SPRINGER BRIEFS IN COMPUTER SCIENCE

Yilei Zhang
Michael R. Lyu

QoS Prediction in Cloud and Service Computing

Approaches and Applications

Springer

SpringerBriefs in Computer Science

More information about this series at http://www.springer.com/series/10028

Yilei Zhang · Michael R. Lyu

QoS Prediction in Cloud and Service Computing

Approaches and Applications

Springer

Yilei Zhang
School of Mathematics and Computer Science
Anhui Normal University
Wuhu, Anhui
China

Michael R. Lyu
The Chinese University of Hong Kong
Hong Kong
China

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-981-10-5277-4 ISBN 978-981-10-5278-1 (eBook)
DOI 10.1007/978-981-10-5278-1

Library of Congress Control Number: 2017946048

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Cloud computing provides shared resources (e.g., infrastructure, platform, and software) as services. Service-oriented architecture (SOA) is the technical foundation of cloud computing, whereby services offered by different cloud providers are discovered and integrated over the Internet. Quality-of-Service (QoS) is widely employed to represent the non-functional performance of services and has been considered as the key factor to differentiate the qualities of service candidates. It becomes important to evaluate the QoS performance of services.

However, QoS evaluation is time- and resource-consuming. Conducting real-world evaluation is difficult in practice. Moreover, in some scenarios, QoS evaluation becomes impossible (e.g., the cloud provider may charge for service invocations, too many services to be evaluated). Therefore, it is crucial to study how to build effective and efficient approaches to predict the QoS performance of services.

In this book, we propose QoS prediction, a novel principle for enabling the QoS-aware approaches. We first formally identify the QoS prediction problem and propose three QoS prediction methods, which utilize the users' past usage experiences. The first prediction method employs the information of neighborhoods for making QoS value prediction and engages matrix factorization techniques to enhance the prediction accuracy. The second method provides time-aware personalized QoS value prediction service. The third method employs time information for efficient online performance prediction.

The predicted QoS values can be employed to a variety of applications in cloud and service computing. We demonstrate the benefits in two QoS-aware applications in this book. The first application employs QoS information to build a Web service search engine, which helps users discover appropriate Web services to fulfill both functional and non-functional requirements. The second application employs dynamic QoS information to build robust Byzantine fault-tolerant cloud systems.

This book is intended for professionals involved in cloud computing and graduate students working on the QoS-related problems. It is assumed that the reader has a basic knowledge of mathematics, as well as a certain background in cloud computing. The reader can get an overview of the QoS prediction research

area. We hope this monograph will be a useful reference for students, researchers, and professionals to understand three basic methodologies of QoS prediction. This book can be used as a starting point for QoS-related research topics. The readers can immediately conduct extensive researches and experiments on the real-world QoS datasets released in this book.

Hong Kong
June 2017

Yilei Zhang
Michael R. Lyu

Acknowledgements

The work described in this book was fully supported by the National Natural Science Foundation of China (Project No. 61332010) and the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 415113 of the General Research Fund).

Contents

Chapter 1
Introduction

Abstract This chapter provides an overview of QoS prediction in cloud and service computing, including backgrounds, related works, and organizations of this book.

1.1 Overview

Cloud computing [6, 22] is a new type of Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on demand [38]. With the exponential growth of cloud computing as a solution for providing flexible computing resources, more and more cloud applications emerge in recent years. The architecture of the Software-as-a-Service (SaaS) systems in the delivering of cloud computing typically involves multiple cloud components communicating with each other over application programming interfaces, usually Web services. [92]. Cloud computing has become a scalable service consumption and delivery platform.

Web services are software systems designed to support interoperable machine-to-machine interaction over a network. The technical foundations of cloud computing include service-oriented architecture (SOA), which is becoming a popular and major framework for building Web applications in the era of Web 2.0 [63], whereby Web services offered by different providers are discovered and integrated over the Internet. Typically, a service-oriented system consists of multiple Web services interacting with each other over the Internet in an arbitrary way. In this book, service refers to Web service in service computing and cloud component which is delivered as a service in cloud computing.

Figure 1.1 shows the system architecture in cloud computing. In a cloud environment, the cloud provider holds a large number of distributed services (e.g., databases, servers, Web services), which can be provided to designers for developing various cloud applications. Designers of cloud applications can choose from a broad pool of distributed services when composing cloud applications. These services are usually invoked remotely through communication links and are dynamically integrated into the applications. The cloud application designers are located in different geographic

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_1

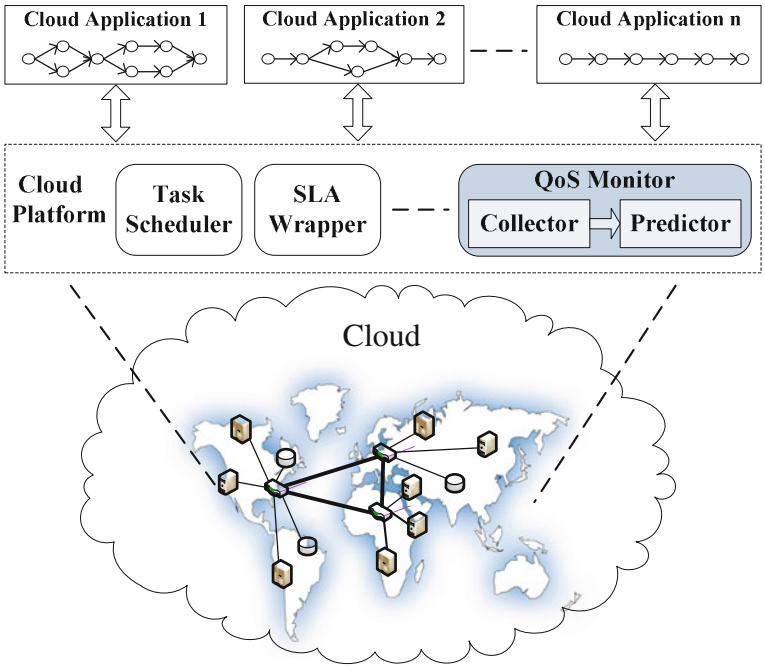


Fig. 1.1 System architecture. ©[2011] IEEE. Reprinted, with permission, from Ref. [104]

and network environments. Since the users invoke services via different communication links, the quality of services they observed are diverse.

Quality-of-Service (QoS) is usually employed to describe the non-functional characteristics of services. It becomes a major concern for application designers when making service selection [37]. Moreover, for the existing cloud applications, by replacing low-quality services with better ones, the overall quality of cloud applications can be improved.

In recent year, a number of research tasks have been focused on optimal service selection [10, 97] and recommendation [108] in distributed systems or service computing. Typically, evaluations on the service candidates are required to obtain their QoS values. In cloud environment, due to their various locations and communication links, different users will have different QoS experiences when invoking even the same service. Personalized QoS evaluation is required for each user at the user-side. However, a service user in general only invoked a limited number of services in the past and only received QoS performance information of these invoked services. In practice, therefore, conducting real-world evaluations on services to obtain their QoS information from the users' perspective is quite difficult, because (1) executing invocations for evaluation purposes becomes too expensive, since cloud providers who maintain and host services (e.g., Amazon EC2, Amazon S3) may charge for

invocations; (2) with the growing number of available services over the Internet, it is time-consuming and impractical to conduct QoS evaluations on all accessible services; (3) component users need to focus on building cloud applications rather than testing a large number of service candidates. Therefore, collecting historical usage records and conducting QoS prediction, which requires no additional invocation, is becoming an attractive approach. Based on the above analysis, in order to provide QoS information to application designers, we need to provide comprehensive investigation on QoS prediction approaches.

Employing the predicted QoS values, a QoS-aware Web service search engine can be enabled. Traditional Web service searching approaches only find the Web services to fulfill users' functionality requirements. However, Web services sharing similar functionalities may possess very different non-functionalities (e.g., response time, throughput, availability, usability, performance, integrity). Web services recommended by the traditional searching approach may not fulfill users' non-functional requirements. In order to find appropriate Web services which can fulfill both functional and non-functional requirements of users efficiently, QoS-aware searching approaches are needed.

Given the predicted QoS information, robust systems can be built based on redundant services by employing QoS-aware fault tolerance framework. Traditional fault tolerance framework [58] usually requires developing several different version of system services. However, due to the cost of development, the fault tolerance strategies are usually employed only for critical systems. In cloud computing, however, users can access multiple functional equivalent services via Internet at a very low cost. These services are usually developed and provided by different organizations, and can be dynamically composed to build a fault tolerance systems. Although some fault tolerance frameworks [52, 56, 107] have been proposed for traditional software systems, they cannot adopt to the highly dynamic cloud environment.

In order to provide accurate QoS prediction approaches, QoS-aware Web service searching mechanisms, and QoS-aware fault-tolerant frameworks for cloud systems, we proposed five approaches to attack these challenges in this book.

1.2 Backgrounds

1.2.1 QoS in Cloud and Service Computing

Cloud computing [6] has been in spotlight recently. Cloud computing has become a scalable service consumption and delivery platform [100]. The technical foundations of cloud computing include service-oriented architecture (SOA) [29]. SOA is becoming a popular and major framework for building Web applications in the era of Web 2.0 [63]. A number of investigations have been carried out focusing on different kinds of research issues such as Web service selection [28, 94, 97, 99], Web service composition [3, 4, 95], SOA failure prediction [9], SOA performance

prediction [105, 106], fault tolerance [52, 103], resiliency quantification [31], service searching [101], resource consistency [79], resource allocation [23], workload balance [87], dynamically resource management [46].

Quality-of-Service (QoS) has been widely employed as a quality measure for evaluating non-functional features of software systems and services [1, 97, 101]. A lot of research works have utilized QoS to describe the characteristics of services in cloud and service computing [42, 61, 64, 65, 72, 86]. Zeng et al. [98] use five QoS properties to compose Web service dynamically. Ardagna et al. [5] employ five QoS properties to conduct flexible service composition processes. Alrifai et al. [3] consider generic and domain-specific QoS for efficient service composition.

QoS performance of services can be measured either from the provider's perspective or from the user's observation. QoS values measured at the service provider side (e.g., price, availability) are usually identical for different users, such as QoS used in the service-level agreement (SLA) [57] (e.g., IBM [48] and HP [73]). While QoS values observed by different users may vary significantly due to the unpredictable communication links and heterogeneous contexts. In this book, we mainly focus on observing QoS data from users' perspective and making use of the QoS data for QoS prediction, service selection, service searching, and fault-tolerant framework building.

Based on the QoS performance of services, several approaches have been proposed for optimizing service selection [8, 10, 13, 27, 84, 97] in improving the whole quality of Web application, Web service composition [3, 5, 13, 14, 98], Web service recommendation [20, 86], reliability prediction [15, 21, 32, 35, 71], etc. Traditionally, reliability of a software system [59] is analyzed without considering the system performance, which is not accurate when applied to modern systems. Moreover, several QoS-aware approaches [24, 60, 72, 93, 97, 98] are proposed in cloud and service computing.

However, there is few real-world QoS data to verify these QoS-aware approaches. To collect the QoS data from the user-side, Zheng et al. [109] proposed a distributed evaluation framework and released the QoS datasets for further extensive research. Different from previous work [2, 89], they conduct large-scale real-world evaluations.

1.2.2 QoS Prediction in Cloud and Service Computing

The QoS-aware approaches usually assume that the QoS values are already known, while in reality a user cannot exhaustively invoke all the services. Although there existed some QoS evaluation approaches and publicly released QoS datasets, it is impossible to conduct personalized evaluation on all accessible services for all users. In this chapter, we focus on predicting missing QoS values by collaborative filtering approach to enable the QoS-aware approaches.

Collaborative filtering approaches are widely adopted in commercial recommender systems [12]. Generally, traditional recommendation approaches can be categorized into two classes: memory-based and model-based. Memory-based approaches, also known as neighborhood-based approaches, are one of the most popular prediction methods in collaborative filtering systems. Memory-based methods employ similarity computation with past usage experiences to find similar users and services for making the performance prediction. The typical example of memory-based collaborative filtering includes user-based approaches [11, 19, 39, 45, 81], item-based approaches [25, 43, 54, 78], and their fusion [34, 90]. Typically, memory-based approaches employ the PCC algorithm [70] for similarity computation.

Model-based approaches employ machine learning techniques to fit a predefined model based on the training datasets. Model-based approaches include several types: the clustering models [96], the latent factor models [74], the aspect models [40, 41, 82, 83]. Lee et al. [50] presented an algorithm for nonnegative matrix factorization that is able to learn the parts of facial images and semantic features of text. It is noted that there is only a small number of factors influencing the service performance in the user-service matrices, and that a user's factor vector is determined by how much each factor applies to that user. For a set of user-service matrices data, three-dimensional tensor factorization techniques are employed for item recommendation [69].

The memory-based approaches employ the information from similar users and services for predicting missing values. When the number of users or services is too small, similarity computation for finding similar users or services is not accurate. When the number of users or services is too large, calculating similarity values for each pair of users or services is time-consuming. In contrast, model-based approaches are very efficient for missing value prediction, since they assume that only a small number of factors influence the service performance.

There is few work of collaborative filtering prediction for QoS values in cloud and service computing, since there lack large-scale real-world QoS datasets for verifying the prediction accuracy. Some approaches [47, 85] employing a movie rating dataset (i.e., MovieLens [70]) for simulation. Shao et al. [80] only conduct a small-scale experiments, which involve 20 Web services for evaluating prediction accuracy.

The existing methods in the literature only consider two dimensions (i.e., user and Web service), while time factor is not included. The periodic features of service QoS values are ignored, which may improve the prediction accuracy significantly. Moreover, the high computational complexity makes it difficult to extend memory-based approaches to handle large amounts of time-aware performance data for timely prediction. There is a lack of fast algorithms to predict the QoS values at runtime to adapt the highly dynamic system environment in cloud and service computing.

In this book, we propose three approaches to address the QoS prediction problems in cloud and service computing, including memory-based prediction [104], time-aware prediction [102], and online prediction [105, 106] approaches. We also conduct large-scale real-world experiments to verify the prediction accuracy and release the QoS datasets for further studies of other researchers.

1.2.3 Web Service Searching

Web service discovery [68] is a fundamental research area in service computing. Several papers in the literature conduct investigations on discovering Web services through syntactic or semantic tag matching in a centralized UDDI repository [66, 88]. However, since UDDI repository is no longer a popular style for publishing Web services, these approaches are not practical now.

Text-based matching approaches have been proposed for querying Web service [33, 91]. These works employ term frequency analysis to perform keywords searching. However, most text descriptions are highly compact, and contain a lot of unrelated information to the Web service functionality. The performances of this approaches are not fine in practice. Plebani et al. [67] extract the information from WSDL files for Web service matching. By comparing with other works [26, 36, 44], it shows better performance in both recall and precision. However, it also does not consider non-functional qualities of Web services. Our searching approach, on the other hand, takes both functional and non-functional features into consideration.

Alrifai et al. [3], Liu et al. [55], and Tao et al. [97] focus on efficiently QoS-driven Web service selection. Their works are all based on the assumption: the Web service candidates are functional identical. Under this assumption, these approaches cannot be directly applied into Web service search engine. In this book, we proposed WSExpress [101], a QoS-aware searching approach which employs both QoS and functionality information, to search appropriate Web services for users.

1.2.4 Fault-Tolerant Cloud Applications

Software fault tolerance techniques (e.g., N-Version Programing [7], distributed recovery block [49]) are widely employed for building reliable systems [58]. Zhang et al. [101] propose a Web service search engine for recommending reliable Web service replicas. Salas et al. [75] propose an active strategy to tolerate faults in Web services. There are many fault tolerance strategies that have been proposed for Web services [17, 18, 30, 76]. Typically, the fault tolerance strategies can be divided into two major types: passive strategies and active strategies. Passive strategies include FT-CORBA [53], FT-SOAP [52]. Active strategies include WS-Replication [75], SWS [51], FTWeb [77].

However, these techniques cannot tolerate Byzantine faults like malicious behaviors. There are some works that focus on Byzantine fault tolerance for Web services as well as distributed systems. BFT-WS [107] is a Byzantine fault tolerance framework for Web services. Based on Castro and Liskov's practical BFT algorithm [16], BFT-WS considers client–server application model running in an asynchronous distributed environment with Byzantine faults. $3f + 1$ replications are employed in the server side to tolerate f Byzantine faults. Thema [62] is a Byzantine fault-tolerant (BFT) middleware for Web services. Thema supports three-tiered application model,

where the $3f + 1$ Web service replicas need to invoke an external Web service for accomplishing their executions. SWS [51] is a survivable Web service framework that supports continuous operation in the presence of general failures and security attacks. SWS applies replication schemes and N-Modular Redundancy concept. Each Web service is replicated into a service group to mask faults.

Different from above approaches, BFTCloud [103] proposed in this book aims to provide Byzantine fault tolerance for voluntary-resource cloud, in which Byzantine faults are very common. BFTCloud selects voluntary nodes based on both their reliability and performance characteristics to adapt to the highly dynamic voluntary-resource cloud environment.

1.3 Book Organization

As shown in Fig. 1.2, the rest of this book is organized as follows:

- Chapter 1
  This chapter briefly reviews some background knowledge and work related to the main methodology that will be explored in this book.
- Chapter 2
  In this chapter, we propose a novel neighborhood-based approach (CloudPred), which is enhanced by character modeling, for providing collaborative and personalized QoS prediction of cloud components. We first present the QoS prediction scenario by a toy example. Then, the QoS prediction problem in cloud computing is formally defined. After that, we present a latent feature learning algorithm to learn the user-specific and service-specific latent features. Based on the latent features, user and service similarity computation approaches are introduced. By identifying similar users and similar services to the active user-service pair, we formulate the CloudPred prediction Algorithm. We conduct extensive experiments to study the prediction accuracy of CloudPred and the impact of various parameters. The experimental results show that CloudPred achieves higher prediction accuracy than other competing methods.

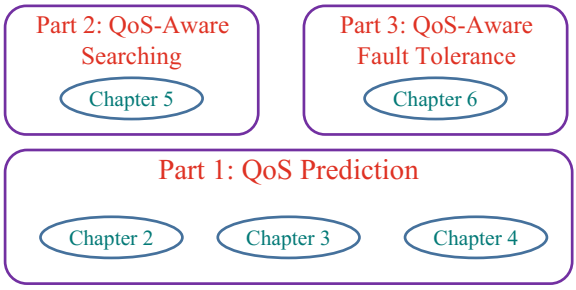


Fig. 1.2 Book structure

- Chapter 3
  In this chapter, we present a model-based time-aware collaborative filtering approach for personalized QoS prediction of Web services. First, we endow a new understanding of user-perspective QoS experiences, which is based on the following observations: (1) during different time intervals, a user has different QoS experiences on the same Web service; (2) in general, the differences are limited within a range. Based on these observations, we formulate the time-aware personalized QoS prediction problem as the tensor factorization problem, and propose an optimization formulation with average QoS constraint. Second, we propose to predict the missing QoS values by evaluating how the user, service, and time latent features are applied to each other. Furthermore, we provide a comprehensive complexity analysis of our approach, which indicates that our approach is efficient and can be applied to large-scale systems. Extensive experiments are conducted to evaluate the prediction accuracy and parameter impacts. The experimental results show the effectiveness and efficiency of our time-aware QoS prediction approach.
- Chapter 4
  In this chapter, we present an online Web service QoS prediction approach by performing time series analysis on user-specific and service-specific latent features. Our online prediction approach includes four phases. In Phase 1, service users monitor the performance of Web service and keep the QoS records in local site. In Phase 2, distributed service users submit local QoS records to the performance center in order to obtain a better QoS prediction service from the performance center. The performance center collects QoS records from different users and generates a set of global QoS matrices. In Phase 3, a set of time-stamped user latent feature matrices and service latent feature matrices are learned from the global QoS matrices. After that, time series analysis are conducted on the latent matrices to build a QoS model in the performance center. By evaluating how each factor applies to the active user and the corresponding service in the QoS model, personalized QoS prediction results can be returned to users on demand. In Phase 4, the system-level QoS performance of service-oriented architecture is predicted by analyzing the service compositional structure and utilizing the service QoS prediction results. The complexity analysis indicates that our approach is efficient and can be applied to large-scale online service-oriented systems. Finally, we conduct a number of experiments to study the performance of our approach and the impacts of algorithm parameters. We also study the effects of integrating service QoS information into the dynamic composition mechanism by a real-world service-oriented system case.
- Chapter 5
  In this chapter, we propose a QoS-aware Web service searching approach to explore the appropriate Web services to fulfill users' functional and non-functional requirements. We first describe the Web service searching scenarios and present the system architecture. Then, we present the QoS model to evaluate the non-functional utility of Web services. After that, functional similarity is introduced to evaluate the functional utility of Web services. Two QoS-aware Web service searching approaches are proposed: the score-based combination and the ranking-based combination. We

further extend the ranking-based approach to online searching scenario. Moreover, three common application scenarios are introduced. Finally, a number of experiments are conducted to study the functional and non-functional performance of our approach. The comprehensive results of experiments show that our approach provides better Web service searching results.

- Chapter 6
  This chapter presents a fault tolerance framework for building robust cloud applications at runtime. Our approach adopts dynamic QoS information to enable automatic system reconfiguration. We first introduce the architecture of our framework in voluntary-resource cloud. Then, we present the work procedures of our approach in detail, including 5 phases: primary selection, replicas selection, request execution, primary updating, and replica updating. After that, we conduct real-world experiments by deploying the prototype of our approach as a middleware in a voluntary-resource cloud environment, which consists of 257 distributed computers located in 26 countries. The experimental results show that our approach guarantees high reliability which enables good performance of cloud systems.
- Chapter 7
  The last chapter summarizes this book and provides some future directions that can be explored.

In order to make each of these chapters self-contained, some critical contents, e.g., model definitions or motivations having appeared in previous chapters, may be briefly reiterated in some chapters.

References

1. N. Ahmed, M. Linderman, J. Bryant, Towards mobile data streaming in service oriented architecture, in *Proceeding of SRDS'10*, pp. 323–327
2. E. Al-Masri, Q.H. Mahmoud, Investigating web services on the world wide web, in *Proceedings of the 17th International Conference on World Wide Web* (ACM, 2008), pp. 795–804
3. M. Alrifai, T. Risse, Combining global optimization with local selection for efficient QoS-aware service composition, in *Proceeding of International Conference on World Wide Web (WWW'09)* (2009), pp. 881–890
4. M. Alrifai, D. Skoutas, T. Risse, Selecting skyline services for QoS-based web service composition, in *Proceeding of WWW'10* (2010), pp. 11–20
5. D. Ardagna, B. Pernici, Adaptive service composition in flexible processes. IEEE Trans. Softw. Eng. **33**(6), 369–384 (2007)
6. M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., A view of cloud computing. Commun. ACM **53**(4), 50–58 (2010)
7. A. Avizienis, The methodology of N-version programming, in *Software Fault Tolerance* (1995), pp. 23–46
8. W. Balke, M. Wagner, Towards personalized selection of web services, in *Proceeding of WWW'03* (2003)
9. C. Bird, N. Nagappan, H. Gall, B. Murphy, P. Devanbu, Putting it all together: using socio-technical networks to predict failures, in *Proceeding of ISSRE'09* (2009), pp. 109–119
10. P. Bonatti, P. Festa, On optimal service selection, in *Proceeding of WWW'05* (2005), pp. 530–538

11. J. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in *Proceedings of UAI'98* (1998), pp. 43–52
12. X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. Kim, P. Compton, A. Mahidadia, Learning collaborative filtering and its application to people to people recommendation in social networks, in *Proceeding of ICDM'10* (2010), pp. 743–748
13. V. Cardellini, E. Casalicchio, V. Grassi, F. Lo Presti, Flow-based service selection for web service composition supporting multiple QoS classes, in *IEEE International Conference on Web Services* (2007), pp. 743–750
14. V. Cardellini, E. Casalicchio, V. Grassi, F. Lo Presti, R. Mirandola, QoS-driven runtime adaptation of service oriented architectures, in *Proceedings of the the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering* (ACM, 2009), pp. 131–140
15. J. Cardoso, A. Sheth, J. Miller, J. Arnold, K. Kochut, Quality of service for workflows and web service processes. Web Semant. Sci. Serv. Agents. World Wide Web **1**(3), 281–308 (2004)
16. M. Castro, B. Liskov, Practical Byzantine fault tolerance. Oper. Syst. Rev. **33**, 173–186 (1998)
17. P. Chan, M. R. Lyu, M. Malek, Reliableweb services: methodology, experiment and modeling, in *IEEE International Conference on Web Services* (IEEE, 2007), pp. 679–686
18. P.P.W. Chan, M.R. Lyu, M. Malek, Making services fault tolerant, in Service Availability (Springer, Berlin, 2006), pp. 43–61
19. W. Chen, J. Chu, J. Luan, H. Bai, Y. Wang, E. Chang, Collaborative filtering for orkut communities: discovery of user latent behavior, in *Proceeding of WWW'09* (2009), pp. 681–690
20. X. Chen, X. Liu, Z. Huang, H. Sun, Regionknn: a scalable hybrid collaborative filtering algorithm for personalized web service recommendation, in *IEEE International Conference on Web Services* (IEEE, 2010), pp. 9–16
21. R.C. Cheung, A user-oriented software reliability model. IEEE Trans. Softw. Eng. **2**, 118–125 (1980)
22. M. Creeger, Cloud computing: an overview. ACM Queue **7**(5), 1–5 (2009)
23. A. Danak, S. Mannor, Resource allocation with supply adjustment in distributed computing systems, in *Proceeding of ICDCS'10* (2010), pp. 498–506
24. V. Deora, J. Shao, W. Gray, N. Fiddian, A quality of service management framework based on user expectations, in *Proceeding of ICSOC'03* (2003), pp. 104–114
25. M. Deshpande, G. Karypis, Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst. (TOIS) **22**(1), 143–177 (2004)
26. X. Dong, A. Halevy, J. Madhavan, E. Nemes, J. Zhang, Similarity search for web services, in *Proceeding 30th International Conference on Very Large Data Bases (VLDB'04)* (2004), pp. 372–383
27. J. El Haddad, M. Manouvrier, G. Ramirez, M. Rukoz, QoS-driven selection of web services for transactional composition, in *Proceeding of ICWS'08* (IEEE, 2008), pp. 653–660
28. J. El Haddad, M. Manouvrier, M. Rukoz, Tqos: transactional and QoS-aware selection algorithm for automatic web service composition. IEEE Trans. Serv. Comput., 73–85 (2010)
29. T. Erl, *Service-oriented Architecture*, vol. 8 (Prentice Hall, New York, 2005)
30. H. Foster, S. Uchitel, J. Magee, J. Kramer, Model-based verification of web service compositions, in *IEEE International Conference on Automated Software Engineering* (IEEE, 2003), pp. 152–161
31. R. Ghosh, F. Longo, V. Naik, K. Trivedi, Quantifying resiliency of IaaS cloud, in *Proceeding of SRDS'10* (2010), pp. 343–347
32. S.S. Gokhale, K.S. Trivedi, Reliability prediction and sensitivity analysis based on software architecture, in *International Symposium on Software Reliability Engineering* (IEEE, 2002), pp. 64–75
33. K. Gomadam, A. Ranabahu, M. Nagarajan, A.P. Sheth, K. Verma, A faceted classification based approach to search and rank web apis, in *Proceeding of 6th International Conference on Web Services (ICWS'08)* (2008), pp. 177–184
34. S. Gong, A collaborative filtering recommendation algorithm based on user clustering and item clustering. J. Softw. **5**(7), 745–752 (2010)

35. V. Grassi, S. Patella, Reliability prediction for service-oriented computing environments. IEEE Internet Comput. **10**(3), 43–49 (2006)
36. Y. Hao, Y. Zhang, J. Cao, WSXplorer: searching for desired web services, in *Proceeding of 19th International Conference on Advanced Information System Engineering (CaiSE'07)* (2007), pp. 173–187
37. B. Hayes, Cloud computing. Commun. ACM **51**(7), 9–11 (2008)
38. T. Henzinger, A. Singh, V. Singh, T. Wies, D. Zufferey, FlexPRICE: flexible provisioning of resources in a cloud environment, in *Proceeding of CLOUD'10* (2010), pp. 83–90
39. J. Herlocker, J. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in *Proceeding of SIGIR'99* (1999), pp. 230–237
40. T. Hofmann, Collaborative filtering via gaussian probabilistic latent semantic analysis, in *Proceeding of SIGIR'03* (2003), pp. 259–266
41. T. Hofmann, Latent semantic models for collaborative filtering. ACM Trans. Inf. Syst. (TOIS) **22**(1), 89–115 (2004)
42. M.C. Jaeger, G. Rojec-Goldmann, G. Muhl, Qos aggregation for web service composition using workflow patterns, in *IEEE International Enterprise Distributed Object Computing Conference* (IEEE, 2004), pp. 149–159
43. M. Jamali, M. Ester, Trustwalker: a random walk model for combining trust-based and item-based recommendation, in *Proceeding of KDD'09* (2009), pp. 397–406
44. Y. Jianjun, G. Shengmin, S. Hao, Z. Hui, X. Ke, A kernel based structure matching for web services search, in *Proceeding of 16th International Conference on World Wide Web (WWW'07)* (2007), pp. 1249–1250
45. R. Jin, J. Chai, L. Si, An automatic weighting scheme for collaborative filtering, in *Proceeding of SIGIR'04* (2004), pp. 337–344
46. G. Jung, M. Hiltunen, K. Joshi, R. Schlichting, C. Pu, Mistral: dynamically managing power, performance, and adaptation cost in cloud infrastructures, in *Proceeding of ICDCS'10* (2010), pp. 62–73
47. K. Karta, An investigation on personalized collaborative filtering for web service selection. *Honours Programme thesis, University of Western Australia, Brisbane, 2005*
48. A. Keller, H. Ludwig, The wsla framework: specifying and monitoring service level agreements for web services. J. Netw. Syst. Manag. **11**(1), 57–81 (2003)
49. K. Kim, H. Welch, Distributed execution of recovery blocks: an approach for uniform treatment of hardware and software faults in real-time applications. IEEE Trans. Comput. **38**(5), 626–636 (2002)
50. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
51. W. Li, J. He, Q. Ma, I. Yen, F. Bastani, R. Paul, A framework to support survivable web services, in *Proceeding of IPDPS'05* (2005), p. 93b, 2005
52. D. Liang, C.-L. Fang, C. Chen, F. Lin, Fault tolerant web service, in *Tenth Asia-Pacific Software Engineering Conference* (IEEE, 2003), pp. 310–319
53. D. Liang, C.-L. Fang, S.-M. Yuan, C. Chen, G.E. Jan, A fault-tolerant object service on corba. J. Syst. Softw. **48**(3), 197–211 (1999)
54. G. Linden, B. Smith, J. York, Amazon. com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)
55. Y. Liu, A.H. Ngu, L.Z. Zeng, Qos computation and policing in dynamic web service selection, in *Proceeding of 13th International Conference on World Wide Web (WWW'04)* (2004), pp. 66–73
56. A. Luckow, B. Schnor, Service replication in grids: ensuring consistency in a dynamic, failure-prone environment, in *IEEE International Symposium on Parallel and Distributed Processing* (IEEE, 2008), pp. 1–7
57. H. Ludwig, A. Keller, A. Dan, R. King, R. Franck, A service level agreement language for dynamic electronic services. Electr. Commer. Res. **3**(1–2), 43–59 (2003)
58. M. Lyu, *Software Fault Tolerance*, (Wiley, 1995)
59. M. Lyu et al., Handbook of Software Reliability Engineering (1996)

60. E.M. Maximilien, M.P. Singh, Conceptual model of web service reputation. ACM Sigmod Rec. **31**(4), 36–41 (2002)
61. D.A. Menascé, Qos issues in web services. IEEE Internet Comput. **6**(6), 72–75 (2002)
62. M. Merideth, A. Iyengar, T. Mikalsen, S. Tai, I. Rouvellou, P. Narasimhan, Thema: Byzantine-fault-tolerant middleware for web-service applications, in *Proceeding of SRDS'05* (2005), pp. 131–140
63. T.O. Reilly, What is Web 2.0: design patterns and business models for the next generation of software. Commun. Strateg. **65**, 17 (2007)
64. J. O'Sullivan, D. Edmond, A. Ter Hofstede, What's in a service? Distrib. Parallel Databases **12**(2–3), 117–133 (2002)
65. M. Ouzzani, A. Bouguettaya, Efficient access to web services. IEEE Internet Comput. **8**(2), 34–44 (2004)
66. M. Paolucci, T. Kawamura, T. R. Payne, K. P. Sycara, Semantic matching of web services capabilities, in *Proceeding of 1st International Semantic Web Conference (ISWC'02)*, (2002), pp. 333–347
67. P. Plebani, B. Pernici, Urbe: web service retrieval based on similarity evaluation. IEEE Trans. Knowl. Data Eng. **21**(11), 1629–1642 (2009)
68. S. Ran, A model for web services discovery with QoS. ACM Sigecom Exch. **4**(1), 1–10 (2003)
69. S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in *Proceeding of WSDM'10* (2010), pp. 81–90
70. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in *Proceeding of CSCW'94*, (1994), pp. 175–186
71. R.H. Reussner, H.W. Schmidt, I.H. Poernomo, Reliability prediction for component-based software architectures. J. Syst. Softw. **66**(3), 241–252 (2003)
72. S. Rosario, A. Benveniste, S. Haar, C. Jard, Probabilistic QoS and soft contracts for transaction-based web services orchestrations. IEEE Trans. Serv. Comput. **1**(4), 187–200 (2008)
73. A. Sahai, A. Durante, V. Machiraju, Towards automated sla management for web services, *Hewlett-Packard Research Report HPL-2001-310 (R. 1)* (2002)
74. R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization. Adv. Neural Inf. Process. Syst. (NIPS) **20**, 1257–1264 (2008)
75. J. Salas, F. Perez-Sorrosal, M. Patiño-Martínez, R. Jiménez-Peris, WS-replication: a framework for highly available web services, in *Proceeding of WWW'06* (2006), pp. 357–366
76. N. Salatge, J. Fabre, Fault tolerance connectors for unreliable web services, in *Proceeding of IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)* (2007), pp. 51–60
77. G. T. Santos, L. C. Lung, C. Montez, Ftweb: a fault tolerant infrastructure for web services, in *IEEE International EDOC Enterprise Computing Conference* (IEEE, 2005), pp. 95–105
78. B. Sarwar, G. Karypis, J. Konstan, J. Reidl, Item-based collaborative filtering recommendation algorithms, in *Proceeding of WWW'01* (2001), pp. 285–295
79. D. Serrano, M. Patiño-Martínez, R. Jiménez-Peris, B. Kemme, An autonomic approach for replication of internet-based services, in *Proceeding of SRDS'08* (2008), pp. 127–136
80. L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, H. Mei, Personalized QoS prediction for web services via collaborative filtering, in *Proceeding of ICWS'07* (2007), pp. 439–446
81. Y. Shi, M. Larson, A. Hanjalic, Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering, in *Proceeding of Recsys'09* (2009), pp. 125–132
82. L. Si, R. Jin, Flexible mixture model for collaborative filtering. ICML **3**, 704–711 (2003)
83. P. Singla, M. Richardson, Yes, there is a correlation:-from social networks to personal behavior on the web, in *Proceeding of WWW'08* (2008), pp. 655–664
84. A. Soydan Bilgin, M. P. Singh, A daml-based repository for QoS-aware semantic web service selection, in *IEEE International Conference on Web Services* (IEEE, 2004), pp. 368–375
85. R.M. Sreenath, M.P. Singh, Agent-based service selection. Web Semant. Sci. Serv. Agents. World Wide Web **1**(3), 261–279 (2004)
86. N. Thio, S. Karunasekera, Automatic measurement of a QoS metric for web service recommendation, in *Software Engineering Conference* (IEEE, 2005), pp. 202–211

87. K. Tsakalozos, M. Roussopoulos, V. Floros, A. Delis, Nefeli: hint-based execution of workloads in clouds, in *Proceeding of ICDCS'10* (2010), pp. 74–85
88. K. Verma, K. Sivashanmugam, A. Sheth, A. Patil, S. Oundhakar, J. Miller, Meteor-s wsdi: a scalable p2p infrastructure of registries for semantic publication and discovery of web services. Inf. Technol. Manag. **6**(1), 17–39 (2005)
89. M. Vieira, N. Antunes, H. Madeira, Using web security scanners to detect vulnerabilities in web services, in *IEEE/IFIP International Conference on Dependable Systems and Networks* (IEEE, 2009), pp. 566–571
90. J. Wang, A. De Vries, M. Reinders, Unifying user-based and item-based collaborative filtering approaches by similarity fusion, in *Proceeding of SIGIR'06* (2006), pp. 501–508
91. Y. Wang, E. Stroulia, Semantic structure matching for assessing web service similarity, in *Proceeding of 1st International Conference on Service Oriented Computing (ICSOC'03)* (2003), pp. 194–207
92. Wikipedia, http://en.wikipedia.org/wiki/cloud_computing
93. G. Wu, J. Wei, X. Qiao, L. Li, A bayesian network based qos assessment model for web services, in *IEEE International Conference on Services Computing* (IEEE, 2007), pp. 498–505
94. P. Xiong, Y. Fan, M. Zhou, QoS-aware web service configuration. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(4), 888–895 (2008)
95. P. Xiong, Y. Fan, M. Zhou, A petri net approach to analysis and composition of web services. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(2), 376–387 (2010)
96. G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, Z. Chen, Scalable collaborative filtering using cluster-based smoothing, in *Proceeding of SIGIR'05* (2005), pp. 114–121
97. T. Yu, Y. Zhang, K. Lin, Efficient algorithms for Web services selection with end-to-end QoS constraints. ACM Trans. Web (TWEB) **1**(1), 6 (2007)
98. L. Zeng, B. Benatallah, A. Ngu, M. Dumas, J. Kalagnanam, H. Chang, QoS-aware middleware for web services composition. IEEE Trans. Softw. Eng. (TSE) **30**(5), 311–327 (2004)
99. L. Zhang, S. Cheng, C. Chang, Q. Zhou, A pattern-recognition-based algorithm and case study for clustering and selecting business services. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **42**(1), 102–114 (2012)
100. L.-J. Zhang, J. Zhang, H. Cai, *Services Computing* (Springer, Berlin, 2007)
101. Y. Zhang, Z. Zheng, M. Lyu, WSExpress: a QoS-aware search engine for Web services, in *Proceeding of ICWS'10* (2010), pp. 91–98
102. Y. Zhang, Z. Zheng, M. Lyu, Wspred: a time-aware personalized QoS prediction framework for web services, in *Proceeding of IEEE Symposium on Software Reliability Engineering (ISSRE'11)* (2011), pp. 210–219
103. Y. Zhang, Z. Zheng, M.R. Lyu, Bftcloud: a byzantine fault tolerance framework for voluntary-resource cloud computing, in *IEEE International Conference on Cloud Computing (CLOUD)* (IEEE, 2011), pp. 444–451
104. Y. Zhang, Z. Zheng, M.R. Lyu, Exploring latent features for memory-based qos prediction in cloud computing, in *IEEE Symposium on Reliable Distributed Systems (SRDS)* (IEEE, 2011), pp. 1–10
105. Y. Zhang, Z. Zheng, M.R. Lyu, Real-time performance prediction for cloud components, in *IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops (ISORCW)* (IEEE, 2012), pp. 106–111
106. Y. Zhang, Z. Zheng, M.R. Lyu, An online performance prediction framework for service-oriented systems. IEEE Trans. Syst. Man Cybern. Syst. (TSMC) **44**(9), 1169–1181 (2014)
107. W. Zhao, BFT-WS: a Byzantine fault tolerance framework for web services, in *Proceeding of EDOC'07* (2008), pp. 89–96
108. Z. Zheng, Y. Zhang, M. Lyu, CloudRank: a QoS-driven component ranking framework for cloud computing, in *Proceeding of SRDS'10* (2010), pp. 184–193
109. Z. Zheng, Y. Zhang, M. Lyu, Distributed QoS evaluation for real-world web services, in *Proceeding of ICWS'10* (2010), pp. 83–90

Chapter 2
Neighborhood-Based QoS Prediction

Abstract With the increasing popularity of cloud computing as a solution for building high-quality applications on distributed components, efficiently evaluating user-side quality of cloud components becomes an urgent and crucial research problem. However, invoking all the available cloud components from user-side for evaluation purpose is expensive and impractical. To address this critical challenge, we propose a neighborhood-based approach, called CloudPred, for collaborative and personalized quality prediction of cloud components. CloudPred is enhanced by feature modeling on both users and components. Our approach CloudPred requires no additional invocation of cloud components on behalf of the cloud application designers. The extensive experimental results show that CloudPred achieves higher QoS prediction accuracy than other competing methods. We also publicly release our large-scale QoS dataset for future related research in cloud computing.

2.1 Overview

In the cloud environment, designers of cloud applications, denoted as component users, can choose from a broad pool of cloud components when creating cloud applications. These cloud components are usually invoked remotely through communication links. Quality of the cloud applications is greatly influenced by the quality of communication links and the distributed cloud components. To build a high-quality cloud application, non-functional Quality-of-Service (QoS) performance of cloud components becomes an important factor for application designers when making component selection [2]. Moreover, for the existing cloud applications, by replacing low-quality components with better ones, the overall quality of cloud applications can be improved.

Different from traditional component-based systems, cloud applications invoke components remotely by Internet connections. User-side QoS experiences of cloud components is thus greatly influenced by the unpredictable communication links. Personalized QoS evaluation is required for each user at the user-side. The most straightforward approach is to evaluate all the candidate components at the user-side. However, this approach is impractical in reality, since invocations of cloud

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_2

components may be charged. Even if the invocations are free, executing a large number of components invocations is time consuming and resource consuming.

Based on the above analysis, it is crucial for the cloud platform to deliver a personalized QoS information service to the application designers for cloud component evaluation. In order to provide personalized QoS values on m cloud components for n users by evaluation, at least $n \times m$ invocations need to be executed, which is almost impossible when n and m are very large. However, without sufficient and accurate personalized QoS values of cloud components, it is difficult for the application designers to select optimal cloud component for building high-quality cloud applications. It is an urgent task for the cloud platform providers to develop an efficient and personalized prediction approach for delivering the QoS information service to cloud application designers.

To address this critical challenge, we propose a neighborhood-based approach, called CloudPred, for personalized QoS prediction of cloud components. CloudPred is enhanced by feature modeling on both users and components. The idea of CloudPred is that users sharing similar characteristics (e.g., location, bandwidth) would receive similar QoS usage experiences on the same component. The QoS value of cloud component c observed by user u can be predicted by exploring the QoS experiences from similar users of u. A user is similar to u if they share similar characteristics. The characteristics of different users can be extracted from their QoS experiences on different components by performing nonnegative matrix factorization (NMF). By sharing local QoS experience among users, our approach CloudPred can effectively predict the QoS value of a cloud component c even if the current user u has never invoked the component c before. The experimental results show that compared with other well-known collaborative prediction approaches, CloudPred achieves higher QoS prediction accuracy of cloud components. Since CloudPred can precisely characterize users features (will be introduced in Sect. 2.3.2), even if some users have few local QoS information, CloudPred can still achieve high prediction accuracy.

In summary, this chapter makes the following contributions:

1. We formally identify the research problem of QoS value prediction in cloud computing and propose a novel neighborhood-based approach, named CloudPred, for personalized QoS value prediction of cloud components. CloudPred learns the characteristics of users by nonnegative matrix factorization (NMF) and explores QoS experiences from similar users to achieve high QoS value prediction accuracy. We consider CloudPred as the first QoS value prediction approach in cloud computing literature.
2. We conduct large-scale experiments to study the prediction accuracy of our CloudPred compared with other approaches. The experimental results show the effectiveness of our approach. Moreover, we also publicly release our large-scale QoS dataset for future research.

The remainder of this chapter is organized as follows: Sect. 2.2 describes the collaborative QoS framework in cloud environment. Section 2.3 presents our CloudPred approach in detail. Section 2.4 introduces the experimental results. Section 2.5 concludes the chapter.

2.2 Collaborative Framework in Cloud

Figure 2.1 shows the system architecture in cloud computing. In a cloud environment, the cloud provider holds a large number of distributed cloud components (e.g., databases, servers, Web services), which can be provided to designers for developing various cloud applications. The cloud application designers, called component users in this chapter, are located in different geographic and network environments. Since users invoke cloud components via different communication links, their usage experiences on cloud components are diverse in several QoS properties including response-time, throughput, etc. In order to provide personalized quality information of different components to application designers for optimal component selection, personalized QoS value prediction is an essential service of a cloud provider.

Within the cloud platform provided by a cloud provider, there are several modules implemented for managing the cloud components. Examples of management modules include *Task Scheduler*, which is responsible for task scheduling, *SLA Wrapper*, which is responsible for service-level negotiation between cloud provider and users, etc. In this chapter, we focus on the design of *QoS Monitor*, which is responsible for monitoring the QoS performance of cloud components from the users' perspective. The *QoS Monitor* consists of two subunits: *Collector*, which is used to collect QoS

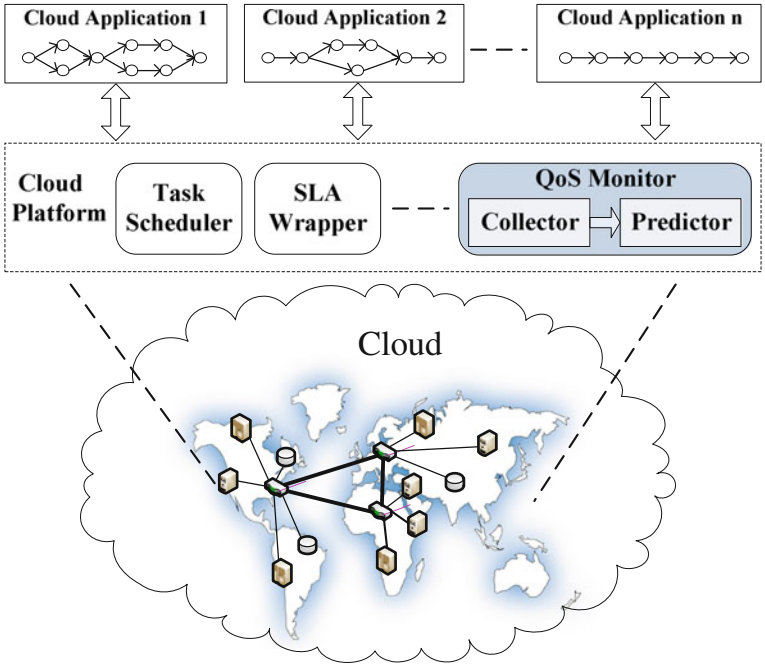


Fig. 2.1 System architecture. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

usage information from various component users, and *Predictor*, which is supposed to provide personalized QoS value prediction for different component users.

The idea of our approach is to share local cloud component usage experience from different component users, to combine this local information to get a global QoS information of all components, and to make personalized QoS value prediction based on both global and local information. As shown in Fig. 2.1, each component user keeps local records of QoS usage experiences on cloud components. Since cloud applications are running on an identical cloud platform, QoS information can be collected by an identical interface on the platform side. If a component user would like to get personalized QoS information service from the cloud provider, authorization should be given to *Collector* for accessing its local QoS records. *Collector* then collects those local QoS records from different component users. Based on the collected QoS information, *Predictor* can perform personalized QoS value prediction and forward the prediction results to component users for optimizing the design of cloud applications. The detailed collaborative prediction approach will be presented in Sect. 2.3.

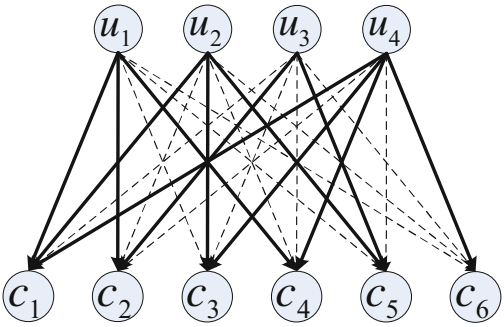
2.3 Collaborative QoS Prediction

We first formally describe the QoS value prediction problem on cloud components in Sect. 2.3.1. Then, we learn the user-specific and component-specific features by running latent features learning algorithm in Sect. 2.3.2. Based on the latent features, similarities between users and components are calculated in Sect. 2.3.3. Finally, the missing QoS values are predicted by applying the proposed algorithm CloudPred in Sect. 2.3.4.

2.3.1 Problem Description

Let us first consider a typical toy example in Fig. 2.2a. In this bipartite graph $G = (U \cup C, E)$, its vertices are divided into two disjoint sets U and C such that each edge in E connects a vertex in U and one in C. Let $U = \{u_1, u_2, \ldots, u_4\}$ be the set of component users, $C = \{c_1, c_2, \ldots, c_6\}$ denote the set of cloud components, and E (solid lines) represent the set of invocations between U and C. This bipartite graph G is modeled as a weighted directed graph. Given a pair (i, j), $u_i \in U$, and $c_j \in C$, edge e_{ij} is included in E if user u_i has invoked component c_j before. The weight w_{ij} on edge e_{ij} corresponds to the QoS value (e.g., response-time in this example) of that invocation. Given the set E, our task is to effectively predict the weight of potential invocations (the broken lines).

The process of cloud component QoS value prediction is illustrated by a user-component matrix as shown in Fig. 2.2b, in which each entry denotes an observed



(a) User-Component Invocation Graph

	c_1	c_2	c_3	c_4	c_5	c_6
u_1	0.98	0.23		0.22		
u_2	0.13		0.27		0.25	
u_3		0.37			0.36	
u_4	0.69		0.22	0.22		0.34

(b) User-Component Matrix

	c_1	c_2	c_3	c_4	c_5	c_6
u_1	0.98	0.23	**0.24**	0.22	**0.22**	**0.53**
u_2	0.13	**0.26**	0.27	**0.25**	0.25	**0.43**
u_3	**0.85**	0.37	**0.38**	**0.36**	0.36	**0.12**
u_4	0.69	**0.23**	0.22	0.22	**0.22**	0.34

(c) Predicted User-Component Matrix

Fig. 2.2 Toy example for QoS prediction. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

weight in Fig. 2.2a. The problem we study in this chapter is then how to precisely predict the missing entries in the user-component matrix based on the existing entries. Once the missing entries are accurately predicted, we can provide users with personalized QoS information, which is valuable for automatic component ranking, component selection, task scheduling, etc.

We observe that although about half of the entries are already known in Fig. 2.2b, every pair of users still have very few commonly invoked components (e.g., u_1 and u_2 only invoke c_1 in common, u_3 and u_4 have no commonly invoked components even if together they invoke all the six components). Since the similarity between two users are calculated by comparing their obtained QoS values on common components, the problem of few common components observed above makes it extremely difficult to precisely calculate similarity between users. Motivated by latent factor model [6], we therefore first factorize the sparse user-component matrix and then use $V^T H$ to approximate the original matrix, where the low-dimensional matrix V denotes the user latent feature space, and the low-dimensional matrix H represents the low-dimensional item latent feature space. The rows in V and H represent different features. Each column in V represents an user, and each column in H denotes a component. The value of a entry in the matrices indicates how the associated feature applies to the corresponding user or component. In this example, we use four dimensions to perform the matrix factorization and obtain:

$$V = \begin{bmatrix} 0.32 & 0.15 & 0.31 & 0.33 \\ 0.23 & 0.15 & 0.26 & 0.28 \\ 0.30 & 0.20 & 0.24 & 0.34 \\ 0.47 & 0.23 & 0.59 & 0.21 \end{bmatrix},$$

$$H = \begin{bmatrix} 0.73 & 0.35 & 0.31 & 0.26 & 0.32 & 0.42 \\ 0.60 & 0.31 & 0.27 & 0.22 & 0.28 & 0.36 \\ 0.69 & 0.37 & 0.32 & 0.27 & 0.33 & 0.45 \\ 0.95 & 0.46 & 0.42 & 0.35 & 0.41 & 0.54 \end{bmatrix},$$

where columns in V and H denote the latent feature vectors of users and components, respectively.

Note that V and H are dense matrices with all entries available. Then, we calculate the similarity between users and components using four-dimensional matrices V and H, respectively. Therefore, all the missing values can be predicted by employing neighborhood-based collaborative method, as shown in Fig. 2.2c.

Now, we formally define the problem of cloud component QoS value prediction as follows: Given a set of users and a set of components, predict the missing QoS value of components when invoked by users based on existing QoS values. More precisely:

> Let U be the set of m users and C be the set of n components. A QoS element is a triplet (i, j, q_{ij}) representing the observed quality of component c_j by user u_i, where $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$ and $q_{ij} \in \mathbb{R}^k$ is a k-dimensional vector representing the QoS values of k^{th} criteria. Let Ω be the set of all pairs $\{i, j\}$ and Λ be the set of all known pairs (i, j) in Ω. Consider a matrix $W \in \mathbb{R}^{m \times n}$ with each entry w_{ij} representing the observed k^{th} criterion value of component c_j by user u_i. Then, the missing entries $\{w_{ij} | (i, j) \in \Omega - \Lambda\}$ should be predicted based on the existing entries $\{w_{ij} | (i, j) \in \Lambda\}$.

Typically the QoS values can be integers from a given range (e.g., $\{0, 1, 2, 3\}$) or real numbers of a close interval (e.g., $[-20, 20]$). Without loss of generality, we can map the QoS values to the interval $[0, 1]$ using the function $f(x) = (x - w_{min})/(w_{max} - w_{min})$, where w_{max} and w_{min} are the maximum and minimum QoS values, respectively.

2.3.2 *Latent Features Learning*

In order to learn the features of the users and components, we employ matrix factorization to fit a factor model to the user-component matrix. This method focuses on filtering the user-component QoS value matrix using low-rank approximation. In other words, we factorize the QoS matrix into two low-rank matrices V and H. The idea behind the factor model is to derive a high-quality low-dimensional feature representation of users and components based on analyzing the user-component matrix. The premise behind a low-dimensional factor model is that there is only a small number of factors influencing QoS usage experiences and that a user's QoS

usage experience vector is determined by how each factor applies to that user and the items.

Consider the matrix $W \in \mathbb{R}^{m \times n}$ consisting of m users and n components. Let $V \in \mathbb{R}^{l \times m}$ and $H \in \mathbb{R}^{l \times n}$ be the latent user and component feature matrices. Each column in V represents the l-dimensional user-specific latent feature vector of a user, and each column in H represents the l-dimensional component-specific latent feature vector of a component. We employ an approximating matrix $\tilde{W} = V^T H$ to fit the user-item matrix W:

$$w_{ij} \approx \tilde{w}_{ij} = \sum_{k=1}^{l} v_{ki} h_{kj}, \tag{2.1}$$

The rank l of the factorization is generally chosen so that $(m + n)l < mn$, since V and H are low-rank feature representations [3]. The product $V^T H$ can be regarded as a compressed form of the data in W.

Note that the low-dimensional matrices V and H are unknown and need to be learned from the obtained QoS values in user-component matrix W. In order to optimize the matrix factorization, we first construct a cost function to evaluate the quality of approximation. The distance between two nonnegative matrices is usually employed to define the cost function. One useful measure of the matrices' distance is the Euclidean distance:

$$F(W, \tilde{W}) = \|W - \tilde{W}\|_F^2 = \sum_{ij} (w_{ij} - \tilde{w}_{ij})^2, \tag{2.2}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm.

In this chapter, we conduct matrix factorization as solving an optimization problem by employing the optimized objective function in [3]:

$$\begin{aligned} \min_{V,H} \quad & f(V, H) = \sum_{(i,j) \in \Lambda} [\tilde{w}_{ij} - w_{ij} \log \tilde{w}_{ij}], \\ s.t. \quad & \tilde{w}_{i,j} = \sum_{k=1}^{l} v_{ki} h_{kj}, \\ & V \geq 0, \\ & H \geq 0. \end{aligned} \tag{2.3}$$

where $V, H \geq 0$ is the nonnegativity constraints leading to allow only additive combination of features.

In order to minimize the objective function in Eq. (2.3), we apply incremental gradient descent method to find a local minimum of $f(V, H)$, where one gradient step intends to decrease the square of prediction error of only one rating, that is, $\tilde{w}_{ij} - w_{ij} \log \tilde{w}_{ij}$. We update the V and H in the direction opposite of the gradient in each iteration:

$$v_{ij} = v_{ij} \sum_k \frac{w_{ik}}{\tilde{w}_{ik}} h_{jk}, \tag{2.4}$$

$$h_{ij} = h_{ij} \sum_k \frac{w_{ik}}{\tilde{w}_{ik}} v_{jk}, \tag{2.5}$$

$$v_{ij} = \frac{v_{ij}}{\sum_k v_{kj}}, \tag{2.6}$$

$$h_{ij} = \frac{h_{ij}}{\sum_k h_{kj}}. \tag{2.7}$$

Algorithm 1 shows the iterative process for latent feature learning. We first initialize matrices V and H with small random nonnegative values. Iteration of the above update rules converges to a local minimum of the objective function given in Eq. (2.3).

Algorithm 1: Latent Features Learning Algorithm

Input: W, l
Output: V, H
Initialize $V \in \mathbb{R}^{l \times m}$ and $H \in \mathbb{R}^{l \times n}$ with small random numbers;
repeat
  for *all* $(i, j) \in \Lambda$ **do**
    $\tilde{w}_{ij} = \sum_k v_{ki} h_{kj}$;
  end
  for *all* $(i, j) \in \Lambda$ **do**
    $v_{ij} \leftarrow v_{ij} \sum_k \frac{w_{ik}}{\tilde{w}_{ik}} h_{jk}$;
    $h_{ij} \leftarrow h_{ij} \sum_k \frac{w_{ik}}{\tilde{w}_{ik}} v_{jk}$;
    $v_{ij} = \frac{v_{ij}}{\sum_k v_{kj}}$;
    $h_{ij} = \frac{h_{ij}}{\sum_k h_{kj}}$;
  end
  for *all* $(i, j) \in \Lambda$ **do**
    $\tilde{w}_{ij} = \sum_k v_{ki} h_{kj}$;
  end
until *Converge*;

2.3.3 Similarity Computation

Given the latent user and component feature matrices V and H, we can calculate the neighborhood similarities between different users and components by employing Pearson correlation coefficient (PCC) [5]. PCC is widely used in memory-based recommendation systems for similarity computation. Due to the high accuracy, we

adopt PCC in this chapter for the neighborhood similarity computation on both sets of users and components. The similarity between two users u_i and u_j is defined by performing PCC computation on their l-dimensional latent feature vectors V_i and V_j with the following equation:

$$S(u_i, u_j) = \frac{\sum_{k=1}^{l}(v_{ik} - \overline{v}_i)(v_{jk} - \overline{v}_j)}{\sqrt{\sum_{k=1}^{l}(v_{ik} - \overline{v}_i)^2}\sqrt{\sum_{k=1}^{l}(v_{jk} - \overline{v}_j)^2}}, \tag{2.8}$$

where $v_i = (v_{i1}, v_{i2}, \ldots, v_{il})$ is the latent feature vector of user u_i and v_{ik} is the weight on the kth feature. $\overline{v}_i$ is the average weight on l-dimensional latent features for user u_i. The similarity between two users $S(i, j)$ falls into the interval $[-1, 1]$, where a larger value indicates higher similarity.

Similar to the user similarity computation, we also employ PCC to compute the similarity between component c_i and item c_j as following:

$$S(c_i, c_j) = \frac{\sum_{k=1}^{l}(h_{ik} - \overline{h}_i)(h_{jk} - \overline{h}_j)}{\sqrt{\sum_{k=1}^{l}(h_{ik} - \overline{h}_i)^2}\sqrt{\sum_{k=1}^{l}(h_{jk} - \overline{h}_j)^2}}, \tag{2.9}$$

where $h_i = (h_{i1}, h_{i2}, \ldots, h_{il})$ is the latent feature vector of component c_i and h_{ik} is the weights on the k^{th} feature. $\overline{h}_i$ is the average weight on l-dimensional latent features for component c_i.

2.3.4 Missing QoS Value Prediction

After computing the similarities between users, we can identify similar neighbors to the current user by ordering similarity values. Note that PCC value falls into the interval $[-1, 1]$, where a positive value means similar and a negative value denotes dissimilar. In practice, QoS usage experience of less similar or dissimilar users may greatly decrease the prediction accuracy. In this chapter, we exclude those users with negative PCC values from the similar neighbor set and only employ the QoS usage experiences of users with Top-K largest PCC values for predicting QoS value of the current user. We refer to the set of Top-K similar users for user u_i as Ψ_i, which is defined as:

$$\Psi_i = \{u_k | S(u_i, u_k) > 0, rank_i(k) \leq K, k \neq i\}, \tag{2.10}$$

where $rank_i(k)$ is the ranking position of user u_k in the similarity list of user u_i, and K denotes the size of set Ψ_i.

Similarly, a set of Top-K similar components for component c_j can be denote as Φ_j by:

$$\Phi_j = \{c_k | S(c_j, c_k) > 0, rank_p(k) \leq K, k \neq j\}, \tag{2.11}$$

where $rank_j(k)$ is the ranking position of component c_k in the similarity list of component c_j, and K denotes the size of set Φ_j.

To predict the missing entry w_{ij} in the user-component matrix, user-based approaches employ the values of entries from Top-K similar users as follows:

$$w_{ij} = \overline{w}_i + \sum_{k \in \Psi_i} \frac{S(u_i, u_k)}{\sum_{a \in \Psi_i} S(u_i, u_a)} (w_{kj} - \overline{w}_k), \tag{2.12}$$

where $\overline{w}_i$ and $\overline{w}_k$ are the average observed QoS values of different components by users u_i and u_k, respectively.

For component-based approaches, entry values of Top-K similar components are employed for predicting the missing entry w_{ij} in the similar way:

$$w_{ij} = \overline{w}_j + \sum_{k \in \Phi_j} \frac{S(i_j, i_k)}{\sum_{a \in \Phi_j} S(i_j, i_a)} (w_{ik} - \overline{w}_k), \tag{2.13}$$

where $\overline{w}_j$ and $\overline{w}_k$ are the average available QoS values of component c_j and c_k by different users, respectively.

In user-component-based approaches, the predicted values in Eqs. (2.12) and (2.13) are both employed for more precise prediction in the following equation:

$$w^*_{ij} = \lambda \times w^u_{ij} + (1 - \lambda) \times w^c_{ij}, \tag{2.14}$$

where w^u_{ij} denotes the predicted value by user-based approach and w^c_{ij} denotes the predicted value by component-based approach. The parameter λ controls how much the hybrid prediction results rely on user-based approach or component-based approach. The proper value of λ can be trained on a small sample dataset extracted from the original one. We summarize the proposed algorithm in Algorithm 2.

2.4 Experiments

In this section, in order to show the prediction quality improvements of our proposed approach, we conduct several experiments to compare our approach with several state-of-the-art collaborative filtering prediction methods.

In the following, Sect. 2.4.1 gives the description of our experimental dataset, Sect. 2.4.2 defines the evaluation metrics, Sect. 2.4.3 compares the prediction quality of our approach with some other methods, and Sects. 2.4.4, 2.4.5, and 2.4.6 study the impact of training data density, Top-K, and dimensionality, respectively.

Algorithm 2: CloudPred Prediction Algorithm

Input: W, l, λ
Output: W^*

Learn V and H by applying Algorithm 1 on W;
for *all* $(u_i, u_j) \in U \times U$ **do**
  calculate the similarity $S(u_i, u_j)$ by Eq. (2.8);
end
for *all* $(c_i, c_j) \in C \times C$ **do**
  calculate the similarity $S(c_i, c_j)$ by Eq. (2.9);
end
for *all* $(i, j) \in \Lambda$ **do**
  construct similar user set Ψ_i by Eq. (2.10);
  construct similar component set Φ_j by Eq. (2.11);
end
for *all* $(i, j) \in \Omega - \Lambda$ **do**
  calculate w_{ij}^u by Eq. (2.12);
  calculate w_{ij}^i by Eq. (2.13);
  $w_{ij}^* = \lambda \times w_{ij}^u + (1 - \lambda) \times w_{ij}^c$;
end

2.4.1 Dataset Description

In real world, invoking thousands of commercial cloud components for large-scale experiments is very expensive. In order to evaluate the prediction quality of our proposed approach, we conduct experiments on our Web service QoS dataset [9]. Web service, a kind of cloud component, can be integrated into cloud applications for accessing information or computing service from a remote system. The Web service QoS dataset includes QoS performance of 5825 openly accessible real-world Web services from 73 countries. The QoS values are observed by 339 distributed computers located in 30 countries from PlanetLab, which is a distributed test bed consisting of hundreds of computers all over the world. In our experiment, each of the 339 computers keeps invocation records of all the 5825 Web services by sending null operating requests to capture the characteristics of communication links. Totally 1,974,675 QoS performance results are collected. Each invocation record is a k-dimensional vector representing the QoS values of k criteria. We then extract a set of 339×5825 user-component matrices, each of which stands for a particular QoS property, from the QoS invocation records. For simplicity, we use two matrices, which represent response-time and throughput QoS criteria, respectively, for experimental evaluation in this chapter. Without loss of generality, our approach can be easily extended to include more QoS criteria.

The statistics of Web service QoS dataset are summarized in Table 2.1. Response-time and throughput are within the range 0–20 s and 0–1000 kbps, respectively. The

Table 2.1 Statistics of WS QoS dataset. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Statistics	Response-time	Throughput
Scale	0–20 s	0–1000 kbps
Mean	0.910 s	47.386 kbps
Num. of users	339	339
Num. of web services	5828	5828
Num. of records	1,974,675	1,974,675

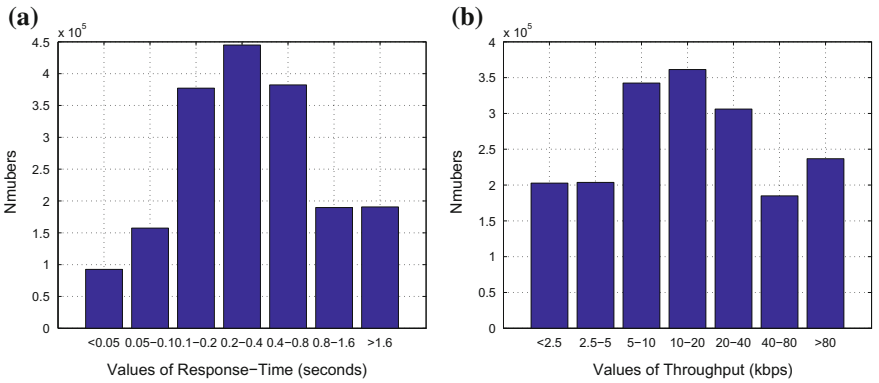


Fig. 2.3 Value distributions. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

means of response-time and throughput are 0.910 s and 47.386 kbps, respectively. Figure 2.3 shows the distributions of response-time and throughput. Most of the response-time values are between 0.1–0.8 s, and most of the throughput values are between 5–40 kbps.

2.4.2 Metrics

We assess the prediction quality of our proposed approach in comparison with other methods by computing mean absolute error (MAE) and root-mean-squared error (RMSE). The metric MAE is defined as:

$$MAE = \frac{\sum_{i,j} |w_{ij} - w^*_{ij}|}{N}, \tag{2.15}$$

and RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i,j} (w_{ij} - w^*_{ij})^2}{N}}, \tag{2.16}$$

where w_{ij} is the QoS value of Web service c_j observed by user u_i, w^*_{ij} denotes the QoS value of Web service c_j would be observed by user u_i as predicted by a method, and N is the number of predicted QoS values.

2.4.3 Performance Comparison

In this section, we compare the prediction accuracy of our proposed approach CloudPred with some state-of-the-art approaches:

1. UPCC (User-based collaborative filtering method using Pearson correlation coefficient): this method employs PCC to calculate similarities between users and predicts QoS value based on similar users [1, 7].
2. IPCC (Item-based collaborative filtering method using Pearson correlation coefficient): this method employs PCC to calculate similarities between Web services and predicts QoS value based on similar items (item refers to component in this chapter) [5].
3. UIPCC (User-item-based collaborative filtering method using Pearson correlation coefficient): this method is proposed by Ma et al. in [4]. It combines UPCC and IPCC approaches and predicts QoS value based on both similar users and similar Web services.
4. NMF (Nonnegative Matrix Factorization): This method is proposed by Lee and Seung in [3]. It applies nonnegative matrix factorization on user-item matrix for missing value prediction.

In this chapter, in order to evaluate the performance of different approaches in reality, we randomly remove some entries from the matrices and compare the values predicted by a method with the original ones. The matrices with missing values are in different sparsity. For example, 10% means that we randomly remove 90% entries from the original matrix and use the remaining 10% entries to predict the removed entries. The prediction accuracy is evaluated using Eqs. (2.15) and (2.16) by comparing the original value and the predicted value of each removed entry. Our proposed approach CloudPred performs matrix factorization in Sect. 2.3.2 and employs both similar users and similar Web services for predicting the removed entries. The parameter settings of our approach CloudPred are Top-K = 10, dimensionality = 20, and $\lambda = 0.5$ in the experiments. Detailed impact of parameters will be studied in Sects. 2.4.4, 2.4.5 and 2.4.6.

The experimental results are shown in Table 2.2. For each row in the table, we highlight the best performer among all methods. From Table 2.2, we can observe that our approach CloudPred obtains better prediction accuracy (smaller MAE and RMSE values) than other methods for both response-time and throughput under different matrix densities. The MAE and RMSE values of dense matrices (e.g., matrix density is 80 or 90%) are smaller than those of sparse matrices (e.g., matrix density is 10 or 20%), since a denser matrix provides more information for predicting the missing values. In general, the MAE and RMSE values of throughput are larger than those

Table 2.2 Performance comparisons (A smaller MAE or RMSE value means a better performance). ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Matrix density (%)	Metrics	Response-time (seconds)				
		IPCC	UPCC	UIPCC	NMF	CloudPred
10	MAE	0.7596	0.5655	0.5654	0.6754	**0.5306**
	RMSE	1.6133	1.3326	1.3309	1.5354	**1.2904**
20	MAE	0.7624	0.5516	0.5053	0.6771	**0.4745**
	RMSE	1.6257	1.3114	1.2486	1.5241	**1.1973**
80	MAE	0.6703	0.4442	0.3873	0.3740	**0.3704**
	RMSE	1.4102	1.1514	1.0785	1.1242	**1.0597**
90	MAE	0.6687	0.4331	0.3793	0.3649	**0.3638**
	RMSE	1.4173	1.1264	1.0592	1.1121	**1.0359**
Matrix density (%)	Metrics	Throughput (kbps)				
		IPCC	UPCC	UIPCC	NMF	CloudPred
10	MAE	31.6722	26.2015	22.6567	19.7700	**19.0009**
	RMSE	65.5220	61.9658	57.4653	57.3767	**51.8236**
20	MAE	35.1780	21.9313	18.1230	15.7794	**15.4203**
	RMSE	66.6028	56.5441	50.0435	50.1402	**44.8975**
80	MAE	29.9146	14.5497	12.4880	12.5107	**10.7881**
	RMSE	64.3079	44.3738	39.6017	39.2029	**36.8506**
90	MAE	29.9404	13.8761	12.0662	11.6960	**10.4722**
	RMSE	63.7149	42.5534	38.0763	36.7555	**35.9225**

of response-time because the scale of throughput is 0–1000 kbps, while the scale of response-time is 0–20 s. Compared with other methods, the improvements of our approach CloudPred are significant, which demonstrates that the idea of combining global and local information for QoS prediction is realistic and reasonable.

2.4.4 Impact of Matrix Density

In Fig. 2.4, we compare the prediction accuracy of all the methods under different matrix densities. We change the matrix density from 10 to 90% with a step value of 10%. The parameter settings in this experiment are Top-K = 10, dimensionality = 20, and $\lambda = 0.5$.

Figure 2.4a, b shows the experimental results of response-time, while Fig. 2.4c, d shows the experimental results of throughput. The experimental results show that our approach CloudPred achieves higher prediction accuracy than other competing methods under different matrix density. In general, when the matrix density is

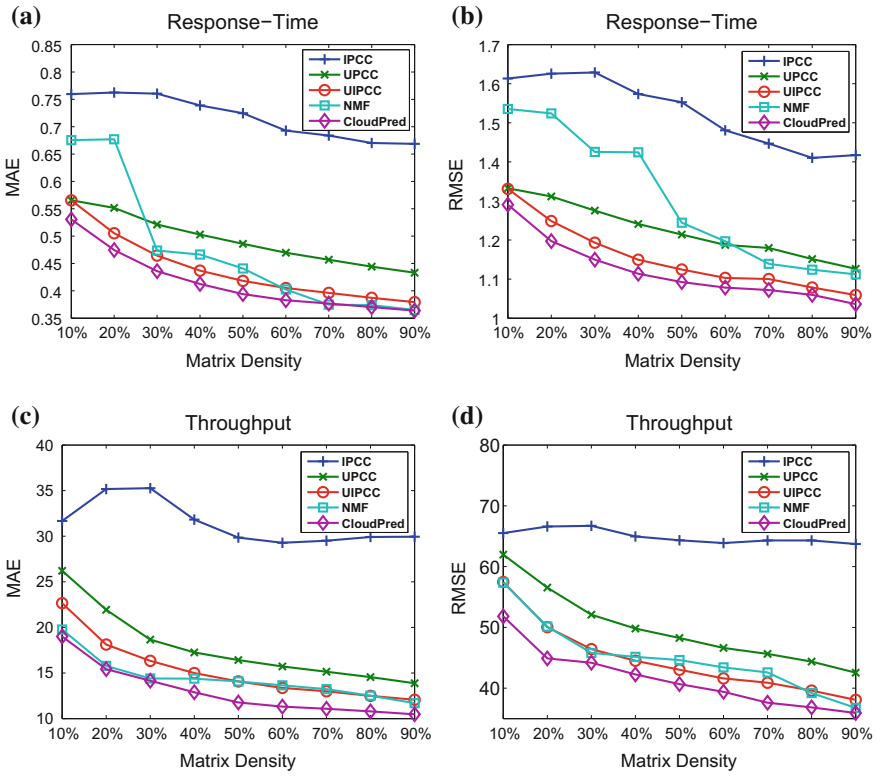


Fig. 2.4 Impact of matrix density. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

increased from 10 to 30%, the prediction accuracy of our approach CloudPred is significantly enhanced. When the matrix density is further increased from 30 to 90%, the enhancement of prediction accuracy is quite limited. This observation indicates that when the matrix is very sparse, collecting more QoS information will greatly enhance the prediction accuracy, which further demonstrates that sharing local QoS information among cloud component users could effectively provide personalized QoS estimation.

In the experimental results, we observe that the performance of IPCC is much worse than that of other methods. The reason is that in our Web service dataset, the number of users, which is 339, is much smaller than the number of components, which is 5258. When some entries are removed from the user-component matrices, the number of common users between two components, on average, is very small, which would greatly impact the accuracy of common user-based similarity computation between components. Therefore, the prediction accuracy of similar item-based method IPCC is greatly decreased by the inaccuracy similarity computation between components.

2.4.5 *Impact of Top-K*

The parameter Top-K determines the size of similar user and similar component sets. In Fig. 2.5, we study the impact of parameter Top-K by varying the values of Top-K from 10 to 50 with a step value of 10. Other parameter settings are dimensionality = 10 and $\lambda = 0.5$.

Figure 2.5a, b shows the MAE and RMSE results of response-time, respectively, while Fig. 2.5c, d shows the MAE and RMSE results of throughput, respectively. The experimental results show that our approach CloudPred achieves best prediction accuracy(smallest MAE and RMSE values) when Top-K is set around 10. Under both sparse matrix, whose density is 10%, and dense matrix, whose density is 90%, all the prediction accuracies decrease when we decrease the Top-K value from 10 to 2 or increase from 10 to 18. This is because too small Top-K value will exclude useful information from some similar users and similar components, while too large Top-K value will introduce noise from dissimilar users and dissimilar components, which will impact the prediction accuracy.

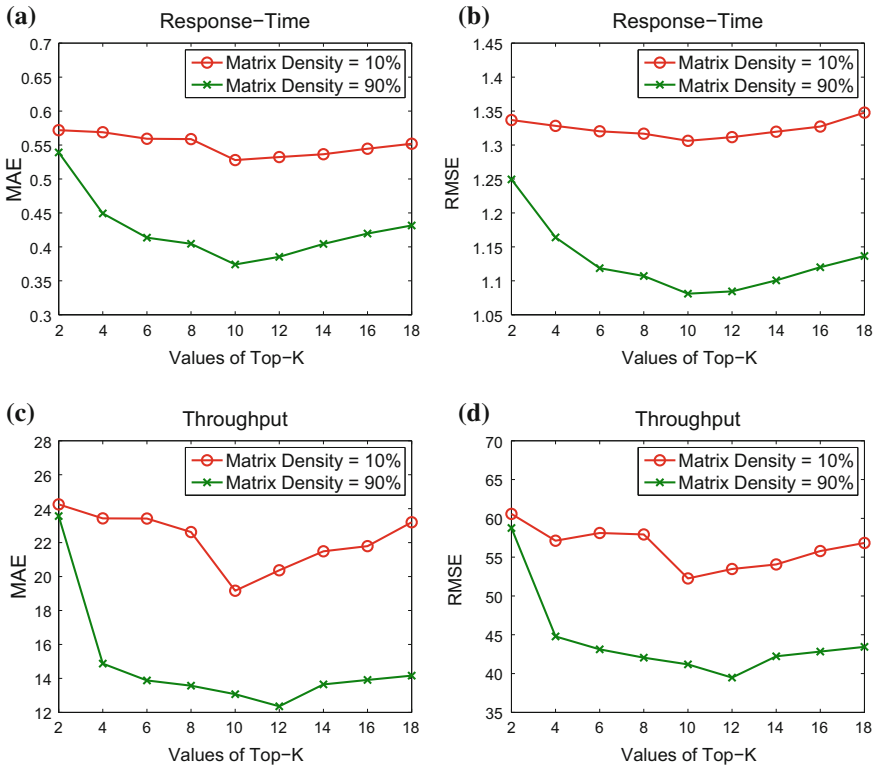


Fig. 2.5 Impact of Top-K. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

2.4.6 Impact of Dimensionality

The parameter dimensionality determines the number of latent features used to characterize user and cloud component. In Fig. 2.6, we study the impact of parameter dimensionality by varying the values of dimensionality from 10 to 50 with a step value of 10. Other parameter settings are Top-K = 10 and $\lambda = 0.5$.

Figure 2.6e, f shows the MAE and RMSE values of response-time, while Fig. 2.6g, h shows the MAE and RMSE values of throughput. When the matrix density is 90%, we observe that our approach CloudPred achieves the best performance when the value of dimensionality is 30, while smaller values like 10 or larger values like 50 can potentially hurt the prediction accuracy. This observation indicates that when the user-component matrices are dense, 10 latent factors is not enough to characterize the features of user and component which are mined from the rich QoS information, while 50 latent factors is too many since it will cause overfitting problem. When the matrix density is 10%, we observed that the prediction accuracy of our approach CloudPred decreases (MAE and RMSE increase) when the value of dimensionality is increased from 10 to 50. This observation indicates that when the user-component matrices are sparse, 10 latent factors is already enough to characterize the features of user and component which are mined from the limited user-component QoS information, while other larger values of dimensionality will cause the overfitting problem.

2.4.7 Impact of λ

The parameter λ determines how much the final prediction results rely on user-based approach or component-based approach. A larger value of λ means user-based approach contributes more to the hybrid prediction. A smaller value of λ means component-based approach contributes more to the hybrid prediction. In Fig. 2.7, we study the impact of parameter λ by varying the values of λ from 0 to 1 with a step value of 0.1. Other parameter settings are dimensionality = 10 and Top-K = 10.

Figure 2.7a, b shows the MAE and RMSE results of response-time, respectively. The experimental results show that the value of λ impacts the recommendation results significantly, which demonstrates that combining the user-based approach and component-based approach improves the recommendation accuracy. The prediction accuracies increase when we increase the value of λ at first. But when λ surpasses a certain threshold, the prediction accuracy decreases with further increase of the value of λ. This phenomenon coincides with the intuition that purely using the user-based approach or purely using the component-based approach cannot generate better results than the hybrid approach. From Fig. 2.7, we observed that when $\lambda \in [0.4, 0.7]$, CloudPred achieves the best performance, while a smaller value or a larger value can potentially degrade the prediction performance. Moreover, the insensitivity of the optimal value of λ shows that the parameter of CloudPred is easy to train.

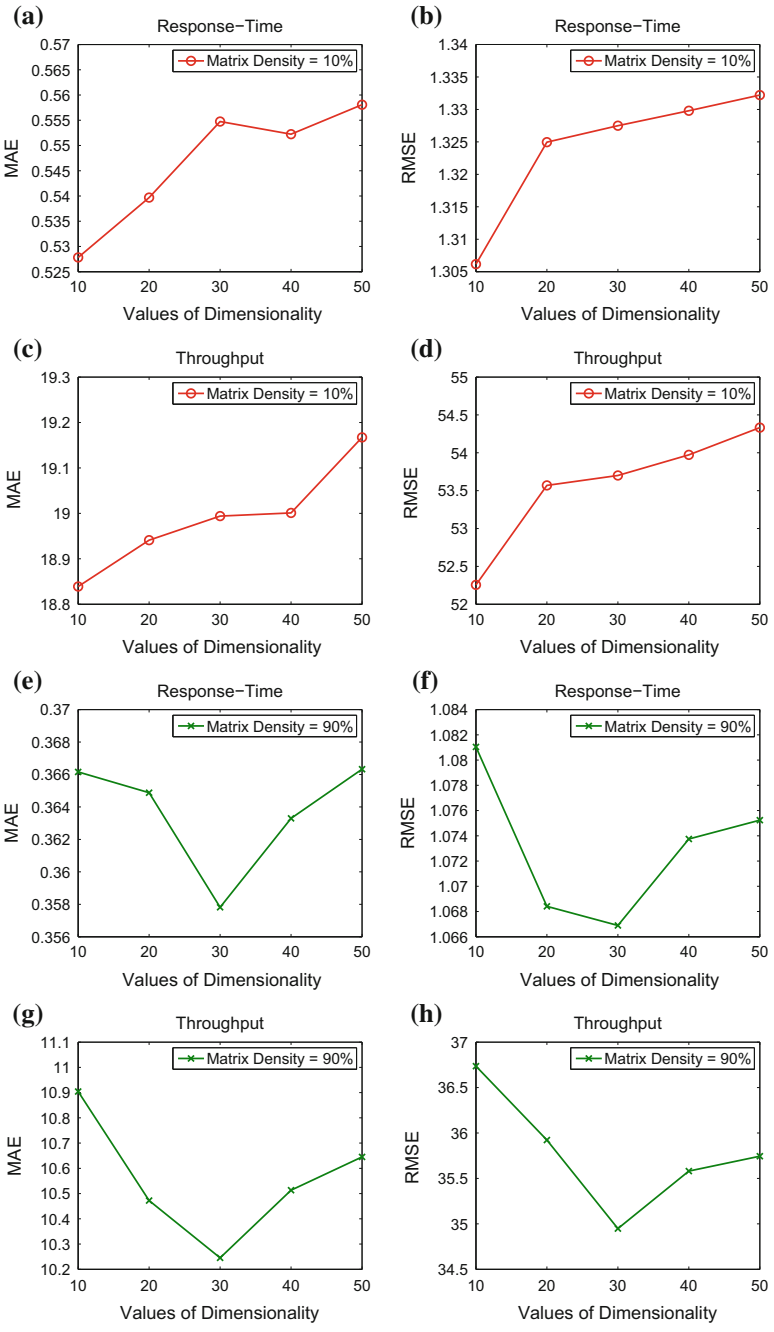


Fig. 2.6 Impact of dimensionality. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

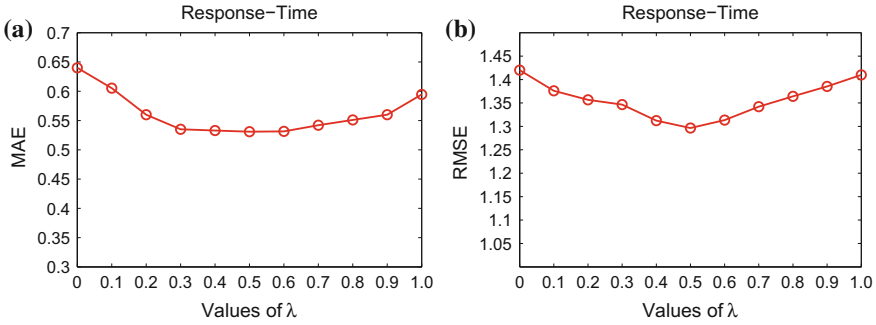


Fig. 2.7 Impact of λ. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

2.5 Summary

Based on the intuition that a user's cloud component QoS usage experiences can be predicted by exploring the past usage experience from both the user and its similar users, we propose a novel neighborhood-based approach, which is enhanced by feature modeling on both user and component, called CloudPred, for collaborative and personalized QoS value prediction on cloud components. Requiring no additional invocation of cloud components, CloudPred makes the QoS value prediction by taking advantage of both local usage information from similar users and similar components and global invocation information shared by all the users. The extensive experimental results show that our approach CloudPred achieves higher prediction accuracy than other competing methods.

Since the Internet environment is highly dynamic, the QoS performances of a cloud component may be variable against time (e.g., due to the network traffic, server workload). In our current approach, the QoS values are observed over a long period, which represent the average QoS performance of cloud components. Since the average QoS performance of cloud components is relatively stable, the predicted QoS values provide valuable information of unused cloud components for the users. In our future work, we will explore an online prediction algorithm to handle the dynamically changing QoS values by fusing with the time information.

Currently, we are collecting QoS information of Web service, which is a kind of cloud component. In the future, we will conduct more experiments to evaluate our approach in commercial clouds which contain multiple kinds of cloud components. For future work, we will investigate more techniques for improving the similarity computation (e.g., clustering models, latent factor models, data smoothing). We will also conduct more investigations on the correlations and combinations of different QoS properties.

References

1. J. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in *Proceedings of UAI'98* (1998), pp. 43–52
2. B. Hayes, Cloud computing. Commun. ACM **51**(7), 9–11 (2008)
3. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
4. H. Ma, I. King, M. Lyu, Effective missing data prediction for collaborative filtering, in *Proceeding of SIGIR'07* (2007), pp. 39–46
5. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in *Proceeding of CSCW'94* (1994), pp. 175–186
6. R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization. Adv. Neural Inf. Process. Syst. (NIPS) **20**, 1257–1264 (2008)
7. L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, H. Mei, Personalized QoS prediction for web services via collaborative filtering, in *Proceeding of ICWS'07* (2007), pp. 439–446
8. Y. Zhang, Z. Zheng, M.R. Lyu, Exploring latent features for memory-based QoS prediction in cloud computing, in *IEEE Symposium on Reliable Distributed Systems (SRDS)* (IEEE, 2011), pp. 1–10
9. Z. Zheng, Y. Zhang, M. Lyu, Distributed QoS evaluation for real-world web services, in *Proceeding of ICWS'10* (2010), pp. 83–90

Chapter 3
Time-Aware Model-Based QoS Prediction

Abstract The exponential growth of Web service makes building high-quality service-oriented applications an urgent and crucial research problem. User-side QoS evaluations of Web services are critical for selecting the optimal Web service from a set of functionally equivalent service candidates. Since QoS performance of Web services is highly related to the service status and network environments which are variable against time, service invocations are required at different instances during a long time interval for making accurate Web service QoS evaluation. However, invoking a huge number of Web services from user-side for quality evaluation purpose is time-consuming, resource-consuming, and sometimes even impractical (e.g., service invocations are charged by service providers). To address this critical challenge, this chapter proposes a Web service QoS prediction framework, called WSPred, to provide time-aware personalized QoS value prediction service for different service users. WSPred requires no additional invocation of Web services. Based on the past Web service usage experience from different service users, WSPred builds feature models and employs these models to make personalized QoS prediction for different users. The extensive experimental results show the effectiveness and efficiency of WSPred. Moreover, we publicly release our real-world time-aware Web service QoS dataset for future research, which makes our experiments verifiable and reproducible.

3.1 Overview

With the growing number of Web services over the Internet, designers of service-oriented applications can choose from a broad pool of functionally identical or similar Web services when creating applications. Web services are usually deployed in remote servers and accessed by users through Internet connections. The quality of a service-oriented application, therefore, is greatly influenced by the quality of the invoked Web services. To build high-quality service-oriented applications, non-functional Quality-of-Service (QoS) performance of Web services becomes a major concern for application designers when making service selections [4]. However, the QoS performance of Web services observed from the users' perspective is usually

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_3

quite different from that declared by the service providers in service-level agreement (SLA), due to:

- QoS performance of Web services is highly related to invocation time, since the service status (e.g., workload, number of clients) and the network environment (e.g., congestion) change over time.
- Service users are typically distributed in different geographic locations. The user-observed QoS performance of Web services is greatly influenced by the Internet connections between users and Web services. Different users may observe quite different QoS performance when invoking the same Web service.

Based on the above analysis, providing time-aware personalized QoS information of Web services is becoming more and more essential for service-oriented application designers to make service selection [4, 7], service composition [1, 2], and automatically late-binding at runtime [3].

In reality, a service user usually only invokes a limited number of Web services in the past and thus only observes QoS values of these invoked Web services. Without sufficient time-aware personalized QoS information, it is difficult for application designers to select optimal Web services at design time and replace low-quality Web services with better ones at runtime. In practice, invoking Web services from users' perspectives for evaluation purpose is quite difficult and includes the following critical drawbacks:

- Executing service invocations to obtain QoS information is too expensive for service users, since service providers may charge for invocations. At the same time, invocations for evaluation purpose consume resources of service users and service providers.
- With the growing number of Web services over the Internet, it is time-consuming to evaluate all the Web services. Moreover, some potentially appropriate Web services may not be discovered by the current user.
- To monitor the QoS performance of Web services continuously, service users need to conduct service invocations periodically, which will introduce a heavy workload to service users.
- Since service users are not experts in service evaluation, it will take a solid effort from service users to evaluate the Web services in-depth. The time-to-market constraints will also limit the amount of resources for service evaluation.

It becomes an urgent task to explore a time-aware personalized prediction approach for efficiently estimating missing QoS information of Web services for different service users. To address this critical challenge, we propose a model-based approach, called WSPred, for time-aware and personalized QoS prediction of Web services. WSPred collects time-aware QoS information from geographically distributed service users and combines the local information to get a global user-service-time tensor. By performing tensor factorization, user-specific, service-specific, and time-specific latent features are extracted from the past QoS experiences of different service users. The unknown QoS values are therefore estimated by analyzing how the user features are applied to the corresponding service features and time features.

We collect a large-scale real-world Web service QoS dataset and conduct extensive experiments to compare the QoS prediction accuracy with several other state-of-the-art approaches. The experimental results show the effectiveness and efficiency of our proposed approach WSPred.

In summary, this chapter makes the following contributions:

- We formally identify the critical problem of time-aware Web service QoS prediction and propose a novel collaborative framework to achieve QoS information sharing among service users. A user-side lightweight middleware is designed for automatically recording and sharing QoS experiences.
- We propose a novel time-aware personalized QoS prediction approach WSPred, which analyzes latent features of user, service, and time by performing tensor factorization. We consider WSPred as the first QoS prediction approach which addresses the difference over time in service computing literature.
- We conduct large-scale real-world experiments to study the prediction accuracy and efficiency of our WSPred compared with other state-of-the-art approaches. Moreover, we publicly release our large-scale Web service QoS dataset for future research. To the best of our knowledge, it is the first multi-user QoS dataset with time series information in the Web service literature.

The remainder of this chapter is organized as follows: Sect. 3.2 describes the collaborative framework for sharing QoS information between service users. Section 3.3 presents our WSPred approach in detail. Section 3.4 introduces the experimental results. Section 3.5 concludes the chapter.

3.2 Collaborative Framework for Web Services

In this section, we present the collaborative framework for QoS prediction of Web services. Figure 3.1 shows the system architecture. Within a service-oriented Web application, several Web services are employed to implement complicated functions.

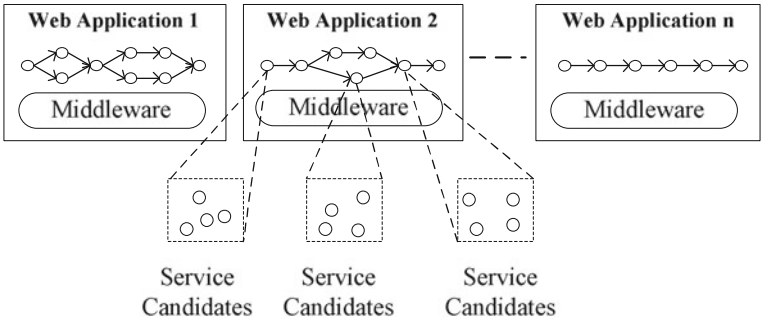


Fig. 3.1 System architecture. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

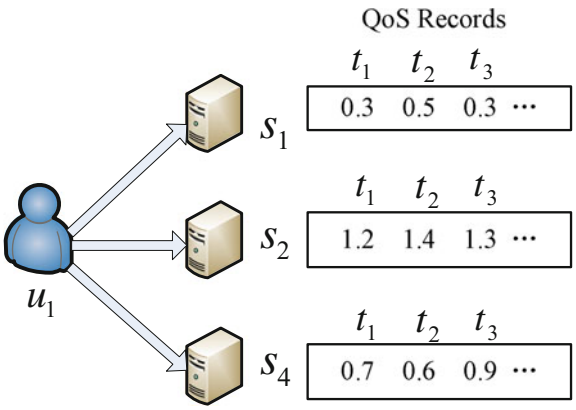


Fig. 3.2 A toy example. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

These Web services are connected with each other in multiple tiers. For each tier, an optimal Web service will be selected from a set of functional equivalent service candidates. Typically the Web service candidates are provided by different organizations and are distributed in different geographic locations and time zones. When invoked through communication links, the user-side usage experiences are influenced by the network environments and the server-side status at invocation time.

The mechanism proposed in this chapter is to (1) share local Web service usage experiences from different service users, (2) combine these pieces of local information together to get global QoS information for all service candidates, (3) extract time-specific user features and service features, and (4) make personalized time-aware QoS value prediction based on these features. As shown in Fig. 3.2, each service user keeps local records of QoS usage experience on Web services and is encouraged to contribute its local records to obtain records from other users. By contributing more individually observed Web service QoS information, a service user can obtain more global QoS information from other users, thus obtaining more accurate Web service QoS prediction values. Given accurate QoS prediction results, service users could select the potentially optimal services for composing service-oriented Web applications. The detailed collaborative prediction approach will be presented in Sect. 3.3.

Since most of the service users are not experts in service testing, to reduce the efforts of service users spent on testing the service QoS performance, we design a user-side lightweight middleware for service users to automatically record QoS values of invocations and to contribute the local records to the server for obtaining more invocation results from other service users. Within the middleware, there are three management components: *Monitor*, *Collector*, and *Predictor*. *Monitor* is responsible for monitoring the QoS performance of Web services when users send invocations. *Collector* is responsible for contributing local QoS information to other users and for collecting shared QoS information from other users. *Predictor* is responsible for providing time-aware personalized QoS value prediction based on local and other users' QoS information collected by *Collector*.

3.3 Time-Aware QoS Prediction

Previous Web service-related techniques such as selection, composition, and orchestration only employ average QoS performance of service candidates at design time. In recent Web service literatures, most of the state-of-the-art techniques can automatically update corresponding Web services with better ones at runtime, which requires time-specific QoS performance of Web services.

In this section, we first formally describe the QoS value prediction problem on Web services in Sect. 3.3.1. Then we propose a latent feature learning algorithm to learn the user-specific, service-specific, and time-specific features in Sect. 3.3.2. The missing QoS values are predicted by applying the proposed algorithm WSPred in Sect. 3.3.3. Finally, the complexity analysis is conducted in Sect. 3.3.4.

3.3.1 Problem Description

Figure 3.2 illustrates a toy example of the QoS prediction problem we study in this chapter. In this figure, user u_1 has used three Web services s_1, s_2, and s_4 in the past. u_1 recorded the observed QoS performance of Web services s_1, s_2, and s_4 with specific invocation time in local site. By integrating all the QoS information from other users, we form a three-dimensional user-service-time tensor as shown in Fig. 3.3. In this example, totally there are 5 users (from u_1 to u_5), 5 services (from s_1 to s_5), and 5 time intervals (from t_1 to t_5). The tensor is split into several slices with each one representing a time interval. Within a slice, each entry denotes an observed QoS value of a Web service from a user during the specific time interval. The problem we study in this chapter is how to efficiently and precisely predict the missing entries in the user-service-time tensor based on the existing entries.

Now we formally define the problem of QoS prediction for Web services as follows: Given a set of users and a set of Web services, based on the existing QoS values from different users, predict the missing QoS values of Web services when invoked by users at different time intervals. More precisely:

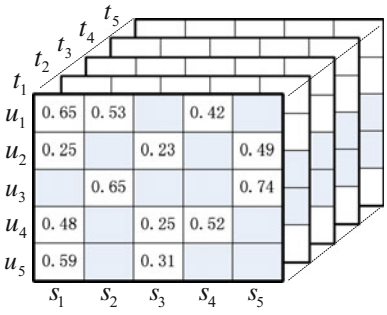


Fig. 3.3 User-service-time tensor. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Let U be the set of m users, S be the set of n Web services, and T be the set of c time intervals. A QoS element is a quartet (i, j, k, q_{ijk}) representing the observed quality of Web service s_j by user u_i at time interval t_k, where $i \in \{1, \cdots, m\}$, $j \in \{1, \cdots, n\}$, $k \in \{1, \cdots, c\}$, and $q_{ijk} \in \mathbb{R}^p$ is a p-dimensional vector representing the QoS values of p criteria. Let Ω be the set of all triads $\{i, j, k\}$ and Λ be the set of all known triads (i, j, k) in Ω. Consider a tensor $Y \in \mathbb{R}^{m \times n \times c}$ with each entry Y_{ijk} representing the observed p^{th} criterion value of service s_j by user u_i at time interval t_k. Then the missing entries $\{Y_{ijk} | (i, j, k) \in \Omega - \Lambda\}$ should be predicted based on the existing entries $\{Y_{ijk} | (i, j, k) \in \Lambda\}$.

Typically, the QoS values can be integers from a given range (e.g., {0, 1, 2, 3}) or real numbers. Without loss of generality, we can map the QoS values to the interval [0, 1] using the following function:

$$f(x) = \begin{cases} 0, \text{ if } x < Y_{min} \\ 1, \text{ if } x > Y_{max} \\ \frac{x - Y_{min}}{Y_{max} - Y_{min}}, \text{ otherwise} \end{cases}$$

where Y_{max} and Y_{min} are the specified upper bound and lower bound of QoS values, respectively.

3.3.2 Latent Features Learning

In order to learn the latent features of users, services, and time, we employ tensor factorization technique to fit a factor model to the user-service-time tensor. The factorized user-specific, service-specific, and time-specific matrices are utilized to make further missing entries prediction. The idea behind the factor model is to derive a high-quality low-dimensional feature representation of users, services, and time by analyzing the user-service-time tensor. The premise behind a low-dimensional factor model is that there is only a small number of factors influencing QoS usage experiences and that a user's QoS usage experience vector is determined by how each factor applies to that user, the corresponding service and the specific time interval. Examples of physical feature are network distance between the user and the server, the workload of the server, etc. Latent features are orthogonal representing the decomposed results of physical factors.

In the chapter, we consider an $m \times n \times c$ QoS tensor consisting of m users, n services, and c time intervals. A low-rank tensor factorization approach seeks to approximate the QoS tensor Y by a multiplication of l-rank factors [6],

$$Y \approx C \times_u U \times_s S \times_t T, \tag{3.1}$$

where $C \in \mathbb{R}^{l \times l \times l}$, $U \in \mathbb{R}^{m \times l}$, $S \in \mathbb{R}^{n \times l}$, and $T \in \mathbb{R}^{c \times l}$ are latent feature matrices. l is the number of latent features. Each column in U, S, and T representing a user, a Web service, and a time interval, respectively. $\times_u$, $\times_s$, and $\times_t$ are tensor-matrix

multiplication operators with the subscript showing in which direction on the tensor to multiply the matrix (i.e., $C \times_u U = \sum_{i=1}^{l} C_{ijk} U_{ij}$). C is set to the diagonal tensor:

$$C = \begin{cases} 1, & \text{if } i = j = k \\ 0, & \text{otherwise} \end{cases}$$

Typically, $l \ll mnc$ since in the real world, each user has invoked only a small portion of Web services, and the tensor Y is usually very sparse. From the above definition, we can see that the low-dimensional matrices U, S, and T are unknown and need to be estimated.

To estimate the quality of tensor approximation, we need to construct a loss function for evaluating the error between the estimated tensor and the original tensor. The distance between two tensors is usually employed to define the loss function:

$$\frac{1}{2}||Y - \hat{Y}||_F^2, \tag{3.2}$$

where $||\cdot||_F^2$ denotes the Frobenius norm. However, due to the reason that there are a large number of missing values, we only factorize the observed entries in tensor Y. Hence, we employ the following loss function instead:

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{c} I_{ijk} (Y_{ijk} - \hat{Y}_{ijk})^2, \tag{3.3}$$

where I_{ijk} is the indicator function that is equal to 1 if user u_i invoked service s_j during the time interval t_k and equal to 0 otherwise. To avoid the overfitting problem, we add three regularization terms to Eq. (3.3) to constrain the norms of U, S, and T. Hence we conduct the tensor factorization as to solve the following optimization problem:

$$\begin{aligned} \min_{U,S,T} \mathscr{L}(Y, U, S, T) = & \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{c} I_{ijk} (Y_{ijk} - \hat{Y}_{ijk})^2 \\ & + \frac{\lambda_1}{2}||U||_F^2 + \frac{\lambda_2}{2}||S||_F^2 + \frac{\lambda_3}{2}||T||_F^2, \end{aligned} \tag{3.4}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$. λ_1, λ_2, and λ_3 define the importance of regularization terms. In other words, the optimal solution is highly rely on the error we evaluated in the first term. λ_1, λ_2, and λ_3 define the degree of accuracy in the first term to avoid overfitting problem. The optimization problem in Eq. (3.4) minimizes the sum-of-squared-errors objective function with quadratic regularization terms.

Figure 3.4 gives a comprehensive illustration of the Web service response-time observed by different service users. We randomly select two service users (User 1

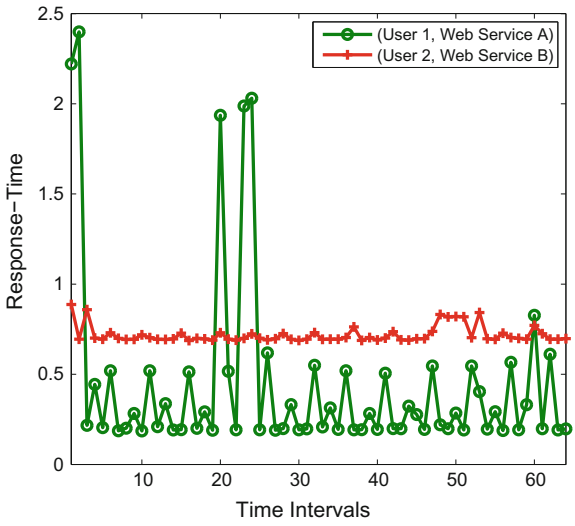


Fig. 3.4 Response-time of two pairs of user-service. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

and User 2) and two real-world Web services (Web Service A and Web Service B) from the experiment described in Sect. 3.4. As shown in Fig. 3.4, during different time intervals, a user has different QoS experiences on the same Web service. In general, the differences are limited within a range (e.g., most of the response-time values of (User 1, Web Service A) are within the range of 0.2–0.6 s and most of the response-time values of (User 2, Web Service B) are within the range of 0.7–0.9 s). This observation indicates that although the QoS values of a particular user-service are different during different time intervals, they fluctuate around the average QoS value of the user-service pair during all time intervals. Based on this observation, we further add a regularization term to Eq. (3.4) to prevent the predicted QoS values from varying a lot against the average QoS value. We define the prediction with average QoS value constraint as the following optimization problem:

$$\begin{aligned}\min_{U,S,T} \mathscr{L}_{\mathscr{A}}(Y,U,S,T) = &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} I_{ijk}(Y_{ijk}-\hat{Y}_{ijk})^2 \\ &+ \frac{\lambda_1}{2}||U||_F^2 + \frac{\lambda_2}{2}||S||_F^2 + \frac{\lambda_3}{2}||T||_F^2 \\ &+ \frac{\eta}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c} I_{ijk}(\hat{Y}_{ijk}-\bar{Y}_{ij})^2,\end{aligned} \tag{3.5}$$

where $\eta > 0$, and $\bar{Y}_{ij}$ denotes the average QoS value of Web service s_j observed by user u_i during all the time intervals. η controls how much the prediction method

should engage the information of average QoS performance. In the extreme case, if we use a very small value of η, we only perform tensor factorization without considering the global QoS information. On the other side, if we use a very large value of η, the average QoS performance will dominate the learning processes.

A local minimum of the objective function given by Eq. (3.5) can be found by performing incremental gradient descent in feature vectors U_i, S_j, and T_k:

$$\begin{aligned}
\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial U_{if}} &= \sum_{j=1}^{n}\sum_{k=1}^{c} I_{ijk}(\hat{Y}_{ijk} - Y_{ijk}) S_j^T T_k + \lambda_1 U_{if} \\
&\quad + \eta \sum_{j=1}^{n}\sum_{k=1}^{c} I_{ijk}(\hat{Y}_{ijk} - \bar{Y}_{ij}) S_j^T T_k, \\
\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial S_{jf}} &= \sum_{i=1}^{m}\sum_{k=1}^{c} I_{ijk}(\hat{Y}_{ijk} - Y_{ijk}) U_i^T T_k + \lambda_2 S_{jf} \\
&\quad + \eta \sum_{i=1}^{m}\sum_{k=1}^{c} I_{ijk}(\hat{Y}_{ijk} - \bar{Y}_{ij}) U_i^T T_k, \\
\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial T_{kf}} &= \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ijk}(\hat{Y}_{ijk} - Y_{ijk}) U_i^T S_j + \lambda_3 T_{kf} \\
&\quad + \eta \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ijk}(\hat{Y}_{ijk} - \bar{Y}_{ij}) U_i^T S_j.
\end{aligned} \tag{3.6}$$

Algorithm 3 shows the iterative process for latent feature learning. We first initialize matrices U, S, and T with small random nonnegative values. Iteration of the update rules derived from Eq. (3.6) converges to a local minimum of the objective function given in Eq. (3.5).

3.3.3 *Missing Value Prediction*

After the user-specific, service-specific, and time-specific latent feature spaces U, S, and T are learned, we can predict the QoS performance of a given service observed by a user during the specific time interval. For the missing entry Y_{ijk} in the QoS tensor, the value predicted by our method is defined as

$$\hat{Y}_{ijk} = I_{ijk} \sum_{f=1}^{l} U_{if} S_{jf} T_{kf}. \tag{3.7}$$

Algorithm 3: Latent Features Learning Algorithm

Input: Y, l, λ, η
Output: U, S, T

Initialize $U \in \mathbb{R}^{l \times m}$, $S \in \mathbb{R}^{l \times n}$, and $T \in \mathbb{R}^{l \times c}$ with small random numbers;
repeat
  for *all* $(i, j, k) \in \Lambda$ **do**
    $\hat{Y}_{ijk} = \sum_{f=1}^{l} U_{if} S_{jf} T_{kf}$;
  end
  for *all* (i, j) **do**
    $\bar{Y}_{ij} = \frac{\sum_{k=1}^{c} I_{ijk} Y_{ijk}}{\sum_{k=1}^{c} I_{ijk}}$;
  end
  for *all* $(i, j, k) \in \Lambda$ **do**
    for $(f = 1; f \leq l; f++)$ **do**
      $U_{if} \leftarrow U_{if} - [(\hat{Y}_{ijk} - Y_{ijk}) S_j^T T_k + \lambda U_{if} + \eta(\hat{Y}_{ijk} - \bar{Y}_{ij}) S_j^T T_k]$;
      $S_{jf} \leftarrow S_{jf} - [(\hat{Y}_{ijk} - Y_{ijk}) U_i^T T_k + \lambda S_{jf} + \eta(\hat{Y}_{ijk} - \bar{Y}_{ij}) U_i^T T_k]$;
      $T_{kf} \leftarrow T_{kf} - [(\hat{Y}_{ijk} - Y_{ijk}) U_i^T S_j + \lambda T_{kf} + \eta(\hat{Y}_{ijk} - \bar{Y}_{ij}) U_i^T S_j]$;
    end
  end
until *Converge*;

3.3.4 Complexity Analysis

The main computation of gradient methods is evaluating the objective function $\mathscr{L}_{\mathscr{A}}$ and their gradients against variables. The computational complexity of evaluating the objective function $\mathscr{L}_{\mathscr{A}}$ is $O(\rho_Y l + \rho_Y c)$, where ρ_Y is the number of nonzero entries in the tensor Y, l is the dimensions of the latent features, and c is the number of time intervals. The computational complexities for the gradients $\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial U}$, $\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial S}$, and $\frac{\partial \mathscr{L}_{\mathscr{A}}}{\partial T}$ in Eq. (3.6) are $O(\rho_Y l + \rho_Y c)$. Therefore, the total computational complexity in one iteration is $O(\rho_Y l + \rho_Y c)$, which indicates that theoretically, the computational time of our method is linear with respect to the number of observed QoS values in the user-service-time tensor Y. Note that because of the sparsity of Y, $\rho_Y << mnc$, which indicates that the computation time grows slowly with respect to the size of Tensor Y. This complexity analysis shows that our proposed approach is very efficient and can be applied to large-scale systems.

3.4 Experiments

In this section, we conduct several experiments to compare our approach with several state-of-the-art collaborative filtering prediction methods. In the following, Sect. 3.4.1 introduces the experimental setup and gives the description of our exper-

imental dataset, Sect. 3.4.2 defines the evaluation metrics, Sect. 3.4.3 compares the prediction quality of our approach with other competing methods, and Sects. 3.4.4 and 3.4.5 study the impact of tensor density and dimensionality, respectively.

3.4.1 Experimental Setup and Dataset Collection

To evaluate our proposed QoS prediction approach in the real world, we implement a tool WSMonitor for monitoring the QoS performance of Web service and collect a large-scale Web service QoS dataset for conducting various experiments.

WSMonitor is implemented and deployed with JDK 6.0, Eclipse 3.3, Axis 2, and Apache 2.2.17. WSMonitor first crawls a set of WSDL files from the Internet and generates a list of openly accessible Web services. For each Web service in the list, WSMonitor automatically generates a java class for service invocation by employing the WSDL2Java tool from the Axis package. Totally, 5871 classes are generated for 5871 Web services. By calling the functions within a class, null operation requests are sent to the corresponding Web service for capturing the QoS performance.

We deploy the WSMonitor on 142 distributed computers located in 22 countries from PlanetLab, which is a distributed test bed consisting of hundreds of computers all over the world. Totally, 4532 publicly available real-world Web services from 57 countries are monitored by each computer continuously. A total of 1339 of the initially selected Web services are excluded in this experiment due to: (1) authentication required and (2) permanent invocation failure (e.g., the Web service is shutdown). In our experiment, each of the 142 computers sends null operation requests to all the 4532 Web services during every time interval. The experiment lasts for 16 h with a time interval lasting for 15 min.

By collecting invocation records from all the computers, finally we include 30,287,611 QoS performance results into the Web service QoS dataset. Each invocation record is a k dimension vector representing the QoS values of k criteria. We then extract a set of $142 \times 4532 \times 64$ user-service-time tensors, each of which stands for a particular QoS property, from the QoS invocation records. For simplicity, we employ two tensors, which represent response-time and throughput QoS criteria respectively, for experimental evaluation in this chapter. Without loss of generality, our approach can be easily extended to include more QoS criteria.

The statistics of Web service QoS dataset are summarized in Table 3.1. Response-time and throughput are within the range of 0–20 s and 0–1000 kbps, respectively. The means of response-time and throughput are 3.165 s and 9.609 kbps, respectively. The distributions of the response-time and throughput values of the user-service-time tensors are shown in Fig. 3.5a, b respectively. Most of the response-time values are between 0.1 and 0.8 seconds, and most of the throughput values are between 0.8 and 3.2 kbps.

Table 3.1 Statistics of WS QoS dataset. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Statistics	Response-time	Throughput
Scale	0–20 s	0–1000 kbps
Mean	3.165 s	9.609 kbps
Num. of users	142	142
Num. of web services	4532	4532
Num. of time intervals	64	64
Num. of records	30,287,611	30,287,611

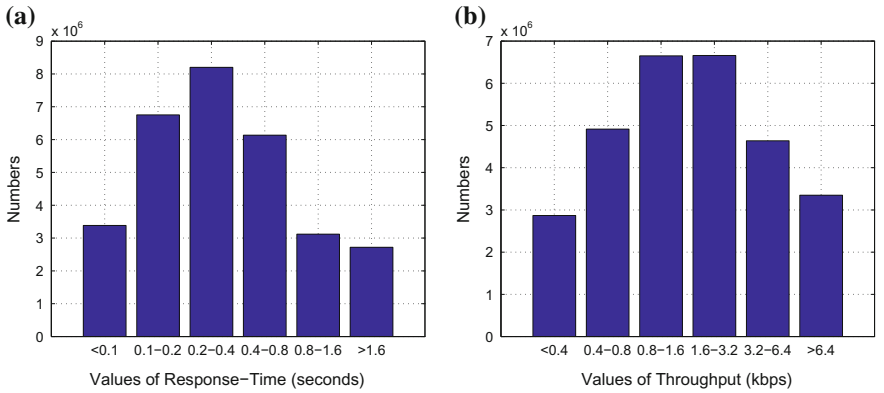


Fig. 3.5 QoS value distributions. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

3.4.2 Metrics

We assess the prediction quality of our proposed approach in comparison with other methods by computing mean absolute error (MAE) and root-mean-squared error (RMSE). The metric MAE is defined as:

$$MAE = \frac{\sum_{ijk} |\hat{Y}_{ijk} - Y_{ijk}|}{N}, \tag{3.8}$$

and RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{ijk} (\hat{Y}_{ijk} - Y_{ijk})^2}{N}}, \tag{3.9}$$

where Y_{ijk} is the QoS value of Web service s_j observed by user u_i at time interval t, $\hat{Y}_{ijk}$ denotes the QoS value of Web service s_j would be observed by user u_i at time interval t_k as predicted by a method, and N is the number of predicted QoS values.

3.4.3 Performance Comparisons

In this section, in order to show the effectiveness of our proposed Web service QoS prediction approach, we compare the prediction accuracy of the following methods:

1. **MF1**—This method considers the user-service-time tensor as a set of user-service matrix slices in terms of time. For each slice, the prediction method proposed by Lee and Seuing in [5] is employed. It applies nonnegative matrix factorization on user-item matrix for missing value prediction.
2. **MF2**—This method first compresses the user-service-time tensor into a user-service matrix. For each entry in the matrix, the value is the average of the specific user-service pair during all the time intervals. After obtaining the compressed user-service matrix, it applies the nonnegative matrix factorization technique proposed by Lee and Seuing [5] on user-item matrix for missing value prediction.
3. **TF**—This is a tensor factorization-based prediction method. It applies tensor factorization on user-service-time tensor to extract user-specific, service-specific, and time-specific characteristics. The missing value is then predicted based on how these characteristics apply to each other.
4. **WSPred**—This method is proposed in this chapter. It is a tensor factorization-based recommendation with average QoS value constraints.

Since memory-based approaches require much more computation time than model-based approaches, we only compare the above four model-based approaches. Since the matrix factorization technique cannot be directly applied to time-aware prediction problem, we extend the prediction approach [5] in two different ways, which derive method MF1 and MF2, respectively.

In order to evaluate the performance of different approaches in reality, we randomly remove some entries from the tensors and compare the values predicted by a method with the original ones. The tensors with missing values are in different densities. For example, 10% means that we randomly remove 90% entries from the original tensor and use the remaining 10% entries to predict the removed entries. The prediction accuracy is evaluated using Eqs. (3.8) and (3.9) by comparing the original value and the predicted value of each removed entry. The values of λ and η are tuned by performing cross-validation on the observed QoS data. Without lost of generality, the parameter settings of all the approaches are $l = 20$ and $\lambda 1 = \lambda 2 = \lambda 3 = \eta = 0.001$ in the experiments conducted in this chapter. Detailed impact of tensor density and dimensionality is studied in Sects. 3.4.4 and 3.4.5.

The QoS value prediction accuracies evaluated by MAE and RMSE are shown in Table 3.2. For each row in the table, we highlight the best performer among all methods. From Table 3.2, we can observe that the tensor factorization-based prediction methods (i.e., TF and WSPred) outperform the matrix factorization-based prediction methods (i.e., MF1 and MF2), since the tensor factorization-based methods use the time-specific features as additional information. We also observe that our approach WSPred constantly performs better (smaller MAE and RMSE values) than the other approaches, including TF, for both response-time and throughput under both dense

Table 3.2 Performance comparisons (A smaller MAE or RMSE value means a better performance). ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Tensor density (%)	Metrics	Response-time (seconds)			
		MF1	MF2	TF	WSPred
5	MAE	3.4137	2.9187	2.9184	**2.5580**
	RMSE	5.3423	5.1024	4.7508	**4.3626**
10	MAE	2.8518	2.8421	2.7888	**2.4990**
	RMSE	5.0667	4.5563	4.5696	**4.2892**
45	MAE	2.4241	2.2679	2.2511	**2.1462**
	RMSE	4.3240	4.2541	4.2071	**3.9200**
50	MAE	2.3959	2.2596	2.2127	**2.1266**
	RMSE	4.2996	4.1490	4.0169	**3.8943**
Tensor density (%)	Metrics	Throughput (kbps)			
		MF1	MF2	TF	WSPred
5	MAE	10.5460	8.8317	8.7997	**8.2761**
	RMSE	46.6735	43.4769	39.5133	**39.0962**
10	MAE	9.9839	8.7522	8.5080	**8.0131**
	RMSE	46.6656	39.7740	39.2792	**38.6251**
45	MAE	8.6773	7.9590	7.9471	**6.9398**
	RMSE	45.0077	39.9388	38.6964	**36.5724**
50	MAE	8.6224	7.8306	7.8045	**6.8558**
	RMSE	44.9407	38.9388	38.6964	**36.5724**

tensors and sparse tensors. This demonstrates the advantage of time-aware prediction algorithm with the constraints of average QoS performance. In Table 3.2, the MAE and RMSE values of dense tensors (e.g., tensor density is 45 or 50%) are smaller than those of sparse tensors (e.g., tensor density is 5 or 10%), since a denser tensor provides more information for predicting the missing values. In general, the MAE and RMSE values of throughput are larger than those of response-time because the scale of throughput is 0–1000 kbps, while the scale of response-time is 0–20 s. Compared with other methods, the improvements of our approach WSPred are significant, which demonstrates that the idea of considering time information for QoS prediction is realistic and reasonable.

3.4.4 Impact of Tensor Density

In Fig. 3.6, we compare the prediction accuracy of all the methods under different tensor densities. We change the tensor density from 5 to 50% with a step value of 5%. The parameter settings in this experiment are $l = 20$ and $\lambda 1 = \lambda 2 = \lambda 3 = \eta = 0.001$.

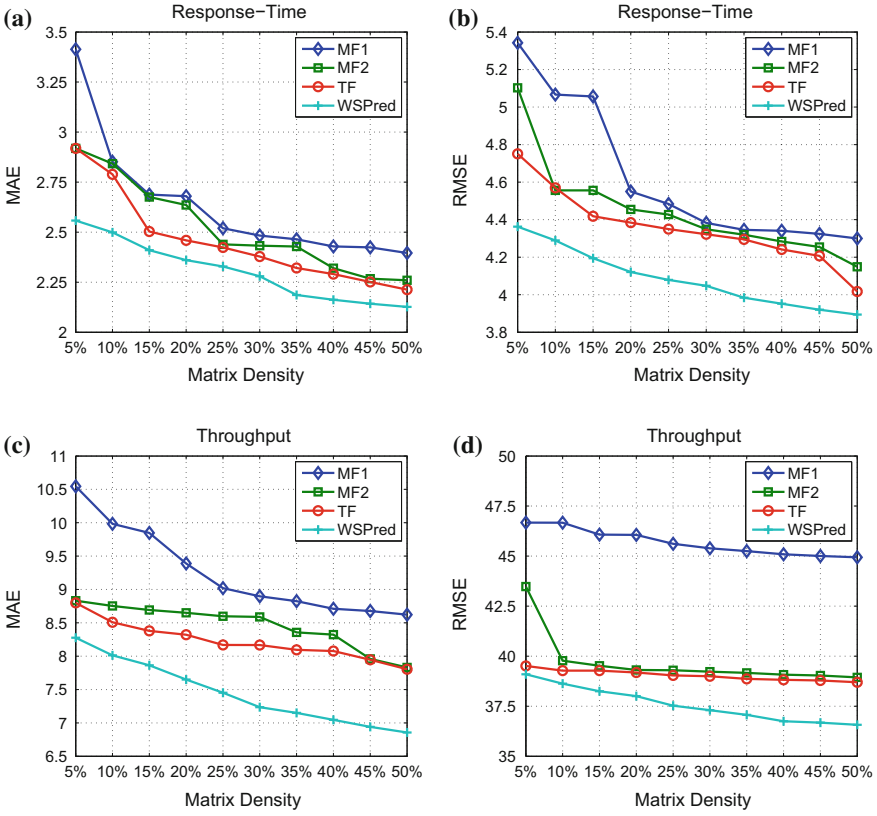


Fig. 3.6 Impact of tensor density. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

Figure 3.6a, b show the experimental results of response-time, while Fig. 3.6c, d show the experimental results of throughput. The experimental results show that our approach WSPred achieves higher prediction accuracy (lower MAE and RMSE values) than other competing methods under different tensor density. In general, when the tensor density is increased from 5 to 20%, the prediction accuracy of our approach WSPred is significantly enhanced. When the tensor density is further increased from 20 to 50%, the enhancement of prediction accuracy is quite limited. This observation indicates that when the tensor is very sparse, collecting more QoS information will greatly enhance the prediction accuracy, which further demonstrates that considering both the difference between time intervals and the average QoS performance could effectively provide personalized QoS estimation.

In the experimental results, we observe that the performance of MF1 is worse than that of other methods. The reason is that MF1 only extracts the user-specific and service-specific features without considering the relationship between QoS performance in time intervals. In general, MF2 performs better than MF1, since MF2 com-

putes the average QoS performance before performing matrix factorization. Applying the features extracted from the original tensor, MF2 predicts the average QoS performance for a particular user-service pair. This observation further demonstrates that the average QoS performance of a particular user-service pair can provide valuable information when predicting the missing QoS value of the user-service pair in a particular time interval.

3.4.5 *Impact of Dimensionality*

The parameter dimensionality l determines the number of latent features applied to characterize user, service, and time. In Figs. 3.7 and 3.8, we study the impact of

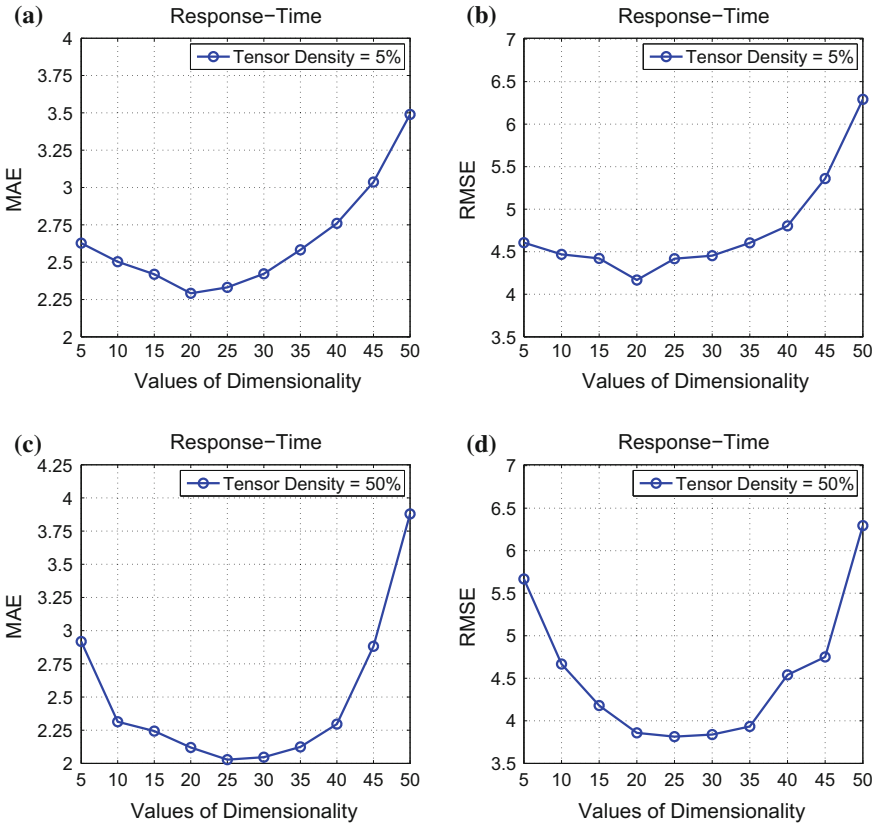


Fig. 3.7 Impact of dimensionality in response-time dataset. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

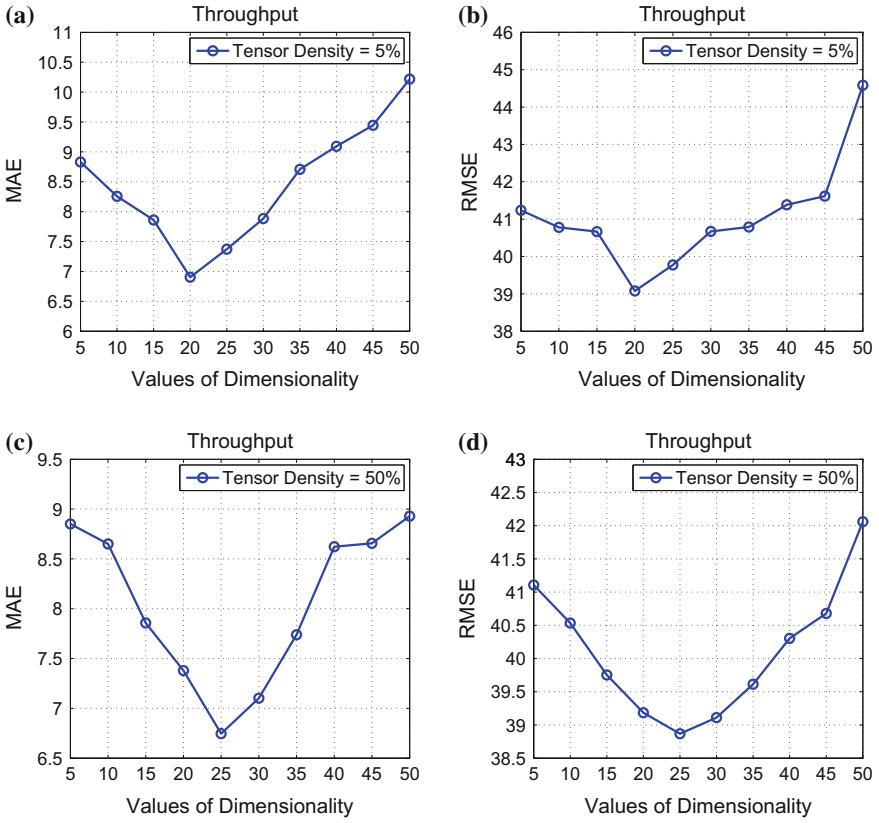


Fig. 3.8 Impact of dimensionality in throughput dataset. ©[2011] IEEE. Reprinted, with permission, from Ref. [8]

parameter dimensionality by varying the values of l from 5 to 50 with a step value of 5. Other parameter settings are $\lambda 1 = \lambda 2 = \lambda 3 = \eta = 0.001$.

Figures 3.7 and 3.8 show the MAE and RMSE values of response-time and throughput, respectively. We observe that in both figures, as l increases, the MAE and RMSE decrease (prediction accuracy increases), but when l surpasses a certain threshold like 20, the MAE and RMSE increase (prediction accuracy decreases) with further increase of the value of l. This observation indicates that too few latent factors are not enough to characterize the features of user, service, and time, while too many latent factors will cause an overfitting problem. There exists an optimal value of l for characterizing the latent features. In both Figs. 3.7 and 3.8, when the tensor density is 50%, we observe that our approach WSPred achieves the best performance when the value of dimensionality is 25, while smaller values like 5 or larger values like 50 can potentially reduce the prediction accuracy. When the tensor density is 5%, we observe that the prediction accuracy of our approach WSPred achieves the best

performance when the value of dimensionality is 20, while smaller values like 5 or larger values like 50 can potentially reduce the prediction accuracy. This observation indicates that when the user-service-time tensor is sparse, 20 latent factors are already enough to characterize the features of user, service, and time which are mined from the limited user-service-time QoS information. On the other hand, when the tensor is dense, more latent factors, like 25, are needed to characterize the latent features since more QoS information can be obtained from the original tensor.

3.5 Summary

Based on the intuition that a user's Web service QoS usage experience can be predicted by using the past usage experience from different users, we propose a novel model-based approach, called WSPred, for time-aware personalized QoS value prediction for Web services. By employing a collaborative framework, WSPred performs feature modeling on user, Web service, and time based on the QoS usage experience collected from both local and global users. Requiring no additional invocation of Web services, WSPred makes the QoS prediction by evaluating how the user-specific, service-specific, and time-specific latent features apply to each other. The extensive experimental results show that our proposed WSPred outperforms the state-of-the-art QoS prediction approaches for Web services.

For future work, we will investigate more techniques for improving the prediction accuracy (e.g., data smoothing, utilizing content information). Currently, we predict the values of different QoS properties independently. In the future, we will also conduct more investigations on the correlations and combinations on the different QoS properties. WSPred predicts missing QoS values based on the past QoS experience and the available QoS information in the current time interval. If no QoS information is available in the current time interval, WSPred purely depends on the past experience. In the future, we will explore an online prediction algorithm to perform time series analysis for prediction and extend WSPred to handle updated QoS information at runtime.

References

1. M. Alrifai, T. Risse, Combining global optimization with local selection for efficient QoS-Aware service composition, in *Proceeding of International Conference on World Wide Web (WWW'09)* (2009), pp. 881–890
2. M. Alrifai, D. Skoutas, T. Risse, Selecting skyline services for QoS-based web service composition, in *Proceeding of WWW'10* (2010), pp. 11–20
3. G. Canfora, M. Di Penta, R. Esposito, M. Villani, QoS-aware replanning of composite Web services, in *Proceeding of ICWS'05* (2005), pp. 121–129
4. J. El Haddad, M. Manouvrier, M. Rukoz, Tqos: transactional and QoS-aware selection algorithm for automatic web service composition. IEEE Trans. Serv. Comput., 73–85 (2010)

5. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
6. S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in *Proceeding of WSDM'10* (2010), pp. 81–90
7. T. Yu, Y. Zhang, K. Lin, Efficient algorithms for Web services selection with end-to-end QoS constraints. ACM Trans. Web (TWEB) **1**(1), 6 (2007)
8. Y. Zhang, Z. Zheng, M. Lyu, Wspred: a time-aware personalized QoS prediction framework for web services, in *Proceeding of IEEE Symposium on Software Reliability Engineering (ISSRE'11)* (2011), pp. 210–219

Chapter 4
Online QoS Prediction

Abstract The exponential growth of Web service makes building high-quality service-oriented systems an urgent and crucial research problem. Performance of the service-oriented systems highly depends on the remote Web services as well as the unpredictability of the Internet. Performance prediction of service-oriented systems is critical for automatically selecting the optimal Web service composition. Since the performance of Web services is highly related to the service status and network environments which are variable over time, it is an important task to predict the performance of service-oriented systems at runtime. To address this critical challenge, this chapter proposes an online performance prediction framework, called OPred, to provide personalized service-oriented system performance prediction efficiently. Based on the past usage experience from different users, OPred builds feature models and employs time series analysis techniques on feature trends to make performance prediction. The results of large-scale real-world experiments show the effectiveness and efficiency of OPred.

4.1 Overview

Low response time is one of the most important requirements of the service-oriented systems, which are widely employed in e-business and e-government. Typically, the response time performance of service-oriented systems involves two parts: local execution time at the system side and the response time of invoking remote Web services. While the local execution time is relatively short, the response time of invoking Web services is usually much longer, which greatly influences the system performance. The reason is that Web services are usually deployed in different geographical locations and invoked via Internet connections. Moreover, the remote Web services may be running on cheap and poor performing servers, leading to a decrease of service performance. In order to build service-oriented systems with good performance, it is important to identify Web services with low response time for composition. Moreover, by identifying the Web services with long response time at runtime, system designers can replace them with better ones to enhance the overall system performance.

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_4

Typically, Web services are considered as black boxes to service users. The user-side observed performance is employed to evaluate the qualities of Web services. Since the service status (e.g., workload, CPU allocations) and the network environment (e.g., congestions, bandwidth) may change over time, response time of Web services varies a lot during different time intervals. In order to identify low response time Web services timely, real-time performance of Web services needs to be continuously monitored.

Based on the above analysis, providing real-time performance information of Web services is becoming more and more essential for service-oriented system designers to build high-quality systems and to maintain the performance of the systems at runtime. However, evaluating the performance of service-orientated systems at runtime is not an easy task, due to the following reasons:

- Since users (SOA systems) and services are typically distributed in different geographical locations, the user-observed performance of Web services is greatly influenced by the Internet connections between users and Web services. Different users may observe quite different performance when invoking the same Web service.
- Real-time performance evaluation may introduce extra transaction workload, which may impact the user experience of using the systems.
- The purpose of performance evaluation is to monitor the current system performance status and allow designers to make adjustments in order to guarantee the performance in the future. This requires frequent performance evaluation, since infrequent evaluation cannot provide useful information to designers for choosing appropriate services in the following time.

It becomes an urgent task to explore an online personalized prediction approach for efficiently estimating the performance of Web services for different service users. Based on the performance information of Web services, the overall performance of a service-oriented system can be estimated by aggregating the performance of services invoked by the system. In this chapter, we propose a service performance estimation framework for providing personalized performance information to the users. The performance of services is predicted by collaborative work of users. We collect time-aware performance information from geographically distributed service users. Due to the fact that a service user usually only invokes a small number of Web services in the past and thus only observes performance of these invoked Web services, the collected performance information is usually sparse. In order to precisely predict the performance of Web service when invoked by users, we employ a set of latent features to characterize the status of Web services and users. Examples of physical feature are network distance between the user and the service server, the workload of the server. Latent features are orthogonal representation of the decomposed results of physical factors. We extract the latent features of users and services in the past time slice from the collected service performance information. By analyzing the trend of the feature changes, we estimate the features of users and services in the current time. Then, the personalized performance of Web service is predicted by evaluating how the features of users apply to features of services.

In summary, this chapter makes the following contributions:

- We propose an online performance prediction framework for estimating the user-observed performance of service-oriented systems. Our approach employs the past usage experiences of different users to efficiently predict the performance of service-oriented systems online.
- We collect a large-scale real-world Web service performance dataset and conduct extensive experiments for evaluating the performance of our proposed approach OPred. Totally, 4532 Web services are monitored by 142 service users and 30,287,611 invocation results are collected. Moreover, we publicly release our large-scale real-world Web service performance dataset for future research.

The rest of this chapter is organized as follows: Sect. 4.2 describes the service-oriented system architecture and introduces the online performance prediction procedures. Sections 4.3 and 4.4 present our online service performance prediction approach OPred in detail. Section 4.5 presents the experimental results. Section 4.6 concludes the chapter.

4.2 Preliminaries

Figure 4.1 shows the architecture of a typical service-oriented system. Within a service-oriented system, several abstract tasks are combined to implement complicated functions. For each abstract task, an optimal Web service is selected from a set of functionally equivalent service candidates. By composing the selected services, a service-oriented system instance is implemented for task execution. The problem of finding functionally equivalent Web service candidates has been discussed by a lot of

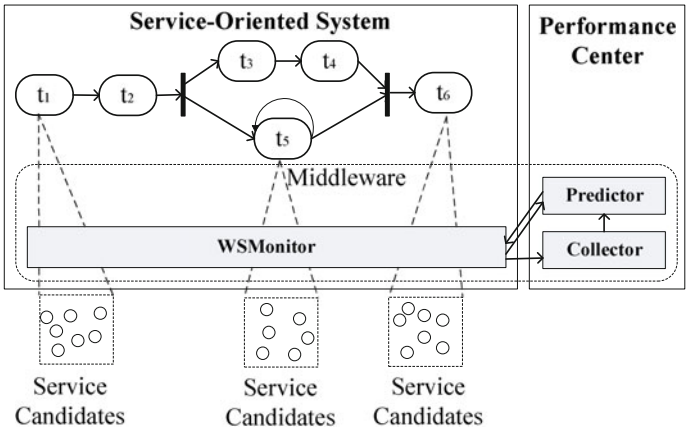


Fig. 4.1 Service-oriented system architecture. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

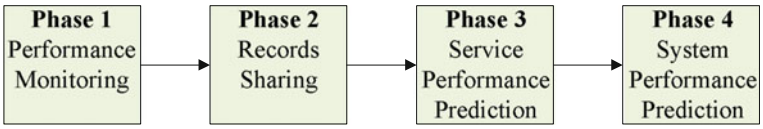


Fig. 4.2 Online performance prediction procedures. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

previous work [10, 15], which is outside the scope of this work. Typically, the Web service candidates are provided by different organizations and distributed in different geographical locations and time zones. When invoked through communication links, the user-side usage experiences are influenced by the network environments and the server-side status at invocation time. Since service-oriented systems are increasingly running on large numbers of dynamic services, users often encounter highly dynamic and uncertain performance of service-oriented systems.

As shown in Fig. 4.2, the online performance prediction mechanism proposed in this chapter contains four phases. In phase 1, each service user keeps local performance records of the Web services. In phase 2, local Web service usage experiences are uploaded to the performance center. Each user is encouraged to contribute its local records to obtain performance prediction service from the performance center. By contributing more individually observed Web service performance records, a service user can obtain more accurate performance prediction results from the performance center. By combining performance records of several users, the performance center can obtain global performance information for all services. In phase 3, by performing time series analysis on the extracted time-specific user features and service features, a performance model is built in the performance center for personalized service performance prediction. The premise behind the performance model is that there is a small number of latent factors influencing the user-observed service performance, and that a user's observed service performance is determined by how each factor applies to that user and the corresponding service at the current time slice. In phase 4, given the service-level performance information, the overall performance of a service-oriented system is predicted based on the analysis of service compositional structures. When the most recent service performance information is available, an online prediction algorithm is applied for quickly updating the performance model, which requires no effort of recalculation for catching the performance trend. The detailed online service performance prediction approach is presented in Sect. 4.3.

In Fig. 4.1, we can observe that the overall execution time of a service-oriented system mainly contains two parts: local computation time at the system side and response time of invoking remote services. The highly dynamic performance of service-oriented systems is mainly due to the highly dynamic response time of the composed services, while the local execution time is relatively stable. To improve the performance of systems at runtime, optimal Web service of each abstract task should be identified timely to replace the bad ones for composition. The overall performance of systems with different compositional options can be compared by estimating the

total response time required for invoking all the composed services. The detailed system-level performance prediction approach will be presented in Sect. 4.4.

Since most of the service users are not experts in service testing, to reduce the efforts of service users spent on testing the service performance, we design a light-weight middleware for service users to automatically record invocation results, contribute the local records to the performance center, and receive performance prediction results from the performance center. Within the middleware, there are three management components: *WSMonitor*, *Collector*, and *Predictor*. *WSMonitor* is deployed on the user-side. *Collector* and *Predictor* are deployed on the performance center. *WSMonitor* is responsible for monitoring the performance of Web services and sending local records to the performance center. *Collector* is responsible for collecting shared performance records from users. *Predictor* is responsible for providing time-aware personalized performance prediction based on users' performance information collected by *Collector*.

4.3 Online Service-Level Performance Prediction

In this section, we propose a collaborative method to predict the performance of services. Previous Web service related techniques such as selection [4, 12, 14, 16], composition [1, 2, 13], and orchestration [5] typically only employ average performance of service candidates at design time. In the recent Web service literature, most of the state-of-the-art techniques can automatically update corresponding Web services with better ones at runtime. Therefore, making personalized time-specific performance prediction of Web services for different users becomes a critical task.

In this section, we first formally describe the online performance prediction problem of Web services in Sect. 4.3.1. Then, we propose a latent feature learning algorithm to learn the time-aware user-specific and service-specific features in Sect. 4.3.2. The performance of services is predicted by applying the proposed online algorithm in Sect. 4.3.3. Finally, the complexity analysis is conducted in Sect. 4.3.4.

4.3.1 Problem Description

Figure 4.3a illustrates a toy example of the performance prediction problem we study in this chapter. In this figure, service user u_1 has used three Web services s_1, s_2, and s_4 in the past. u_1 recorded the observed performance of Web services s_1, s_2, and s_4 with time stamp in the local site. By integrating all the performance information from different users, we can form a set of matrices as shown in Fig. 4.3b with each matrix representing a time slice. In this example, there are totally 5 users (from u_1 to u_5) and 5 services (from s_1 to s_5). Within a matrix, each entry denotes the observed performance (e.g., response time) of a Web service by a user during a specific time slice. A missing entry denotes that the corresponding user did not invoke the service

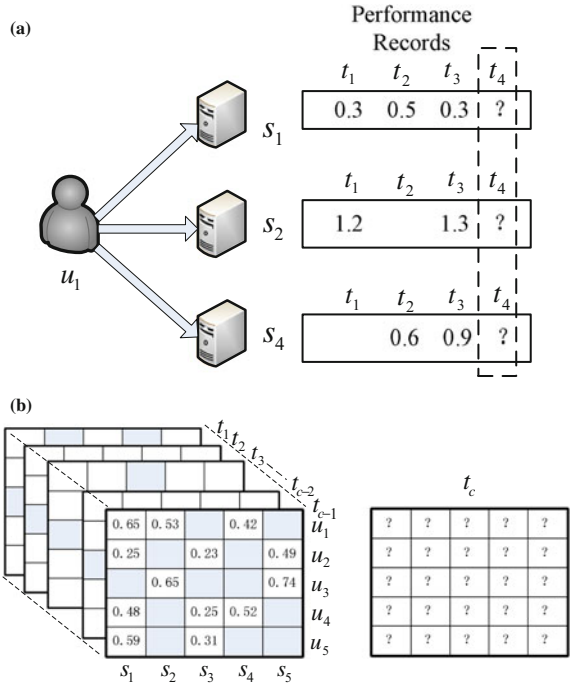


Fig. 4.3 Toy example of performance prediction. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

in the time slice. The problem we study in this chapter is how to efficiently and precisely predict performance of services observed by a user in the next time slice based on the previously collected performance information.

Let U be the set of m users and S be the set of n Web services. In each time slice t, the observed response time from all users is represented as a matrix $R(t) \in \mathbb{R}^{m \times n}$ with each existing entry $r_{ui}(t)$ representing the response time of service i observed by user u in time slice t. Given the set of matrices $\Psi = \{R(k)|k < t_c\}$, matrix $R(t_c)$ should be predicted representing the expected response time of services in time slice t_c.

Without loss of generality, we can map the response time values to the interval [0, 1] using the following function:

$$f(x) = \begin{cases} 0, \text{ if} x < r_{min} \\ 1, \text{ if} x > r_{max} \\ \frac{x - r_{min}}{r_{max} - r_{min}}, \text{ otherwise} \end{cases}$$

where r_{max} and r_{min} are the upper bound and lower bound of the response time values, respectively, which can be defined by users.

4.3.2 Time-Aware Latent Feature Model

In order to learn the latent features of users and services, we employ a matrix factorization technique to fit a feature model to user-service matrix in each time slice. The factorized user-specific and service-specific features are utilized to make further performance prediction. The idea behind the feature model is to derive a high-quality low-dimensional feature representation of users and services by analyzing the user-service matrices. It is noted that there is only a small number of features influencing performance experiences, and that a user's performance experience vector is determined by how each feature is applied to that user and the corresponding service. Examples of physical features are network distance between the user and the server, the workload of the server. Latent features are orthogonal representation of the decomposed results of physical features. Consider the matrix $R(t) \in \mathbb{R}^{m \times n}$ consisting of m users and n services. Let $p(t) \in \mathbb{R}^{l \times m}$ and $q(t) \in \mathbb{R}^{l \times n}$ be the latent user and service feature matrices in time slice t. Each column in $p(t)$ represents the l-dimensional user-specific latent feature vector of a user, and each column in $q(t)$ represents the l-dimensional service-specific latent feature vector of a service. We employ an approximating matrix to fit the user-service matrix $R(t)$, in which each entry is approximated as:

$$\hat{r}_{ui}(t) = p_u^T(t) q_i(t) \tag{4.1}$$

where l is the rank of the factorization which is generally chosen so that $(m+n)l < mn$, since $p(t)$ and $q(t)$ are low-rank feature representations [7]. This matrx factorization procedure (i.e., decompose the user-service matrix $R(t)$ into two matrices $p(t)$ and $q(t)$) has clear physical meanings: Each column of $q(t)$ is a factor vector including the values of the l factors for a Web service, while each column of $p(t)$ is the user-specific coefficients for a user. In Eq. (4.1), the user-observed performance on service i at time t (i.e., $\hat{r}_{ui}(t)$) corresponds to the linear combination of the user-specific coefficients and the service factor vector.

In order to optimize the matrix factorization in each time slice, we first construct a cost function to evaluate the quality of approximation. The distance between two nonnegative matrices is usually employed to define the cost function. In this chapter, due to the reason that there are a large number of missing values in practice, we only factorize the observed entries in matrix $R(t)$. Hence, we conduct the matrix factorization as to solve the following optimization problem:

$$\begin{aligned} &\min \mathscr{L}(p_u(t), q_i(t)) \\ &= \frac{1}{2} \sum_{u=1}^{m} \sum_{i=1}^{n} I_{ui} (r_{ui}(t) - g(\hat{r}_{ui}(t)))^2 \\ &+ \frac{\lambda_1}{2} ||p(t)||^2 + \frac{\lambda_2}{2} ||q(t)||^2, \end{aligned} \tag{4.2}$$

where $\lambda_1, \lambda_2 > 0$, I_{ui} is the indicator function that is equal to 1 if user u invoked service i during the time slice t and equal to 0 otherwise. $(r_{ui}(t) - g(\hat{r}_{ui}(t)))^2$ evaluates

the prediction error. To avoid the overfitting problem, we add two regularization terms to Eq. (4.2) to constrain the norms of $p(t)$ and $q(t)$ where $||\cdot||^2$ denotes the Frobenius norm. λ_1 and λ_2 defines the importance of regularization terms. In other words, the optimal solution is highly relying on the error we evaluated in the first term. λ_1 and λ_2 defines the degree of accuracy in the first term to avoid overfitting problem. The optimization problem in Eq. (4.2) minimizes the sum-of-squared-errors objective function with quadratic regularization terms. $g(x)$ is the logistic function $g(x) = 1/(1 + exp(-x))$, which maps $\hat{r}_{ui}(t)$ to the interval [0, 1]. By solving the optimization problem, we can find the most appropriate latent feature matrices $p(t)$ and $q(t)$ to characterize the users and services, respectively.

A local minimum of the objective function given by Eq. (4.2) can be found by performing incremental gradient descent in feature vectors $p(t)$ and $q(t)$:

$$\begin{aligned}\frac{\partial L}{p_u(t)} &= I_{ui}(g(\hat{r}_{ui}(t)) - r_{ui}(t))g'(\hat{r}_{ui}(t))q_i(t) \\ &\quad +\lambda_1 p_u(t), \qquad (4.3)\\ \frac{\partial L}{q_i(t)} &= I_{ui}(g(\hat{r}_{ui}(t)) - r_{ui}(t))g'(\hat{r}_{ui}(t))p_u(t) \\ &\quad +\lambda_2 q_i(t). \qquad (4.4)\end{aligned}$$

Algorithm 4 shows the iterative process for time-aware latent feature learning. We first initialize matrices $p(t)$ and $q(t)$ with small random nonnegative values. Update iterations derived from Eqs. (4.3) and (4.4) allow the objective function given in Eq. (4.2) converge to a local minimum.

Algorithm 4: Time-Aware Latent Features Learning.

Input: $R(t), l, \lambda_1, \lambda_2$
Output: $p(t), q(t)$

```
Initialize p(t) ∈ R^{l×m} and q(t) ∈ R^{l×n} with small random numbers;
Load the performance records from matrix R(t);
Calculate the objective function value ℒ(p_u(t), q_i(t)) by Eq. (4.1) and
  Eq. (4.2);
repeat
    Calculate the gradient of feature vectors ∂L/p_u(t) and ∂L/q_i(t) according Eq. (4.3)
    and Eq. (4.4), respectively;
    Update the latent user and service feature matrices p(t) and q(t);
    p_u(t) ← p_u(t) − ∂L/p_u(t);
    q_i(t) ← q_i(t) − ∂L/q_i(t);
    Update the objective function value ℒ(p_u(t), p_i(t)) by Eq. (4.1) and
    Eq. (4.2);
until Converge;
```

4.3.3 Service Performance Prediction

After the user-specific and service-specific latent feature spaces $p(t)$ and $q(t)$ are learned in each time slice t, we can predict the performance of a given service observed by a user during the next time slice. The service performance prediction is conducted in two phases: offline phase and online phase. In the offline phase, the performance information collected from all the service users is used for statically modeling the trends of user features and service features. By employing a time series analysis, the features of users and services in the next time slice are calculated based on the evolutionary algorithm. The predicted features are further applied for calculating the predicted performance of services in the next time slice. In the online phase, the newly observed service performance information by users at runtime is integrated into the feature model built in the offline phase. By employing the incremental calculation algorithm, the feature model is updated efficiently to catch the latest trend for ensuring the prediction accuracy.

4.3.3.1 Phase 1: Offline Evolutionary Algorithm

Given the latent feature vectors of users and services in time slices before t_c, the latent feature vectors in time slice t_c can be predicted by precisely modeling the trends of features. Intuitively, older features are less correlated with a service's current status or a user's current characteristics. To characterize the latent features at time slice t_c, the prediction calculation should rely more on the information collected in the latest time slices than that collected in older time slices. In order to integrate the information from different time slices, we therefore employ the following temporal relevance function [8]:

$$f(k) = e^{-\alpha k}, \tag{4.5}$$

where k is the amount of time that has passed since the corresponding information was collected. $f(k)$ measures the relevance of information collect from different time slices for making prediction on latent features at time t_c. Note that $f(k)$ decreases with k. By employing the temporal relevance function $f(k)$, we can assign a weight for each latent feature vector depending on the collecting time when making prediction. In the temporal relevance function, α controls the decaying rate. By setting α to 0, the evolutionary nature of the information is ignored. A constant temporal relevance value of 1 is assigned to latent feature vectors in all the time slices, which means latent feature vectors in time slice t_c are predicted simply by averaging the vectors before time slice t_c. Since $e^{-\alpha}$ is a constant value, the value of temporal relevance function can be recursively computed: $f(k+1) = e^{-\alpha} f(k)$, in which $e^{-\alpha}$ denotes the constant decay rate.

By analyzing the collected performance data, we obtain two important observations: (1) Within a relatively longtime period such as one day or one week, the service

performance observed by a user may vary significantly due to the highly dynamic service side status (e.g., workloads of weather forecasting service may increase sharply when weekends are coming.) and user-side environment (e.g., network latency would increase during the office hours). (2) Within a relatively short-time period such as one minute or one hour, a service performance observed by a user is relatively stable. The above two observations indicate that the feature information of latent feature vectors in time slice t_c can be predicted by utilizing the feature information collected before t_c. Moreover, the performance curve in terms of time should be smooth, which means more recent information is placed with more emphasis for predicting the performance in time slice t_c. Therefore, we estimate the feature vectors in time slice t_c by computing the weighted average of feature vectors in the past time slice:

$$\hat{p}_u(t_c) = \frac{\sum_{k=1}^{w} p_u(t_{c-k}) f(k)}{\sum_{k=1}^{w} f(k)}, \tag{4.6}$$

$$\hat{q}_i(t_c) = \frac{\sum_{k=1}^{w} q_i(t_{c-k}) f(k)}{\sum_{k=1}^{w} f(k)}, \tag{4.7}$$

where $\hat{p}_u(t_c)$ and $\hat{q}_i(t_c)$ are the predicted user feature vector and service feature vector in time slice t_c, respectively. w controls the information of how many past time slices are used for making prediction. In Eqs. (4.6) and (4.7), large weight values are assigned to the feature vectors in recent slices, while small weight values are assigned to the feature vectors in old slices.

Given the predicted latent feature vectors $\hat{p}_u(t_c)$ and $\hat{q}_i(t_c)$, we can predict the service performance value observed by a user in time slice t_c. For the user u and the service i, the predicted performance value $\hat{r}_{ui}(t_c)$ is defined as

$$\hat{r}_{ui}(t_c) = \hat{p}_u^T(t_c)\hat{q}_i(t_c) \tag{4.8}$$

4.3.3.2 Phase 2: Online Incremental Algorithm

In this phase, we propose an incremental algorithm for efficiently updating the feature model built in phase 1 at runtime as new performance data are collected in each time slice. In time slice t_{c-1}, $\hat{p}_u(t_{c-1})$ and $\hat{q}_i(t_{c-1})$ are predicted based on the data collected during the time slice t_{c-2-w} and t_{c-2}. During the time slice t_{c-1}, there would be some services invoked by several different users. Therefore, newly observed service performance values are available and collected from users. The new performance data are stored in a user-service matrix $R(t_{c-1})$ representing information in time slice t_{c-1}. By performing matrix factorization on $R(t_{c-1})$, latent feature vectors $p_u(t_{c-1})$ and $q_i(t_{c-1})$ in time slice t_{c-1} are learned from the real performance data. According to Eqs. (4.6) and (4.7), the feature vector prediction needs to be recomputed repeatedly at each time slice using all the vectors in previous w time slices, which is highly computationally expensive. In order to predict the feature vectors in time slice t_c more efficiently, we rewrite the Eqs. (4.6) and (4.7) as follows:

$$\hat{p}_u(t_c) = e^{-\alpha}\left(\frac{p_u(t_{c-1})}{\sum_{k=1}^{w} f(k)} + \hat{p}_u(t_{c-1}) - \frac{p_u(t_{c-1-w}) f(w)}{\sum_{k=1}^{w} f(k)}\right), \tag{4.9}$$

$$\hat{q}_i(t_c) = e^{-\alpha}\left(\frac{q_i(t_{c-1})}{\sum_{k=1}^{w} f(k)} + \hat{q}_i(t_{c-1}) - \frac{q_i(t_{c-1-w}) f(w)}{\sum_{k=1}^{w} f(k)}\right), \tag{4.10}$$

where $e^{-\alpha}$, $f(w)$ and $\sum_{k=1}^{w} f(k)$ are constant values. $p_u(t_{c-1-w})$ and $q_i(t_{c-1-w})$ are feature vectors calculated in time slice t_{c-1-w} and can be stored with only constant memory space. $p_u(t_{c-1})$ and $q_i(t_{c-1})$ can be quickly calculated in time slice t_{c-1}, since the computation complexity of matrix factorization is very low. Note that in Eqs. (4.9) and (4.10), we obtain a recursive relation between $[p_u(t_{c-1}), q_i(t_{c-1})]$ and $[p_u(t_c), q_i(t_c)]$, which means the feature model in time slice t_{c-1} can be efficiently updated for predicting the feature vectors in new time slice t_c.

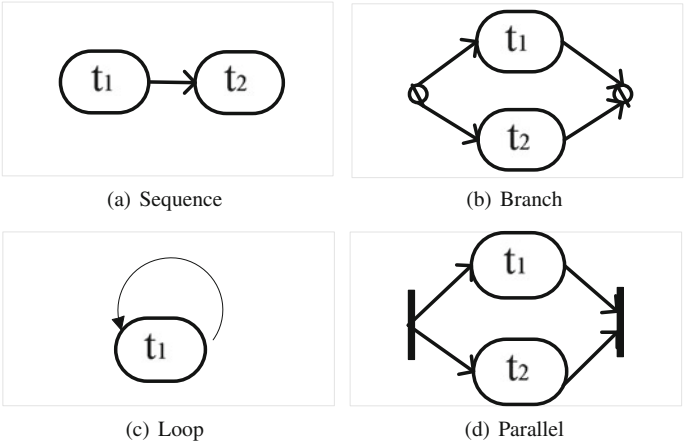
In the online phase, it could be possible that a new user or service is found. Since there is no prior information about the user or the service in the previous time slices, it is difficult to precisely predict the corresponding features by employing the online incremental algorithm. To address the cold start problem, we employ average performance for prediction. More precisely, the prediction for a new user or a new service is set as follows:

$$\hat{r}_{ui}(t) = \begin{cases} \bar{r}_i(t), & \text{if new user and old service} \\ \bar{r}_u(t), & \text{if old user and new service} \\ \bar{r}(t), & \text{if new user and new service} \end{cases}$$

where $\bar{r}_i(t)$ is the average-predicted performance of service i observed by all users in time slice t, $\bar{r}_u(t)$ is the average-predicted performance of all services observed by user u in time slice t, and $\bar{r}(t)$ is the average-predicted performance of all user-service pairs in time slice t.

4.3.4 Computation Complexity Analysis

The offline phase includes learning latent features in w time slices and running an evolutionary algorithm. The main computation is evaluating the objective function $\mathscr{L}$ and its gradients against the variables. Since the matrix $R(t)$ is typically sparse, the computational complexity for evaluating the objective function $\mathscr{L}$ in each time slice is $O(\rho_r l)$, where ρ_r is the number of nonzero entries in the matrix $R(t)$ and l is the dimension of the latent features. The computational complexities for the gradients $\frac{\partial \mathscr{L}}{\partial p_u(t)}$ and $\frac{\partial \mathscr{L}}{\partial q_i(t)}$ in Eqs. (4.3) and (4.4) are $O(\rho_r l)$. Therefore, the total computational complexity in one iteration is $O(\rho_r l w)$, where w is the number of time slices. In



(a) Sequence (b) Branch

(c) Loop (d) Parallel

Fig. 4.4 Basic compositional structures. ©[2014] IEEE. Reprinted, with permission, from Ref.[18]

the online phase, the main computation is factorizing the new performance matrix in time slice t. The computational complexity of online incremental algorithm is $O(\rho_r l)$.

The analysis indicates that theoretically, the computational time of offline algorithm is linear with respect to the number of observed performance entries in one time slice and the total number of time slices whose information is used for prediction. Note that because of the sparsity of $R(t)$, $\rho_r << mn$, which indicates that the computation time grows slowly with respect to the size of matrix $R(t)$. The computational time of the online algorithm is linear with the amount of newly observed performance information, which indicates that our proposed approach can efficiently integrate the performance model with new information and make online prediction timely. This complexity analysis shows that our proposed approach is very efficient and can be applied to large-scale systems.

4.4 System-Level Performance Prediction

In this section, we first present the aggregated response time calculation methods for basic compositional structures. Then, by analyzing the service flow, the system-level response time can be predicted in a hierarchical way. The overall performance of a system consists of service response time and local execution time. Local execution time refers to the computation time between service invocations in local system. Since the variance of system performance at runtime is mainly due to the highly varying service response time, local execution time, which is relatively constant at runtime, is not included in the defined system-level performance.

Table 4.1 Calculation of aggregated response time. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

Structure	Calculation method	Meaning of notation
Sequence	$r = \sum_{i=1}^{n} r_i$	n: number of sequential subtasks
		r_i: response time of the ith subtask
Branch	$r = \sum_{i=1}^{n} p_i r_i$	n: number of branches
		r_i: response time of the ith branch
		p_i: probability of the ith branch to be executed
Loop	$r = \sum_{i=1}^{n} p_i r_i i$	n: maximum looping times
		r_i: response time of the ith subtask
		p_i: probability of executing the subtask for i times
Parallel	$r = \max_{i=1}^{n} r_i$	n: number of branches
		r_i: response time of the ith branch

Typically, as shown in Fig. 4.4, there are four types of basic compositional structures, i.e., sequence, branch, loop, and parallel. The response time of each structure can be calculated by aggregating the response time of its subtasks as shown in Table 4.1.

For predicting the overall execution time of a service flow, we first decompose the system structure to a set of basic compositional structures in a hierarchical way. Then, the end-to-end system execution time is calculated in a bottom-up way. Taking Fig. 4.5 as an example, first the execution time of basic compositional structures T_1 and T_2 is calculated by employing the aggregation methods of sequence and loop, respectively. Then, the execution time of T_3 is calculated by employing aggregation method for branch compositional structure. Finally, the overall system execution time is calculated by employing aggregation method for sequence on t_1, t_2, T_3, and t_6.

With the aggregation approach, designers of service-oriented systems can estimate the performance of systems at design time. At runtime, the user-observed system-level performance can be efficiently predicted automatically. Once the system performance is decreased at runtime, by analyzing the system structure in a top-down way,

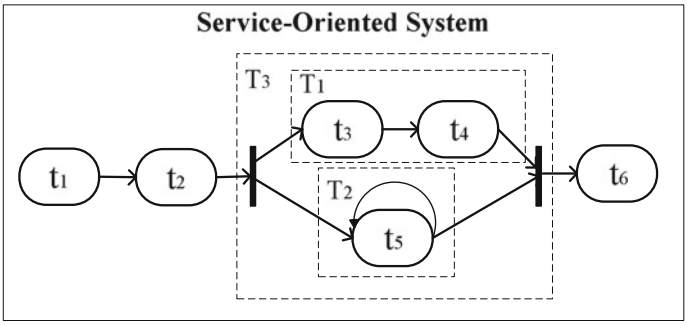


Fig. 4.5 Performance composition example. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

bad performance services can be quickly identified. With the predicted service performance information, dynamical service composition techniques can be employed to improve the system performance by replacing the long response time services with better ones.

4.5 Experiments

In this section, we conduct two experiments to evaluate our online performance prediction approach. In the first experiment, by comparing with several state-of-the-art service performance prediction methods, we present the effectiveness and efficiency of our approach. In the second experiment, we study the service flow of a real-world service-oriented system. We also study the performance improvement by integrating the predicted performance information of our approach into the dynamic composition mechanism.

In the following, Sect. 4.5.1 introduces the experimental setup and gives the description of the experimental dataset. Section 4.5.2 defines the evaluation metrics. Section 4.5.3 compares the prediction quality of our approach with other competing approaches. Sections 4.5.4, 4.5.5, and 4.5.6 study the impact of data density, dimensionality, and parameter α and w, respectively. Section 4.5.7 compares the computational time of different approaches. Section 4.5.8 studies the system-level performance prediction.

4.5.1 Experimental Setup and Dataset Collection

To evaluate the service-level performance prediction quality of our proposed approach in the real world, we implement a tool WSMonitor for collecting the performance

information of Web services. WSMonitor is deployed as a middleware on the user-side, which can continuously monitor the user-experienced performance of invoked services. By sharing the user-side observed performance to performance center, it can obtain performance prediction service from performance center at runtime.

WSMonitor is implemented and deployed with JDK 6.0, Eclipse 3.3, Axis 2, and Apache 2.2.17. Within WSMonitor, there are several modules including *WSDL Crawler*, *Code Generator*, and *Performance Monitor*. *WSDL Crawler* first crawls a set of WSDL files from the Internet and generates a list of openly accessible Web services. For each Web service in the list, *Code Generator* automatically generates a java class for service invocation by employing the WSDL2Java tool from the Axis package. Totally, 5871 classes are generated for 5871 Web services. By calling the functions generated by Code Generator, *Performance Monitor* is able to send operation requests to Web services and record the corresponding response time with time stamps.

We deploy the WSMonitor on 142 distributed computers located in 22 countries from PlanetLab, which is a distributed test bed consisting of hundreds of computers all over the world. Each computer acts as a service user by invoking the listed Web services from time to time. Totally, 4532 publicly available real-world Web services from 57 countries are monitored by each computer continuously. About 1339 of the initially selected Web services are excluded in this experiment due to: (1) authentication required and (2) permanent invocation failure (e.g., the Web service is shutdown). In our experiment, each of the 142 computers sends operation requests to all the 4532 Web services in every time slice. The experiment lasts for 16 hours with one time slice lasting for 15 min.

By collecting performance records from all the computers, finally 30,287,611 performance results are included into the Web service response time dataset. The response time of all the 4532 Web services observed by all the 142 service users during 64 time slices can be presented as a set of 142×4532 user-service matrices, each of which stands for a particular time slice.

The statistics of Web service response time dataset are summarized in Table 4.2. Response-time is within the range of 0–20 s, whose mean is 3.165 s. The distribution of the response-time values of all the matrices is shown in Fig. 4.6a. From Fig. 4.6a, we can observe that most of the response-time values are between 0.1 and 0.8 s.

Table 4.2 Statistics of Web service response time dataset. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

Statistics	Response time
Scale	0–20 s
Mean	3.165 s
Num. of users	142
Num. of Web services	4532
Num. of time slices	64
Num. of records	30,287,611

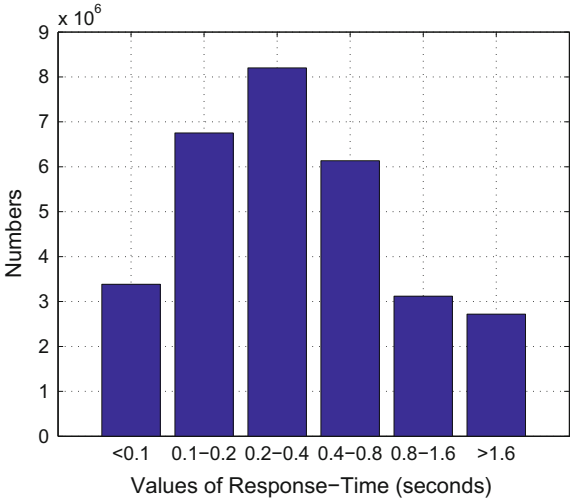


Fig. 4.6 Response time value distribution. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

4.5.2 Metrics

We assess the prediction quality of our proposed approach in comparison with other methods by computing mean absolute error (MAE) and root-mean-squared error (RMSE). The metric MAE is defined as:

$$MAE = \frac{\sum_{uit} |\hat{r}_{ui}(t) - r_{ui}(t)|}{N}, \tag{4.11}$$

and RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{uit} (\hat{r}_{ui}(t) - r_{ui}(t))^2}{N}}, \tag{4.12}$$

where $r_{ui}(t)$ is the response time value of Web service i observed by user u in time slice t, $\hat{r}_{ui}(t)$ denotes the predicted response time value of Web service i which would be observed by user u in time slice t, and N is the number of predicted response time values in the experiments.

4.5.3 Comparison

In this section, in order to show the effectiveness and efficiency of our proposed online Web service performance prediction approach, we compare the service-level prediction accuracy of the following methods:

- **UPCC**—This is a neighborhood-based method which employs Pearson correlation coefficient to calculate similarities between users. It predicts response time of services based on the observed performance from similar users [3, 11]. Since UPCC cannot perform online prediction for the next time slice, we extend the traditional UPCC by using the average performance from similar users for prediction.
- **IPCC**—This is a neighborhood-based method which employs Pearson correlation coefficient to calculate similarities between services. It predicts response time of services based on the performance of similar services [9]. Similar to UPCC, we make an extension to IPCC in order to compare the online prediction quality with other methods.
- **MF**—This method first compresses the set of user-service matrices into an average user-service matrix. For each entry in the matrix, the value is the average of the specific user-service pair during all the time slices. After obtaining the compressed user-service matrix, it applies the nonnegative matrix factorization technique proposed by Lee and Seuing [7] on user-service matrix for missing value prediction. The predicted values are used as the response time of the corresponding user-service pair in the next time slice.
- **TF**—This is a tensor factorization-based prediction method. It combines the set of user-service matrices as a tensor with a third dimension representing the time. Then, it applies tensor factorization on the user-service-time tensor to extract user-specific, service-specific, and time-specific characteristics. The missing value is then predicted based on how these characteristics apply to each other.
- **WSPred**—This is a tensor factorization-based prediction method [17]. Different from method **TF**, it adds average performance value constraints when extracting the latent characteristics.
- **OPred**—This method is proposed in this chapter. Firstly, the user features and service features are extracted in each time slice by employing matrix factorization. Then, the user features and service features in the new time slice are predicted by performing time analysis on the feature trends. Finally, the response time of user-service pairs is predicted by evaluating how the predicted features of users and services are applied to each other.

In order to evaluate the performance of different approaches in reality, we randomly remove some entries from the performance matrices to obtain observation matrices and compare the values predicted by a method with the original ones. The observation matrices with missing values are in different densities. For example, 10% means that we randomly remove 90% entries from the original matrices and use the remaining 10% entries for prediction. Note that under a certain density, we employ different approaches to predict the values by using the same observation matrix. The prediction accuracy is evaluated using Eqs. (4.11) and (4.12) by comparing the original values and the predicted values in the corresponding matrices. The values of λ_1, and λ_2 are tuned by performing cross-validation on the observed performance data. Without lost of generality, the parameter settings of all the approaches are $l = 20$, $w = 8$, $\alpha = 0.2$ and $\lambda 1 = \lambda 2 = 0.001$ in the experiments conducted in this

chapter. Detailed impacts of parameters are studied in Sects. 4.5.4, 4.5.5, and 4.5.6, respectively.

The service performance prediction accuracies evaluated by MAE and RMSE are shown in Table 4.3. A smaller MAE or RMSE value means a better performance. From Table 4.3, we can observe that the time-aware prediction methods (i.e., TF and OPred) outperform the non-time-aware prediction methods (i.e., UPCC, IPCC, and MF), since the time-aware methods employ the time-specific features as additional information for performance prediction. We also observe that our approach OPred constantly performs better than TF under both dense data and sparse data. This is because OPred assigns different weights on the performance information collected in different time slices. The prediction results rely more on recent user and service features than older ones. By setting $f(x)$ in Eq. (4.5) to a constant value (e.g., $f(x) = 1$), OPred is reduced to TF. WSPred further improves TF by employing a regularization term to prevent the predicted values from varying a lot against the average performance value. WSPred catches the periodic features of service performance. OPred proposed in this chapter captures not only the periodic features but also the non-periodic features of service performance. Therefore, OPred can predict the performance trend more precisely than WSPred. Moreover, WSPred is not an online approach and requires more computational time than OPred. The computational time

Table 4.3 Performance comparisons (A smaller MAE or RMSE value means a better performance). ©[2014] IEEE. Reprinted, with permission, from ref. [18]

Data density (%)	RMSE	Response time (seconds)					
		UPCC	IPCC	MF	TF	WSPred	OPred
5	Mean	5.312	5.289	5.329	4.751	4.362	**4.330**
	Best	5.263	5.276	5.321	4.747	4.358	**4.327**
10	Mean	5.043	4.972	5.079	4.567	4.287	**4.151**
	Best	4.962	4.946	5.063	4.563	4.283	**4.148**
45	Mean	4.425	4.371	4.337	4.208	3.923	**3.855**
	Best	4.388	4.342	4.318	4.202	3.918	**3.851**
50	Mean	4.352	4.354	4.298	4.016	3.899	**3.809**
	Best	4.331	4.336	4.274	4.012	3.894	**3.808**
Data density (%)	MAE	Response time (seconds)					
		UPCC	IPCC	MF	TF	WSPred	OPred
5	Mean	3.720	3.213	3.387	2.915	2.559	**2.417**
	Best	3.687	3.207	3.381	2.911	2.555	**2.413**
10	Mean	3.264	2.841	2.873	2.786	2.495	**2.376**
	Best	3.243	2.812	2.851	2.782	2.488	**2.374**
45	Mean	2.627	2.455	2.436	2.253	2.141	**2.029**
	Best	2.613	2.431	2.423	2.247	2.137	**2.026**
50	Mean	2.619	2.417	2.391	2.211	2.130	**2.011**
	Best	2.609	2.404	2.384	2.207	2.125	**2.008**

Table 4.4 Performance improvement of OPred. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

Competing approach	Performance improvement of OPred (%)
UPCC	22–36
IPCC	16–25
MF	15–28
TF	9–17
WSPred	1–6

is compared in Sect. 4.5.7. In Table 4.3, the MAE and RMSE values of dense data (e.g., data density is 45 or 50%) are smaller than those of sparse data (e.g., data density is 5 or 10%), since denser data provide more information for prediction. Performance improvement of OPred is shown in Table 4.4. Our online approach OPred improves the prediction accuracy by 22–36%, 16–25%, 15–28%, 9–17%, and 1–6% relative to UPCC, IPCC, MF, TF, and WSPred, respectively. The improvements are significant, which indicate the prediction effectiveness of OPred.

4.5.4 Impact of Data Density

In Fig. 4.7, we compare the prediction accuracy of all the methods under different data densities. We change the data density from 5 to 50% with a step value of 5%. The parameter settings in this experiment are $l = 20$, $w = 8$, $\alpha = 0.2$, and $\lambda 1 = \lambda 2 = 0.001$.

In Fig. 4.7a, b, the experimental results show that our approach OPred achieves higher prediction accuracy (smaller MAE and RMSE values) than other competing methods under different data density. In general, when the data density is increased from 5 to 20%, the prediction accuracy of our approach OPred is significantly enhanced. When the data density is further increased from 20 to 50%, the enhancement of prediction accuracy will decrease. This observation indicates that when the data are very sparse, collecting more performance information will greatly enhance the prediction accuracy.

4.5.5 Impact of Dimensionality

The parameter dimensionality l determines the number of latent features applied to characterize users and services. In Fig. 4.8, we study the impact of parameter dimensionality by varying the values of l from 5 to 50 with a step value of 5. Other parameter settings are $w = 8$, $\alpha = 0.2$, and $\lambda 1 = \lambda 2 = 0.001$.

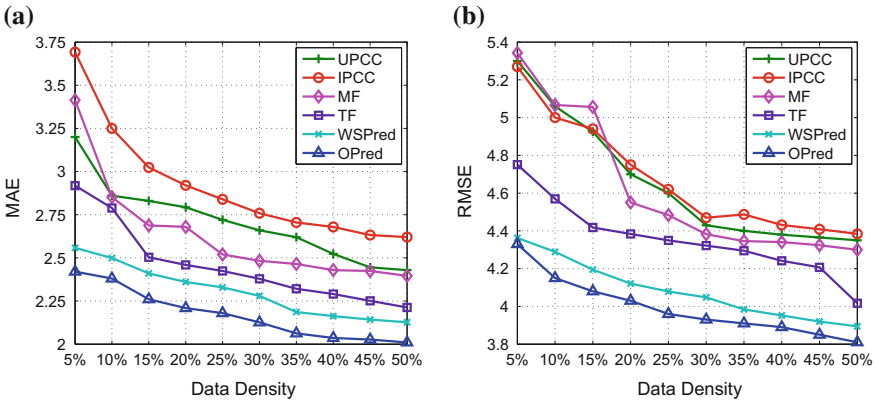


Fig. 4.7 Impact of data density. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

In Fig. 4.8, we observe that as l increases, the MAE and RMSE decrease (prediction accuracy increases), but when l surpasses a certain threshold like 20, the MAE and RMSE increase (prediction accuracy decreases) with further increase of the value of l. This observation indicates that too few latent factors are not enough to characterize the features of user and service, while too many latent factors will cause an overfitting problem. There exists an optimal value of l for characterizing the latent features. When the data density is 50%, we observe that our approach OPred achieves the best performance when the value of dimensionality is 25, while smaller values like 5 or larger values like 50 can potentially reduce the prediction accuracy. When the data density is 5%, we observe that the prediction accuracy of our approach OPred achieves the best performance when the value of dimensionality is 20, while smaller values like 5 or larger values like 50 can potentially reduce the prediction accuracy. This observation indicates that when the service performance

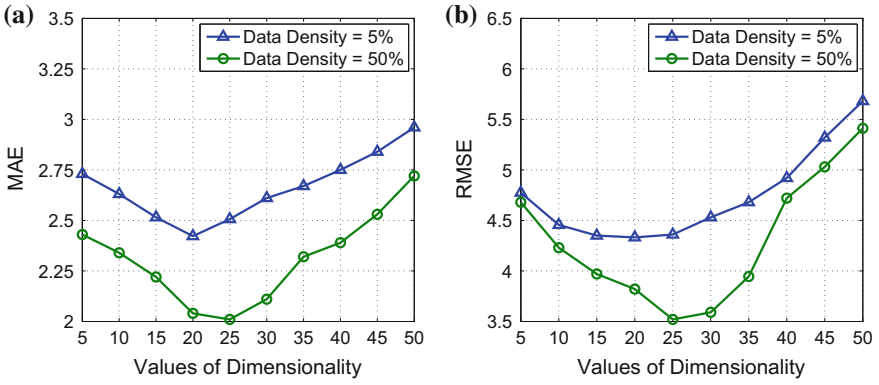


Fig. 4.8 Impact of dimensionality. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

data are sparse, 20 latent factors are already good enough to characterize the features of user and service, which are mined from the limited performance information. On the other hand, when the data are dense, more latent factors, like 25, are needed to characterize the latent features since more performance data are available.

4.5.6 Impact of α and w

The parameter α controls the decaying rates of weights assigned to different time slices. A larger value of α gives more weights to the recent time slices. w controls the information of how many past time slices are used for making prediction. In Fig. 4.9, we vary the values of w from 1 to 20 with a step value of 1. Other parameter settings are $\lambda 1 = \lambda 2 = 0.001$.

Figure 4.9 shows the impacts of α and w on MAE and RMSE. We observe that as w increases, the values of MAE and RMSE decrease (prediction accuracy increases) at first, but when w pass a certain threshold, the MAE and RMSE converge. This phenomenon coincides with the intuition that employing past performance information from more time slices can increase prediction accuracy. When w surpasses a certain threshold, the MAE and RMSE decrease little with further increase of the value of w. The reason is that when w is large enough, small weight values are assigned to the information of older time slices, which contribute little to the prediction accuracy. This observation indicates that too large w is unnecessary. The thresholds are different under different values of α. Since a larger value of α gives more weights to the recent time slices, the threshold is smaller than those under smaller values of α. In Fig. 4.9, OPred achieves the best performance when $\alpha = 0.2$. The observation confirms with the intuition that with a large value of α useful information from older

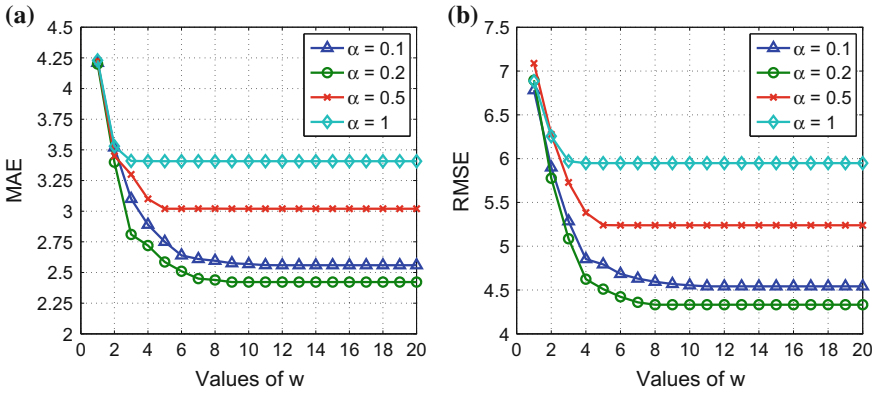


Fig. 4.9 Impact of α and w. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

time slice will be lost, and with a small value of α noisy data will cause the decrease of prediction accuracy.

4.5.7 Computational Time Comparisons

In Sect. 4.3.4, we theoretically analyze the computation time of OPred. In this section, we compare the computation efficiencies of different approaches. In our experiments, one time slice lasts for 15 min. We compare the average computational time of a prediction approach with the length of a time slice. The data used for performance prediction are the same for all approaches. From Table 4.5, we observe that the computational time of OPred takes less than 2% of a time slice. This observation is consistent with the time complexity analysis in Sect. 4.3.4 and shows that our proposed approach OPred is efficient and can be applied to large-scale systems in real world. TF and WSPred use more than 10% of a slice time to conduct prediction, since they are not online approaches and need to rebuild the model whenever new data are available. TF performs better than WSPred because WSPred contains an extra term in the objective function representing the average performance constraints. MF performs better than TF and WSPred because time factor is not considered when predicting the performance values. UPCC and IPCC perform worst since they are neighborhood-based approaches and take a lot of time to find the relationship between users and services.

4.5.8 System-Level Performance Case Study

In this section, we evaluate our approach OPred by using a sample service-oriented system. Figure 4.10 shows a typical online shopping system. It allows customers to browse and order products from the shopping Web site. In this shopping system, the designer integrates three Web services for providing users access to various

Table 4.5 Average computational time comparisons. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

Approach	Computational time (m)	Percentage of a time slice (%)
UPCC	10.095	67.3
IPCC	9.735	64.9
MF	1.575	10.5
TF	1.860	12.4
WSPred	2.055	13.7
OPred	0.240	1.6

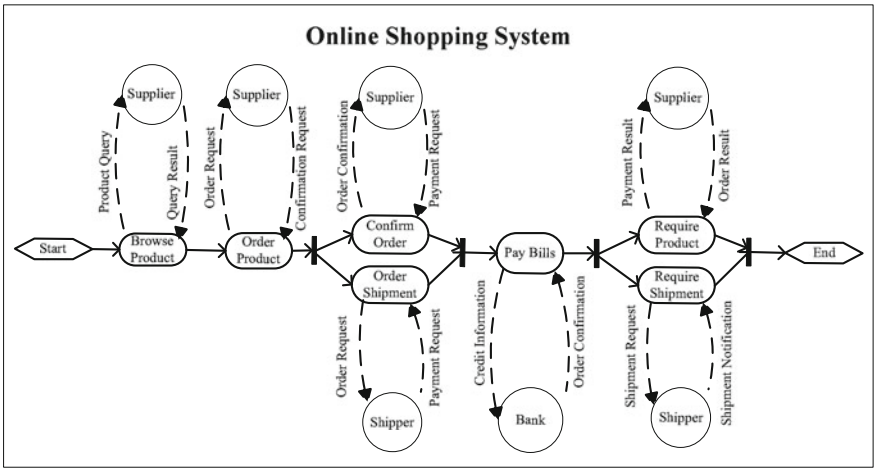


Fig. 4.10 Online shopping system. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

product suppliers, banks, and shippers. This example is taken from the online services provided by a gift Web site [6].

The service flow is illustrated in Fig. 4.10. By sending product queries to suppliers, the shopping system can obtain plenty of product information, which allows customers to browse various products on the Web site. Once a customer decides to buy a product, the shopping system sends an order request with product information to the corresponding supplier. The supplier then reserves a product for the customer and replies the shopping system with an order confirmation request. At this point, the shopping system needs to send an order confirmation to the supplier and an order request to a shipper service. Once the shopping system receives payment requests from both the product supplier and a shipper service, it proceeds to launch a payment transaction via a credit card payment service (e.g., PayPal). In the task of paying bills, customer's credit card information is transferred to the bank, and an invoice is sent back by the bank. Finally, the product supplier is notified of a bank invoice to complete the purchase. At the same time, a request is sent to the shipper to arrange the shipment of the product. Once the product is aboard, the shipper notifies the shopping system with estimated arrival date of the shipment.

After, we find a set of functional identical Web services from the performance dataset for each abstract task in the shopping system. The predicted service performance results are used to predicting the end-to-end performance of shopping system by employing the compositional methods in Sect. 4.4. As discussed before, by calculating system performance, poor services can be identified in a hierarchical way. Then, the identified services can be replaced with better ones to maintain the

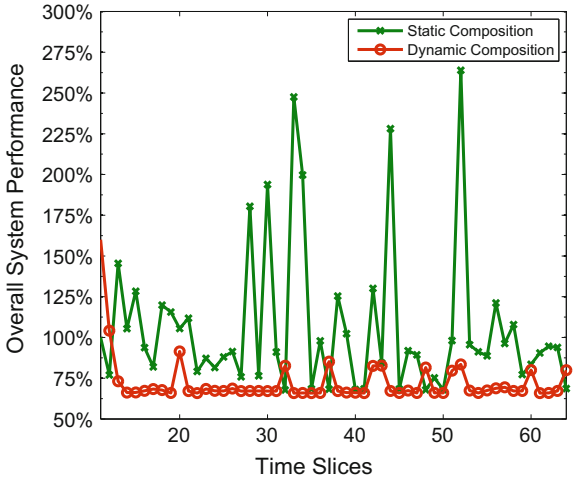


Fig. 4.11 System performance improvement of dynamically service composition. ©[2014] IEEE. Reprinted, with permission, from Ref. [18]

overall system performance at runtime. In Fig. 4.11, we compare the system performance of static composition and dynamical composition. In static composition, for each abstract task we randomly choose a service from the set of functional identical candidates. The set of selected services is fixed in all time slices. In dynamical composition, the predicted service performance of OPred is employed to select the optimal services for task executions in each time slice. In this book, we focus on dynamic selecting atomic services. The comparison begins from time slice 11 since the performance information of the first 10 time slices is used as training data for OPred. The system performance of static composition method in time slice 11 is chosen as baseline. Other performance is compared with baseline in percentage (a smaller number means better performance). From Fig. 4.11, we can observe that the system performance of static composition is unstable at runtime. This is because the performance of some selected services is unstable, which impacts the system overall performance. For dynamic composition, since OPred can precisely predict service performance, the service-oriented system can be updated by integrating potentially optimal services at runtime. The system performance of dynamical composition maintains stable in a good level, which indicates the effectiveness of OPred.

4.6 Summary

Based on the intuition that a user's current Web service performance usage experience can be predicted by using the past usage experience from different users, we propose a novel online service performance prediction approach, called OPred, for

personalized performance prediction at runtime. Using the past Web service usage experience from different users, OPred builds feature models and employs time series analysis techniques on feature trends to make personalized performance prediction for different service users. The predicted service performance is critical for identifying poor services and maintaining the system performance timely. The extensive experimental results show that our proposed OPred outperforms the state-of-the-art performance prediction approaches in terms of prediction accuracy. The case study on a typical shopping system shows the effectiveness of OPred.

For future work, we will investigate more techniques for improving the prediction accuracy (e.g., data smoothing, utilizing content-aware information). We will conduct experiments on more real-world service-oriented systems to evaluate the effectiveness and efficiency of OPred when applied to different domains.

References

1. M. Alrifai, T. Risse, Combining global optimization with local selection for efficient QoS-aware service composition, in *Proceedings of International Conference on World Wide Web (WWW'09)* (2009), pp. 881–890
2. M. Alrifai, D. Skoutas, T. Risse. Selecting skyline services for qos-based web service composition, in *Proceedings of WWW'10* (2010), pp. 11–20
3. J. Breese, D. Heckerman, C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering, in *Proceedings of UAI'98* (1998), pp. 43–52
4. J. El Haddad, M. Manouvrier, M. Rukoz, Tqos: Transactional and qos-aware selection algorithm for automatic web service composition. IEEE Trans. Serv. Comput., 73–85 (2010)
5. D. Fensel, F. Facca, E. Simperl, I. Toma, Web service modeling ontology. Seman. Web Serv., 107–129 (2011)
6. Gift Website, http://bjqad.com/yawen/mall/index.asp
7. D. Lee, H. Seung, Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
8. N.N. Liu, M. Zhao, E. Xiang, Q. Yang, Online evolutionary collaborative filtering, in *Proceedings of the Fourth Conference on Recommender Systems (RecSys'10)* 2010, pp. 95–102
9. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: an open architecture for collaborative filtering of netnews, in *Proceedings of CSCW'94* (1994), pp. 175–186
10. N. Salatge, J. Fabre, Fault tolerance connectors for unreliable web services, in *Proceeding of IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)* (2007), pp. 51–60
11. L. Shao, J. Zhang, Y. Wei, J. Zhao, B. Xie, H. Mei, Personalized qos prediction forweb services via collaborative filtering, in *Proceedings of ICWS'07* (2007), pp. 439–446
12. P. Xiong, Y. Fan, M. Zhou, Qos-aware web service configuration. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **38**(4), 888–895 (2008)
13. P. Xiong, Y. Fan, M. Zhou, A petri net approach to analysis and composition of web services. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(2), 376–387 (2010)
14. T. Yu, Y. Zhang, K. Lin, Efficient algorithms for web services selection with end-to-end QoS constraints. ACM Trans. Web (TWEB) **1**(1), 6 (2007)
15. L. Zeng, B. Benatallah, A. Ngu, M. Dumas, J. Kalagnanam, H. Chang, QoS-aware middleware for web services composition. IEEE Trans. Software Eng. (TSE) **30**(5), 311–327 (2004)
16. L. Zhang, S. Cheng, C. Chang, Q. Zhou, A pattern-recognition-based algorithm and case study for clustering and selecting business services. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **42**(1), 102–114 (2012)

17. Y. Zhang, Z. Zheng, M. Lyu, Wspred: a time-aware personalized qos prediction framework for web services, in *Proceedings of IEEE Symposium on Software Reliability Engineering (ISSRE'11)* (2011), pp. 210–219
18. Y. Zhang, Z. Zheng, M.R. Lyu, An online performance prediction framework for service-oriented systems. IEEE Trans. Syst. Man Cybern. Syst. (TSMC) **44**(9),1169–1181 (2014)

Chapter 5
QoS-Aware Web Service Searching

Abstract Web services are becoming prevalent nowadays. Finding desired Web services is becoming an emergent and challenging research problem. In this chapter, we present WSExpress (Web Service Express), a novel Web service search engine to expressively find expected Web services. WSExpress ranks the publicly available Web services not only by functional similarities to user queries, but also by non-functional QoS characteristics of Web services. WSExpress provides three searching styles, which can adapt to the scenario of finding an appropriate Web service and the scenario of automatically replacing a failed Web service with a suitable one. WSExpress is implemented by Java language, and large-scale experiments employing real-world Web services are conducted. Totally, 3738 Web services (15,811 operations) from 69 countries are involved in our experiments. The experimental results show that our search engine can find Web services with the desired functional and non-functional requirements. Extensive experimental studies are also conducted on a well-known benchmark dataset consisting of 1000 Web service operations to show the recall and precision performance of our search engine.

5.1 Overview

As shown in Fig. 5.1, with a set of standard protocols, i.e., SOAP (Simple Object Access Protocol), WSDL (Web Services Description Language), and UDDI (Universal Description, Discovery and integration), Web services provided by different organizations can be discovered and integrated to develop service-oriented applications [3]. With the growing number of Web services in the Internet, many alternative Web services can provide similar functionalities to fulfill users' requests. Syntactic or semantic matching approaches based on services' tags in UDDI repository are usually employed to discover suitable Web services [9]. However, discovering Web services from UDDI repositories suffers several limitations. First, since UDDI repository is no longer a popular style for publishing Web services, most of the UDDI repositories are seldom updated. This means that a significant part of information in these repositories is out of date. Second, arbitrary tagging methods used in different UDDI repositories add to the complexity of searching Web services of interest.

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_5

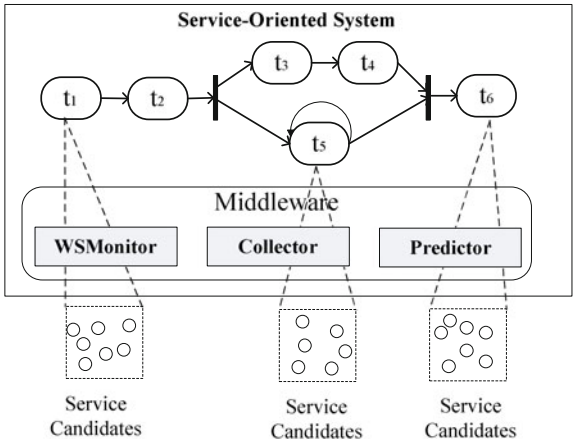


Fig. 5.1 Service-oriented system architecture. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

To address these problems, an automated mechanism is required to explore existing Web services. Considering that WSDL files are used for describing Web services and can be obtained in several ways other than UDDI repositories, several WSDL-based Web service searching approaches are proposed such as *Binding Point*, *Grand Central*, *Salcentral*, and *Web Service List*. However, these engines only simply exploit keyword-based search techniques which are obviously insufficient for catching the Web services' functionalities. First, keywords cannot represent Web services' underlying semantics. Second, since a Web service is supposed to be used as part of the user's application, keywords cannot precisely specify the information user needs and the interface acceptable to the user. In this chapter, we employ not only keywords but also operation parameters to comprehensively capture Web service's functionality.

In addition, Web services sharing similar functionalities may possess very different non-functionalities (e.g., response time, throughput, availability, usability, performance, integrity). In order to effectively provide personalized Web service ranking, it is requisite to consider both functional and non-functional characteristics of Web services. Unfortunately, the Web service search engines mentioned above cannot distinguish the non-functional differences between Web services.

QoS-driven Web service selection is a popular research problem [1, 6, 10]. A basic assumption in the field of selection is that all the Web services in the candidate set share identical functionality. Under this assumption, most of the selection approaches can only differentiate among Web services' non-functional QoS characteristics, regardless of their functionalities. While these QoS-driven selection approaches are directly employed to Web service search engines, several problems will arise. One is that Web services whose functionalities are not exactly equivalent to the user searching query are completely excluded from the result list. Another problem is that Web services in the result list are ordered only according to their QoS metrics, while combining both functional and non-functional attributes is a more reasonable method.

To address the above issues, we propose a new Web service discovering approach by paying respect to functional attributes as well as non-functional features of Web services. A search engine prototype, WSExpress, is built as an implementation of our approach. Experimental results show that our search engine can successfully discover user-interested Web services within top results. In particular, the contributions of this chapter are threefold:

- Different from all previous work, we propose a brand new Web service searching approach considering both functional and non-functional qualities of the service candidates.
- We conduct a large-scale distributed experimental evaluation on real-world Web services. 3738 Web services (15,811 operations) located in 69 countries are evaluated both on their functional and non-functional aspects. The evaluation results show that we can recommend high-quality Web services to the user. The precision and recall performance of our functional search is substantially better than the approach in previous work [7].
- We publicly release our large-scale real-world Web service WSDL files and associated QoS datasets for future research. To the best of our knowledge, our dataset is the first publicly available real-world dataset for functional and non-functional Web service searching research.

The rest of this chapter is organized as follows: Sect. 5.2 introduces Web service searching backgrounds. Section 5.3 presents the system architecture. Section 5.4 presents our QoS-aware searching approach. Section 5.5 describes our experimental results. Section 5.6 concludes the chapter.

5.2 Motivation

Figure 5.2 shows a common Web service query scenario. A user wants to find an appropriate Web service which contains operations that can be integrated as part of the user's application. The user needs to specify the functionality of a suitable operation by filling the fields of keywords, input and output. Also the user may have some special requirements on service quality, such as the maximum price. These personal requirements can be represented by setting the QoS constraint field. The criticality of different quality criteria for a user can be defined by setting the QoS weight field.

A lot of Web services can be accessed over the Internet. Each service candidate provides one or more operations. Generally, these operations can be described in the structure shown in Fig. 5.2. Each operation includes a name, the parameters of input and output elements, and the descriptions about the functionality of this operation as well as the Web services it belongs to in its associated WSDL document. The service quality associated with this operation is represented by several criteria values, e.g., Q1, Q2 in Fig. 5.2.

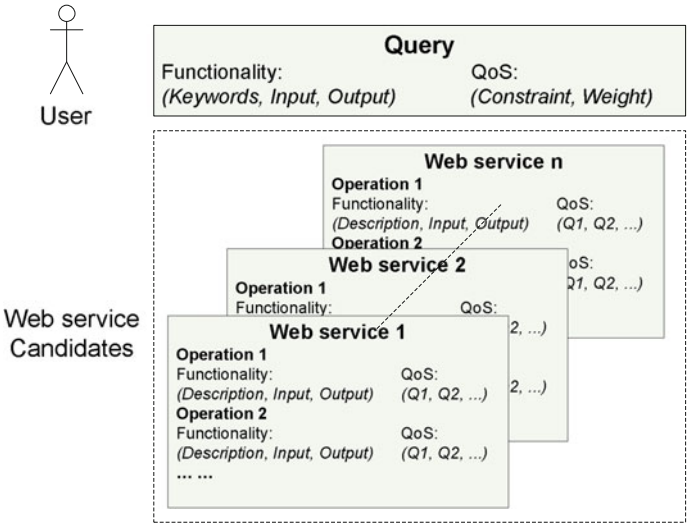


Fig. 5.2 Web service query scenario. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

Table 5.1 User query examples. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

User query	Functionality			QoS	
	Keywords	Input	Output	Constraint (C1, C2, C3)	Weight (W1, W2, W3)
Query 1	Car	Name, type	Price	(0.5, 0.5, 0.2)	(0.4, 0.4, 0.2)
Query 2	Weather	City, country	Weather	(0.6, 0.3, 0.3)	(0.3, 0.4, 0.3)

Table 5.1 shows Web service query examples. In query 1, a user wants to find a Web service that can provide appropriate operations for displaying prices of different types and brands of cars. The input information provided by the user for that particular operation is the types and names of cars. This query is structured into three parts: *keywords*, *input*, and *output*. The *keywords* part defines in which domain is the query about. In this example, the user concerns about the domain "car." The *input* part contains "name" and "type" since they can be provided by the user. The *output* part is set as "price" to specify the information the user wants to obtain from an appropriate operation.

In Table 5.2, we enumerate three possible results for the user's search query. Web service 1 provides one operation *CarPrice* and this operation's functionality is almost the same as what the user specifies in the query. In addition, the service quality meets the user's requirements. Web service 2 provides operation *AutomobileInformation*. Operation *AutomobileInformation* can provide many information details including the price of the automobiles after invoked with "name" and "model" as input. However, some service quality criteria, such as the service price (Q1) and the response time (Q2), are beyond the user's tolerance. Operation *VehicleRecommend* provided

Table 5.2 Web service examples. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

Service ID	Operation name	Input	Output	QoS (Q1, Q2, Q3)
WS 1	CarPrice	Name, type	Price	(0.8, 0.6, 0.6)
WS 2	AutomobileInformation	Name, model	Price, color, company	(0.2, 0.4, 0.6)
WS 3	VehicleRecommend	Name, model, usage	Rent, prime cost, provider	(0.6, 0.8, 0.5)

by Web service 3 recommends suitable vehicles for the user to rent. Although its target is to suggest the most suitable vehicle and vehicle rental companies to the user, it can also be invoked for obtaining the prices of cars due to the prime cost information provided. Besides, operation *VehicleRecommend*'s service quality fits the user's constraints and preferences quite well. Among these three Web services, the most suitable one is Web service 1, and another acceptable one is Web service 3, but Web service 2 is not highly suggested due to its service quality. Thus, a reasonable order of the recommendation list for the user's query is Web service 1, Web service 3, and Web service 2.

5.3 System Architecture

Now, we describe the system architecture of our QoS-aware Web service search engine. As shown in Fig. 5.3, after accepting a user's query specification, our search engine should be able to provide a practical Web service recommendation list. The search engine consists of three components: non-functional evaluation, ki functional evaluation, and QoS-aware Web service ranking.

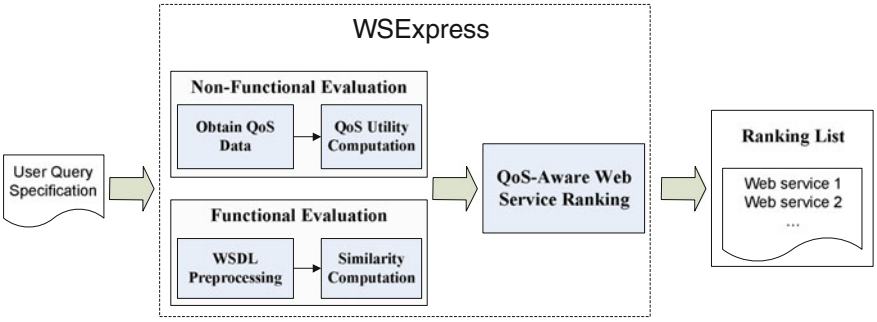


Fig. 5.3 System architecture. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

There are two phases in the non-functional evaluation component. In phase 1, the search engine obtains QoS criteria values of all the available Web services. In phase 2, the search engine computes the QoS utilities of different Web services according to the constraints and preferences specified in the QoS part of the user's query.

The functional evaluation component contains two phases. In phase 1, the search engine carries out a preprocessing work on the WSDL files associated with the Web services. This work aims at removing noise and improving accuracy of functional evaluation. In phase 2, the search engine evaluates the Web service candidates' functional features. These features are described by similarities between the functionality specified in the query and the functionality of operations provided by those Web services.

Finally, the search engine combines both functional and non-functional features of Web services in the QoS-aware Web service ranking component. A practical and reasonable Web service recommendation list is then provided as a result to the user's search query.

5.4 QoS-Aware Web Service Searching

5.4.1 QoS Model

In our QoS model, we describe the quantitative non-functional properties of Web services as quality criteria. These criteria include generic criteria and business specific criteria. Generic criteria are applicable to all Web services like response time, throughput, availability, and price, while business criteria such as penalty rate are specified to certain kinds of Web services.

By assuming m criteria are employed for representing a Web service quality, we can describe the service quality using a QoS vector $(q_{i,1}, q_{i,2}, \ldots, q_{i,m})$, where $q_{i,j}$ represents the jth criterion value of Web service i.

Some QoS criteria values of Web services, such as penalty rate and price, can be obtained from the service providers directly. However, other QoS attributes' values like response time, availability, and reliability need to be generated from all the users' invocation records due to the differences between network environments. In this chapter, we use the approach proposed in [12] to collect QoS performance on real-world Web services.

We put all the Web services' QoS vectors together and form a QoS matrix Q. Each row in Q represents a Web service, while each column represents a QoS criterion value.

$$Q = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,t} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,t} \\ \vdots & \vdots & \vdots & \vdots \\ q_{s,1} & q_{s,2} & \cdots & q_{s,t}. \end{pmatrix} \tag{5.1}$$

A utility function is used to evaluate the multi-dimensional quality of a Web service. The utility function maps a QoS vector into a real value for evaluating the Web service candidates. To represent user priorities and preferences, Two steps are involved into the utility computation: (1) The QoS criteria values are normalized to enable a uniform measurement of the multi-dimensional Quality-of-Service independent of their units and ranges. (2) The weighted evaluation on criteria is carried out for representing user's constraints, preference, and special requirements.

5.4.1.1 Normalization

In this step, each criterion value is transformed to a real value between 0 and 1 by comparing it with the maximum and minimum values of that particular criterion. For some criterion, the possible absolute value could be very large or infinite. A pair of maximum and minimum values are specified for every criterion, respectively. Let $q_{i,u}$ be the upper bound value and $q_{i,l}$ be the lower bound value for the ith criterion, respectively. Every QoS value is transformed according to the following equations:

$$f(x) = \begin{cases} r_{min}, \text{ if} x < r_{min} \\ r_{max}, \text{ if} x > r_{max} \\ x, \text{ otherwise.} \end{cases}$$

The normalized value of $q_{i,j}$ can be represented by $q'_{i,j}$ as follows:

$$q'_{i,j} = \frac{q_{i,j} - q_{i,0}}{q_{i,n} - q_{i,0}}. \tag{5.2}$$

Thus, the QoS matrix Q is transformed into a normalized matrix Q' as follows:

$$Q' = \begin{pmatrix} q'_{1,1} & q'_{1,2} & \cdots & q'_{1,t} \\ q'_{2,1} & q'_{2,2} & \cdots & q'_{2,t} \\ \vdots & \vdots & \vdots & \vdots \\ q'_{s,1} & q'_{s,2} & \cdots & q'_{s,t}. \end{pmatrix} \tag{5.3}$$

5.4.1.2 Utility Computation

Some Web services need to be excluded from the candidate set due to their inconsistency with the user's QoS constraints. The QoS constraints set the worst quality user can accept. These constraints are usually set according to the application developers' experience or computed by some QoS-driven composition algorithm. Web service with any QoS criterion grade unsatisfying user constraint may cause problem while integrated into user's application. For example, if a service fails to return the result within a given period of time, another service may exit with a error code time out while

waiting for the result. Assume a user's constraint vector is $C = (c_1, c_2, \ldots, c_m)$, in which c_i sets the minimum normalized ith criterion grade. We will only consider those Web services whose criteria grades are all larger than the constraints. In other words, we delete the rows which fail to satisfy the constraints from Q' and produce a new matrix Q^*:

$$Q^* = \begin{pmatrix} q^*_{1,1} & q^*_{1,2} & \cdots & q^*_{1,t} \\ q^*_{2,1} & q^*_{2,2} & \cdots & q^*_{2,t} \\ \vdots & \vdots & \vdots & \vdots \\ q^*_{s,1} & q^*_{s,2} & \cdots & q^*_{s,t}. \end{pmatrix} \tag{5.4}$$

For the sake of simplicity, we only consider positive criteria whose values need to be maximized (negative criteria can be easily transformed into positive attributes by multiplying -1 to their values).

Finally, a weight vector $W = (w_1, w_2, \ldots, w_m)$ is used to represent user's priorities on preferences given to different criteria with $w_k \in \mathbb{R}_0^+$ and $\sum_{k=1}^{m} w_k = 1$. The final QoS utilities vector $U = (u_1, u_2, \ldots)$ of Web service candidates are therefore can be computed as follows:

$$U = Q^* * W^T \tag{5.5}$$

in which u_i is the ith Web service QoS utility value within the range $[0, 1]$.

5.4.2 Similarity Computation

Web services provide reusable functionalities. The functionalities are described by the input and output parameters defined in WSDL file.

Now, we describe a similarity model for computing similarities between a user query and Web service operations. In this model, a vector (*Keywords*, *Input*, *Output*) is used to represent the functionality part of a user query as well as the functionality part of Web service operations. Particularly, the keywords of a Web service operation are abstracted from the descriptions in its associated WSDL file. Three phases are involved in the similarity search: WSDL preprocessing, clustering parameters, and similarity computation.

5.4.2.1 WSDL Preprocessing

In order to improve the accuracy of similarity computation for operations and user query in our approach, we first need to preprocess the WSDL files. There are two steps as follows:

1. Identify useful terms in WSDL files. Since the descriptions, operation names, and input/output parameters' names are made manually by the service provider, there

are a lot of misspelled and abbreviated words in real-world WSDL files. This step replaces such kind of words with normalized forms.
2. Perform word stemming and remove stop words. A stem is the basic part of the word that never changes even when morphologically infected. This process can eliminate the difference between inflectional morphemes. Stop words are those with little substantive meaning.

5.4.2.2 Similarity Computation

Now, we describe how to measure the similarities of Web service operations to a user's query. The functionality part of a user's query R_f consists of three elements $R_f = (r^k, r^{in}, r^{out})$. The *keywords* element is a vector $r^k = (r_1^k, r_2^k, \ldots, r_l^k)$, where r_i^k is the ith keyword. Moreover, the *input* element $r^{in} = (r_1^{in}, r_2^{in}, \ldots, r_m^{in})$ and the *output* element $r^{out} = (r_1^{out}, r_2^{out}, \ldots, r_n^{out})$, where r_i^{in} and r_i^{out} are the ith terms of input element and output element, respectively. A Web service operation also consists of three elements $OP_f = (K, In, Out)$. The *keywords* element of operation i is a vector of words $K^i = (k_1^i, k_2^i, \ldots, k_{l'}^i)$. The *input* and the *output* elements are vectors $In^i = (in_1^i, in_2^i, \ldots, in_{m'}^i)$ and $Out^i = (out_1^i, out_2^i, \ldots, out_{n'}^i)$, respectively. Thus, user queries and Web service operations are described as sets of terms. By applying the TF/IDF (Term Frequency/Inverse Document Frequency) measure [8] into these sets, we can measure the cosine similarity s_i between Web service operation i and a user's query.

Vector similarity (VS) measures the cosine of the angle between two corresponding vectors and sets it as the similarity of the two vectors. In similarity search for Web service, the two vectors measured are Web service operation and user query:

$$s_i = \frac{\sum_{i=1}^{t} r_i \cdot t_i}{\sqrt{\sum_{i=1}^{t} r_i^2} \cdot \sqrt{\sum_{i=1}^{t} t_i^2}}. \tag{5.6}$$

Pearson correlation coefficient (PCC), another popular similarity measurement approach, was introduced in a number of recommender systems for similarity computation, since it can be easily implemented and can achieve high accuracy. The similarity between an operation and a user's query can be calculated by employing PCC as follows:

$$s_i = \frac{\sum_{i=1}^{t} (r_i - \bar{r}) \cdot (t_i - \bar{t})}{\sqrt{\sum_{i=1}^{t} (r_i - \bar{r})^2} \cdot \sqrt{\sum_{i=1}^{t} (t_i - \bar{t})^2}} \tag{5.7}$$

where $\bar{r}$ is average TF/IDF value of all terms in a operation vector and $\bar{t}$ is average TF/IDF value of all terms in a user's query vector. The PCC similarity value s_i is in the interval of -1 and 1, and a larger value means indicate a higher similarity.

5.4.3 QoS-Aware Web Service Searching

With an increasing number of Web services being made available in the Internet, users are able to choose functionally appropriate Web services with high non-functional qualities in a much larger set of candidates than ever before. It is highly necessary to recommend to the user a list of service candidates which fulfill both the user's functional and non-functional requirements.

5.4.3.1 Utility Computation

A final rating score r_i is defined to evaluate the conformity of each Web service i to achieve the search goal.

$$r_i = \lambda \cdot \frac{1}{\log(p_{s_i} + 1)} + (1 - \lambda) \cdot \frac{1}{\log(p_{u_i} + 1)}, \tag{5.8}$$

where p_{s_i} is the functional rank position and p_{u_i} is the non-functional rank position of Web service i among all the service candidates. Since the absolute values of similarity and service quality indicate different features of Web service and include different units and range, rank positions rather than absolute values is a better choice to indicate the appropriateness of all candidates. $\frac{1}{\log(p+1)}$ calculates the appropriateness value of a candidate in position p for a query. $\lambda \in [0, 1]$ defines how much the functionality factor is more important than the non-functionality factor in the final recommendation.

λ can be a constant to allocate a fixed percentage of the two parts' contributions to the final rating score r_i. However, it is more realistic if λ is expressed as a function of p_{s_i}:

$$\lambda = f(p_{s_i}) \tag{5.9}$$

λ is smaller if the position in similarity rank is lower. This means a Web service is inappropriate if it cannot provide the required functionality to the users no matter how well it serves. The relationship between searching accuracy and the formula of λ will be identified to extend the search engine prototype in our future work.

5.4.3.2 Rank Aggregation

After receiving the users' query, the functional component of WSExpress computes the similarity s_i in Sect. 5.4.2 between search query R_f and operations of Web service i, while the non-functional component of WSExpress employs R_q to compute the QoS utility u_i in Sect. 5.4.1 of each Web service i.

Now our goal is to consider user's preferences on both functional and non-functional features and provide a rank list by combing evaluation results of the two

aspects of service candidates. Given the user's preference on functional and non-functional aspect, we can provide a personalized rank list by assigning each service candidate a certain score based on its positions in similarity ranking and QoS utility ranking. In other words, we aggregate the rankings of similarity and QoS utility according user defined preference.

We formally describe the optimal rank aggregation problem in the following. Given a set $S = \{s_1, s_2, \ldots\}$ of service candidates, an ranking list $l =< l(1), l(2), \ldots >$ is an permutation of all service candidates, where $l(i)$ denotes the service at position i of l. Given two ranking lists l_p, l_q of similarity and QoS utility, respectively, the optimal rank list l_o, which is an aggregation of l_p and l_q, should be recommended to users.

Given the similarity values or QoS utility scores of candidates, we assume that there is an uncertainty of ranking list l_p or l_q. In other words, any service $s_j \in S$ is assumed to be possible for ranked in the top position of l. But different services may have different likelihood values. Under this assumption, we define the top one probability of Web service s_j as follows:

$$P(s_j) = \frac{f(r_j)}{\sum_{k=1}^{m} f(r_k)}, \tag{5.10}$$

where $f(x)$ can be any monotonically increasing and strictly positive function, $P(s_j) > 0$ and $\sum P(s_j) = 1$. For simplicity, we take the exponential function for $f(x)$ [2]. Note that the top one probabilities $P(s_j)$ form a probability distribution over the set of services S. The top one probability indicates the probability of a service being ranked in the top position of a user's ranking list. By Eq. (5.10), a Web service with high similarity value or QoS utility value is assigned to a high probability value.

In order to estimate the quality of recommended Web service list, we need to define the distance between two ranking lists []. Ranking list distance evaluates the similarity of two lists. A distance value is smaller if more items are ordered in the similar positions. Given two ranking lists l_1 and l_2 over the Web service set S, the distance between l_1 and l_2 is defined by

$$d(l_1, l_2) = -\sum_{j=1}^{m} P(s_{1j})P(s_{2j}), \tag{5.11}$$

where s_1j is the service in the jth position of l_1 and s_2j is the service in the jth position of l_2.

We therefore define the Web service recommendation as the following optimization problem:

$$\min_{l_o} \mathcal{L}(l_p, l_q) = \lambda d(l_o, l_p) + (1 - \lambda) d(l_o, l_q), \tag{5.12}$$

where $d(l_o, l_p)$ is the distance between the optimal ranking list and the functionality ranking list, $d(l_o, l_q)$ is the distance between the optimal ranking list and the non-

functionality ranking list, and λ controls the trade-off between functionality and non-functionality.

Intuitively, Web services recommended by the final ranking list are functional comply with the users' requirements and with high QoS level. Our goal is to find a rank of all candidate in U that minimize the objective value function Eq. 5.12. One possible approach to solve the problem is check all the possible ranking lists in the solution space and select the optimal ranking which minimize the objective value function Eq. 5.12. The size of solution space is $O(n!)$ for n candidates. In fact, this is a NP-complete problem, which can be proved by transforming into a NP-complete problem of finding minimum cost perfect matching in the bipartite graph. Therefore, we propose a greedy algorithm to find a suboptimal solution as follows:

Algorithm 5: Greedy Rank Aggregation

Input: a candidate set S, two ranking lists l_p and l_q
Output: a optimal rank aggregation l_o^*

for *each service* s_j *in* S **do**
  $P_1(s_j) = \frac{f(u_j)}{\sum_{k=1}^{m} f(u_k)}$;
  $P_2(s_j) = \frac{f(sim_j)}{\sum_{k=1}^{m} f(sim_k)}$;
  $AP(s_j) = \lambda P_1 s_j + (1-\lambda) P_2(s_j)$;
end
Generate a ranking list l_o^* of all the service candidates according to their AP values;
$d_{l_o^*, l_p} = -\sum_{j=1}^{m} P(s_{pj}) P(s_{oj})$;
$d_{l_o^*, l_q} = -\sum_{j=1}^{m} P(s_{qj}) P(s_{oj})$;
$\mathcal{L}^*(l_p, l_q) = \lambda d(l_o^*, l_p) + (1-\lambda) d(l_o^*, l_q)$;
for *each candidate* s *in* S **do**
  change the position of s higher or lower;
  $d_{l_o, l_p} = -\sum_{j=1}^{m} P(s_{pj}) P(s_{oj})$;
  $d_{l_o, l_q} = -\sum_{j=1}^{m} P(s_{qj}) P(s_{oj})$;
  $\mathcal{L}(l_p, l_q) = \lambda d(l_o, l_p) + (1-\lambda) d(l_o, l_q)$;
  if $\mathcal{L}^*(l_p, l_q) < \mathcal{L}(l_p, l_q)$ **then**
    $l_o^* = l_o$;
  end
end
return l_o^*;

5.4.4 Online Ranking

In this section, we propose an online service recommendation algorithm. Since the QoS performance of Web services is dynamic at runtime, the ranking list should adopt the updated QoS information. Therefore, the Optimal Rank Algorithm is extended to integrate QoS information dynamically. In our ranking aggregation approach, a nice property is that before aggregation the functional utility and non-functional utility are calculated independently. For functional similarity search, the ranking list remains the same in different time intervals. The QoS ranking list is changing from time to time. Therefore, the optimal recommendation list should be adopted to the new QoS value accordingly. The online service recommendation algorithm is described as follows:

Algorithm 6: Online Service Recommendation

Input: a candidate set S, an optimal ranking list l_o, functional ranking lists l_p, a new QoS matrix Q

Output: a new optimal rank aggregation l_o^*

Conduct normalization on the new QoS matrix Q according to 5.2;
Compute the QoS utility vector U according to 5.5;
for *each service* s_j *in* S **do**
  $P_2(s_j) = \frac{f(sim_j)}{\sum_{k=1}^{m} f(sim_k)}$;
end
$l_o^* = l_o$;
for *each candidate* s *in* S **do**
  change the position of s higher or lower;
  $d_{l_o,l_p} = -\sum_{j=1}^{m} P(s_{pj})P(s_{oj})$;
  $d_{l_o,l_q} = -\sum_{j=1}^{m} P(s_{qj})P(s_{oj})$;
  $\mathcal{L}(l_p, l_q) = \lambda d(l_o, l_p) + (1-\lambda)d(l_o, l_q)$;
  if $\mathcal{L}^*(l_p, l_q) < \mathcal{L}(l_p, l_q)$ **then**
    $l_o^* = l_o$;
  end
end
return l_o^*;

5.4.5 Application Scenarios

5.4.5.1 Searching Styles

To attack the above problem, we propose a novel search engine which can provide the user with brand new searching styles. We define a user search query in the form of a vector $R = (R_f, R_q)$, which contains functionality part R_f and non-functionality

part R_q for representing the user's ideal Web service candidate. $R_q = (C, W)$ defines the user's non-functional requirements, where C and W set the user's constraints and preferences on QoS criteria separately as mentioned in Sect. 5.4.1. Our new searching procedure consists of three styles in the following discussion.

Keywords Specified In this searching style, the user only needs to simply enter the keywords vector r^k and QoS requirements R_q. The keywords should capture the main functionality the user requires in the search goal. In Table 5.1 as an example, since the user needs price information of cars, it is reasonable to specify "car" or "car, price" as the keywords vector.

Interface Specified In order to improve the searching efficiency, we design the "interface specified" searching style. In this style, the user specifies the expected functionality by setting the input vector r^{in} and/or output vector r^{out} as well as QoS requirements R_q. The input vector r^{in} represents the largest amount of information the user can provide to the expected Web service operation, while the output vector represents the least amount of information that should be returned after invoking the Web service operation.

Similar Operations For a more accurate and advanced Web service searching, we design the "similar operation" searching style by combining above two styles. This style is especially suitable in the following two situations. In the first situation, the user has already received a Web service recommendation list by performing one of the above searching styles. The user decides the Web service to explore in detail, checks the inputs and outputs of its operations, and even tries some of the operations. After carefully inspecting a Web service the user may find that this Web service is not suitable for the applications. However, the user does not want to repeat the time-consuming inspecting process for other service candidates. This style enables the user to find similar Web service operations by only modifying a small part of the previous query to exclude these inappropriate features. In the second situation, the user already integrates a Web service into the application for a particular functionality. However, due to some reason this Web service becomes unaccessible. Without requesting an extra query process, the search engine can automatically find other substitutions.

Now, we discuss in detail how the functional evaluation component operates in different scenarios.

- If only the keywords vector in the functionality part of the user query is defined, the similarity is computed in Sect. 5.4.2 using the keywords vector r^k of the query and the keywords vector K extracted from the descriptions, operation names, and parameter names.
- If the input and output vectors in the functionality part of the user query are defined, the input similarity and output similarity are computed in Sect. 5.4.2 using the input/output vector r^{in}/r^{out} of the query and the input/output vector In/Out of an operation. The functional similarity is a combination of input and output similarities.
- If the whole functionality part of a query is available. The functional similarity of an operation is a combination of the above two kinds of similarities, which is computed using R_f and OP_f.

5.5 Experiments

The aim of the experiments is to study the performance of our approach compared with other approaches (e.g., the one proposed by [7]). We conduct two experiments in Sects. 5.5.1 and 5.5.2, respectively. Firstly, we show that the Top-k Web services returned by our approach have much more QoS gain than other approaches. Secondly, we demonstrate that our approach can achieve highly relevant results as good as other similarity-based service searching approaches even there is no available QoS values.

5.5.1 QoS Recommendation Evaluation

In this section, we conduct a large-scale real-world experiment to study the QoS performance of the Top-k Web services returned by our searching approach.

To obtain real-world WSDL files, we developed a Web crawling engine to crawl WSDL files from different Web resources (e.g., UDDI, Web service portal, and Web service search engine). We obtain totally 3738 WSDL files from 69 countries. Totally, 15,811 operations are contained in these Web services. To measure the non-functional performance of these Web services, 339 distributed computers in 30 countries from PlanetLab are employed to monitor these Web services. The detailed non-functional performance of Web service invocations is recorded by these service users (distributed computer nodes). Totally, 1,267,182 QoS performance results are collected. Each invocation record is a k-dimensional vector representing the QoS values of k criteria. For simplicity, we use two matrices, which represent response-time and throughput QoS criteria, respectively, for experimental evaluation in this chapter. Without loss of generality, our approach can be easily extended to include more QoS criteria.

The statistics of Web service QoS dataset are summarized in Table 5.3. Response-time and throughput are within the range 0–20 s and 0–1000 kbps, respectively. The means of response-time and throughput are 0.910 s and 47.386 kbps, respectively. Figure 5.4 shows the distributions of response-time and throughput. Most of the response-time values are between 0.1 and 0.8 s, and most of the throughput values are between 5 and 40 kbps.

Table 5.3 Statistics of WS QoS dataset, ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

Statistics	Response-time	Throughput
Scale	0–20 s	0–1000 kbps
Mean	0.910 s	47.386 kbps
Num. of users	339	339
Num. of Web services	3738	3738
Num. of records	1,267,182	1,267,182

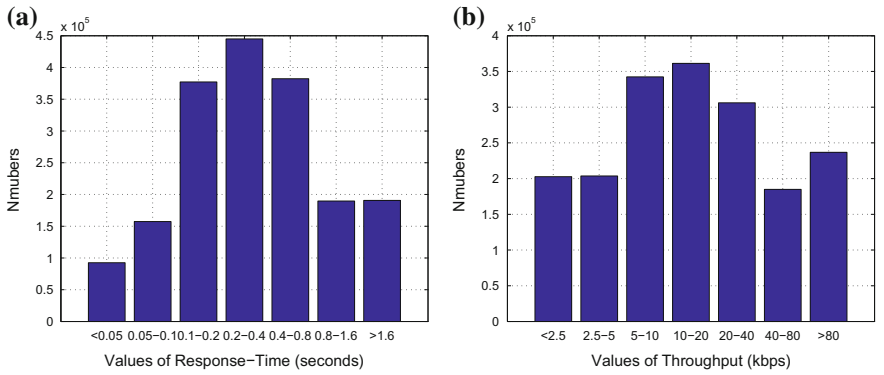


Fig. 5.4 Value distributions. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

In most of the searching scenarios, users tend to look at only the top items of the returned result list. The items in the higher position, especially the first position, is more important than the items in lower positions in the returned result list. To evaluate the qualities of Top-k returned results in a ranked list, we employ the Normalized Discounted Cumulative Gain (NDCG), a standard IR measure [4] approach as performance evaluation metric. Let $s_1, s_2, \ldots, s_p$ be a ranked list of Web services produced by a searching approach. Let u_i be the associated QoS utility value of Web service s_i, which ranked in position p_i. Discounted Cumulative Gain (DCG) and NDCG of at rank k are defined, respectively, as

$$DCG_k = u_i + \sum_{i=2}^{k} \frac{u_i}{\log_2 p_i}, \tag{5.13}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k} \tag{5.14}$$

where IDCG is the maximum possible gain value that is obtained with the optimal reorder of k Web services in the list $s_1, s_2, \ldots, s_p$. For example, consider the following QoS utility values which are ordered according to the position of associated Web services in a ranked Web service list:

$u = [0.3, 0.2, 0.3, 0, 0, 0.1, 0.2, 0.2, 0.3, 0]$

The perfect ranking would have QoS utility values of each rank of

$u = [0.3, 0.3, 0.3, 0.2, 0.2, 0.2, 0.1, 0, 0, 0]$

which would give ideal DCG utility values.

To study the performance of our approach, we compared our WSExpress Web service searching engine with the URBE [7], a keywords matching approach, employing our real-world dataset described above. Totally, 5 query domains are studied in this experiment. Each domain contains 4 user queries. Figure 5.5 shows the NDCG values of Top-k recommended Web services. The Top-k NDCG values of our WSExpress

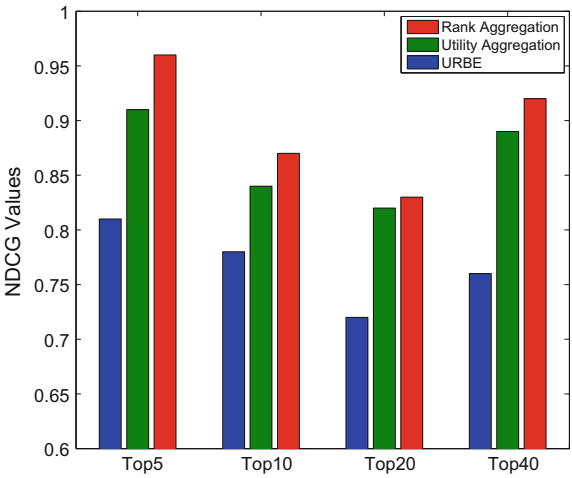


Fig. 5.5 NDCG of Top-K Web services, ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

engine are considerably higher than URBE (i.e., 0.767 of WSExpress compared with 0.200 of URBE for Top5 and 0.697 of WSExpress compared with 0.303 of URBE for Top10). This means that, given a query, our search engine can recommend high-quality Web services in the first positions.

Table 5.4 shows the NDCG values of Top-k recommended Web services in the five domains. In most of the queries, NDCG values of WSExpress are much higher than URBE. In some search scenarios such as query 2, the NDCG values of WSExpress and URBE for Top5 are identical, since in this particular case the most functional appropriate Web services have the most appropriate non-functional properties. In other words, these Top5 Web services have highest QoS utilities and similarity values. However, while more top Web services are considered, such as Top10, the NDCG values of WSExpress are becoming much higher than URBE.

5.5.2 Functional Matching Evaluation

In this experiment, we study the relevance of the recommended Web services to the user's query without considering non-functional performance of the Web services. By comparing our approach with URBE, we observe that the Top-k Web services in our recommendation list are highly relevant to the user's query even without any available QoS values.

The benchmark adopted for evaluating the performance of our approach is the OWLS service retrieval test collection OWLS-TC v2 [5]. This collection consists of more than 570 Web services and 1000 operations covering seven application domains (i.e., education, medical care, food, travel, communication, economy, and weaponry). The benchmark includes WSDL files of the Web services, 32 test queries, and a set

Table 5.4 NDCG values (A larger NDCG value means a better performance). ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

Domain	Query ID	Top5		Top10		Top20		Top40	
		URBE	WSExpress	URBE	WSExpress	URBE	WSExpress	URBE	WSExpress
Business	1	0.437	0.661	0.444	0.599	0.439	0.633	0.527	0.659
	2	0.653	0.653	0.668	0.721	0.657	0.666	0.634	0.645
	3	0.402	0.502	0.456	0.512	0.502	0.544	0.574	0.603
	4	0.200	0.767	0.303	0.697	0.399	0.667	0.496	0.699
Education	5	0.603	0.742	0.604	0.753	0.598	0.664	0.631	0.717
	6	0.621	0.732	0.571	0.715	0.574	0.675	0.598	0.696
	7	0.645	0.688	0.579	0.671	0.560	0.643	0.632	0.662
	8	0.509	0.642	0.562	0.642	0.575	0.633	0.600	0.672
Science	9	0.423	0.538	0.478	0.549	0.495	0.572	0.502	0.578
	10	0.573	0.731	0.525	0.717	0.546	0.693	0.602	0.702
	11	0.632	0.819	0.613	0.823	0.583	0.757	0.628	0.774
	12	0.622	0.754	0.593	0.728	0.582	0.681	0.597	0.734
Weather	13	0.214	0.574	0.245	0.551	0.243	0.559	0.259	0.581
	14	0.713	0.825	0.701	0.814	0.687	0.802	0.725	0.824
	15	0.431	0.581	0.346	0.566	0.465	0.566	0.530	0.606
	16	0.475	0.611	0.485	0.519	0.501	0.529	0.525	0.543
Media	17	0.409	0.516	0.419	0.485	0.403	0.496	0.589	0.530
	18	0.393	0.519	0.373	0.488	0.450	0.527	0.532	0.567
	19	0.544	0.740	0.554	0.683	0.512	0.642	0.551	0.683
	20	0.504	0.678	0.473	0.613	0.451	0.559	0.497	0.602

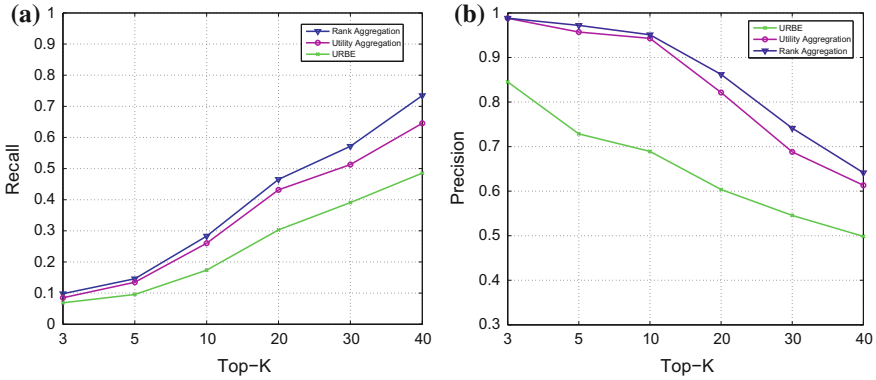


Fig. 5.6 Recall and precision performance. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

of relevant Web services associated with each of the queries. Since the QoS feature is not considered in this experiment, we set the QoS utility value of each Web service as 1.

Top-k recall ($Recall_k$) and *Top-k precision* ($Precision_k$) are adopted as metrics to evaluate the performance of different Web search approaches. $Recall_k$ and $Precision_k$ can be calculated by

$$Recall_k = \frac{|Rel \cap Ret_k|}{|Rel|}, \tag{5.15}$$

$$Precision_k = \frac{|Rel \cap Ret_k|}{|Ret_k|}, \tag{5.16}$$

where Rel is the relevant set of Web services for a query and Ret_k is a set of Top-k Web services search results.

Since user tends to check only top few Web services in common search scenario, an approach with high Top-k precision values is very practical in reality. Figure 5.6 shows the experimental results of our WSExpress approach and the URBE approach. In Fig. 5.6a, the Top-k recall values of WSExpress are higher than URBE. In Fig. 5.6b, the Top-k precision values of WSExpress are considerably higher than URBE, indicating that more relevant Web services are recommended in high positions by our approach.

5.5.3 *Online Recommendation*

In this chapter, we propose an online Web service recommendation approach. Different from the previous ranking approach, it adopts the real time QoS information

Table 5.5 Statistics of online QoS dataset. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

Statistics	Response-time	Throughput
Scale	0–20 s	0–1000 kbps
Mean	3.165 s	9.609 kbps
Num. of users	142	142
Num. of Web services	4532	4532
Num. of time intervals	64	64
Num. of records	30,287,611	30,287,611

to recommend Web services. In this section, we evaluate the performance of online recommendation approaches.

In this experiment, we deploy 142 distributed computers located in 22 countries from PlanetLab. Totally, 4532 publicly available real-world Web services from 57 countries are monitored by each computer continuously. In our experiment, each of the 142 computers sends null operation requests to all the 4532 Web services during every time interval. The experiment lasts for 16 hours with a time interval lasting for 15 min. By collecting invocation records from all the computers, finally we include 30,287,611 QoS performance results into the Web service QoS dataset. Each invocation record is a k dimension vector representing the QoS values of k criteria. We then extract a set of $142 \times 4532 \times 64$ user-service-time tensors, each of which stands for a particular QoS property, from the QoS invocation records. For simplicity, we employ two tensors, which represent response-time and throughput QoS criteria, respectively, for experimental evaluation in this chapter. Without loss of generality, our approach can be easily extended to include more QoS criteria.

The statistics of Web service QoS dataset are summarized in Table 5.5. Response-time and throughput are within the range of 0–20 s and 0–1000 kbps, respectively. The means of response-time and throughput are 3.165 s and 9.609 kbps, respectively. The distributions of the response-time and throughput values of the user-service-time tensors are shown in Fig. 5.7a, b respectively. Most of the response-time values are between 0.1 and 0.8 s and most of the throughput values are between 0.8 and 3.2 kbps.

The experimental results are shown in Fig. 5.8. Each time interval lasts for 15 minutes. The parameter setting is Top-K=5. From Fig. 5.8, we observe that in each time interval, online recommendation approach has a higher NDCG value than URBE, which means Web services with better QoS performance are recommended compared with URBE. Since URBE cannot adopt the dynamic QoS information for recommendation in time, the NDCG values of approach URBE decrease significantly as the time passed. After about 30 time intervals, the NDCG value is below 0.3 which means QoS performance of the recommended Web services has a high probability that cannot fulfill the users' non-functional requirements. In our online rank aggregation approach, we employ the latest QoS information of Web services

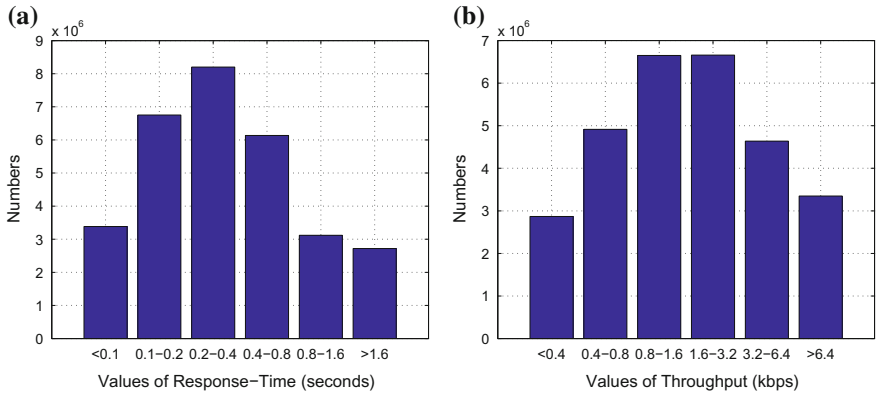


Fig. 5.7 QoS value distributions of online dataset. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

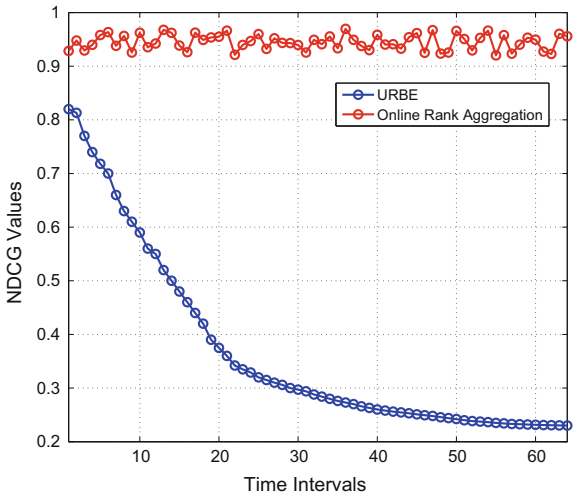


Fig. 5.8 NDCG of online recommendation. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

for recommendation. Therefore, the NDCG values are maintained in a high level, which indicates that we can always recommend appropriate Web services with high QoS performance to the users.

5.5.4 Impact of λ

In our method, the parameter λ controls the user's preference on functionality and non-functionality. A larger value of λ means functionality is preferred. In Fig. 5.9,

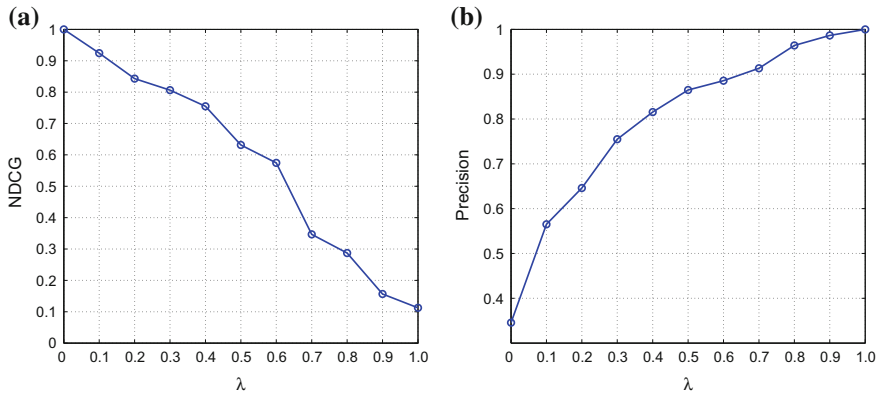


Fig. 5.9 Impact of λ. ©[2010] IEEE. Reprinted, with permission, from Ref. [11]

we study the impact of λ by varying the values of *lambda* from 0 to 1 with a step value of 0.1. Other parameter setting is Top-K = 10.

Figure 5.9a shows the NDCG values and Fig. 5.9b shows the Precision values. From Fig. 5.9a, we observe that λ impacts the NDCG performance significantly, which demonstrates that incorporating the QoS information greatly improves the non-functional quality of recommended Web services. In general, when the value of λ is increased from 0 to 1, the NDCG value is decreased. This observation indicates that if functionality is preferred, the QoS performance of recommended Web services is decreased. If $\lambda = 0$, we only employ the QoS information for Web service recommendation; therefore, the NDCG value is 1. If $\lambda = 1$, we only employ the functional similarity information for Web service recommendation; therefore, the NDCG value is very small. From Fig. 5.9b, we observe that λ also impacts the precision significantly, which demonstrates that incorporating the functional similarity information greatly improves the recommendation accuracy. In general, when the value of λ is increased from 0 to 1, the precision value is increased. This observation indicates that if functionality is preferred, the functional requirements can be fulfilled well. If $\lambda = 0$, we only employ the QoS information for Web service recommendation; therefore, the precision value is vary small. If $\lambda = 1$, we only employ the functional similarity information for Web service recommendation; therefore, the precision value is 1. In other cases, we fuse the information of QoS and functionality for Web service recommendation.

A proper value of λ is highly related to the preference of the user. The user defines the importance of functionality and non-functionality. A proper value of λ can be defined by analyzing the impact of λ on a small sample dataset.

5.6 Summary

In this chapter, we present a novel Web service search engine WSExpress to find the desired Web service. Both functional and non-functional characteristics of Web services are captured in our approach. We provide user three searching styles in the WSExpress to adapt different searching scenarios. A large-scale real-world experiment in distributed environment and an experiment on benchmark OWLS-TC v2 are conducted to study the performance of our search engine prototype. The results show that our approach outperforms related works.

In the future work, we will conduct data mining in our dataset to identify for which formulas of λ our search approach can achieve optimized performance. Clustering algorithms for similarity computation will be designed for improving functional accuracy of searching result. Finally, the non-functional evaluation component will be extended to dynamically collect quality information of Web services.

References

1. M. Alrifai, T. Risse, Combining global optimization with local selection for efficient QoS-aware service composition, in *Proceedings of International Conference on World Wide Web (WWW'09)* (2009), pp. 881–890
2. Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, H. Li, Learning to rank: from pairwise approach to listwise approach, in *Proceedings of 24th international conference on Machine learning* (2007), pp. 129–136
3. F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, S. Weerawarana, Unraveling the web services web: an introduction to soap, wsdl, and uddi. IEEE Internet Comput. **6**(2), 86–93 (2002)
4. K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)
5. M. Klusch, B. Fries, K. Sycara, Automated semantic web service discovery with owls-mx, in *Proceeding of the 5th International Conference on Autonomous Agents and Multiagent systems (AAMAS '06)* (2006), pp. 915–922
6. Y. Liu, A.H. Ngu, L.Z. Zeng, Qos computation and policing in dynamic web service selection, in *Proceedings of 13th International Conference on World Wide Web (WWW'04)* (2004), pp. 66–73
7. P. Plebani, B. Pernici, Urbe: web service retrieval based on similarity evaluation. IEEE Trans. Knowl. Data Eng. **21**(11), 1629–1642 (2009)
8. G. Salton, *The SMART Retrieval System—Experiments in Automatic Document Processing*. (Prentice-Hall Inc., 1971)
9. K. Verma, K. Sivashanmugam, A. Sheth, A. Patil, S. Oundhakar, J. Miller, Meteor-s wsdi: a scalable p2p infrastructure of registries for semantic publication and discovery of web services. Inf. Technol. Manage. **6**(1), 17–39 (2005)
10. T. Yu, Y. Zhang, K. Lin, Efficient algorithms for web services selection with end-to-end QoS constraints. ACM Trans. Web (TWEB) **1**(1), 6 (2007)
11. Y. Zhang, Z. Zheng, M. Lyu, WSExpress: a QoS-aware search engine for web services, in *Proceedings of ICWS'10* (2010), pp. 91–98
12. Y. Zhang, Z. Zheng, M.R. Lyu, Exploring latent features for memory-based qos prediction in cloud computing, in *IEEE Symposium on Reliable Distributed Systems (SRDS)* (IEEE 2011), pp. 1–10

Chapter 6
QoS-Aware Byzantine Fault Tolerance

Abstract Cloud computing is becoming a popular and important solution for building highly reliable applications on distributed resources. However, it is a critical challenge to guarantee the system reliability of applications especially in voluntary-resource cloud due to the highly dynamic environment. In this chapter, we present Byzantine fault-tolerant cloud (BFTCloud), a Byzantine fault tolerance framework for building robust systems in voluntary-resource cloud environments. BFTCloud guarantees robustness of systems when up to f of totally $3f + 1$ resource providers are faulty, including crash faults and arbitrary behaviors faults. BFTCloud is evaluated in a large-scale real-world experiment which consists of 257 voluntary-resource providers located in 26 countries. The experimental results show that BFTCloud guarantees high reliability of systems built on the top of voluntary-resource cloud infrastructure and ensure good performance of these systems.

6.1 Overview

Currently, most of the clouds are deployed on two kinds of infrastructures. One is well-provisioned and well-managed infrastructure [8] managed by a large cloud provider (e.g., Amazon, Google, Microsoft, and IBM). The other one is voluntary-resource infrastructure which consists of numerous user-contributed computing resources [2]. With the exponential growth of cloud computing [1, 3] as a solution for providing flexible computing resource on demand [4], more and more cloud applications emerge in recent years. How to build high-reliable cloud applications, which are usually large-scale and very complex, becomes an urgent and crucial research problem.

Typically, cloud applications consist of a number of cloud modules. The reliability of cloud applications is greatly influenced by the reliability of cloud modules. Therefore, building high-reliable cloud modules becomes the premise of developing high-reliable cloud applications. Traditionally, testing schemes [7] are conducted on the software systems of cloud modules to make sure that the reliability threshold has been achieved before releasing the software. However, reliability of a cloud module not only relies on the system itself, but also heavily depends on the node it has

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_6

deployed and the unpredictable Internet. Traditional testing has limited improvement on the reliability of a cloud module under voluntary-resource cloud infrastructure due to:

- Computing resources, denoted as nodes in the cloud, are frangible. Different from the powerful and performance-guaranteed nodes managed by large cloud providers, user-contributed nodes are usually highly dynamic, much cheaper, less powerful, and less reliable. The reliability of a cloud module deployed on these nodes is mainly determined by the robustness of nodes rather than the software implementation.
- Communication links between modules are not reliable. Unlike nodes in well-provisioned cloud infrastructure, which are connected by high-speed cables, nodes in voluntary-resource cloud infrastructure are usually connected by unpredictable communication links. Communication faults, such as time out, will greatly influence the reliability of cloud applications.

Based on the above analysis, in order to build reliable cloud applications on the voluntary-resource cloud infrastructure, it is extremely urgent to design a fault tolerance mechanism for handling different faults. Typically, the reliability of cloud applications is effected by several types of faults, including node faults like crashing, network faults like disconnection, and Byzantine faults [6] like malicious behaviors (i.e., sending inconsistent response to a request [9]). The user-contributed nodes, which are usually cheap and small, make malicious behaviors increasingly common in voluntary-resource cloud infrastructure. However, traditional fault tolerance strategies cannot tolerate malicious behaviors of nodes.

To address this critical challenge, we propose a novel approach, called Byzantine fault-tolerant cloud (BFTCloud), for tolerating different types of failures in voluntary-resource clouds. BFTCloud uses replication techniques for overcoming failures since a broad pool of nodes are available in the cloud. Moreover, due to the different geographical locations, operating systems, network environments, and software implementation among nodes, most of the failures happened in voluntary-resource cloud are independent of each other, which is the premise of Byzantine fault tolerance mechanism. BFTCloud can tolerate different types of failures including the malicious behaviors of nodes. By making up a BFT group of one primary and $3f$ replicas, BFTCloud can guarantee the robustness of systems when up to f nodes are faulty at runtime. The experimental results show that compared with other fault tolerance approaches, BFTCloud guarantees high reliability of systems built on the top of voluntary-resource cloud infrastructure and ensures good performance of these systems.

In summary, this chapter makes the following contributions:

1. We identify the Byzantine fault tolerance problem in voluntary-resource cloud and propose a Byzantine fault tolerance framework, named BFTCloud, for guaranteeing the robustness of cloud application. BFTCloud uses dynamical replication techniques to tolerate various types of faults including Byzantine faults. We consider BFTCloud as the first Byzantine fault-tolerant framework in cloud computing literature.

2. We have implemented the BFTCloud system and test it on a voluntary-resource cloud, Planet-lab, which consists of 257 user-contributed computing resources distributed in 26 countries. The prototype implementation indicates that BFTCloud can be easily integrated into cloud nodes as a middleware.
3. We conduct large-scale real-world experiments to study the performance of BFTCloud on reliability improvement compared with other approaches. The experimental results show the effectiveness of BFTCloud on tolerating various types of faults in cloud.

The rest of this chapter is organized as follows: Sect. 6.2 describes the system architecture of BFTCloud. Section 6.3 presents our BFTCloud fault tolerate mechanism in detail. Section 6.4 introduces the experimental results. Section 6.5 concludes the chapter.

6.2 System Architecture

We begin by using a motivating example to show the research problem in this chapter. As shown in Fig. 6.1, cloud applications usually consist of a number of modules. These modules are deployed on distributed cloud nodes and connected with each other through communication links. Each module is supposed to finish a certain task (e.g., product selection, bill payment, and shipping addresses confirming) for a cloud application (e.g., shopping agency). A cloud module will form a sequence of requests (e.g., browsing products and choosing products) for the task (e.g., product selection) and send the requests to a group of nodes in the voluntary-resource cloud for execution.

Figure 6.2 shows the system architecture of BFTCloud in voluntary-resource cloud environment. Under the voluntary-resource cloud infrastructure, end-users contribute a larger number of computing resources which can be provided to cloud

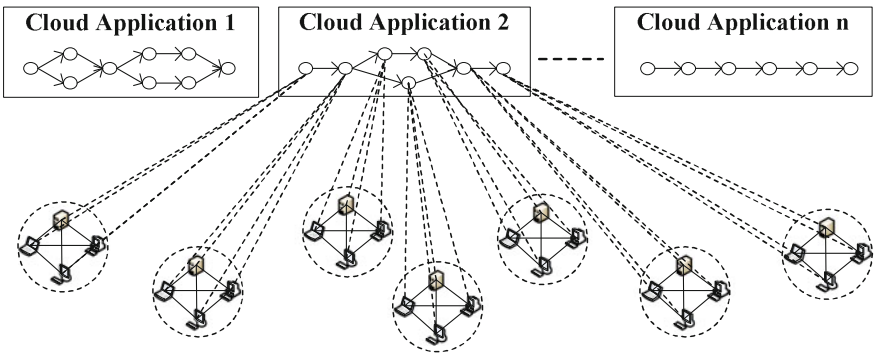


Fig. 6.1 Architecture of cloud applications. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

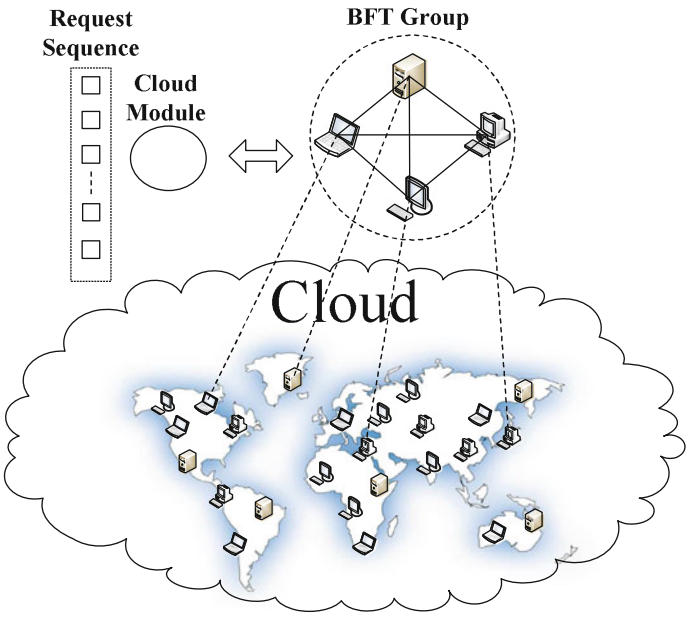


Fig. 6.2 Architecture of BFTCloud in voluntary-resource cloud. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

modules for request execution. Typically, computing resources in the voluntary-resource cloud are heterogeneous and less reliable, and malicious behaviors of resource providers cannot be prevented. Byzantine faults could be very common in a user-contributed cloud environment. In order to guarantee the robustness of the module, the replication technique is employed for request execution upon the user-contributed nodes. After a cloud module generated a sequence of requests, it first needs to choose a BFT group from the pool of cloud nodes for request execution. Since cloud nodes are located in different geographic locations with heterogeneous network environments, and the failure probabilities of nodes are diverse, a monitor is implemented on the cloud module side as a middleware for monitoring the QoS performance and failure probability of nodes. By considering the QoS performance and failure probability, the cloud module first chooses a node as primary and sends the current request to the primary. After that, a set of replicas are selected according to their failure probability and QoS performance to both the cloud module and the primary. The primary and replicas form a BFT group for executing requests from the cloud module. After the BFT group returns responses to the current request, the cloud module will judge whether the responses can be committed. Then, the cloud module will send the next request or resend the current request to the BFT group. If some nodes of the BFT group are identified as faulty, the cloud module will update the BFT group to guarantee the system reliability. The detailed approach will be presented in Sect. 6.3.

6.3 System Design

In this section, we present BFTCloud, a practical framework for building robust systems with Byzantine fault tolerance under voluntary-resource cloud infrastructure. We first give an overview on the work procedures of BFTCloud in Sect. 6.3.1. Then, we describe the five phases of BFTCloud in Sect. 6.3.2 to Sect. 6.3.6, respectively.

6.3.1 System Overview

Figure 6.3 shows the work procedures of BFTCloud. The input of BFTCloud is a sequence of requests with specified QoS requirements (e.g., preferences on price, capability, bandwidth, workload, response latency, and failure probability) sent by the cloud module. The output of BFTCloud is a sequence of committed responses corresponding to the requests. BFTCloud consists of five phases described as follows:

1. **Primary Selection**: After accepting a request from the cloud module, a node is selected from the cloud as the primary. The primary is selected by applying the primary selection algorithm with respect to the QoS requirements of the request.
2. **Replica Selection**: In this phase, a set of nodes are selected as replicas by applying a replica selection algorithm with respect to the QoS requirements of the request. The primary then forwards the request to all replicas for execution. The selected replicas together with the primary make up a BFT group.
3. **Request Execution**: In this phase, all members in the BFT group execute the request locally and send back their responses to the cloud module. After collecting responses from the BFT group within a certain period of time, the cloud module will judge the consistency of responses. If the BFT group respond consistently, the current request will be committed and the cloud module will send the next request. If the BFT group responds inconsistently, the cloud module will trigger a fault tolerance procedure to tolerate up to f faulty nodes and trigger the primary updating procedure and/or replica updating procedure to update the

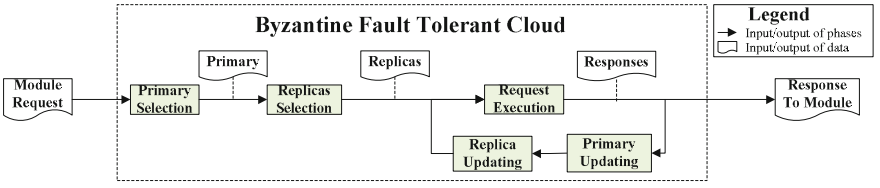


Fig. 6.3 Work procedures of BFTCloud. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

group members. If more than f nodes are identified as faulty, the cloud module will resend the request to the refresh BFT group and enter into the request execution phase again.
4. **Primary Updating**: In this phase, faulty primary in the BFT group will be identified and replaced by a newly selected primary.
5. **Replica Updating**: In this phase, faulty replicas in the BFT group will be identified and updated according to the information obtained from the request execution phase. The replica updating algorithm will be applied to replace the faulty replicas with other suitable nodes in the cloud.

6.3.2 Primary Selection

Under the voluntary-resource cloud infrastructure, a cloud module will send the request directly to a node which it believes to be the primary. Therefore, the primary plays an important role in a BFT group. The responsibilities of primary include accepting requests from the cloud module, selecting appropriate replicas to form a BFT group, forwarding the request to all replicas, and replacing faulty replicas with newly selected nodes. Since failures happened on primary will greatly decrease the overall performance of a BFT group, the requirements on primary attributes (e.g., capability, bandwidth, and workload) are more strict than those on replicas. In order to select an optimal primary, we propose a primary selection algorithm.

We model the primary selection problem under voluntary-resource cloud infrastructure as follows:

> Let N be the set of nodes available in the cloud and Q be the set of m dimension vectors. For each node n_i in N, there is a $q_i = (q_{i1}, q_{i2}, \cdots, q_{im})$ in Q representing the QoS values of m criteria. Given a priority vector $W = (w_1, w_2, \cdots, w_m)$ on the m QoS criteria, the optimal primary should be selected from the set N.

Note that $w_k \in \mathbb{R}^+$ and $\sum_{k=1}^{m} w_k = 1$. Typically, the QoS values of can be integers from a given range (e.g., 0, 1, 2, 3 or real numbers of a close interval (e.g., $[-20, 20]$). Without loss of generality, we can map a QoS value to the interval $[0, 1]$ using the function $f(x) = (x - q_{min})/(q_{max} - q_{min})$, where q_{max} and q_{min} are the maximum and minimum QoS values of the corresponding criterion, respectively.

The proposed primary selection algorithm is shown in Algorithm 7. After accepting the priority vector from the cloud module, a rating value r_i is computed for each node $n_i \in N$ as follows:

$$r_i = \sum_{k=1}^{m} q_{ik} \times w_k, \tag{6.1}$$

Algorithm 7: Primary Selection Algorithm

Input: N, Q, W
Output: n^*

```
n* = null;
r_max = 0;
for all n_i ∈ N do
    r_i = Σ_{k=1}^{m} q_ik × w_k;
    if r_i > r_max then
        n* = n_i;
        r_max = r_i;
    end
end
return n*;
```

where r_i fall into the interval [0, 1]. The cloud module will choose the node n^*, which has the highest rating value, as the primary:

$$n^* = \arg\max_{n_i \in N} r_i. \tag{6.2}$$

6.3.3 Replica Selection

After the primary is selected in Sect. 6.3.2, a set of replicas should be chosen to form a BFT group. Since replicas in a BFT group need to communicate with both the primary and the cloud module, the QoS performance of a node should be considered from both the cloud module perspective and the primary perspective. Let q_i be the QoS vector of node n_i observed by the cloud module and q_i' be the QoS vector of node n_i observed by the primary. Then, the combined QoS vector q_i'' is calculated by a set of transformation rules as follows:

- minimum: $q_{ik}'' = min(q_{ik}, q_{ik}')$, for QoS criterion like bandwidth.
- average: $q_{ik}'' = avg(q_{ik}, q_{ik}')$, for QoS criterion like response time.
- equality: $q_{ik}'' = q_{ik} = q_{ik}'$, for QoS criterion like price.

Without loss of generality, the rule set can be easily extended to include more rules for calculating complex QoS criterion values.

Given the combined QoS vector q_i'', we can evaluate how appropriate the node n_i is as a replica of the BFT group. A score s_i is assigned to each node $n_i \in N$ as follows:

$$s_i = \sum_{k=1}^{m} q_{ik}'' \times w_k, \tag{6.3}$$

where s_i falls into the interval [0, 1]. After ordering the scores, we can select the nodes ranked in high positions as replicas of the BFT group.

In order to decide the replication degree, we first consider the failure probability of a BFT group in its entirety. Since the BFTCloud guarantees the execution correctness when up to f nodes are faulty, a BFT group is faulty if and only if more than f nodes are faulty. We define the failure probability of a BFT group σ as follows:

$$P_\sigma = P(|F| > f), \tag{6.4}$$

where F is the set of failure nodes in σ.

In order to reduce the cost of request execution, the replication degree f should be as small as possible, and the failure probability of a BFT group σ should be guaranteed under a certain threshold at the same time. Let P_0 be the threshold of P_σ defined by the cloud module. The replication degree decision problem can be formulated as an optimization problem:

$$\begin{aligned} \min_f \quad & f = \frac{|\sigma| - 1}{3}, \\ & P_\sigma = \sum_{F \in \Omega} \prod_{n_i \in F} P_i \prod_{n_j \in \sigma \setminus F} (1 - P_j), \\ & P_\sigma < P_0, \\ & \Omega = \{F | f < |F|\}. \end{aligned} \tag{6.5}$$

where P_i is the failure probability of node n_i, and Ω is the set of events that more than f nodes of the BFT group σ are fault. Note that a solution to this problem decides the replication degree and the replicas of BFT group σ at the same time. We summarize the replica selection algorithm in Algorithm 8.

6.3.4 Request Execution

After the BFT group members are determined, requests can be sent to the BFT group for execution. The cloud module first forms a request sequence and sends the sequence of requests to the primary. The primary will order the requests and forward the ordered requests to all the BFT group members. Each member of the BFT group will execute the sequence of requests and send the corresponding responses back to the cloud module. The cloud module collects all the received responses from the BFT group members and makes a judgement on the consistence of responses. A action strategy will be chose according to the consistence of responses as follows:

- **Case 1**: The cloud module receives $3f + 1$ consistent responses from the BFT group. In this case, the cloud module will commit the current request since there is no fault happens in the current BFT group.

Algorithm 8: Replica Selection Algorithm

Input: N, Q, Q', W, P_0
Output: σ

$\sigma = null$;
$f = 0$;
$P_\sigma = P^*$;
for *all* $n_i \in N$ **do**
  $q_i'' \leftarrow (q_i, q_i')$ by applying the set of transformation rules;
  $s_i = \sum_{k=1}^{m} q_{ik}'' \times w_k$;
end
Generate a permutation $< n_1', n_2', \cdots >$ of the set N such that $s_1' \geq s_2' \geq \cdots$;
while $P_\sigma > P_0$ **do**
  $f = f + 1$;
  $\sigma = \{n_1', n_2', \cdots, n_{3f}'\}$;
  $P_\sigma = 0$;
  for *all* $F \in \Omega$ **do**
    $P_\sigma = P_\sigma + \prod_{n_i \in F} P_i \prod_{n_j \in \sigma \setminus F} (1 - P_j)$;
  end
end
return σ;

- **Case 2**: The cloud module receives between $2f + 1$ to $3f$ consist responses. In this case, the cloud module can still commit the current request since there are less than $f + 1$ faults happened. To commit the current request and identify the faulty nodes, the cloud module assembles a commit certificate and sends the commit certificate to all the BFT group members. Each member will acknowledge the cloud module with a local-commit message once it receives the commit certificate from the cloud module. If more than $2f + 1$ local-commit messages are received, the cloud module will commit the current request and invoke a replica updating procedure to replace all the faulty BFT group members with new members. If less than $2f + 1$ local-commit messages are received, the cloud module will resend the commit certificate until it receives local-commit messages from more than $2f + 1$ members.
- **Case 3**: The cloud module receives less than $2f + 1$ response messages. In this case, either the primary is faulty or more than $f + 1$ replicas are faulty. The cloud module will then resend the current request again but to all members this time. Each replica forwards the request to the node it believes to be the primary. If the replica receives a request from the primary within a given time and the proposed sequence number is consistency with that sent by the cloud module, the replica will execute the request and send response to the cloud module. If the replica does not receive an ordered request from the primary within a given time, or the request sequence number is not consistent with the request sent by

the cloud module, the primary must be faulty. The replica will send a primary election proposal to all replicas to trigger a primary updating procedure.

- **Case 4**: The cloud module receives more than $2f + 1$ responses, but fewer than $f + 1$ responses are consistency. This indicates inconsistent ordering of requests by the primary. The cloud module will send a proof of misbehavior of primary to all the replicas and trigger a primary updating procedure.

6.3.5 Primary Updating

When the primary is faulty, primary updating procedures will be triggered in the request execution phase. The procedures of primary updating phase are as follows:

1. A replica which suspects the primary to be faulty sends an primary election proposal to all the other replicas. However, it still participates in the current BFT group as it may be only a network problem between the replica and the primary. Other replicas, once receiving a primary election proposal, just store it since the primary election proposal could arrive from a faulty replica as well.
2. If a replica receives $f + 1$ primary election proposals, it indicates that the primary is really faulty. It will send a primary selection request to the cloud module. The cloud module then will start the primary selection phase and return a new primary which is one of the current replicas. The replica then sends a primary updating message to all the other replicas, which includes the new primary name and $f + 1$ primary election proposals. Other replicas which receive such primary updating message again confirm that at least $f + 1$ replicas sent a primary election proposal, and then resend the primary updating message together with the proof to the new primary.
3. If the newly selected primary receives $2f + 1$ primary updating messages, it sends a new BFT group setup message to all the replicas, which again includes all the primary updating messages as proof.
4. A replica which received and confirmed the new BFT group setup message will send out a BFT group confirm message to all replicas.
5. If a replica receives $2f + 1$ BFT group confirm messages, it starts performing as a member in the new BFT group.

6.3.6 Replica Updating

In the voluntary-resource cloud environment, nodes are highly dynamic and fragile. Different types of faults (e.g., response time out, unavailable, and arbitrary behavior) may happen to the nodes after a period of time. Under voluntary-resource cloud infrastructure, the failure probability of a BFT group increases sharply as the fraction of faulty nodes increases. The failure probability of a BFT group under the condition

that a set of replicas are already faulty is:

$$\begin{aligned} P_\sigma &= P(|F| > f|F^*) \\ &= P(|F \setminus F^*| > f - f^*), \end{aligned} \tag{6.6}$$

where F^* is the set of replicas which are faulty already.

To ensure the failure probability of a BFT group below a certain threshold, we need to replace the members once they are identified to be faulty. Moreover, due to the highly dynamic voluntary-resource cloud environment, the QoS performance of nodes are changed rapidly. Updating replicas timely could keep the performance of a BFT group stable.

Let T be the set of new nodes which will be added to the current BFT group. F^* is the set of nodes which will be removed from the current BFT group. Let σ' be the new BFT group with updated replicas. We have $\sigma' = \sigma \setminus F^* \cup T$, where T consists of nodes which are in the top $|T|$ positions ordered by score in Eq. (6.3).

The new BFT group σ', which can tolerate up to f' nodes failure, should satisfy $P_{\sigma'} > P_0$. Therefore, the replica updating problem is reduced to a replication degree decision problem, which can be further formalized as an optimization problem as follows:

$$\begin{aligned} &\min_{f'} f' = \frac{|\sigma'| - 1}{3}, \\ &P_{\sigma'} = \sum_{F' \in \Lambda} \prod_{n_i \in F'} P_i \prod_{n_j \in \sigma' \setminus F'} (1 - P_j), \\ &P_{\sigma'} < P_0, \\ &\Lambda = \{F' | f < |F'|\}. \end{aligned} \tag{6.7}$$

where Λ is the set of events that more than f' nodes of the BFT group σ' are fault. We summarize the replica updating algorithm in Algorithm 9.

6.4 Experiments

In this section, in order to study the performance improvements of our proposed approach, we conduct several experiments to compare our BFTCloud with several other fault tolerance approaches.

In the following, Sect. 6.4.1 describes the system implementation of BFTCloud and the experimental settings, and Sect. 6.4.2 compares the performances of BFT-Cloud with some other fault tolerance methods.

6.4.1 Experimental Setup

We have implemented our BFTCloud approach by Java language and deployed it as a middleware in a voluntary-resource cloud environment. The cloud infrastructure is constructed by 257 distributed computers located in 26 countries from Planet-lab, which is a distributed test bed consisting of hundreds of computers all over the world. Each computer, named as node in the cloud infrastructure, can participate several BFT groups as a primary or replica simultaneously.

Algorithm 9: Replica Updating Algorithm

Input: N, Q, Q', W, P_0, σ, F^*
Output: σ'

$\sigma' = \sigma \setminus F^*$;
$T = null$;
$f' = \lceil \frac{3f - |F^*|}{3} \rceil$;
$P'_\sigma = P^*$;
for *all* $n_i \in N \setminus \sigma$ **do**
  $q''_i \leftarrow (q_i, q'_i)$ by applying the set of transformation rules;
  $s_i = \sum_{k=1}^{m} q''_{ik} \times w_k$;
end
Generate a permutation $< n'_1, n'_2, \cdots >$ of the set $N \setminus \sigma$ such that $s'_1 \geq s'_2 \geq \cdots$;
$T = \{n'_1, n'_2, \cdots, n'_{3f' - |\sigma'|}\}$;
$\sigma' = \sigma' \cup T$;
while $P'_\sigma > P_0$ **do**
  $f' = f' + 1$;
  $T = \{n'_1, n'_2, \cdots, n'_{3f' - |\sigma'|}\}$;
  $\sigma' = \sigma' \cup T$;
  $P'_\sigma = 0$;
  for *all* $F \in \Lambda$ **do**
    $P'_\sigma = P'_\sigma + \prod_{n_i \in F'} P_i \prod_{n_j \in \sigma' \setminus F'} (1 - P_j)$;
  end
end
return σ';

Based on the voluntary-resource cloud infrastructure, we conduct large-scale experiments to study the performance improvements of BFTCloud compared with other approaches. In our experiments, each node in the cloud is configured with a random malicious failure probability, which denotes the probability of arbitrary behavior happens in the node. Note that the failure probability of a node observed by other nodes is not necessarily equal to the malicious failure probability since other types of faults (e.g., node crashing and disconnection) may also occur. Each

Table 6.1 Average sending times per request. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

Benchmark (KB)	BFTCloud	BFTRandom	Zyzzyva	NoFT
0/0	1.3428	1.7096	2.9167	1.0725
4/0	1.3035	1.7248	3.1002	1.1042
0/4	1.3820	1.7340	3.2058	1.3055

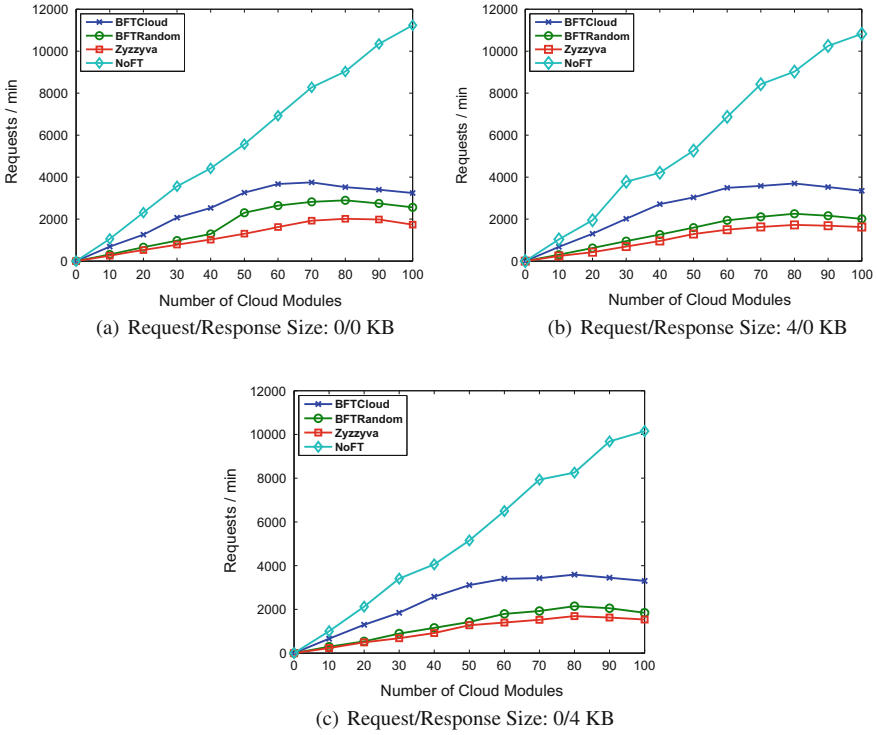
node keeps the QoS information of all the other nodes and updates the information periodically. For simplicity, we use response time for QoS evaluation in this chapter. Without loss of generality, our approach can be easily extended to include more QoS criteria. We also employed 100 computers from Planet-lab to perform as cloud modules.

6.4.2 Performance Comparison

In this section, we compare the performance of our approach BFTCloud with other fault tolerance approaches in the voluntary-resource cloud environment. We have implemented four approaches:

- NoFT: No fault tolerance strategy is employed for task execution in the voluntary-resource cloud.
- Zyzzyva: A state-of-the-art Byzantine fault tolerance approach proposed in [5]. The cloud module sends requests to a fixed primary and a group of replicas. There is no mechanism designed for adopting the dynamic voluntary-resource cloud environment.
- BFTCloud: The Byzantine fault tolerance framework proposed in this chapter. The cloud module employs Algorithm 1–3 to mask faults and adopt the highly dynamic voluntary-resource environment.
- BFTRandom: The framework is the same with BFTCloud. However, this approach just randomly selects nodes in primary selection, replica selection, primary updating, and replica updating phases.

In Fig. 6.4, we compare the throughput of all approaches in terms of different number of cloud modules by executing null operations. We change the number of cloud module from 0 to 100 with a step value of 10. The requests are sent by a variable number of cloud modules in each experiment (0–100). We conducts experiments on three benchmarks [5] with different request and response size. The sizes of request/response messages are 0/0, 4/0, and 0/4 KB in Fig. 6.4a, b, and c, respectively. The parameter settings in this experiment are $P_0 = 0.5$ and $timeout = 500$ ms, where $timeout$ defines the maximum waiting time for a message. From Fig. 6.4, we can observe that our approach BFTCloud can commit more requests per minute than Zyzzyva and BFTRandom under different sizes of request/response messages. The



(a) Request/Response Size: 0/0 KB

(b) Request/Response Size: 4/0 KB

(c) Request/Response Size: 0/4 KB

Fig. 6.4 Throughput comparison for 0/0, 4/0, and 0/4 benchmarks as the number of cloud modules varies. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

reason is that BFTCloud always chooses nodes with low failure probabilities as BFT group members. Therefore, the high reliability of BFT group guarantees that in most cases, a request can be committed without being resent. Note that NoFT achieves the highest throughput among all approaches since NoFT employs no fault tolerance mechanism. Every request will be committed once the cloud module received a reply. However, NoFT cannot guarantee the correctness of committed requests, which will be discussed in Table 6.2. Table 6.1 shows the average sending times of a request by the cloud module before it is committed. A request can be committed with much fewer sending times in BFTCloud than request in Zyzzyva, since BFT group members in BFTCloud are carefully selected and the probability of successfully executing a request is higher than that in Zyzzyva. Moreover, BFTCloud always chooses nodes with good QoS performance as BFT group members which makes requests and responses are transmitted more quickly than other approaches. In general, BFTCloud achieves high throughput of committed requests which demonstrates that the idea of considering failure probability and QoS performance when selecting nodes is realistic and reasonable.

Table 6.2 Correct rate of committed requests. ©[2011] IEEE. Reprinted, with permission, from Ref. [10]

size (KB)	BFTCloud	BFTRandom	Zyzzyva	NoFT
0/0	0.9855	0.9468	0.8726	0.2589
4/0	0.9840	0.9259	0.8925	0.2107
0/4	0.9794	0.9278	0.8621	0.2216

In Table 6.2, we evaluate the correctness of committed requests of different approaches. The experimental result shows that among all the committed requests, the percentage of correctly committed requests is highest in BFTCloud. This is because BFTCloud can guarantee a low probability P_0 that more than f BFT group members are faulty. While Zyzzyva cannot guarantee the failure probability of BFT group since the primary and replicas in Zyzzyva are fixed. Most of the requests are not correctly committed in NoFT despite high throughput of NoFT, since no fault tolerance mechanism is employed.

6.5 Summary

In this chapter, we propose BFTCloud, a Byzantine fault tolerance framework for building reliable systems in voluntary-resource cloud infrastructure. In BFTCloud, replication techniques are employed for improving the system reliability of cloud applications. To adapt to the highly dynamic voluntary-resource cloud environment, BFTCloud select voluntary nodes based on their QoS characteristics and reliability performance. Faulty voluntary resources will be replaced with other suitable resources once they are identified. The extensive experimental results show the effectiveness of our approach BFTCloud on guaranteeing the system reliability in cloud environment.

In the future, we will conduct more experimental analysis on open-source cloud applications and conduct more investigations on different QoS properties of voluntary resources. We will conduct more experiments to study the impact of parameters and investigate the optimal values of parameters in different experimental settings.

References

1. M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., A view of cloud computing. Commun. ACM **53**(4), 50–58 (2010)
2. A. Chandra, J. Weissman, Nebulas: using distributed voluntary resources to build clouds, in *Proceedings of HOTCLOUD'09* (2009)
3. M. Creeger, Cloud computing: an overview. ACM Queue **7**(5), 1–5 (2009)

4. T. Henzinger, A. Singh, V. Singh, T. Wies, D. Zufferey, FlexPRICE: flexible provisioning of resources in a cloud environment, in *Proceedings of CLOUD'10* (2010), pp. 83–90
5. R. Kotla, L. Alvisi, M. Dahlin, A. Clement, E. Wong, Zyzzyva: speculative byzantine fault tolerance, in *Proceedings of SOSP'07* (2007), pp. 45–58,
6. L. Lamport, R. Shostak, M. Pease, The Byzantine generals problem. ACM Trans. Program. Lang. Syst. (TOPLAS) **4**(3), 382–401 (1982)
7. M. Lyu et al. *Handbook of Software Reliability Engineering*. (1996)
8. L. Tang, J. Dong, Y. Zhao, L. Zhang, Enterprise cloud service architecture, in *Proceedings of CLOUD'10* (2010), pp. 27–34
9. Wikipedia, http://en.wikipedia.org/wiki/byzantine_fault_tolerance
10. Y. Zhang, Z. Zheng, M.R. Lyu, Bftcloud: a byzantine fault tolerance framework for voluntary-resource cloud computing, in *IEEE International Conference on Cloud Computing (CLOUD)* (IEEE 2011), pp. 444–451

Chapter 7
Conclusion and Discussion

Abstract This chapter concludes this book and discusses the future work.

7.1 Conclusion

This book is aiming at advancing quality engineering in cloud and service computing. This book consists of three parts: The first part deals with service QoS prediction, the second part focuses on QoS-aware Web service searching, and the third part concentrates on QoS-aware fault-tolerant systems in cloud computing.

In the first part, we present three QoS prediction approaches for services. We first propose a neighborhood-based collaborative QoS prediction approach, which is enhanced by character modeling, for services. The second method is a model-based time-aware collaborative filtering approach, which utilizes time information to capture the periodicity features of service QoS values. Finally, we propose an online QoS prediction approach, which employs time series analysis to adapt to the highly dynamic service computing environment. The online prediction approach consists of an offline evolutionary algorithm and an online incremental algorithm for precisely predicting the QoS values of services at runtime. The experimental results and the system-level case study show the efficiency and effectiveness of our approach.

In the second part, we propose a QoS-aware Web service search engine. In order to provide better searching results to users for fulfilling their Web service requirements, we systematically fuse the functional approach and non-functional approach to achieve better performance. Moreover, we conduct experiments on the real-world Web services. The collected WSDL files and QoS datasets are released for the Web service research community.

In the third part, we conduct a fault tolerance study on cloud applications. By taking the advantage of multiple functional equivalent services over the Internet, we design a Byzantine fault tolerance framework to build robust systems in voluntary-resource cloud environments. Our fault tolerance framework employs dynamic QoS information of services to select the most suitable services for system integration. The experimental results show the effectiveness of this framework.

Y. Zhang and M.R. Lyu, *QoS Prediction in Cloud and Service Computing*,
SpringerBriefs in Computer Science, DOI 10.1007/978-981-10-5278-1_7

In general, the goal of our work is to predict and utilize the QoS information in cloud and service computing as accurate and effective as possible. Our released QoS datasets enable the extensive research of other researchers.

7.2 Discussion

There are several research directions that can be conducted in the future.

For the service QoS prediction, we plan to conduct more research on the correlation of multiple QoS characteristics since the different QoS characteristics are considered independently in current stage. The relationship between different QoS properties may provide some useful information for improving the prediction accuracy. Another direction worthy of investigation is how to explore the relationship between user information and service information to enhance the prediction accuracy.

For the QoS-aware Web service searching, we plan to design a clustering algorithm, which improves the accuracy of functional similarity computation. Currently, we only use the average QoS performance of Web services. However, due to the dynamic network environment and service status, we plan to extend the non-functional evaluation module to adopt dynamic QoS information of Web services.

For the QoS-aware fault tolerance framework in cloud computing, we can conduct more studies on the correlation of different types of failure, since failures may not be independent of each other. Moreover, failures of different services in the cloud application may have correlation with each other.