Computers and Mathematics with Applications 61 (2011) 2638–2646

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

# Local martingale difference approach for service selection with dynamic QoS

Xiaofeng Di [a,*], Yushun Fan [a], Yimin Shen [b]

[a] Department of Automation, Tsinghua University, Beijing, 100084, China
[b] Chengdu Electromechanical College, Chengdu, 611730, China

ARTICLE INFO

*Article history:*
Received 5 November 2010
Received in revised form 2 March 2011
Accepted 2 March 2011

*Keywords:*
Martingale
Service selection
Stopping time
QoS

ABSTRACT

Users in Service-oriented architecture (SOA) seek the best Quality of service (QoS) by service selection from the candidates responding in succession. In case the QoS changes dynamically, choosing one service and stop the searching is problematic for a service user who makes the choice online. Lack of accurate knowledge of service distribution, the user is unable to make a good decision. The Local Martingale Difference (LMD) approach is developed in this paper to help users to achieve optimal results, in the sense of probability. The stopping time is proved to be bounded to ensure the existence of an optimal solution first. Then, a global estimation over the time horizon is transformed to a local determination based on current martingale difference to make the algorithm feasible. Independent of any predetermined threshold or manual intervention, LMD enables users to stop around the optimal time, based on the information collected during the stochastic process. Verified to be efficient by comparison with three traditional methods, LMD is adaptable in vast applications with dynamic QoS.



## 1. Introduction

Service-oriented architecture (SOA)-based application differ from traditional ones by its dynamic behavior, i.e., the existing services may compose dynamically at runtime to complete a complex job according to users' needs [1], and hence simplify integration among various systems [2].

The services implementing the same functionality make up a group of functionally equivalent candidates for services user, which are identified by their non-functional characteristics, i.e., Quality of Service (QoS) properties [3–5]. The user has to evaluate the available candidates at runtime according to the favorite QoS indices, and selects the best index to achieve the highest possible payoff.

The literature has proceeded to the origin, influence, and related methodologies of dynamic QoS though static QoS is still the main aspect of SOA study [6–8]. Wu et al. concluded that enough information is necessary in wireless cellular networks for guaranteeing QoS such as call dropping probability (CDP) and call blocking probability (CBP), but unfortunately, the information is hard to be obtained [9]. Tansupasiri et al. noticed that QoS should be dynamically adaptable to user requirements [10]. Shen et al. classified time to a distinctive QoS index set, decision-dependent index, to reflect its influence on user decision, and proposed a mixed strategy to solve the conflict among users led by the dynamic influence of time [11]. Díaz et al. went so far as to identify time on the systems where it plays an important role, as a special kind of user requirements other than quality of service [12].

* Corresponding author.
*E-mail addresses:* dixf06@mails.tsinghua.edu.cn, xiaofeng.di06@gmail.com (X. Di).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.03.010

Though dynamic QoS index grows in a significant way, the waiting time for the providers' dependence is usually ignored in real service selection scenarios. The reason widely accepted is that the runtime of services usually exceeds the time for choosing the provider by far, so it will be 'worth the wait' for the best possible service [13]. This might be correct for large-scale tasks, especially for the business processes containing human tasks.

However, the above reason (and the conclusion) is no longer correct because of the emergence of new architectures for intelligent computing. For example, the trends of services with corpuscular size, resources with improving ability and capability, and extended scope with more available providers are inevitable in cloud computing architecture, in order to meet the needs of services easy to deploy, and pay only for what is used [14]. In this case, the time for completing a single task tends to be shortened. Meanwhile, the time of waiting for service providers becomes more time consuming than before because users wish to choose the most suitable services from the providers coming in swarms, according to the properties of their needs, especially under the environment of the communication with delay (e.g. wireless circumstance [15]). Therefore, the proportion of the waiting time to the total time increases.

Therefore, online selecting approaches are proposed to save the waiting time for selecting services. For instance, Shen and Fan developed an iterative online methodology that enable the user to renew the chosen services during the arriving process of providers to obtain better QoS gradually, instead of waiting until all candidates respond to the request [16].

But the current online approaches apply traditional methods to determine when to stop waiting and kick off service. Yasmine Charif-Djebbar and Nicolas Sabouret suggested the mediator agent be provided with a timeout value over which, if no more messages arrive, it proceeds to service selection [17]. Three kinds of thresholds were suggested in Shen and Fan's paper to give the user a rule to stop waiting [16].

It merits noticing that the above methods depend on predetermined thresholds. But in fact, it is hard for the user to predict the best threshold because of the little information about the number of providers that will respond to the request, at what time and in which QoS level. If the user stops at an early stage because of a wrong threshold, the user might miss the coming provider with better QoS. On the contrary, waiting the provider to arrive might end up with a notable waste of valuable time, and hence decreases the payoff of the user. Thus the user sinks in a dilemma under such complex situation, and will have a small possibility to obtain a QoS level higher than it should be.

A local martingale method named Local Martingale Difference (LMD) is proposed in this paper to equip the service user with a condition to judge whether waiting for next provider is profitable. The LMD approach is a self-adaptable methodology for service selection. It makes the decision based on all the dynamic information gathered during the process of waiting, without any manual predetermined threshold or accurate parameters. The user is enable to make the optimal decision (in the sense of probability) utilizing the approach, without any advanced knowledge.

This paper is organized as following. Section 2 defines the problem solved in this paper with related mathematical background. Then, the dynamic QoS structure is explained in Section 3. On this basis, the stopping time is proved to be bounded in Section 4 to ensure the existence of the optimal solution. The statistical approach for calculating stopping time is given in Section 5. Section 6 evaluates the proposed method by quantitative experiments. Finally, Section 7 concludes the paper.

## 2. Mathematical preliminaries and problem definition

The coming service providers construct a discrete stochastic process, in which a user wants to find an optimal moment of stopping waiting. To achieve this, the process in which the decision is performed needs analysis. Mathematical preliminaries are given below so as to facilitate the discussion [18].

- Let $(\Omega, A, P)$ denote a probability space, where $\Omega$ is a space, $A$ is a $\sigma$-field in $\Omega$, and $P$ is the probability measure of $A$.
- Discrete Filtration is defined as an increasing sequence $\mathcal{F} = \{\mathcal{F}_n\}$ of $\sigma$-fields in $\Omega$;
- A process $X$ is adapted to filtration $\mathcal{F}$, i.e. $\mathcal{F}$-adaptable, if and only if $X_n$ is $\mathcal{F}_n$-measurable for every $n \in Z_+$.
- $\{\mathcal{F}_n = \sigma\{X_{n'}; n' \leq n\}, n \in Z_+\}$ is called $X$-induced filtration. Here $\sigma(C)$ denotes the smallest $\sigma$-field in $\Omega$ containing an arbitrary class $C$ of subsets of $\Omega$.
- A submartingale is defined as a process $X$ with $X_n \leq \mathrm{E}[X_{n+1}|\mathcal{F}_n]$ for every $n \in Z_+$. Here $\mathcal{F}_n$ is $X$-induced filtration.
- A random time $N \in Z_+ + \{\infty\}$ is said to be a $\mathcal{F}$-stopping time if $\{N \leq n\}$ is $\mathcal{F}_n$-measurable for every $n \in Z_+$, that is, if the process $\{X_n = 1(N \leq n)\}$ is adapted to $\mathcal{F}$. (Here and in similar cases, the prefix $\mathcal{F}$ of $\mathcal{F}$-stopping time is omitted when there is no risk for confusion.)
- Say that a stopping time $N$ is bounded if $N \leq n$ for some $n \in Z_+$.
- Normal distribution has reproduction property, i.e., a linear function of any independent normally distributed variables is itself normally distributed [19].

Let us focus on a kind of service $S$ whose quality is identified by $m$ independent QoS indices $Q = \{Q_1, Q_2, \ldots, Q_m\}$. When the $i$th provider $S_i$ with the corresponding QoS value $\{q_{i1}, q_{i2}, \ldots, q_{im}\}$ arrives, the evaluation function (i.e. payoff function) is

$$v_n(S_i) = \sum_{j=1}^{m} \alpha_j q_{ij}^{(n)}. \tag{1}$$

Here $\{\alpha_j | j = 1, 2, \ldots, m\}$ are constants that a user chooses according to his fancy. The superscript $(n)$ of $q_{ij}$ indicates $q_{ij}$ changes when waiting is extended, e.g. its QoS value is $q_{ij}^{(n)}$ when the $n$th provider comes.

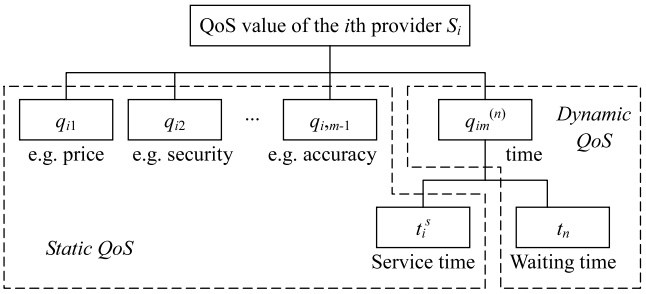


**Fig. 1.** A typical dynamic QoS structure.

At this moment $t_n$, the maximum value of $v_n(S_i)$ among all the providers is denoted as $v_n$, i.e.

$$v_n = \max_{i=1}^{n} v_n(S_i). \tag{2}$$

What the service user looks for is a positive integer $N$ so that $\{v_n\}_{n=1}^{\infty}$ reach its maximum when $n = N$, i.e. $v_N = \max_{n=1}^{\infty} v_n$.

Since $v_n$ is a discrete stochastic process, the optimal Problem of Selecting services Online in a Dynamic QoS environment (PSODQ) can only be defined in the sense of conditional mathematical expectation, as follows.

**Definition 1.** Let $\{\mathcal{F}_i | i = 1, 2, \ldots\}$ be $\{v_n\}$-induced filtration, then PSODQ is to find out the stopping time $N$ so that $\mathrm{E}(v_n|\mathcal{F}_N) < v_N$ for all $n > N$.

It needs illustrating first that the positive integer $N$ in Definition 1 is surely a stopping time because $\mathrm{E}(v_n|\mathcal{F}_N)$ and $v_N$ are $\mathcal{F}_N$-adaptable for all $n > N$. At stopping time $N$, the evaluation function $v_n$ of service QoS reaches its upper bound.

In order to achieve the stopping time in Definition 1, the stochastic process $\{v_n\}$ and its corresponding QoS structure is studied in the next section.

## 3. Dynamic QoS structure

Assuming the first $m - 1$ QoS values $q_{ij}$ $(j = 1, 2, \ldots, m - 1)$ of the $i$th provider $S_i$ are static QoS values depending on $S_i$ itself, such as unit price, quality level, packaging, shipment, etc. (Hence the superscript $(n)$ of $q_{ij}^{(n)}$ is omitted for $j = 1, 2, \ldots, m - 1$.) And the last value $q_{im}^{(n)}$ represents a special QoS value, the total time of the service including service time $t_i^s$ and waiting time $t_n$, i.e.

$$q_{im}^{(n)} = t_i^s + t_n. \tag{3}$$

Here $t_i^s$ is a static QoS value, similar as $\{q_{i1}, q_{i2}, \ldots, q_{i,m-1}\}$, while $t_n$ increases to $t_{n+1}, t_{n+2}, \ldots$, along with the time moving forward.

This constructs a dynamic QoS structure, as shown in Fig. 1.

$$\text{Let } q_i = \sum_{j=1}^{m-1} \alpha_j q_{ij} + \alpha_m t_i^s. \tag{4}$$

Substituting Eqs. (3) and (4) into the equation above, we have

$$v_n(S_i) = q_i + \alpha_m t_n.$$

Then the Eq. (2) is

$$v_n = \max_{i=1}^{n} q_i + \alpha_m t_n.$$

Assuming $q_{ij} (i = 1, 2, \ldots;\ j = 1, 2, \ldots, m - 1)$ follows a normal distribution $N(\mu_j, \sigma_j)$, and $t_i^s$ follows the normal distribution $N(\mu_m, \sigma_m)$, independently. Since $\alpha_j$ $(j = 1, 2, \ldots, m)$ is a constant, and $q_{ij}$ $(j = 1, 2, \ldots, m - 1)$ and $t_i^s$ are independent static QoS values determined by provider itself, $q_i$ is normal distributed because of the reproduction property of normal distribution [19]. So let us assume it follows $N(\mu, \sigma)$.

And the service providers are assumed to arrive according to Poisson process with arrival rate $\lambda$, i.e. the time difference series $\{t_{n+1} - t_n\}$ follows negative exponential distribution. It is always correct that $\alpha_m < 0$ because a shorter time means a better payoff for most service users.

All the assumptions in this paper are listed above, which include only reasonable ones like normal distribution and Poisson process. But all the parameters such as $\lambda$, $\mu$, or $\sigma$ are unknown, which challenges the solution in the paper and will be solved in Section 5.

## 4. Bounded stopping time and solution existence

It seems that the user has to predict $\mathrm{E}(v_n|F_N)$ for every $n > N$ at time $N$, according to Definition 1. Although $\mathrm{E}(v_n|F_N)$ is predictable since $\mathrm{E}(v_n|F_N)$ is $F_N$-adaptable, it is still not practical for the user to predict all the $v_n$ because all the $n > N$ compose an infinite set. Thus, it is necessary for us to solve such infinite determination problem within finite steps while the global optimization is achieved.

A stopping time can be reached in a finite time if it is bounded. In fact, the stopping time in Definition 1 possesses this characteristic, which is approved as follows.

**Lemma 1.** *$\mathrm{E}(v_{n+2}|\mathcal{F}_N) - \mathrm{E}[v_{n+1}|\mathcal{F}_N] \leq \mathrm{E}[v_{n+1}|\mathcal{F}_N] - \mathrm{E}[v_n|\mathcal{F}_N]$ for all positive integers $n \geqslant N$.*

**Proof.**

$$\begin{aligned}\mathrm{E}[v_{n+1}|\mathcal{F}_N] - \mathrm{E}[v_n|\mathcal{F}_N] &= E\left(\max_{i=1}^{n+1} q_i + \alpha_m t_{n+1} - \max_{i=1}^{n} q_i - \alpha_m t_n|\mathcal{F}_N\right)\\ &= E\left(\max\left(0, q_{n+1} - \max_{i=1}^{n} q_i\right) + \alpha_m(t_{n+1} - t_n)|\mathcal{F}_N\right).\end{aligned}$$

Because Poisson process increases independently, the above expression equals

$$E\left[\max\left(0, q_{n+1} - \max_{i=1}^{n} q_i\right)|F_N\right] + \frac{\alpha_m}{\lambda}.$$

So,

$$\begin{aligned}&\mathrm{E}(v_{n+2}|F_N) - \mathrm{E}[v_{n+1}|F_N] - \{\mathrm{E}[v_{n+1}|F_N] - \mathrm{E}[v_n|F_N]\}\\ &= E\left[\max\left(0, q_{n+2} - \max_{i=1}^{n+1} q_i\right)|F_N\right] + \frac{\alpha_m}{\lambda} - E\left[\max\left(0, q_{n+1} - \max_{i=1}^{n} q_i\right)|F_N\right] - \frac{\alpha_m}{\lambda}\\ &= E\left[\max\left(0, q_{n+2} - \max_{i=1}^{n+1} q_i\right) - \max\left(0, q_{n+1} - \max_{i=1}^{n} q_i\right)|F_N\right]\\ &\leq E\left[\max\left(0, q_{n+2} - \max_{i=1}^{n} q_i\right) - \max\left(0, q_{n+1} - \max_{i=1}^{n} q_i\right)|F_N\right].\end{aligned}$$

The above expression equals 0 because $q_{n+1}$ and $q_{n+2}$ follows same distribution.
Therefore, $\mathrm{E}(v_{n+2}|F_N) - \mathrm{E}[v_{n+1}|F_N] \leq \mathrm{E}[v_{n+1}|F_N] - \mathrm{E}[v_n|F_N]$. □

**Theorem 1.** *$\{v_n\}$ is a submartingale $\longleftrightarrow$. For all $N$ existing where $n > N$ so that $\mathrm{E}(v_n|F_N) \geqslant v_N$.*

**Proof** (*Counterevidence*)**.** If $\{v_n\}$ is not a submartingale, then there must exist $N \geqslant 1$ so that $\mathrm{E}(v_{N+1}|F_N) < v_N$. Then for any $n > N$, we have

$$\mathrm{E}(v_n|F_N) - v_N = \mathrm{E}(v_n|F_N) - \mathrm{E}(v_{n-1}|F_N) + \sum_{i=N}^{n-2}[\mathrm{E}(v_{i+1}|F_N) - \mathrm{E}(v_i|F_N)].$$

From Lemma 1, the above expression

$$\begin{aligned}&\leq 2[\mathrm{E}(v_{n-1}|F_N) - \mathrm{E}(v_{n-2}|F_N)] + \sum_{i=N}^{n-3}[\mathrm{E}(v_{i+1}|F_N) - \mathrm{E}(v_i|F_N)]\\ &\leq 3[\mathrm{E}(v_{n-2}|F_N) - \mathrm{E}(v_{n-3}|F_N)] + \sum_{i=N}^{n-4}[\mathrm{E}(v_{i+1}|F_N) - \mathrm{E}(v_i|F_N)]\\ &\leq \cdots\\ &\leq (n-N)[\mathrm{E}(v_{N+1}|F_N) - \mathrm{E}(v_N|F_N)] < 0.\end{aligned}$$

This brings a contradiction.
Conversely, assume that $\{v_n\}$ is a submartingale. Then $\mathrm{E}(v_n|F_{n-1}) < v_{n-1}$. Hence for any $n > N$:
$\mathrm{E}(v_n|F_N) = \mathrm{E}(\mathrm{E}(v_n|F_{n-1})|F_N) \geqslant \mathrm{E}(v_{n-1}|F_N)$.
Consequently, $\mathrm{E}(v_{n-1}|F_N) \geqslant \mathrm{E}(v_{n-2}|F_N) \geqslant \cdots \geqslant \mathrm{E}(v_N|F_N) = v_N$.
Thus, $\mathrm{E}(v_n|F_N) \geqslant v_N$ holds.
The following corollary is the negative proposition of Theorem 1. □

**Corollary 1.** *$\{v_n\}$ is not a submartingale $\longleftrightarrow$. $N$ exist so that $\mathrm{E}(v_n|F_N) < v_N$ for all $n > N$.*

**Lemma 2.** $\lim_{n\to\infty} \mathrm{E}(v_n) = -\infty$.

**Proof.** Denote the intensity function and distribution function of the normal distributed variable $q_i$ as $f$ and $F$. Then

$$E\left(\max_{i=1}^{n} q_i\right) = \int_{-\infty}^{\infty} x\mathrm{d}F^n = \int_{-\infty}^{\infty} nxfF^{n-1}\mathrm{d}x.$$

Because $f$ and $F$ are non-negative, then the above expression

$$\begin{aligned}
&< \int_0^{\infty} nxfF^{n-1}\mathrm{d}x \\
&= \int_0^{\sqrt{n}} nxfF^{n-1}\mathrm{d}x + \int_{\sqrt{n}}^{\infty} nxfF^{n-1}\mathrm{d}x \\
&< \int_0^{\sqrt{n}} n\sqrt{n}fF^{n-1}\mathrm{d}x + \int_{\sqrt{n}}^{\infty} x^3fF^{n-1}\mathrm{d}x \\
&< \sqrt{n}F^{n-1}\,|_{\mu}^{\sqrt{n}} + \int_{\sqrt{n}}^{\infty} x^3f\,\mathrm{d}x \\
&< \sqrt{n} + \int_{\sqrt{n}}^{\infty} x^3\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x \\
&< \sqrt{n} + \int_{\sqrt{n}}^{\infty} x^3\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x-\mu}{\sqrt{2}\sigma}}\mathrm{d}x \\
&= \sqrt{n} - \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{x-\mu}{\sqrt{2}\sigma}}\left(\frac{x^3}{\sqrt{2}\sigma} + \frac{3x^2}{2\sigma^2} + \frac{3x}{\sqrt{2}\sigma^3} + \frac{3}{2\sigma^3}\right)|_{\sqrt{n}}^{\infty} \\
&= o(n).
\end{aligned}$$

Since $\alpha_m$ is negative, $\alpha_m t_n$ is -O($n$) because of the property of Poisson process. Therefore,

$$\mathrm{E}(v_n) = E\left(\max_{i=1}^{n} q_i + \alpha_m t_n\right) \xrightarrow{n\to\infty} -\infty. \quad \square$$

**Theorem 2.** *The stopping time N in Definition 1 is bounded.*

**Proof.** According to Definition 1 and Corollary 1, it is only needed to prove that $\{v_n\}$ is not a submartingale.

(Counterevidence) If $\{v_n\}$ is a submartingale, then for any $n$, we have:

$$\mathrm{E}(v_n) = \mathrm{E}[\mathrm{E}(v_n|F_{n-1})] \geqslant \mathrm{E}(v_{n-1}) \geqslant \cdots \geqslant \mathrm{E}(v_1).$$

Since $\mathrm{E}(v_1)$ is a finite number, $\{\mathrm{E}(v_n)\}$ is a monotonically increasing sequence of $n$, then $\lim_{n\to\infty} \mathrm{E}(v_n) \neq -\infty$, inconsistent with Lemma 2. Therefore, Theorem 2 holds by counterevidence. $\square$

Theorem 2 ensures the existence and availability of the optimal solution of the problem, according to the definition of bounded stopping time. The next section proposes the statistic approach to obtain the solution.

## 5. Local martingale difference approach

The service user has to make a decision whether next coming provider is worth waiting for to obtain a better $v_n$ after the first $n$ providers come. It seems that the user has to predict $\mathrm{E}(v_n|F_N)$ for every $n > N$ at time $N$, according to Definition 1.

Although $\mathrm{E}(v_n|F_N)$ is predictable since $\mathrm{E}(v_n|F_N)$ is $F_N$-adaptable, it is still unacceptable for the user to predict all the $v_n$ because all the $n$ compose an infinite set. Thus, it is necessary for us to solve such an infinite determination problem within finite steps while the global optimization is achieved.

For this problem, it can be indicated from the proof of Theorem 1 that the following sets are equal to each other.

$$\{N|\mathrm{E}(v_{N+1}|F_N) < v_N,\ N \in Z_+\} = \{N|\mathrm{E}(v_n|F_N) < v_N, N \in Z_+,\ \text{for all } n \geqslant N\}$$

Therefore the minimum elements of the above sets are equal too. So the minimum $N$ that holds $\mathrm{E}(v_{N+1}|F_N) < v_N$ is exactly the minimum $N$ that holds $\mathrm{E}(v_n|F_N) < v_N$ for all $n > N$, so that is the optimal stopping time of Definition 1. Hence it can be determined that the optimal $v_N$ arrives, in the sense of mathematical expectation, once $\mathrm{E}(v_{N+1}|F_N) < v_N$ is detected.

In other words, the user needs only to wait until the first $\mathrm{E}(v_{N+1}|F_N) < v_N$ emerges, which solves the problem by simplifying the global problem of verifying all $n > N$ to that of checking the local martingale difference $\mathrm{E}(v_{n+1}|F_n) - v_n$ only. This is the key point of the proposed approach in this paper.

Denote $u_n = \max_{i=1}^{n} q_i$, then the martingale difference $\mathrm{E}(v_{n+1}|F_n) - v_n$

$$\begin{aligned}
&= \mathrm{E}(u_{n+1}|F_n) - u_n + \mathrm{E}(\alpha_m t_{n+1} - \alpha_m t_n) \\
&= \int_{u_n}^{\infty} (x - u_n) \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + \frac{\alpha_m}{\lambda} \\
&= \int_{u_n}^{\infty} (x - \mu) \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + (\mu - u_n) \int_{u_n}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x + \frac{\alpha_m}{\lambda} \\
&= \frac{\sigma}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(u_n-\mu)^2}{2\sigma^2}} + (\mu - u_n)[1 - F(u_n)] + \frac{\alpha_m}{\lambda}. \qquad (5)
\end{aligned}$$

However, the parameters $\lambda$, $\mu$, and $\sigma$ are all unknown, so it is unclear whether the above expression is positive or negative. A statistical method is utilized here to solve the problem. Namely, unbiased estimators are established based on the sample sequence $\{q_i\}$, and applied in the expression (5), instead of the unknown parameters. The derived estimation of the martingale difference is then used for determination.

Here the unbiased estimators are chosen as follows.

$$\hat{\lambda} = \frac{n}{t_n}, \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} q_i}{n}, \qquad \hat{\sigma} = \frac{\sum_{i=1}^{n} (q_i - \hat{\mu})^2}{n-1}. \qquad (6)$$

Therefore, the service user calculates the martingale difference (5) according to unbiased estimators (6) when each service provider comes. Whenever the difference turns negative, the optimal service provider, in the sense of expected value, is obtained while the stopping time defined in Definition 1 is reached.

The algorithm to implement the above methodology is listed as below.

1° A user publishes the service requirement in the SOA.
2° The user waits until the next, i.e. the $n$th service provider responding for the requirement.
3° The QoS values $\{q_{n1}, q_{n2}, \ldots, q_{nm}\}$ including the time $t_n$ and $t_n^s$ that the coming provider announces are collected.
4° Calculate $q_n = \sum_{j=1}^{m-1} \alpha_j q_{nj} + \alpha_m t_n^s$.
5° Estimate the parameters $\hat{\lambda} = \frac{n}{t_n}$, $\hat{\mu} = \frac{\sum_{i=1}^{n} q_i}{n}$, $\hat{\sigma} = \frac{\sum_{i=1}^{n} (q_i - \hat{\mu})^2}{n-1}$.
6° Calculate the martingale difference $\frac{\sigma}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(u_n-\hat{\mu})^2}{2\hat{\sigma}^2}} + (\hat{\mu} - u_n)[1 - F(u_n)] + \frac{\alpha_m}{\hat{\lambda}}$, where $F(u_n)$ is obtained from any normal distribution table.
7° If the calculated martingale difference is positive, turn to 1°.
8° Or else, stop waiting, and start binding the service with the provider with maximum $q_i$ so far.

The methodology for online service selection described above is called the local martingale determination (LMD) approach. LMD is acquainted with the providers' distribution gradually during the stochastic process, and provide the users a condition to determine whether to stop waiting.

The performance of LMD is simulated in the next section.

## 6. Simulations

In order to verify the efficiency of LMD, 300 service providers with the same functionalities are generated according to Poisson process with rate $\lambda = 0.2/\mathrm{s}$. The QoS function $q_n$ can be simplified to a combined variable in the experiment because of the reproduction property of normal distribution [19]. The expectation value and standard deviation of the normal distributed QoS variable $q_n$ is set to $\mu = 100$ s and $\sigma = 300$ s respectively. $\alpha_m$, the weight of time in the evaluation function $v_n$, is set to $-0.2/\mathrm{s}$.

Fig. 2 shows a typical process of an experiment, in which the red curve reflects the evaluation function $v_n$ along with the time. The martingale difference, the blue curve in Fig. 2., intersects downward the horizontal 0-line for the first time at the time 407.74 s, hence it turns negative after that. This event triggers LMD to stop waiting and achieve a result 707.32 which is such a satisfying payoff since it is only 4.03% lower than the global optimal value 737.02.

To compare the proposed approach against traditional ones, three determination methods suggested in [16], as described below, are executed in same simulation environment.

- WEN (Waiting until Enough Number of providers). A threshold defining the maximum number of providers is chosen in advance. The user waits until providers more than the threshold arrive.
- WET (Waiting until Enough Time). The user waits until the waiting time exceeds a threshold set before he starts waiting.
- WEP (Waiting until Enough Payoff). The user sets a QoS value acceptable. And then, skips all providers unable to reach that level until a provider achieves it finally.

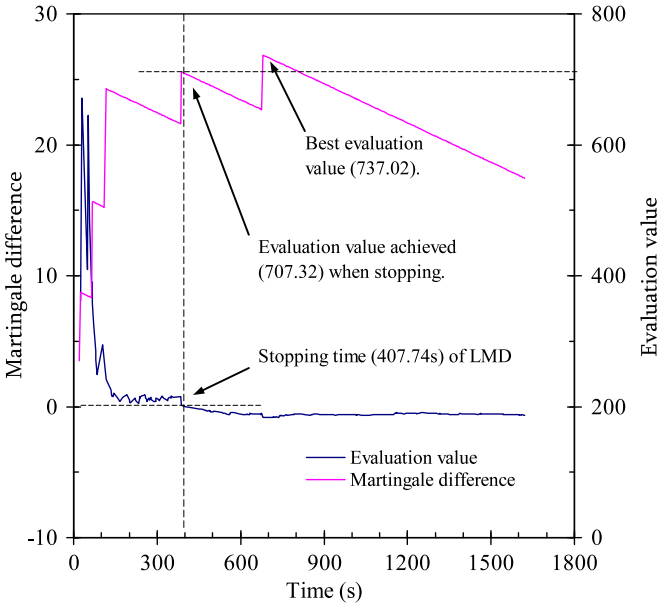


**Fig. 2.** The performance of LMD in a typical experiment.

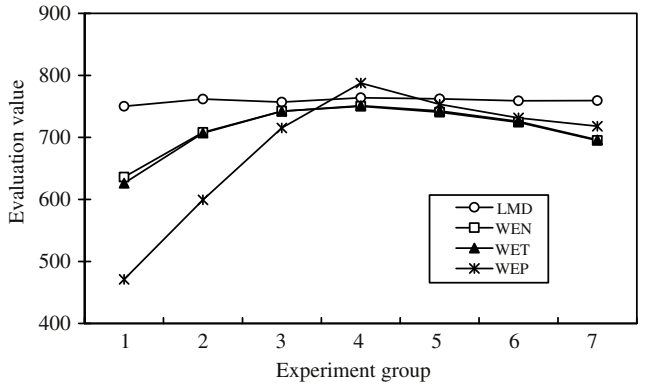


**Fig. 3.** Performance of the four methods under various thresholds.

**Table 1**
Thresholds set for experiment groups.

| Experiment group | Threshold for the methods | | |
|---|---|---|---|
| | WEN | WET (s) | WEP |
| 1 | 20 | 100 | 300 |
| 2 | 40 | 200 | 450 |
| 3 | 80 | 400 | 600 |
| 4 | 120 | 600 | 750 |
| 5 | 160 | 800 | 900 |
| 6 | 200 | 1000 | 1050 |
| 7 | 240 | 1200 | 1200 |

8 groups of experiments are designed to compare LMD with the 3 methods described above. The thresholds for the experiment groups are listed in Table 1.

1000 simulation experiments are implemented for each group of threshold combinations.

The stopping time of the three methods other than the LMD depends on thresholds set before experiments. So they gain different evaluation in the 7000 experiments in the 7 groups with various thresholds. The average evaluation value of the 1000 simulations of each group is calculated separately, and listed in Fig. 3.

**Table 2**
The comparison between LMD and the best possible payoff.

| Experiment group | Best payoff | Payoff of LMD | LMD/best (%) |
|---|---|---|---|
| 1 | 838.0976234 | 750.1281238 | 89.50 |
| 2 | 845.6486432 | 761.8814627 | 90.09 |
| 3 | 841.7189020 | 756.8386247 | 89.92 |
| 4 | 848.7235743 | 764.0301300 | 90.02 |
| 5 | 845.4925235 | 762.0501038 | 90.13 |
| 6 | 848.5182703 | 759.1493571 | 89.47 |
| 7 | 839.4017429 | 759.5354315 | 90.49 |
| Average | 843.9430399 | 759.0876048 | 89.95 |

It can be concluded from Fig. 3. that

(1) The evaluation values of the three traditional methods vary in experiment groups, which imply that their performance depends on the chosen thresholds. WEN, WET, and WEP reach their best payoff when their thresholds are set around 120, 600 s, and 750 respectively. On the contrary, the evaluation value of LMD keeps around 750 stably, independent of any thresholds, so that is superior to the other methods in most cases.
(2) LMD overcomes WEN and WET in all the experimental groups, with an evaluation value up to 20% greater than that of WEN or WET.
(3) LMD performs better than WEP in most cases. Only when the threshold chances to fall into a narrow neighborhood around 750 can WEP slightly surpass LMD. But it is worth noticing that there is little possibility in which the user of WEP happens to choose the right threshold because such multi-group experiments are impossible to be handled practically in advance.

It is difficult for the three traditional methods to predict a good threshold without a priori knowledge. On one hand, if a too small threshold is chosen, the process will be terminated too early to return a satisfying provider. On the other hand, the payoff will decrease gradually during the useless waiting process. Therefore, independent of any threshold, LMD has extreme superiority over traditional ways.

Similar with Fig. 2., the best evaluation value of all experiments are recorded to evaluate the performance of LMD. Table 2 compares average payoff of LMD in each experimental group with that of the actual best result.

It can be concluded from Table 2 that LMD stabilizes its performance around 90% of the optimal payoff. As a probability-based approach without a priori knowledge, LMD keeps so close to the best practical result that it represents an outstanding efficiency in solving online service selecting problem.

## 7. Conclusion

After sending an inquiry on a specific service, a user in SOA has to evaluate the responding candidates one by one and make an online decision of whether to continue to wait for a possible provider with a better QoS. Different from the traditional methods, the LMD proposed in the paper requires neither predetermined threshold nor accurate parameters of the providers' distribution. So LMD becomes the first self-adaptable methodology in the domain since it makes the decision based on the dynamic information all gathered during the process of waiting, without any manual intervention.

Another contribution of this paper is giving consideration to both the static and dynamic indices of the QoS, and develops the stochastic process methods suitable for any linear QoS structure under a multi-index environment. It is expected to benefit applications extensively, such as service selecting, resource scheduling, project management, etc.

LMD not only ensures the stopping time is bounded (i.e. the feasibility in finite steps), according to Theorem 2, but also simplifies the global estimation over the time horizon to a local determination based on current martingale difference. Therefore, the user is able to gain the stochastic optimal solution within a limited time.

The future work might focus on the following aspects.

(1) The assumption of a normal or Poisson population in the paper is popularly observed thus adoptable in most real environments. It might needs deduction research for other populations when the idea of LMD is applied in special cases.
(2) The algorithm's performance depends partially on the accuracy of the estimators in (6). It has been noticed during the experiments that when a larger $\alpha_m$ is set, the process tends to be terminated earlier (i.e. at smaller $N$) because of the heavier time punishment. In this case, the bias of the estimators becomes greater since not enough samples have been collected. Therefore, a better estimator for the expression (5) is necessary to be discovered to improve the LMD further, especially in the case where time is so important that $\alpha_m$ is high enough.

## Acknowledgements

This work is supported by National Natural Science Foundation of China (Key Program, No. 61033005), the National High Technology Research and Development Program of China (863 Program, No. 2009AA010308) and the National Key Technology R&D Program (No. 2008BAH32B03-1). The author would like to thank Hamzeh Sheikhhasan for his precious comments and for the time he spent in reviewing the paper.

## References


[1] W.T. Tsai, Chun Fan, Yinong Chen, Raymond Paul, Jen-Yao Chung, Architecture classification for SOA-based applications, in: Proceedings of the Ninth IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing, IEEE Computer Society, Washington DC, USA, 2006, pp. 295–302.

[2] Boualem Benatallah, Hamid R. Motahari Nezhad, Service oriented architecture: overview and directions, in: E. B¨orger, A. Cisternino (Eds.), Software Engineering, in: LNCS, vol. 5316, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 116–130.

[3] Sun Meng, Farhad Arbab, QoS-driven service selection and composition using quantitative constraint automata, Fundamenta Informaticae 95 (2009) 103–128.

[4] Du, Qixing, Chi Chi-Hung, Chen Shuo, Deng Jiamning, Modeling service quality for dynamic QoS publishing, in: Proceedings of IEEE International Conference on Services Computing, 2008, 2, Honolulu, USA, IEEE Comp Soc, TCSC, 2008, pp. 307–314.

[5] Pengcheng Xiong, Yushun Fan, Mengchu Zhou, Web service configuration under multiple quality-of-service attribute, IEEE Transactions on Automation Science and Engineering 6 (2) (2009) 311–321.

[6] Wu Bang-Yu, Chi Chi-Hung, Xu Shi-Jie, Gu Ming, Sun Jia-Guang, QoS requirement generation and algorithm selection for composite service based on reference vector, Journal of Computer Science and Technology 24 (2) (2009) 357–372.

[7] F. Curbera, R. Khalaf, N. Mukhi, Quality of service in SOA environments, an overview and research agenda, IT-Information Technology 50 (2) (2008) 99–107.

[8] Pengcheng Xiong, Yushun Fan, Mengchu Zhou, QoS-aware web service configuration, IEEE Transactions on System, Man and Cybernetics, Part A 38 (4) (2008) 888–895.

[9] Chen-Feng Wu, Liang-Teh Lee, Der-Fu Tao, An HMM prediction and throttling-based call admission control scheme for wireless multimedia networks, Computers and Mathematics with Applications 54 (3) (2007) 364–378.

[10] T. Tansupasiri, K. Kanchanasut, C. Barakat, P. Jacquet, Using active networks technology for dynamic QoS, Computer Networks: The International Journal of Computer and Telecommunications Networking 50 (11) (2006) 1692–1709.

[11] Yimin Shen, Jing Zhang, Yushun Fan, Multi-index cooperative mixed strategy for service selection problem in service-oriented architecture, Journal of Computers 3 (8) (2008) 69–76.

[12] Gregorio Díaz, María-Emilia Cambronero, M. Llanos Tobarra, Valentín Valero, Fernando Cuartero, Analysis and verification of time requirements applied to the web services composition, in: M. Bravetti, M. Nues, G. Zavattaro (Eds.), WS-FM 2006, in: LNCS, vol. 4184, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 178–192.

[13] Hong Qing Yu, Stephan Reiff-Marganiec, Automated context-aware service selection for collaborative systems, in: P. van Eck, J. Gordijn, R. Wieringa (Eds.), CAiSE 2009, in: LNCS, vol. 5565, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 261–274.

[14] Milad Pastaki Rad, et al., A survey of cloud platforms and their future, in: O. Gervasi (Ed.), ICCSA 2009, Part I, in: LNCS, vol. 5592, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 788–796.

[15] Gang Li, Hongmei Sun, Huahao Gao, Haiyan Yu, Yue Cai, A survey on wireless grids and clouds, in: 2009 Eighth International Conference on Grid and Cooperative Computing, GCC. Lanzhou, China, 2009, pp. 261–267.

[16] Yi-Min Shen, Yu-Shun Fan, Online selection approach for service composition in enterprises coordination, Computer Integrated Manufacturing Systems 14 (4) (2008) 799–805.

[17] Yasmine Charif-Djebbar, Nicolas Sabouret, Dynamic web service selection and composition: an approach based on agent dialogues, in: A. Dan and W. Lamersdorf (Eds.), ICSOC, LNCS, vol. 4294, 2006, pp. 515–521.

[18] Olav Kallenberg, Foundations of modern probability — Probability and Its Applications, second edition, Springer-Verlag, New York, Berlin, Heidelberg, 2002, pp. 46–126.

[19] Harald Cramér, Mathematical Methods of Statistics, Princeton University Press, US, 1999, 9th printing, pp. 213.