A Study of Different QoS Management Techniques in Cloud Computing

Mandeep Devgan, Kanwalvir Singh Dhindsa

Abstract*— Cloud Services are becoming a major system for constructing distributed systems. Service-oriented architecture (SOA) is widely working in electronic business, electronic -government, automotive systems, multimedia services, process control, finance, and a lot of other domains. Quality-of-Service (QoS) is usually employed for describing the non-functional characteristics of Cloud services and employed as an important differentiating point of different Cloud services. With the prevalence of Cloud services on the Internet, Cloud service QoS management is becoming more and more important. This paper first study a distributed QoS evaluation technique for Cloud services. In this technique, users in different geographic locations collaborative with each other to evaluate the target Cloud services and share their observed Cloud service QoS information. Based on this Cloud service evaluation technique, several large-scale distributed evaluations are conducted on many real-world Cloud services and the detailed evaluation results are released for future research. Cloud service evaluation is time and resource consuming. Moreover, in some scenarios, Cloud service evaluation may not be possible (e.g., the Cloud service invocation is charged, too many service candidate, etc.). Therefore, Cloud service QoS prediction approaches are becoming more and more attractive. In order to prediction the Cloud service QoS as accurate as possible, this paper studies three prediction methods. The first prediction method employs the information of neighborhoods for making missing value prediction. The second method discusses matrix factorization techniques to enhance the prediction accuracy. The third method predicts the ranking of the target Cloud services instead of QoS values. The predicted Cloud service QoS values can be employed to build fault-tolerant service-oriented systems. In the area of service computing, the cost for developing multiple redundant components is greatly reduced, since the functionally equivalent Cloud services are provided by different organizations and are accessible via Internet. Hence, based on the predicted QoS values, this paper study two methods for building fault tolerance Cloud services. Firstly, this paper studies an adaptive fault tolerance strategy for Cloud services. Then, this paper presents an optimal fault tolerance strategy selection technique for Cloud services.

***Index Terms*—QoS, Evaluation, Prediction, Active User, Ranking.**

I. INTRODUCTION

Cloud services are self-contained and self-describing computational Cloud components designed to support User-to-Service interaction by programmatic Cloud method calls [10]. Cloud services are becoming a major technique for building loosely-coupled distributed systems. Examples of service-oriented systems span a variety of diversified application domains, such as e-commerce, automotive systems [9], multimedia services [8], etc.

As shown in Fig.1, in the service-oriented environment, complex distributed systems are dynamically composed by discovering and integrating distributed Cloud services, which are provided by different organizations. The distributed Cloud services are usually employed by more than one service users (i.e., the service-oriented systems). The performance of the service oriented systems is highly relying on the performance of the employed Cloud services. Quality-of-Service (QoS) is usually engaged for describing the non-functional characteristics of Cloud services. QoS management of Cloud services refers to the activities in QoS specification, evaluation, prediction, aggregation, and control of resources to meet end-to-end user and application requirements. With the prevalence of Cloud services on the Internet, investigating Cloud service QoS is becoming more difficult and in recent years, a number of QoS-aware approaches have been comprehensively studied for Cloud services. However, there is still a lack of real-world Cloud service QoS datasets for validating new QoS-driven techniques and models. Without convincing and sufficient real-world Cloud service QoS datasets, characteristics of real-world Cloud service QoS cannot be fully mined and the performance of various recently developed QoS-based approaches cannot be justified. To collect sufficient Cloud service QoS data, evaluations from different geographic locations under various network conditions are usually required. However, it is not an easy task to conduct large-scale distributed Cloud service evaluations in reality. Effective and efficient Cloud service distributed evaluation mechanism is consequently required. The Cloud service evaluation approaches attempt to obtain the Cloud service QoS values by monitoring the target Cloud service. However, in some scenarios, a comprehensive Cloud service evaluation may not be possible (e.g., when the Cloud service invocation is charged; there are too many service candidates, etc.). Therefore, Cloud service QoS prediction approaches, which require no additional real-world Cloud service invocations, are becoming more and more attractive. Cloud service QoS prediction aims at making personalized QoS value prediction for the service users by employing the partially available information (e.g., QoS information of other users, characteristics of the current user, historical QoS performance of the target Cloud services, etc.). To predict the Cloud service QoS values as accurate as possible, comprehensive investigations on the prediction approaches are needed. Employing the evaluated/predicted Cloud service QoS values, QoS-aware fault-tolerant service-oriented systems can be built using redundant Cloud services in the Internet. Due to the cost of developing redundant components, traditional software fault tolerance is usually employed only for critical systems. In the area of service-oriented computing, however, the cost for developing multiple redundant components is greatly reduced, since the functionally equivalent Cloud services are provided by different organizations and are accessible via Internet.

Manuscript Received on July, 2013.

Mandeep Devgan, Department of Information Technology, Chandigarh Engineering College, Landran, Mohali, India.

Kanwalvir Singh Dhindsa, Department of CSE/IT, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib.

These Cloud services can be employed as alternative components for building fault-tolerant service-oriented systems. Although a number of fault tolerance strategies [22] have been developed for Cloud services, the highly dynamic Internet environment requires smarter and more adaptive fault tolerance strategies. Dynamic selection and reconfiguration of the optimal fault tolerance strategy becomes a necessity in service computing. Based on the above analysis, in order to improve QoS management of Cloud services, this paper need to provide efficient Cloud service QoS evaluation mechanisms, accurate Cloud service QoS prediction approaches, and robust QoS-aware fault tolerance strategies for Cloud services. This paper studies six approaches to attack these challenging research problems.

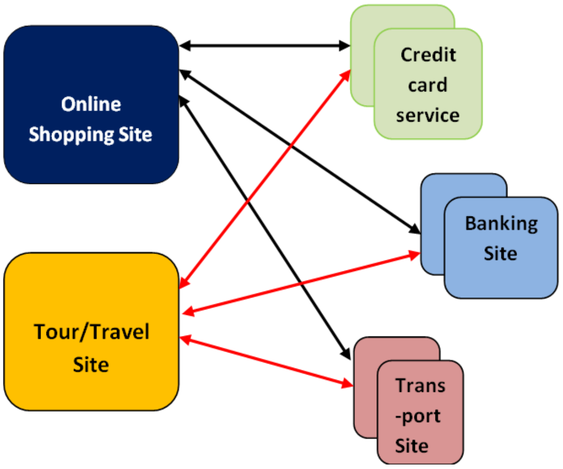

Fig.1. Example of Service Oriented System

II. QOS EVALUATION OF CLOUD SERVICES

In the field of service computing [16], Cloud service QoS have been discussed in a number of research investigations for presenting the non-functional characteristics of the Cloud services [16]. Cardellini et al. [15] employ five generic QoS properties (i.e. execution price, execution duration, reliability, availability, and reputation) for dynamic Cloud service composition. Ardagna et al. [4] use five QoS properties (i.e., execution time, availability, price, reputation, and data quality) when making adaptive service composition in flexible processes. Alrifai et al. [2] study an efficient service composition approach by considering both generic QoS properties and domain-specific QoS properties. QoS measurement of Cloud services has been used in the Service Level Agreement (SLA) [17], such as IBMs WSLA technique and the work from HP [7]. In SLA, the QoS data are mainly for the service providers to maintain a certain level of service to their clients and the QoS data are not available to others. This paper mainly focus on encouraging the service users to share their individually-obtained QoS data of the Cloud services, making efficient and effective Cloud service evaluation and selection.

A. *System Architecture*

Since the service providers may not deliver the QoS they declared and some QoS properties (e.g., *response-time* and *failure probability*) are highly related to the locations and network conditions of service users, Cloud service evaluation can be performed at the client-side to obtain more accurate QoS performance [7]. However, several challenges have to be solved when conducting Cloud service evaluation at the client-side: (1) It is difficult for the service users to make professional evaluation on the Cloud services themselves, since the service users are usually not experts on the Cloud service evaluation, which includes WSDL file analysis, test case generation, evaluation mechanism implementation, test result interpretation and so on; (2) It is time-consuming and resource-consuming for the service users to conduct a long-duration evaluation on many Cloud service candidates themselves; and (3) The common time-to-market constraints limit an in-depth and accurate evaluation of the target Cloud services. To address these challenges, this paper studies a distributed evaluation technique for Cloud services, together with its prototyping system [12], as shown in Fig. 2. This technique employs the concept of *user-collaboration*, which has contributed to the recent success of Bit Torrent [10] and Wikipedia (www.wikipedia.org). In this technique, users in different geographic locations share their observed QoS performance of Cloud services by contributing them to a centralized server. Historical evaluation results saved in a data center are available for other service users. In this way, QoS performance of Cloud services becomes easy to be obtained for the service users. As shown in Fig. 3, the developed distributed evaluation technique includes a centralized server with a number of distributed clients. The overall procedures can be explained as follows.

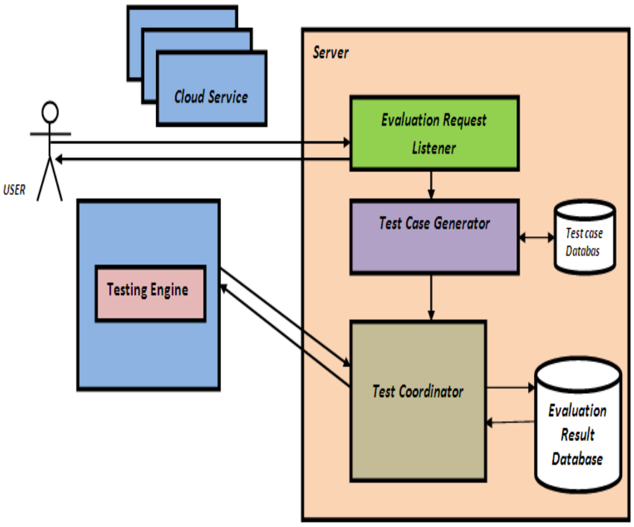

Fig. 2. Distribute Evaluation Technique

III. QOS PREDICTION OF CLOUD SERVICES

Collaborative filtering methods are widely adopted in recommender systems [12]. There types of collaborative filtering approaches are widely studied: neighborhood-based (memory based), model-based, and ranking-based. The most analyzed examples of memory-based collaborative filtering include user-based approaches [20], item-based approaches, and their fusion [18]. User-based approaches predict the ratings of active users based on the ratings of their similar users, and item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active users. User-based and item-based approaches often use the PCC algorithm [15] and the VSS algorithm [11] as the similarity computation methods. PCC-based collaborative filtering generally can achieve higher performance than VSS, since it considers the differences in the user rating style.

Wang et al. combined user-based and item-based collaborative filtering approaches for movie recommendation. In the model-based collaborative filtering approaches, training datasets are used to train a predefined model. Examples of model-based approaches include the clustering model , aspect models [18] and the latent factor model [14]. Kohrs *Part 2. Background Review* 14 and Merialdo [14] present an algorithm for collaborative filtering based on hierarchical clustering, which tries to balance robustness and accuracy of predictions, especially when few data are available. Hofmann [17] study s an algorithm based on a generalization of probabilistic latent semantic analysis to continuous valued response variables. Recently, several matrix factorization methods [19] have been developed for collaborative filtering. These methods focus on fitting the user-item matrix with low-rank approximations, which is engaged to make further predictions. The premise behind a low-dimensional factor model is that there is only a small number of factors influencing the values in the user-item matrix, and that a user's factor vector is determined by how each factor applies to that user. The neighborhood-based methods utilize the values of similar users or items (local information) for making value prediction, while model-based methods, like matrix factorization models, employ all the value information of the matrix (global information) for making value prediction. The neighborhood-based and model-based collaborative filtering approaches usually try to predict the missing values in the user-item matrix as accurately as possible. However, in the ranking-oriented scenarios, accurate missing value prediction may not lead to accuracy ranking. Therefore, ranking-oriented collaborative filtering approaches are becoming more and more attractive. Liu et al. study a ranking-oriented collaborative filtering approach to rank movies. Yang et al. [10] study another ranking-oriented approach for ranking books in digital libraries. There is limited work in the literature employing collaborative filtering methods for Cloud service QoS value prediction. One of the most important reasons that obstruct the research is that there is no large-scale real-world Cloud service QoS datasets available for studying the prediction accuracy. Without convincing and sufficient real-world Cloud service QoS data, the characteristics of Cloud service QoS information cannot be fully mined and the performance of the developed algorithms cannot be justified. A few approaches [18] mention the idea of applying neighborhood-based collaborative filtering methods for Cloud service QoS value prediction. However, these approaches simply employ a movie rating dataset, i.e., Movie Lens [5], for experimental studies, which is not convincing enough. Shao et al. [16] study a user-based PCC method for the Cloud service QoS value prediction. However, only 20 Cloud services are studied in this paper. This paper studies various approaches to address the problem of Cloud service QoS prediction, including neighborhood-based [11], model-based [11], and ranking based approaches [19].

IV. FAULT TOLERANT CLOUD SERVICE

Software fault tolerance is widely employed for building reliable stand-alone systems as well as distributed system [13]. The major software fault tolerance techniques includes recovery block [17], N-Version Programming (NVP) [6], N self-checking programming [4], distributed recovery block [13], and so on. In the area of service-oriented computing, the cost of developing redundant components are greatly reduced, since the functionally equivalent Cloud services can be employed for building diversity-based fault-tolerant service-oriented systems [15]. A number of service fault tolerance strategies have been developed in the recent literature [13]. The major fault tolerance strategies for Cloud services can be divided into passive strategies and active strategies. Passive strategies have been discussed in FT-SOAP [12], FT-CORBA [18], and in work. Active strategies have been investigated in FT Cloud, The ma, WS-Replication, SWS, and Perpetual. Work employs a rigorous development process to build reliable connector, which is a critical component. The connector is implemented as a Cloud service using the original WSDL description of the Cloud service replicas. Within the connector, lots of fault tolerance strategies can be implemented (e.g., active or passive replication strategies). FT Cloud study is a *WS Dispatcher* to make parallel Cloud service invocations and to return the final result to the users. Work study s a survivable Cloud Service technique named SWS. In SWS, each Cloud services are replicated and deployed onto a set of nodes to form a Cloud service group. All the replicas are invoked to process the same user request independently. Value faults can thus be tolerated by majority voting. Moreover, SWS supports continuous operation in the presence of Byzantine faults. Ye et al. study a middleware, PWSS, to support client transparent active replication strategy. When a client sends a request *r*, *r* is first sent to a PWSS. The PWSS then multicasts *r* to all the other PWSSs. After agreeing a total order on threads execution, all the replicas process the client's request and return the response a PWSS which first receive the client. This PWSS then return a result to the client's invocation after running a voting strategy on all the responses it received. Thema is a Byzantine Fault Tolerant (BFT) middleware for Cloud services which supports three-tiered application model. 3*f* +1 Cloud service replicas in the server-side need to invoke an external Cloud service for accomplishing their executions. Different from these previous works, this paper, will study an adaptive fault tolerance strategy for Cloud services, and study a QoS-aware selection technique for fault tolerant Cloud services.

V. QOS PREDICTION OF CLOUD SERVICES: NEIGHBORHOOD BASED

With the number increasing of Cloud services, Quality-of-Service (QoS) is usually employed for describing non-functional characteristics of Cloud services. Among different QoS properties of Cloud services, some QoS properties are user-dependent and have different values for different users (e.g., *response time*, *in- vocation failure probability*, etc.). Obtaining values of the user dependent QoS properties is a challenging task. Real-world Cloud service evaluation in the client-side [21] is usually required for measuring performance of the user-dependent QoS properties of Cloud services. Client-side Cloud service evaluation requires real-world Cloud service invocations and encounters the following drawbacks:

- Firstly, real-world Cloud service invocations impose costs for the service users and consume resources of the service providers. Some Cloud service invocations may even be charged.

- Secondly, there may exist too many Cloud service candidates to be evaluated and some suitable Cloud services may not be discovered and included in the evaluation list by the service users.
- Finally, most service users are not experts on Cloud service evaluation and the common time-to-market constraints limit an in-depth evaluation of the target Cloud services.

However, without sufficient client-side evaluation, accurate values of the user-dependent QoS properties cannot be obtained. Optimal Cloud service selection and recommendation are thus difficult to achieve. To attack this critical challenge, this paper study a neighborhood-based collaborative filtering approach for making personalized QoS value prediction for the service users. Collaborative filtering is the method which automatically predicts values of the current user by collecting information from other similar users or items. Well-known neighborhood-based collaborative filtering methods include user-based approaches [11] and item-based approaches. Due to their great successes in modeling characteristics of users and items, collaborative filtering techniques have been widely employed in famous commercial systems, such as Amazon1, Ebay2, etc. This paper systematically combines the user-based approach and item-based approach for predicting the QoS values for the current user by employing historical Cloud service QoS data from other similar users and similar Cloud services. Similar service users are defined as the service users who have similar historical QoS experience on the same set of commonly-invoked Cloud services with the current user. Different from traditional Cloud service evaluation approaches, this approach predicts user-dependent QoS values of the target Cloud services without requiring real-world Cloud service invocations. The Cloud service QoS values obtained by this approach can be employed by other QoS driven approaches (e.g., Cloud service selection , fault-tolerant Cloud service, etc.).

A. User Collaborative QoS Collection

To make accurate QoS value prediction of Cloud services without real-world Cloud service invocations, this paper need to collect past Cloud service QoS information from other service users. However, it is difficult to collect Cloud service QoS information from different service users due to: (1) Cloud services are distributed over the Internet and are hosted by different organizations. (2) Service users are usually isolated from each other. (3) The current Cloud service architecture does not provide any mechanism for the Cloud service QoS information sharing. Inspired by the recent success of *YouTube*3 and *Wikipedia* (4) this paper studies the concept of *user-collaboration* for the Cloud service QoS information sharing between service users. The idea is that, instead of contributing videos (*YouTube*) or knowledge (*Wikipedia*), the service users are encouraged to contribute their individually observed past Cloud service QoS data. Fig. 3 shows the procedures of this user-collaborative QoS data collection mechanism, which are introduced as follows: 1. A service user contributes past Cloud service QoS data to a centralized server Web service recorder. In the following of this

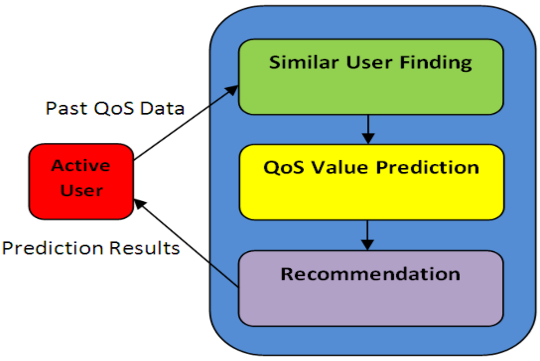

Fig.3. Procedure of QoS Prediction

Web service recorder selects similar users from the training users for the active user. *Training users* represent the service users whose QoS values are stored in the Web service recorder server and employed for making value predictions for the active users. 3. Web service recorder predicts QoS values of Cloud services for the active user. 4. Web service recorder makes Cloud service recommendation based on the predicted QoS values of different Cloud services. 5. The service user receives the predicted QoS values as well as the recommendation results, which can be employed to assist decision making (e.g., service selection, composite service performance prediction, etc.).

VI. QOS PREDICTION OF CLOUD SERVICES: MODEL BASED

The neighborhood-based QoS prediction approach has several drawbacks, including (1) the computation complexity is too high, and (2) it is not easy to find similar users/items when the user-item matrix is very sparse. To address these drawbacks, this paper study a neighborhood-integrated matrix factorization (NIMF) approach for Cloud service QoS value prediction in this part. The idea is that client-side Cloud service QoS values of a service user can be predicted by taking advantage of the social wisdom of service users, i.e., the past Cloud service usage experiences of other service users. By the collaboration of different service users, the QoS values of a Cloud service can be effectively predicted in this approach even the current user did not conduct any evaluation on the Cloud service and has no idea on its internal design and implementation details. In this part, firstly, this paper study a neighborhood-integrated matrix factorization (NIMF) approach for personalized Cloud service QoS value prediction. This approach explores the social wisdom of service users by systematically fusing the neighborhood based and the model-based collaborative filtering approaches to achieve higher prediction accuracy compared with the neighborhood based prediction approach. Secondly, this paper studies real-world Cloud service QoS dataset for future research. To the best of this knowledge, the scale of this released Cloud service QoS dataset is the largest in the field of service computing. Based on this dataset, extensive experimental investigations can be conducted to study the QoS value prediction accuracy of this approach.

VII. QOS PREDICTION OF CLOUD SERVICES: RANKING BASED

The neighborhood-based and model-based QoS prediction approaches aim at predicting the Cloud service QoS values for different service users. These predicting approaches are also named rating-based approaches. The predicted QoS values can be employed to rank the target Cloud services. In some cases (e.g., Cloud service search, Cloud service ranking), the users only need the quality ranking of the target Cloud services instead of the detailed QoS values. Ranking-based QoS prediction approaches aim at predicting the quality ranking of the target Cloud services instead of the detailed QoS values. The major challenge for making QoS-driven Cloud service quality ranking is that the Cloud service quality ranking of a user cannot be transferred directly to another user, since the user locations are quite different. Personalized Cloud service quality ranking is therefore required for different service users. The most straightforward approach of personalized Cloud service ranking is to evaluate all the Cloud services at the user-side and rank the Cloud services based on the observed QoS performance. However, this approach is impractical in reality, since conducting Cloud services evaluation is time consuming and resource consuming. Moreover, it is difficult for the service users to evaluate all the Cloud services themselves, since there may exist a huge number of Cloud services in the Internet.

VIII. CONCLUSIONS

The paper consists of three parts: the first part studies cloud service QoS evaluation, the second part focuses on cloud service QoS prediction, and the third part concentrates on QoS-aware fault-tolerant cloud services. All of the approaches studied in this paper are aiming at improving QoS management of cloud services. In the first part, we discussed a distributed QoS evaluation mechanism for cloud services. In order to speed up cloud service evaluation, the service users are encouraged to collaborate with each other and share their individually obtained evaluation results. Employing this evaluation mechanism, several real-world cloud service evaluations are conducted. The obtained cloud service QoS values are released as archival research datasets for other researchers. In the second part, we studied three QoS prediction approaches for cloud services. We first combine the user-based and item-based collaborative filtering approaches to achieve higher prediction accuracy. After that, a neighborhood integrated model based approach is discussed. The results show that this model-based approach provides higher prediction accuracy than neighborhood-based approaches.

REFERENCES

1. E. Al-Masri and Q. H. Mahmoud. Investigating cloud services on the World Wide Web. In *Pro. 17th Int'l Conf. World Wide cloud (WWW'08)*, pages 795–804, 2008.
2. M. Alrifai and T. Risse. Combining global optimization with local selection for efficient qos-aware service composition. In *Proc. 18th Int'l Conf. World Wide cloud (WWW'09)*, pages 881–890, 2009.
3. Apache. Axis2. In *http://ws.apache.org/axis2*, 2008.
4. D. Ardagna and B. Pernici. Adaptive service composition in flexible processes. *IEEE Trans. Software Engeering*, 33(6):369–384, 2007.
5. K. J. arvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
6. A. Avizienis. The methodology of n-version programming.*Software Fault Tolerance, M. R. Lyu (ed.), Wiley, Chichester*, pages 23–46, 1995.
7. B. Benatallah, M. Dumas, Q. Z. Sheng, and A. H. H. Ngu. Declarative composition and peer-to-peer provisioning of dynamic cloud services. In *Proc. 18th Int'l Conf. Data Eng. (ICDE'02)*, 2002.
8. A. S. Bilgin and M. P. Singh. A daml-based repository for qos-aware semantic cloud service selection. In *Proc. 2nd Int'l Conf. cloud Services (ICWS'04)*, pages 368–375, 2004.
9. P. A. Bonatti and P. Festa. On optimal service selection. In *Proc. 14th Int'l Conf. World Wide cloud (WWW'05)*, pages 530–538, 2005.
10. C. Bram. Incentives build robustness in bittorrent. In *Proc. First Workshop Economics of Peer-to-Peer Systems*, pages 1–5, 2003.
11. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. 14th Annual Conf. Uncertainty in Arti_cial Intelli- gence (UAI'98)*, pages 43–52, 1998.
12. R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
13. C.-L.Hwang and K.Yoon. Multiple criteria decision making. *Lecture Notes in Economics and Mathematical Sys- tems*, 1981.
14. J. Canny. Collaborative filtering with privacy via factor analysis. In *Proc. 25th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SI- GIR'02)*, pages 238–245, 2002.
15. V. Cardellini, E. Casalicchio, V. Grassi, F. Lo Presti, and R. Mirandola. Qos-driven runtime adaptation of service oriented architectures. In *Proc. 7th Joint Meet- ing European Software Engineering Conf. and ACM SIGSOFT Symp. Foundations of Software Engineering*
16. *(ESEC/FSE'09)*, pages 131–140, 2009.
V. Cardellini, E. Casalicchio, V. Grassi, and F. L. Presti. Flow-based service selection for cloud service composition supporting multiple qos classes. In *Proc. 5th Int'l Conf. cloud Services (ICWS'07)*, pages 743–750, 2007.
17. J. Cardoso, J. Miller, A. Sheth, and J. Arnold. Modeling quality of service for workflows and cloud service processes. *Journal of cloud Semantics*, 1:281–308, 2002.
18. P. P. Chan, M. R. Lyu, and M. Malek. Reliable cloud services: Methodology, experiment and modeling. In *Proc. 5th Int'l Conf. cloud Services (ICWS'07)*, pages 679–686,2007.
19. P. P.-W. Chan, M. R. Lyu, and M. Malek. Making services fault tolerant. In *Proc. 3rd Int'l Service Avail. Symp. (ISAS'06)*, pages 43–61, 2006.
20. X. Chen, X. Liu, Z. Huang, and H. Sun. Regionknn: A scalable hybrid collaborative filtering algorithm for personalized cloud service recommendation. In *Proc. 8th Int'l Conf. cloud Services (ICWS'10)*, pages 9–16, 2010.
21. X. Chen and M. R. Lyu. Message logging and recovery in wireless corba using access bridge. In *The 6th Int'l Symp. Autonomous Decentralized Systems*, pages 107–114, 2003.
22. *R. C. Cheung. A user-oriented software reliability model. IEEE Trans. Software Engeering, 6(2):118–125, 1980. B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman. Planetlab: An overlay testbed for broad-coverage services. ACM SIGCOMM Computer Communication Review, 33(3):3–12, July 2003.*