Context-aware Cloud Service Selection based on Comparison and Aggregation of User Subjective Assessment and Objective Performance Assessment

Lie Qu, Yan Wang and Mehmet A. Orgun
Macquarie University
Sydney, NSW 2109, Australia
{*lie.qu, yan.wang, mehmet.orgun*}*@mq.edu.au*

Ling Liu
Georgia Institute of Technology
Atlanta, Georgia 30332, USA
ling.liu@cc.gatech.edu

Athman Bouguettaya
RMIT University
Melbourne, VIC 3001, Australia
athman.bouguettaya@rmit.edu.au

***Abstract*—Due to the diversity and dynamics of cloud services, it is usually hard for potential cloud consumers to select the most suitable cloud service. In prior studies, cloud service selection is usually based on either objective performance assessment or cloud users' subjective assessment (e.g., subjective ratings). However, either assessment way has its limitation in reflecting the quality of cloud services. This causes a problem that some vital performance aspects which concern potential cloud consumers are not taken into account in cloud service selection.**

This paper proposes a novel context-aware cloud service selection model based on the comparison and aggregation of subjective assessment extracted from cloud user feedback and objective assessment from quantitative performance testing. In this model, objective assessment provided by some professional testing parties is used as a benchmark to filter out potentially biased subjective assessment from cloud users, then objective assessment and subjective assessment are aggregated to evaluate the overall performance of cloud services according to potential cloud users' personalized requests. Moreover, our model takes the contexts of objective assessment and subjective assessment into account. By calculating the similarity between different contexts, the benchmark level of objective assessment is dynamically adjusted according to context similarity, which makes the following comparison and aggregation process more accurate and effective. After aggregation, the final results can quantitatively reflect the overall quality of cloud services. Finally, our proposed model is evaluated through the experiments executed in different conditions.

***Keywords*-Cloud service selection; Subjective or objective assessment; Context similarity;**

I. Introduction

Many individuals and organizations have started to consume cloud services in their daily work because of many advantages, such as unlimited resources, flexibility, low-cost and especially the pay-as-you-go model. Cloud computing is service-oriented. Compared to traditional web services, cloud services can provide more complex functions, e.g., users can develop their own applications in a PaaS (Platform-as-a-Service) cloud.

Due to the diversity and dynamics of cloud services, selecting the most suitable cloud service has become a major issue for potential cloud consumers. Prior to cloud service selection, an evaluation of cloud services should be applied first. There are two types of approaches which can be used to conduct such an evaluation. The first type of approaches is based on objective performance assessment from ordinary QoS (Quality-of-Service) value (e.g., service response time, availability and throughput) monitoring [22] [2] [17] and predesigned benchmark testing [10] [1] [9]. As cloud services are highly virtualized, some methods and tools for traditional IT computation measurements can be appropriately applied in cloud environments. By combining these methods and tools according to cloud features, many metrics can be quantitatively assessed (e.g., the speed of CPU and storage in virtual machines). The second type of approaches is based on user subjective assessment which is usually extracted from user ratings for each concerned aspect of cloud services [18] [13]. In this type of approaches, cloud services are usually treated like traditional web services, thus some rating-based reputation systems [15] [11] can be utilized for cloud service selection.

Nevertheless, these two types of cloud service evaluation approaches have their own limitations. For example, considering a health center processing a large amount of sensitive customer data every day, the security and privacy of customer data have a crucial impact on the center's reputation. If the center plans to move its work into cloud environments in order to reduce daily costs, a suitable cloud provider which has very good reputation on data security and privacy needs to be selected. In addition, as the health center is not a professional IT organization, comprehensive and high-quality after-sales services are highly desired. Moreover, a variety of encryption approaches need to be frequently applied due to the sensitivity of customer data. Hence, the speed of data encryption and decryption is a big concern for the center. In this example, neither of the two types of cloud service evaluation approaches introduced above can be used to reflect all the concerned aspects (e.g., data security and privacy, after-sales services and cryptographic calculation speed) of the health center. That is because, firstly, objective performance assessment can only be carried out for the performance aspects which can be easily quantified. Conversely, objective assessment is not appropriate for those aspects which are quite hard to quantify, such as data security, privacy and after-sales services. On the other hand, subjective assessment has the risk of inaccuracy since users' subjective feelings are very likely to contain bias and not reflect the real situations of cloud performance. In addition, as cloud users who give subjective assessment are usually spread throughout the world, for any cloud service, the subjective feelings of a cloud user in a context (e.g., morning in Sydney) may be much different from those of another user in a different context (e.g., afternoon in Paris). Furthermore, there may be malicious users who give unreasonable subjective assessment to deceive others and/or benefit themselves in some cases. As a result, the accuracy of overall subjective assessment for cloud services can be significantly affected. Hence, a cloud service selection model which can be used to not only aggregate different performance aspects of cloud services but also filter unreasonable user subjective assessment is highly desirable.

To overcome the aforementioned drawbacks, this paper proposes a novel context-aware cloud service selection model based on the comparison and aggregation of subjective assessment extracted from cloud user feedback and

objective assessment from quantitative performance testing. In this model, according to a potential cloud consumer's requirements, an objective assessment provided by some professional testing party is first applied as a benchmark to filter out biased or unreasonable subjective assessments. In order to guarantee the accuracy of such filtering, our work considers two assessment features (i.e., *location* and *time*) in contexts, which can commonly affect both objective assessment and subjective assessment. In this paper, the process of filtering is based on the context similarity between objective assessment and subjective assessment, i.e., the more similar the context, the more reliable subjective assessment, so that the benchmark level is dynamically adjusted according to the corresponding context similarity. After such filtering, the final aggregated results can reflect the overall performance of cloud services according to potential users' personalized requirements.

After introducing the related work of cloud service selection in Section 2, some preliminaries are presented in Section 3. Then, the details of our model are discussed in Section 4. Section 5 presents the experimental results. Finally, this paper is concluded in Section 6.

II. Related Work

Approaches to cloud service selection can be categorized into two types based on whether objective assessment or subjective assessment has been utilized. Objective assessment is usually acquired from service QoS monitoring and benchmark testing, and subjective assessment is usually acquired from ratings in user feedback.

In the literature of objective assessment based cloud service selection, Zheng *et al*. [22] introduce a QoS prediction framework for optimal cloud service selection based on users' ranking similarity, and propose two personalized QoS ranking prediction approaches for potential users. However, their approaches can only rank different QoS properties independently, thus cannot reflect the overall performance of cloud services. Chen *et al*. [2] propose a cloud service evaluation model using QoS ontology for only three dimensions (i.e., resource utilization, service performance and cost) of service performance. Rehman *et al*. [17] propose a multi-criteria model for IaaS (Infrastructure-as-a-Service) cloud service selection using QoS history which is divided into several timeslots. The optimal cloud service in each timeslot is selected first, and then aggregated to find the overall optimal service. Although their work considers multiple criteria, the performance aspects which are hard to quantify are not taken into account. In the literature, as cloud services are web-based, some service selection approaches designed for common web services [19] [20] can also be applied in cloud environments through modification according to cloud features.

It should be noted that QoS-based objective assessment for cloud services is insufficient to evaluate the complex and flexible functions of cloud services. To this end, many predesigned benchmark testing scenarios are employed for the specific quantitative performance aspects of cloud services. In [10], Li *et al*. propose a systematic comparator called *CloudCmp*, which can be employed to compare three specific performance aspects (i.e., elastic computing, persistent storage and intra-cloud and wide area networking) of public clouds through a set of standard benchmark tools. In [1], Binning *et al*. propose a benchmarking approach for cloud services based on the metrics of scalability, cost, peak-load and fault tolerance. In [9], Lenk *et al.* highlight the significance of third-party performance testing of cloud services, as the performance indicators provided by cloud providers may not be enough to judge the real performance of cloud services. They propose a new performance measurement which considers the types of services running on IaaS clouds. Recently, some organizations (e.g., CloudHarmony[1] and CloudSleuth[2]) have started to offer third-party cloud monitoring and testing services. Compared to the performance indicators provided by cloud providers, such third party testing may be more reliable due to no direct profits involved.

In the literature of subjective assessment based cloud service selection, Rehman *et al*. [18] propose a framework for dynamically monitoring and predicting cloud performance based on user feedback. However, their work only considers cloud users' subjective assessment. There is no mechanism to check the reliability of users' feedback. In [13], Noor *et al*. propose a framework for trust management in cloud environments, and introduce a credibility model that has the ability to detect malicious user feedback based on majority consensus and feedback density. However, their work does not consider the case that plenty of users might collude to behave maliciously. As cloud services can be considered as common web services, many rating-based reputation systems can be employed in cloud environments for service selection. Srivastava *et al*. [15] present a method to compare functionally equivalent services on the basis of the customers' perception of the QoS attributes rather than the actual attribute values. In [11], Li *et al.* propose several trust vector based service evaluation approaches, where a trust vector is calculated to reflect both the current trustworthiness of a service and its trust trend. Such trust values or vectors are all evaluated from ratings which represent the subjective assessment of services given by service consumers.

In order to consider multiple performance aspects, some studies model the cloud service selection problem as a multi-criteria decision-making (MCDM) problem, which can be solved by Analytic Hierarchy Process (AHP) [12]. Godse *et al*. [6] focus on the selection of SaaS (Software-as-a-Service) clouds using AHP based on five factors (i.e., functionality, architecture, usability, vendor reputation and cost). Their approach is mainly based on user subjective assessment. Another AHP-based cloud service selection approach is proposed by Garg *et al.* [5]. It should be noted that, in these AHP-based approaches, all performance aspects should be standardized before processing AHP. However, such standardization for some attributes (e.g., elasticity and reliability) cannot be easily achieved in practice. Moreover, none of these approaches consider the credibility of the assessment before processing AHP.

In our prior work [14], we propose a cloud service selection model based on both objective assessment and subjective assessment. In this model, objective assessment and subjective assessment are first normalized in fuzzy numbers, and then objective assessment is used as a benchmark to filter out biased subjective assessment since objective assessment is usually more objective and accurate through scientific analysis and statistics. However, such filtering process does not take the contexts of assessment which can commonly affect assessment into account. That makes that the filtering process may be carried out inaccurately in some cases, thus affecting the overall result of cloud service selection. This paper proposes a solution to overcome this drawback in our prior model of cloud service selection.

III. Preliminaries

Before introducing our context-aware cloud service selection model in Section 4, we first briefly introduce our preliminary work [14] in this section.

A. The Proposed Framework

In our prior work [14], a framework is proposed for cloud service selection based on both cloud user feedback and

[1] cloudharmony.com

[2] cloudsleuth.net

objective performance benchmark testing. This framework is composed of four components, namely, (1) *cloud selection service*, (2) *benchmark testing service*, (3) *user feedback management service*, and (4) *assessment aggregation service*, where *cloud selection service* is in the higher layer of the framework to command the others in the lower layer.

1) Cloud Selection Service: The *cloud selection service* is responsible for accepting and preliminarily processing requests from potential cloud consumers and issuing commands to the lower layer services. When a potential cloud user submits a request for selecting the most suitable cloud service, the *cloud selection service* first chooses those cloud services which can meet all the basic and minimum requirements (e.g., the type of services, the specification of virtual machines and costs) of the potential user from a candidate list of cloud services. Then, according to the user's further requirements, it sends requests to the *benchmark testing service* and the *user feedback management service* for accessing the related records of all alternative clouds. These records are then sent to the *assessment aggregation service*, which returns the final aggregated score of every alternative cloud service to the *cloud selection service*.

2) Benchmark Testing Service: The *benchmark testing service* is in charge of QoS (e.g., response time and availability) monitoring and performance benchmark testing. When the *benchmark testing service* receives the commands from the *cloud selection service*, the results of QoS monitoring for all the alternative clouds are gathered, and some specific performance aspects (e.g., cryptographic calculation speed) are tested on each alternative cloud through a variety of predesigned testing scenarios according to the potential user's requirements. Then the *benchmark testing service* sends both the QoS monitoring records and testing results to the *assessment aggregation service*. Each performance aspect monitored and tested by the *benchmark testing service* can be considered as an *objective attribute* of a cloud service. All these objective attributes are expressed in quantitative forms (e.g., $2.15s$ for response time and 29.87 benchmark scores for CPU performance).

3) User Feedback Management Service: The *user feedback management service* is responsible for collecting and managing the feedback from the users who are consuming cloud services. For each performance aspect of a cloud service, a user gives his/her subjective assessment according to his/her perception. Each aspect that users can assess can be considered as a *subjective attribute* of the cloud service. These subjective attributes are expressed by linguistic variables (e.g., "good", "fair" and "poor").

In the framework, some subjective attribute and some objective attribute can represent the same performance aspect of a cloud service. For example, the response time of a cloud service can be accurately monitored and measured under different circumstances. By analyzing these quantitative results, an objective assessment of response time can be achieved for the cloud service. Meanwhile, a user consuming this cloud service can also give subjective assessment of response time by sensing how long the cloud responds to his/her requests. Such attributes (e.g., response time) are named as *associated attributes* in the framework. Figure 1 illustrates an example. Assume there are s subjective attributes, o objective attributes and u pairs of associated attributes for a cloud service ($u \leqslant s, u \leqslant o$), where *privacy, after-sales services, availability* and *response time* are its *subjective attributes* extracted from users' feedback. On the other hand, *availability, response time* and *cryptographic calculation* are its *objective attributes* extracted from objective performance assessment. And *availability* and *response time* are considered as its *associated attributes*.

4) Assessment Aggregation Service: The *assessment aggregation Service* is in charge of aggregating the values of subjective attributes and objective attributes from the *benchmark testing service* and the *user feedback management service*, and calculating the final aggregated score for each alternative cloud service according to the importance weights that are assigned to every attribute by a potential cloud user in the form of linguistic variables (e.g., "low", "medium", and "high"). By using these weights, a potential cloud user can also determine whether to put more trust on subjective assessment or objective assessment, so that the final score based on aggregating all these attributes can reflect the various needs of potential cloud users.

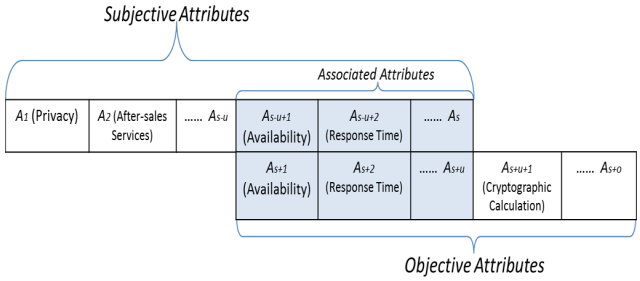

Figure 1. The Relationship of Subjective Attributes and Objective Attributes [14]

B. The Cloud Service Selection Model

In our prior work [14], the overall assessment for a cloud service depends on s subjective attributes and o objective attributes, where there are u pairs of associated attributes. All the subjective attributes and the objective attributes are denoted as $\{A_i\}$, where $i = 1, 2, \cdots, s+o$. $\{A_i\}$ $(i = 1, \cdots, s)$ denotes the subjective attributes, where $\{A_i\}$ $(i = s-u+1, \cdots, s)$ denotes the subjective associated attributes; $\{A_i\}$ $(i = s+1, \cdots, s+o)$ denotes the objective attributes, where $\{A_i\}$ $(i = s+1, \cdots, s+u)$ denotes the objective associated attributes. The corresponding objective associated attribute of the subjective associated attribute A_i is A_{i+u} for each $i = s-u+1, \cdots, s$. Suppose that there are m alternative clouds denoted by $\{C_j\}$, where $j = 1, \cdots, m$. For an alternative cloud j, there are n subjective assessments from the *user feedback management service* denoted as $\{F_{jk}\}$ and one objective assessment from the *benchmark testing service* denoted as T_j, where $k = 1, \cdots, n$.

As introduced in Section 3.1, objective assessment and subjective assessment are expressed in different forms (i.e., linguistic variables and quantitative terms). In order to deal with the uncertainty of linguistic variables, trapezoidal fuzzy numbers are used to represent these linguistic variables through some mapping [3]. Meanwhile, trapezoidal fuzzy numbers can also be used to represent quantitative terms. In addition, some operations including *addition* $\oplus$, *multiplication* $\otimes$, *division* $/$ and a defuzzification method $d(\widetilde{*})$ for converting a fuzzy number to a crisp number are defined for trapezoidal fuzzy numbers. As a result, both objective assessments and subjective assessments can be normalized for further comparison and aggregation.

The detailed procedure of our prior cloud service selection model [14] consists of five steps:

Step 1 (Converting the values of subjective attributes into ratings): Through a mapping from linguistic variables to fuzzy numbers [3], the values of subjective attributes in all subjective assessments given by different cloud users are converted into trapezoidal fuzzy numbers. For example, "good" can be represented by $(5, 7, 7, 10)$, where the value 7 represents "absolute good" and the values in the intervals $[5, 7]$ and $[7, 10]$ represent the "fuzziness of good". And "very good" can be represented by $(7, 10, 10, 10)$. These fuzzy numbers are considered as the fuzzy ratings for every subjective attribute. Let $\widetilde{r}_{ijk}$ and r_{ijk} denote the fuzzy rating and the crisp rating computed by $d(\widetilde{*})$ for the subjective

attribute A_i of the alternative cloud C_j from the subjective assessment F_{jk} respectively, where $i = 1, \cdots, s$.

Step 2 (Converting the values of objective attributes into ratings): Quantitative terms can also be represented by trapezoidal fuzzy numbers. For example, "equal to 30" can be represented by $(30, 30, 30, 30)$, and "approximately equal to 30" can be represented by $(28, 30, 30, 32)$. In this step, the quantitative values of every objective attribute from the objective assessments for all alternative cloud services are first represented using fuzzy numbers. Then, all these fuzzy numbers are converted into fuzzy ratings by comparing every fuzzy value to the best fuzzy value of each objective attribute in all alternative cloud services [14]. Let $\widetilde{r}_{ijk}$ and r_{ijk} denote the fuzzy rating and the crisp rating for the objective attribute A_i of the alternative cloud C_j from the objective assessment T_j respectively, where $i = s+1, \cdots, s+o$.

So far, all the values in both subjective assessments and objective assessments have been converted into fuzzy ratings. As a result, a fuzzy rating matrix is formed for every alternative cloud service. For the alternative cloud C_j, its fuzzy rating matrix is

$$\widetilde{M}_j = \begin{bmatrix} \widetilde{r}_{1j1} & \widetilde{r}_{2j1} & \cdots & \widetilde{r}_{(s+o)j1} \\ \widetilde{r}_{1j2} & \widetilde{r}_{2j2} & \cdots & \widetilde{r}_{(s+o)j2} \\ \cdots & \cdots & \cdots & \cdots \\ \widetilde{r}_{1jn} & \widetilde{r}_{2jn} & \cdots & \widetilde{r}_{(s+o)jn} \end{bmatrix}.$$

Step 3 (Filtering out unreasonable subjective assessments): For a fuzzy rating matrix $\widetilde{M}_j$ of the alternative cloud C_j, the Euclidean distance between the ratings of the corresponding subjective associated attributes and the objective associated attributes is computed for each subjective assessment as follow:

$$ED_{jk} = \sqrt{\sum_{i=s-u+1}^{s} (r_{ijk} - r_{(i+u)jk})^2}.$$

Here, the objective assessment is taken as a benchmark to filter out unreasonable subjective assessments. If the distance exceeds a **fixed threshold** (e.g., 60% of the maximum Euclidean distance), the subjective assessment offering such values of the subjective attributes are removed from the fuzzy rating matrix. By this way, even plenty of cloud users are colluded to provide unfair subjective assessments, such unfair assessments can also be filtered out.

Step 4 (Computing the importance weight for each attribute): According to the potential cloud user's requirement, an importance weight in the form of linguistic variables is assigned to each subjective or objective attribute. Through these weights, the potential user can also determine how much to trust subjective or objective assessment.

A fuzzy weight in the form of trapezoidal fuzzy numbers is assigned to each attribute, denoted as $\widetilde{W}_i$, where $i = 1, \cdots, s+o$, through another mapping from linguistic variables to fuzzy numbers [3]. For example, "medium" can be represented by $(2, 5, 5, 8)$, and "high" can be represented by $(5, 7, 7, 10)$. W_i is the normalized weight for the attribute A_i, which is computed as follow:

$$W_i = \frac{d(\widetilde{W}_i)}{\sum_{i=1}^{s+o} d(\widetilde{W}_i)}, \text{where} \quad i = 1, \cdots, s+o.$$

Step 5 (Aggregating all attributes): $\widetilde{M}'_j$ is the fuzzy rating matrix for the alternative cloud C_j after the filtering process in Step 3. Its final score $\overline{S_j}$ is computed as follows:

$$\widetilde{S_j} = \widetilde{M}_j \otimes \begin{bmatrix} W_1 \\ W_2 \\ \cdots \\ W_{s+o} \end{bmatrix} = \begin{bmatrix} \widetilde{f}_{j1} \\ \widetilde{f}_{j2} \\ \cdots \\ \widetilde{f}_{jn'} \end{bmatrix}, \quad \overline{S_j} = \frac{1}{n'} \left(\sum_{k=1}^{n'} d(\widetilde{f}_{jk})\right),$$

where n' is the number of the rest of subjective assessments after Step 3. Finally, according to the final scores, all the alternative cloud services are ranked for selection by the potential cloud user.

IV. Context-aware Cloud Service Selection

In this section, we first introduce what the contexts are in cloud service selection, and then point out the details of the drawback in our prior work [14] introduced in Section 3. After that, an improved model is proposed based on the similarity of assessment contexts.

A. Contexts in Cloud Service Selection

The definition of contexts usually varies in different application environments. In our cloud service selection model based on both objective assessment and subjective assessment, the context of an assessment for a cloud service refers to a group of values of the features of the assessment, which can affect the result of the assessment.

To give an example of the impact of a context, according to the objective statistics from CloudSleuth, the response time of a cloud service varies significantly under different worldwide QoS monitoring centers, and generally increases with the increasing distances between the cloud provider and these monitoring centers because of the increasing length of the network routes of cloud service delivery. Meanwhile, the monitoring results of response time can also be affected by the time of a day, in other words, how busy the cloud service and the network accessed by the monitoring centers for monitoring can vary at different times of a day. Therefore, both objective assessment and subjective assessment can be affected according to different assessment contexts. At the current stage of our work, we consider two assessment features (i.e., *location* and *time*) in our context-aware cloud service selection model.

In our prior cloud service selection model [14], assessment contexts are not taken into account. However, in order to have more accurate comparison between objective assessment and subjective assessment, the similarity between the contexts of objective assessment and subjective assessment should be considered. More similar contexts indicate the subjective assessments are given in the more similar situation with that of the given objective assessment, thus such subjective assessments are considered more reliable. Furthermore, in our prior model [14], a fixed threshold is used as the benchmark value for the objective assessment to filter out unreasonable subjective assessments. This threshold reflects how much the objective assessment is trusted. If the threshold is high, more subjective assessments are retained for the following aggregation process, which means more subjective assessments are considered reasonable, otherwise fewer subjective assessments are retained, which means fewer subjective assessments are considered reasonable. However, determining such a suitable fixed threshold is very difficult. Because the fixed threshold means the subjective assessments with different contexts are treated equally. If the threshold is determined too high, more noisy subjective assessments will be left in the final aggregated results. Conversely, if the threshold is too low, only a few subjective assessments are left so that only these few subjective assessments can affect the final aggregated results, and, as a consequence, the final aggregated results cannot reflect most users' subjective assessment. An intuitive solution to overcome this drawback is to adjust the threshold dynamically according to the context similarity between objective assessment and subjective assessment. The more similar the contexts, the more reliable subjective assessments, thus the threshold should be set higher for retaining more of such subjective assessments. On the contrary, if the contexts are less similar, then the threshold should be set lower to filter out more subjective assessments which are given in more different situations. Next, the details of computing such context similarity will be introduced.

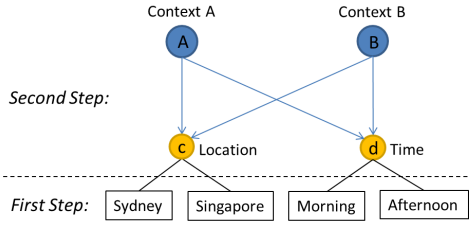

Figure 2. An Example of Two Contexts

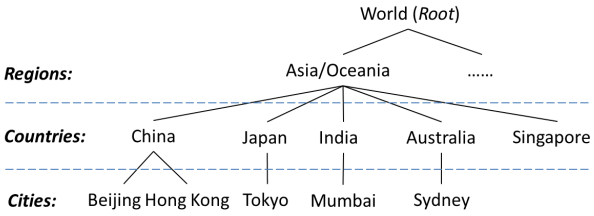

Figure 3. A Geographical Hierarchy

B. Context Similarity

In [16], Tavakolifard *et al.* introduce a general idea of the calculation of context similarity based on the bipartite SimRank algorithm [8] for trust transferability among similar contexts in electronic transactions. In order to compute the similarity of assessment contexts in our context-aware cloud service selection model, we follow Tavakolifard *et al.*'s idea and propose a concrete approach for context similarity measurement. In details, our approach consists of two steps:

The ***first step*** is to compute the similarity between two values from the same assessment feature.

The ***second step*** is to model all contexts and their relevant assessment features as a graph and compute the overall similarity between contexts.

Figure 2 illustrates an example of two contexts A (*Sydney, morning*) and B (*Singapore, afternoon*) belonging to two assessments respectively. Each context contains two values for two assessment features (i.e., *location* and *time*) respectively. *Sydney* and *Singapore* are the values of the feature *location* for both contexts respectively. Likewise, *morning* and *afternoon* are the values of the feature *time*.

In [16], Tavakolifard *et al.* only introduced how to compute overall context similarity (i.e., the second step) through the bipartite SimRank algorithm and did not present details on computing the similarity between two values from the same assessment feature (i.e., the first step). For each assessment feature, a specific comparator needs to be designed for computing similarity among the values of each feature. In our context-aware cloud service selection model, two features are considered in assessment contexts. Next, we first present a modified version of the bipartite SimRank algorithm [8] according to our model, and then introduce the design of the comparators for *location* and *time*.

1) Modified Bipartite SimRank: The original bipartite SimRank algorithm is modified to take different context comparators into account in our model. Let A and B denote two contexts and, $s(A, B)$ denote the similarity between A and B. If $A = B$, then $s(A, B) \in [0, 1]$ is defined to be 1. Let c and d denote assessment features for contexts A and B, and $s(c, d) \in [0, 1]$ denote the similarity between features c and d. Let $V_c(A)$ and $V_c(B)$ denote the values of the feature c in the contexts A and B respectively. Likewise, $V_d(A)$ and $V_d(B)$ denote the values of the feature d in the contexts A and B respectively. If $c = d$, then $s(c, d) = Cmp_c(V_c(A), V_c(B)) = Cmp_d(V_d(A), V_d(B)) \in [0, 1]$, where Cmp_c and Cmp_d are the comparators for the features c and d.

Now, A, B and c, d can be formed to a directed graph pointing from contexts to features. If we take Figure 2 as an example, we have that $A = (Sydney, morning)$, $B = (Singapore, afternoon)$, $c = location$, $d = time$, $V_c(A) = Sydney$, $V_c(B) = Singapore$, $V_d(A) = morning$ and $V_d(B) = afternoon$. In the directed graph, $I(v)$ and $O(v)$ denote the set of in-neighbors and out-neighbors of v respectively, where v is a node in the graph. $I_i(v)$ denotes an individual in-neighbor of v for $1 \leq i \leq |I(v)|$, and $O_i(v)$ denotes an individual out-neighbor of v for $1 \leq i \leq |O(v)|$.

Now we have the recursive equations: for $A \neq B$,

$$s(A, B) = \frac{C}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)), \quad (1)$$

and for $c \neq d$,

$$s(c, d) = \frac{C}{|I(c)||I(d)|} \sum_{i=1}^{|I(c)|} \sum_{j=1}^{|I(d)|} s(I_i(c), I_j(d)), \quad (2)$$

where $C \in (0, 1)$ is a constant which can be considered as either a confidence level or a decay factor. In the full version [7] of Jeh *et al.*'s paper [8] proposing bipartite SimRank, they argue that the constant C can be viewed as the bases of exponential functions whose only purpose is to map distances to finite intervals. Although the values of similarity can be affected by C, the relative results of similarity is still retained. Hence, for the sake of efficiency, we follow Jeh *et al.*'s setting to set $C = 0.8$ in our model. In addition, Jeh *et al.* have proven that a simultaneous solution $s(*, *) \in [0, 1]$ to the recursive equations (1) and (2) always exists and is unique.

2) Design of Comparators: According to each assessment feature, a corresponding comparator needs to be designed and applied in the above modified bipartite SimRank algorithm. In our model, two assessment features are considered, i.e., *location* and *time*.

Similarity of Locations:

The effect for both objective assessment and subjective assessment of cloud services is usually caused by the delay of the Internet communication between the locations of where the assessments are given and the target cloud service. In order to precisely model such an effect, the Internet topology between these parties should be first determined. However, such a topology should be created by some domain experts, and is out of the scope of this paper. For the sake of simplicity, in this paper, we use geographical locations instead of the Internet locations. That is because the distance between two nodes in the Internet is commonly determined by their geographical locations. We introduce a similarity measurement based on a hierarchical ontology structure [21] for the assessment feature *location* in our model.

According to the real monitoring data from CloudSleuth, we establish a geographical hierarchy according to the order of $regions \rightarrow countries \rightarrow cities$. Figure 3 illustrates the Aisa/Oceania part of the hierarchy. In order to measure the similarity between any two nodes in the hierarchy, we apply Zhang *et al.*'s hierarchy-based approach of similarity measurement [21]. Let D denote the depth of the deepest common ancestor of two nodes n and n'. For example, the deepest common ancestor of *Beijing* and *Tokyo* is *Asia/Oceania* in Figure 3, thus $D(Beijing, Tokyo) = 1$. The smaller D represents the deepest common ancestor of the two nodes is on the upper layer of the hierarchy, which means the two nodes are fallen into a more general classification, thus are less similar. Conversely, a larger D means the two nodes are fallen into a more concrete classification, thus are more similar. Hence, a monotonically increasing hyperbolic tangent function is defined to model this trend:

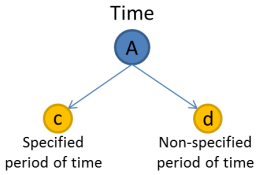

Figure 4. Similarity of Time

$$Cmp(n, n') = \frac{e^{\alpha D(n,n')} - e^{-\alpha D(n,n')}}{e^{\alpha D(n,n')} + e^{-\alpha D(n,n')}}, \quad (3)$$

where $Cmp(n, n')$ represents the similarity comparator returning the similarity value between n and n'; $\alpha \in (0, 1)$ is a constant. Here, we follow Zhang *et al.*'s setting to set $\alpha = 0.4$, which is considered optimal according to their experimental results.

Similarity of Time:

In practice, the reasons why the different times of a day can affect both objective assessment and subjective assessment of cloud services are quite diverse and complicated, where the main reason for such an effect is how busy the networks used by users to access cloud services are. However, the extent of how busy networks are varies frequently according to different users' situations, thus it is also quite hard to quantitatively measure such changes.

Hence, in our model of context-aware cloud service selection, we divide 24 hours of a day into two time intervals. When a potential cloud user asks for cloud service selection, he/she needs to specify in what period of time he/she hopes to frequently employ the selected cloud service. The assessments given within that period of time are considered more reliable for the potential user, and the assessments given within the non-specified period of time are considered less reliable for the user. Therefore, in our model every subjective assessment contains a time stamp to identify the time when the assessment is given. We assume that such assessments are required to represent users' subjective judgement at that time only. To this end, in our future work we plan to design an incentive mechanism for giving cloud users incentives to provide subjective assessments regularly. Due to the incentive mechanism, most cloud users will give such subjective assessments with time stamps.

In our model, the assessment feature *time* has two states, i.e., specified and non-specified. The similarity between these two states can be computed through the basic bipartite SimRank algorithm [8]. Figure 4 illustrates the graph of similarity of the two states. Then, the similarity between *specified period of time* and *non-specified period of time* can be computed through Equations (1) and (2).

It should be noted that, except *location* and *time*, there are some other assessment features which can also affect the assessment results for some reasons (e.g., the Internet service providers). The similarity among the values of such features should be computed through specific designed comparator. And the modified bipartite SimRank algorithm introduced above can be applied with any further comparator.

C. The Proposed Model

In our context-aware cloud service selection model, we assume there are plenty of *benchmark testing agents* spread around the world providing *benchmark testing services*. When a potential cloud user asks for selecting the most suitable cloud service, according to his/her situation, he/she needs to specify which agents should be selected to offer objective assessments for all alternative cloud services. Then, the cloud service selection will be processed independently according to each agent. For each *benchmark testing agent*, the *cloud selection service* asks the *user feedback management service* to provide the subjective assessments for all alternative cloud services from the cloud users all over the world. Then, all the subjective assessments are classified according to their contexts (i.e., *location* and *time*). As the nodes in the deepest level of our geographical hierarchy are cities, the location of each subjective assessment is set as the nearest city to the real location specified in the assessment in the hierarchy. And due to time differences among cloud users all over the world, the time specified in every assessment is converted into one standard time.

Assume that there are l locations shown in all the subjective assessments. As there are only two states for the assessment feature *time* in our model, i.e., *specified period of time* and *non-specified period of time*, all the subjective assessments are classified into $2l$ groups. Then, according to the potential user's requirement, the *benchmark testing agent* provides an objective assessment with contextual information (e.g., objective performance assessment in the *morning* of *Sydney*).

The process of the comparison and aggregation of the objective assessment and the subjective assessments is the same as that of our prior work without the consideration of assessment contexts [14], except the importance weight setting and changing a fixed threshold to a group of dynamical thresholds. Such thresholds are computed as follows:

Step 1: The potential user first sets the importance weights on how much to trust objective assessment or subjective assessment through linguistic variables. Then, through a mapping [3], linguistic weights are converted into fuzzy weights, which are denoted as $\widetilde{W_o}$ and $\widetilde{W_s}$ for objective assessment and subjective assessment respectively. Then, the potential user sets the importance weight for each objective or subjective attribute, denoted as $\widetilde{W_i}$, where $i = 1, \cdots, s+o$. After that, W_i is the normalized weight of each attribute, which is computed as follow:

$$\begin{aligned} W_i &= \frac{d(\widetilde{W_s})}{d(\widetilde{W_s}) + d(\widetilde{W_o})} \times \frac{d(\widetilde{W_i})}{\sum_{i=1}^{s} d(\widetilde{W_i})}, \ i = 1, \cdots, s, \\ W_i &= \frac{d(\widetilde{W_o})}{d(\widetilde{W_s}) + d(\widetilde{W_o})} \times \frac{d(\widetilde{W_i})}{\sum_{i=s+1}^{s+o} d(\widetilde{W_i})}, \ i = s+1, \cdots, s+o. \end{aligned} \quad (4)$$

Step 2: Let g_o denote the context of the objective assessment, and g_v denote the context of each group of the subjective assessments in the total $2l$ groups, where $1 \leq v \leq 2l$. Through the approach introduced in Section 4.2, the similarity between each g_v and g_o is computed and denoted as $s_v(g_v, g_o)$.

Step 3: In order to offset the effect caused by the constant C in the modified bipartite SimRank algorithm, let $s_o(g_o, g_o)$ denote the similarity between the contexts of the objective assessment and itself, and E_{dis} denote the theoretical maximum Euclidean distance between corresponding objective associated attributes and subjective associated attributes according to our model. The filtering threshold R_v for the subjective assessment group with the context g_v is computed as follow:

$$R_v = (1 - \frac{d(\widetilde{W_o})}{d(\widetilde{W_s}) + d(\widetilde{W_o})}) \times \frac{s_v(g_v, g_o)}{s_o(g_o, g_o)} \times E_{dis}, \quad (5)$$

where $v = 1, \cdots, 2l$. From the above equation, we can see when the potential user trusts objective assessment more, R_v will become smaller, so that more subjective assessments are considered unreasonable and will be filtered out. In addition, when the context similarity $s_v(g_v, g_o)$ becomes lower, R_v will become smaller. That means the subjective assessments are given in a more different situation with that of the objective assessment, thus such subjective assessments are considered less reliable and will be filtered out more rigorously. Finally, the rest of the subjective assessments after such filtering and the objective assessment are aggregated

to reflect the overall performance of a cloud service more accurately.

V. Experiments

A. Experiment Setup

In our experiments, there are three subjective attributes, i.e., *cloud provider reputation on privacy (A_1), after-sales services (A_2)*, *service response time (A_3)*, and two objective attributes, i.e., *service response time (A_4)* and *CPU performance (A_5)*, where *service response time* A_3 and A_4 are the associated attribute pair.

In order to evaluate our context-aware cloud service selection model, two kinds of data are required, i.e., subjective ratings from cloud users, and objective results of QoS monitoring and benchmark testing. In our experiments, we collect the data of *response time* A_4 from CloudSleuth and the data of benchmark scores of *CPU performance* A_5 from CloudHarmony for 59 real cloud services. To the best of our knowledge, there is no data set of cloud user ratings published for these 59 cloud services. Hence, we simulate user ratings of the attributes A_1, A_2 and A_3 according to the collected objective data (i.e., A_4 and A_5). In details, the ratings of A_1 and A_2 are randomly generated, and the normal ratings of A_3 are generated according to the ranking of the real data of *response time* in A_4. Then, some biased ratings are added into the normal ratings of A_3 to simulate the ratings from the users who are in different contexts with that of objective assessments. Here, a bias level denoted as BL is set to represent how much the biased ratings deviate from the normal synthetic ratings of A_3, where $BL = 1, \cdots, 8$ since a rating scale of 1-9 is employed in our model. Moreover, a biased rating percentage denoted as BRP is set to represent how many biased ratings there are in all the subjective ratings.

We assume that all the subjective ratings are from the cloud users belonging to two different contexts. The one context is (*Sydney, afternoon*) which is also the context of the objective assessment in our experiments, and the other context is (*Hong Kong, morning*). According to the algorithm introduced in Section 4.2, the similarity of the two contexts is 0.4714. Thus, two thresholds are computed for the comparison of subjective assessment and objective assessment according to different importance weights (i.e., $\widetilde{W_o}$ and $\widetilde{W_s}$).

B. Evaluation Metric

In our experiments, we first generate 1000 normal ratings for the attributes A_1, A_2 and A_3 respectively through the way introduced in Section 5.1, and then replace some proportion of normal ratings with biased ratings. Here, the original normal rating matrix is denoted as M_o, and the corresponding processed rating matrix including biased ratings is denoted as M_b.

As M_o is generated according to the objective assessment, the ratings in M_o are considered very accurate. Thus, the final aggregated result for each alternative cloud service without filtering between subjective assessment and objective assessment is considered very accurate to represent the overall performance of each cloud service. Here, $R(M_o)$ denotes the ranking of all the 59 cloud services based on such aggregated results without filtering according to M_o. $R_f(M_b)$ denotes the ranking of the cloud services based on our prior cloud service selection model [14] according to M_b without the consideration of assessment contexts; $R_c(M_b)$ denotes the ranking of the cloud services based on our context-aware cloud service selection model according to M_b with dynamic threshold filtering. $R_{sim}(*,*)$ denotes the similarity between two ranking lists. If $R_{sim}(R(M_o), R_c(M_b)) > R_{sim}(R(M_o), R_f(M_b))$, that means our context-aware model is more effective than our prior model [14].

In our experiments, $R_{sim}(*,*)$ is calculated through the *Kendall tau rank distance* [4] which is a common metric to measure the distance between two rankings through counting the number of pairwise disagreements between the two rankings. Here, we use the function $corr()$ provided in $Matlab$ to compute the normalized *Kendall tau distance* which lies in the interval $[-1, 1]$, where 1 means two rankings are in the same order, and -1 means two rankings are in the opposite order.

C. Experimental Results

In our experiments, the importance weight for each attribute is randomly selected. According to our experiments, the importance weights do not affect the trend of our experimental results, that is, our context-aware cloud service selection model is more effective. Due to the limitation of space, Table 1 only shows a part of experimental results for the 59 real cloud services based on two settings of importance weights. A larger value indicates better ranking accuracy. In order to more accurately simulate the ratings from real cloud users in our experiments, every value in Table 1 is the average ranking similarity computed based on every 100 different groups of M_o and M_b. And each group of data is generated independently. Thus, the generality of experimental data can be kept in our experiments. Table 1 shows that, among different experimental conditions (i.e., different BLs and BRPs), our context-aware cloud service selection model performs better than our prior model [14] without the consideration of contexts. And our context-aware model can achieve approximately 1.5% to 9% improvements.

Table 1 shows that our context-aware cloud service selection model based on dynamic threshold filtering can more effectively deal with the effect of biased subjective ratings than our prior cloud service selection model [14] in different conditions (i.e., different BLs and BRPs) except the conditions that $BL = 1, 2$ or 3. That is because, in the real world, cloud users' subjective assessment for a cloud service cannot perfectly match the objective assessment of the cloud service due to users' different preferences. However, users' subjective assessment should not be far off from objective assessment. For this consideration, in our experiments, every individual synthetic normal subjective rating does not perfectly match the objective assessment, and may have a random small deviation (up to 3). If the deviation (i.e., BL) between biased ratings and normal ratings is too small, such biased ratings are very likely to be considered as normal ratings since such a small deviation should not be detected as the deviation between biased ratings and normal ratings. That leads to the fact that our experimental results in the conditions of $BL = 1, 2$ or 3 may be opposite since such biased ratings with small deviations cannot be detected in our experimental setting. However, in the other conditions of any BRP and $BL = 4, \cdots, 8$, the trend of our experimental results is the same. Figure 5 illustrates such an example when $BRP = 20\%$.

VI. Conclusion

This paper has proposed a novel model of context-aware cloud service selection based on comparison and aggregation of subjective assessment from cloud users and objective assessment from quantitative QoS monitoring and benchmark testing. Our model takes the contexts of both subjective assessment and objective assessment into account, and uses objective assessment as a benchmark to filter out unreasonable subjective assessment. The process of such filtering is based on a group of dynamic thresholds which are determined by the similarity between the contexts of subjective assessment and objective assessment. Our experimental

Importance Weights	BRP	Ranking Similarity \ BL	4	5	6	7	8
$W_s = High, W_o = High$ $W_1 = VeryHigh$	20%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	84.5235	88.9254	92.9106	92.9663	92.9837
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**90.9944**	**93.3161**	**94.8105**	**94.8335**	**94.9709**
$W_2 = High$ $W_3 = Medium$ $W_4 = VeryHigh$ $W_5 = High$	50%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	84.4224	88.9060	92.7962	92.6966	92.8795
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**93.7033**	**93.6175**	**96.4285**	**96.3141**	**96.4009**
	70%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	84.2947	89.0061	92.9147	92.8943	92.7497
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**93.2850**	**93.8473**	**95.4045**	**95.3478**	**95.4597**
$W_s = High, W_o = VeryHigh$ $W_1 = Medium$ $W_2 = High$ $W_3 = High$ $W_4 = VeryHigh$ $W_5 = Medium$	20%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	92.4678	93.1859	93.9019	93.9888	93.9055
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**94.3121**	**95.1874**	**95.2651**	**95.3754**	**95.3366**
	50%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	92.3825	94.1011	93.8963	93.9423	93.8958
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**94.7589**	**96.3560**	**95.9770**	**96.0603**	**96.0587**
	70%	$R_{sim}(R(M_o), R_f(M_b)) \times 100$	91.4525	93.1057	92.9219	92.9775	92.9704
		$R_{sim}(R(M_o), R_c(M_b)) \times 100$	**93.4980**	**95.3238**	**95.0511**	**95.1287**	**95.2288**

Table I
ACCURACY COMPARISON BASED ON RANKING SIMILARITY

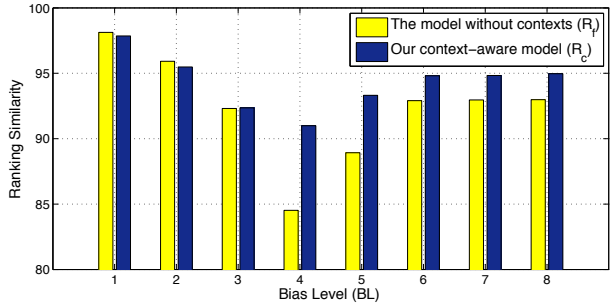

Figure 5. Ranking Similarity when BRP = 20%

results show that our context-aware model performs better than our prior cloud selection model which has no consideration of assessment contexts. Hence, the final aggregated results of cloud services based on our context-aware model can more accurately reflect the overall performance of cloud services.

For future work, we plan to design an incentive mechanism for our context-aware cloud service selection model. Through this mechanism, cloud users are encouraged to regularly provide honest subjective assessment for cloud services. Furthermore, this incentive mechanism should be secure against attacks from malicious parties. In addition, we plan to focus on designing more accurate context comparators for more different assessment features.

REFERENCES

[1] C. Binnig, D. Kossmann, T. Kraska, and S. Loesing. How is the weather tomorrow?: towards a benchmark for the cloud. In *ACM SIGMOD Conference - DBTest Workshop*, pages 9:1–9:6, 2009.

[2] G. Chen, X. Bai, X. Huang, M. Li, and L. Zhou. Evaluating services on the cloud using ontology QoS model. In *IEEE International Symposium on Service-Oriented System*, pages 312–317, 2011.

[3] S.-Y. Chou, Y.-H. Chang, and C.-Y. Shen. A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes. *European Journal of Operational Research*, 189(1):132–145, 2008.

[4] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

[5] S. K. Garg, S. Versteeg, and R. Buyya. SMICloud: A framework for comparing and ranking cloud services. In *Utility and Cloud Computing*, pages 210–218, 2011.

[6] M. Godse and S. Mulik. An approach for selecting Software-as-a-Service (SaaS) product. In *IEEE CLOUD*, pages 155–158, 2009.

[7] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. http://dbpubs.stanford.edu/pub/2001-41, 2001.

[8] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

[9] A. Lenk, M. Menzel, J. Lipsky, S. Tai, and P. Offermann. What are you paying for? performance benchmarking for Infrastructure-as-a-Service offerings. In *IEEE CLOUD*, pages 484–491, 2011.

[10] A. Li, X. Yang, S. Kandula, and M. Zhang. CloudCmp: comparing public cloud providers. In *Internet Measurement Conference*, pages 1–14, 2010.

[11] L. Li and Y. Wang. A trust vector approach to service-oriented applications. In *IEEE International Conference on Web Services*, pages 270–277, 2008.

[12] K. Muralidhar, R. Santhanam, and R. L. Wilson. Using the analytic hierarchy process for information system project selection. *Information & Management*, 18(2):87–95, 1990.

[13] T. H. Noor and Q. Z. Sheng. Credibility-based trust management for services in cloud environments. In *International Conference on Service Oriented Computing*, pages 328–343, 2011.

[14] L. Qu, Y. Wang, and M. A. Orgun. Cloud service selection based on the aggregation of user feedback and quantitative performance assessment. In *IEEE International Conference on Services Computing*, pages 152–159, 2013.

[15] A. Srivastava and P. G. Sorenson. Service selection based on customer rating of quality of service attributes. In *IEEE International Conference on Web Services*, pages 1–8, 2010.

[16] M. Tavakolifard, S. J. Knapskog, and P. Herrmann. Trust transferability among similar contexts. In *Q2SWinet*, pages 91–97, 2008.

[17] Z. ur Rehman, O. K. Hussain, and F. K. Hussain. Multi-criteria IaaS service selection based on QoS history. In *IEEE International Conference on Advanced Information Networking and Applications*, pages 1129–1135, 2013.

[18] Z. ur Rehman, O. K. Hussain, S. Parvin, and F. K. Hussain. A framework for user feedback based cloud service monitoring. In *International Conference on Complex, Intelligent and Software Intensive System*, pages 257–262, 2012.

[19] L.-H. Vu, M. Hauswirth, and K. Aberer. QoS-based service selection and ranking with trust and reputation management. In *OTM Conferences (1)*, pages 466–483, 2005.

[20] T. Yu, Y. Zhang, and K.-J. Lin. Efficient algorithms for web services selection with end-to-end QoS constraints. *ACM Transactions on the Web*, 1(1), 2007.

[21] H. Zhang, Y. Wang, and X. Zhang. Transaction similarity-based contextual trust evaluation in e-commerce and e-service environments. In *IEEE International Conference on Web Services*, pages 500–507, 2011.

[22] Z. Zheng, X. Wu, Y. Zhang, M. R. Lyu, and J. Wang. QoS ranking prediction for cloud services. *IEEE Trans. Parallel Distrib. Syst.*, 24(6):1213–1222, 2013.