2013) and so on. To solve the objective function in MF, Gradient Descent (GD) technique has been employed to search the local minima. This straight manner has been criticized for its low efficiency problem. For a robust QoS prediction system, it is unacceptable to consume endless computation cost to generate high accuracy. In Yin et al. (2014), the framework works less efficiently to gain more accuracy. To our best knowledge, limited works have been done to boost the solving process for QoS prediction. In this paper, we not only