A Survey of Time-Aware Dynamic QoS Forecasting Research, Its Future Challenges and Research Directions

Yang Syu[1](✉), Chien-Min Wang[1], and Yong-Yi Fanjiang[2]

[1] Institute of Information Science, Academia Sinica, Taipei City, Taiwan
{yangsyu, cmwang}@iis.sinica.edu.tw,
a29066049@gmail.com
[2] Department of Computer Science and Information Engineering,
Fu Jen Catholic University, New Taipei City, Taiwan
yyfanj@csie.fju.edu.tw

Abstract. The problem of time-aware (time series-based) dynamic quality of service (QoS) forecasting has attracted increased attention over the past decade. Developed forecasting approaches have been used to obtain the future values of dynamic QoS attributes for the support of the proactive decisions of various QoS-based applications (e.g., QoS-aware service selection and composition). Thus far, however, a comprehensive investigation and overview of the current research on this topic has yet to be produced. This paper proposes and introduces six assessment criteria which are then applied to the existing literature to produce a comprehensive comparison. Based on this analysis, we describe potential future challenges and research directions in this research area, focusing on gaps in the current literature. This survey provides a clear understanding of the current status of this research area with this paper; additionally, we also technically point out what have to be done by the researchers in this area for the advance of this research topic.

Keywords: Time-aware dynamic QoS attributes · Time series forecasting
Web services

1 Introduction

Many software systems rely on external Cloud/Web services to provide required functions or information, and new mobile applications increasingly incorporate existing services (e.g., the RESTful Web APIs exposed by Facebook, Google, and Instagram). The quality of these services (called quality of service, QoS) is an important concern to developers/engineers in service-oriented software engineering (SOSE) and researchers in service computing (SC). Among current QoS-related research topics and concerns, particular attention has focused on the prediction of dynamic QoS properties, such as service response time.

In both SOSE and SC, QoS is mostly just a general and conceptual term covering many different concrete properties and attributes, such as availability, throughputs, reputation (these are domain-independent, general QoS properties), and various

J. E. Ferreira et al. (Eds.): SCC 2018, LNCS 10969, pp. 36–50, 2018.
https://doi.org/10.1007/978-3-319-94376-3_3

domain-specific QoS attributes. A detailed introduction and explanation to QoS in the context of Web services can be found in Sect. 2 of [1]. For disparate QoS properties and attributes, a binary classification categorizes them into *static QoS* and *dynamic QoS*. The values of the former type of QoS are fixed or rarely changed. However the values of dynamic QoS could vary with one of the QoS dynamic factors discussed below. Because the real values of dynamic QoS are uncertain and unknown, a prediction method is needed, and several attempts have been made.

Our survey of the relevant literature finds that, when calling/using the same service, two factors have been identified as causing QoS value variation: *disparate service consumers* and *different invoking times* [2]. When invoking a service, it is likely to receive different dynamic QoS values for different service consumers because of the heterogeneity of their internal and external environments, as demonstrated and discussed in [3]. Furthermore, actual QoS values could even be changeable and dynamic in the case of the same user using a single service at different times. We identify and consider the predictions for these two factors as two different research areas (i.e., *consumer-aware* dynamic QoS prediction and *time-aware* dynamic QoS forecasting). This paper focuses on the later, namely, the research efforts targeting time-aware QoS variations and their forecasting.

Our survey identifies a number of relevant findings, but there is currently no systematic review or analysis of these works to provide an informative comparison of the properties and inadequacies of the various proposed approaches. This paper first conducts a detailed review and analysis of current time-aware (time series-based) dynamic QoS forecasting research, comparing various approaches in terms of the six identified criteria. Based on these observations, we then propose and discuss future challenges and research directions that must be considered and addressed to further advance this research area.

The first contribution of this paper is a detailed comparison and discussion of current time-aware QoS forecasting studies. This comparison allows the reader to quickly understand the current status of this research area, the pros and cons of each reviewed work, and the concerns that must be considered and addressed for this research topic. For our comparison, six common characteristics are identified as criteria for comparison: (1) year published, (2) problem specifications defined and addressed, (3) forecasting methods proposed/employed and methods used for performance comparison, (4) measures considered for performance evaluation, (5) dynamic QoS datasets tested, and (6) QoS-aware applications combined. These criteria can be viewed as key research concerns for this topic and researchers working on this problem should consider them carefully. The second contribution is to identify future research directions and challenges in this area, including (1) the collection of a large-scale, real-world dynamic QoS time series dataset from modern Web/cloud services, (2) a formal and detailed analysis of dynamic QoS data for understanding their intrinsic characteristics and specialties, (3) an investigation on the use of innovative machine-learning (ML) techniques (e.g., Markov chain models and recurrent neural networks, RNNs) and different statistical strategies (e.g.,, cross-approach selection and combination strategies), and (4) the efficient integration of a QoS forecasting approach with its application.

The remainder of this paper is organized as follows. Section 2 presents a survey of the existing research, including an explanation for the assessment criteria identified and adopted, along with a comparison table showing the existing studies in terms of the proposed criteria, and a description and discussion of the cited works. Section 3 lists and discusses potential future work and research challenges in this area. Conclusion are drawn in Sect. 4.

2 A Survey of Time-Aware Dynamic QoS Forecasting

This section reviews current research on time-aware (time series-based) dynamic QoS forecasting to produce a detailed comparison table. Published works are reviewed and carefully analyzed to extract and identify a number of common properties as the fundamental criteria for comparison and discussion. Below, we first introduce the identified criteria and then apply them to the collected studies for comparison, followed by a discussion of the comparison and the cited works.

2.1 Criteria

Each study to be compared is identified in terms of *the names of its authors* and *the year* it was published. We consider the publication date because SC research and the service industry are developing rapidly, and results from only a few years ago may already be out of date. For example, research performed on an obsolete time-aware dynamic QoS dataset is less relevant than from a recently collected dataset. However, some recent works have performed experiments on ancient QoS datasets [2, 4, 5]. In this case, it would be ideal to consider both the time that a work was reported and when its experimental QoS dataset was made.

Even though the works cited in this paper all focus on time-aware dynamic QoS forecasting, their detailed problem specifications still differ in many ways. Many of the surveyed works consider their studied problem as a case of time series forecasting of which there are many variations, such as univariate prediction and multivariate prediction as well as one-step-ahead forecasting and n-step-ahead forecasting (i.e., the length of forecasting horizon). Thus, one cannot say that the reviewed works all study exactly the same problem. Due to its complexity, we divide the second criterion, *the problem specifications*, into three sub-concerns to accurately differentiate between the reviewed works. The first sub-criterion indicates whether *the formal (mathematical) problem definition* has been presented in the original paper. The second sub-criterion is that the problem defined to study or the experimental results reported comprises *both modeling and forecasting stages or only covers the later*. The last sub-concern is *the forms of the studied problem in terms of time series research*. Due to the diversity of problem specifications, if a formal definition for the studied problem is not provided in the original paper (according to the first sub-criteria), it is difficult to determine its detailed specification (the forms of the studied problem, namely, the third sub-criterion) and to reproduce its (experimental) results for impartial comparison.

The third criterion regards the solutions developed and considered in each work. This criterion is also divided into two sub-criteria, namely, *the approach proposed by* *the authors in a work* and *the methods that are considered for a comparison with the proposed approach*. Note that, among the reviewed works, a few empirical studies just employ a number of well-developed, long-standing (e.g., time series and/or machine learning) techniques to compare their performance on a time-aware QoS dataset. For these works, there would not be any proposed approach for them in our comparison.

Most time-aware QoS forecasting research focuses on numerical QoS attributes (e.g., response times, throughputs, and availability), and determining the performance of the proposed approaches and methods considered for comparison, requiring *a way to quantitatively evaluate and present their results*. Many of the reviewed studies just intuitively adopt some common measures of accuracy in time series research as their metrics, such as mean absolute errors (MAEs) and mean absolute percentage errors (MAPEs). However, some QoS-specific metrics have also been considered, such as the rate at which QoS violations can be correctly predicted. Aside from accuracy, time cost is also a common concern of many existing studies, primarily the time consumed for training/fitting a predictor (predictor generation time, PGT) or for yielding a forecasting value (forecast production time, FPT).

In addition to prediction approaches, another indispensible element for time-aware dynamic QoS forecasting research is *QoS time series datasets used for practical evaluations (experiments)*. Most of the work reviewed for our study consider a real-world, time-aware QoS dataset, with some collecting original real world QoS data, while others rely on publicly available datasets. Note that, as previously mentioned, the age of the QoS dataset is an important concern.

Our final consideration is *whether the applications are combined and used with the proposed QoS forecasting approaches*. For QoS forecasting approaches, actually, it is not meaningful in practice; to be useful, a QoS forecasting approach should be integrated with a QoS-aware application that needs future QoS values to proactively and reliably perform make decisions and/or perform operations, such as design-time QoS-aware service selection/composition and runtime service adaptation (i.e., dynamic binding). In the column for this criterion, we list the application in each reviewed work, and the entry would be empty if the study in question only focuses on the forecasting problem.

2.2 Comparison

Given the criteria proposed and discussed in the previous section, the results of our survey on existing time-aware dynamic QoS forecasting approaches are presented in Table 1, ordered by date of publication.

2.3 Discussion

Based on Table 1, this section discusses current time-aware dynamic QoS forecasting research in terms of the six proposed criteria. Note that, in this section, we also discuss what concerns must be considered, addressed, and presented (i.e., the contents that must be shown for the reader to read and understand) in a paper to be considered referable and valuable research.

Table 1. The survey of time-aware dynamic QoS forecasting research in terms of the criteria proposed.

Author	Year published	Problem specifications			Approaches		Evaluations	Time-aware dynamic QoS datasets	QoS-aware applications combined
		Formal definitions	Stages covered	Problem forms	Models/methods proposed	Models/methods compared			
Yang et al. [4]	2017	Provided	Forecasting	Univariate, one-step-ahead, and single-predictor mode	None	Baseline approaches (Average, three Naïve methods, Drift, and six Regression models), TS approaches (ARIMA family, ES family, two SETAR models, and GARCH family), and ML approaches (ANNs and GP)	Accuracy (MAE and MAPE) and time (PGT and FPT)	Two response-time time series in the dataset provided by [6]	None
Fanjiang et al. [2]	2016	Provided	Forecasting	Univariate, one-step-ahead, and single-predictor mode	GP	Baseline approaches (Random search, Average, three Naïve methods, Drift, and six Regression models), TS approaches (ARIMA models and ES models), and one ML approach (ANNs)	Accuracy (MAE and MAPE) and improvement (RPAI)	One response-time time series in the dataset provided by [6]	None
Ye et al. [7]	2016	Provided	Forecasting	Multivariate, N-step-ahead (long-term), and single-predictor mode	Multivariate ARIMA or Multivariate Holt-Winters, depending on their univariate RMSE modeling accuracy	VAR, (univariate) ARIMA models, and (univariate) Holt-Winters method	Accuracy (RMSE)	A set of short (28 observations only) response-time and throughput time series for 100 real-world cloud services	QoS-aware cloud service composition (selection) based on long-term QoS time series similarity
Nourikhah et al. [5]	2015	None	Forecasting	Difficult to define	ARFIMA	Naïve, Average, and ARIMA	Accuracy (MAE and MAPE) and improvement (RPAI)	The dataset provided by [6]	None
Rehman et al. [8]	2014	None	Modeling and Forecasting	Difficult to define	None	ARIMA and ES	Accuracy (MAE, RMSE, and MASE) and residual diagnosis	Amazon cloud service data provided by CloudClimate (no longer available)	None
Leitner et al. [9]	2013	None	Forecasting	Difficult to define	None	ML approaches (DTs for nominal attributes and ANNs for numerical attributes) and TS approach (ARIMA models)	Accuracy (training/modeling data correction, MAE, and SD of absolute errors for numerical data and precisions and recalls for nominal attributes) and time (PGT and FPT)	The historical execution data of the sub-flow of a real-world business process	Theoretically, proactive detection for SLA violation (i.e., SLA prediction)

(*continued*)

Table 1. (*continued*)

Author	Year published	Problem specifications			Approaches		Evaluations	Time-aware dynamic QoS datasets	QoS-aware applications combined
		Formal definitions	Stages covered	Problem forms	Models/methods proposed	Models/methods compared			
YunNi et al. [10]	2013	None	Forecasting	Difficult to define	ARMA	None	None (overall dependability of a composite service)	The QoS time series collected from eight services	Dependability prediction for WS-BPEL-based composite services
Amin et al. (a). [11]	2012	None	Forecasting	Difficult to define	ARIMA-GARCH (both sequentially combined)	ARIMA	Accuracy (MAPE), improvement (RPAI), and time (PGT and FPT)	The dataset provided by [6]	Simulated proactive detection for QoS violation (i.e., SLA prediction)
Amin et al. (b). [12]	2012	None	Forecasting	Difficult to define	ARIMA or SETARMA (selected based on linearity using statistical test)	ARIMA	Accuracy (MAPE), improvement (RPAI), and time (PGT and FPT)	The QoS time series collected from 800 real-world services	None
Senivongse et al. [13]	2011	None	Forecasting	Difficult to define	ANNs	None	None (not revealed)	The QoS time series collected from 39 real-world/self-made services	QoS-aware service selection with varying granularity
Yilei et al. [14]	2011	Provided	Forecasting	Time-aware and consumer-aware QoS prediction (tensor filling)	Tensor Factorization	Three other less sophisticated tensor factorization approaches	Accuracy (MAE and RMSE)	A 142 × 4532 × 64 (consumes × services × times) QoS data tensors	None
Zadeh et al. [15]	2010	Not well-defined	Modeling and Forecasting	Difficult to define	ANNs	None	Accuracy (MAE, MSE, and percentage of correctly predicted trends)	Not clearly mentioned	Theoretically, designed for replacing QoS monitoring
Cavallo et al. [6]	2010	None	Forecasting	Difficult to define	None	Average, Last Observation (Naive), Linear Regression, and ARIMA	Accuracy (MAPE and SD of percentage errors)	The QoS time series collected from ten real-world services	Simulated detection for QoS violation

(*continued*)

Table 1. (*continued*)

Author	Year published	Problem specifications			Approaches		Evaluations	Time-aware dynamic QoS datasets	QoS-aware applications combined
		Formal definitions	Stages covered	Problem forms	Models/methods proposed	Models/methods compared			
Godse et al. [16]	2010	None	Forecasting	Difficult to define	ARIMA	None	None (only an unclear program execution result was presented)	The QoS time series collected from two real-world services	Theoretically, designed to be combined with QoS-aware service selection
Mu et al. [17]	2009	Provided (but difficult to understand)	Forecasting	Difficult to define	Structural Equation Modeling	None	Accuracy (MAPE)	The QoS time series data collected from self-made services	QoS-aware service selection

The oldest study reviewed in Table 1 dates back almost ten years (i.e., Mu et al. [17] in 2009), and new studies have been published each year since.

Regarding the problem specifications, except for a few of recent works (i.e., Yang et al. [4] in 2017 and Fanjiang et al. [2] and Ye et al. [7] in 2016), most studies did not formally define or present their addressed problem (or they are difficult to understand due to the poor definition or presentation, such as those in Zadeh et al. [15] and Mu et al. [17]). We consider that this lack of formal definition is a series defect for such quantitative research because the lack of clear definitions makes it difficult to infer the research assumptions or to understand the work in details sufficient to reproduce the experimental results for an impartial comparison with newly proposed forecasting approaches. Second, regarding the covered problem stages, only one work (Zadeh et al. [15]) reported the modeling accuracy of their proposed/applied forecasting approach (ANNs). Even though forecasting performance is definitely the most important concern of any time series forecasting task, it could be helpful to also report corresponding modeling/training performance to understand the fitting/modeling capability of the considered approaches/models and for analysis of the original data and forecasting results (e.g., exploring the relationships between the modeling and forecasting performance of an approach/model to assist the development of strategies for cross-approach/model selection or combination). Finally, in terms of the problem forms, lack of formal definitions in most of the reviewed papers (as indicated in the first subcriterion) makes it difficult to fully realize the problem specifications in forms of time series forecasting. Among papers without any formal definitions, some provided a brief text description of their problem form (e.g., they performed one-step-ahead forecasting). However, we consider such information to be incomplete and partial for the inference of a full specification. According to our survey, most of the reviewed studies were assumed to perform univariate, one-step-ahead forecasting, which is the most basic form for the problems in time series research. One exception is Ye et al. [7], in which the authors demonstrated that the correlations between the different dynamic QoS attributes of the same service can sometimes help to improve forecasting accuracy (i.e., performing multivariate prediction), and cloud service consumers raise the need for long-term QoS information (which can be obtained by using N-step-ahead forecasting). Yilei et al. [14] is another exception in that the study problem definition also includes another dynamic QoS factor, namely, different consumers (consumer-aware). The difference in the problem studied in Yilei et al. [14] also results in a different type of the employed solution, as discussed below.

Aside from the tensor factorization used in Yilei et al. [14] and the structural equation modeling adopted in Mu et al. [17], the approaches considered here can be categorized into three different classes, namely, the baseline approaches (including Random Search, Average, Naïve, Drift, and Regressions methods), the statistical time-series methods (Auto-Regressive Integrated Moving Average, Auto-Regressive Fractionally Integrated Moving Average, Exponential Smoothing, Generalized Autoregressive Conditional Heteroskedasticity, and Self-Exciting Threshold Auto-Regressive models), and the machine-learning techniques (Genetic Programming, Decision Trees, and Artificial Neural Networks). In time series research, the baseline approaches were mostly used as the bottom for a comparison with sophisticated approaches that the proposed approach must be at least superior to (i.e., more accurate than) the baseline approaches; otherwise, it is useless and meaningless [18]. According to our survey, in most studies that have compared TS and/or ML approaches with the baseline approaches (including Yang et al. [4], Fanjiang et al. [2], Nourikhah et al. [5], and Cavallo et al. [6]), the TS and ML approaches generally surpass the baseline approaches in terms of forecasting accuracy. Regarding the comparison between the TS approaches and the ML techniques, recent studies (e.g., Yang et al. [4] and Fanjiang et al. [2]) demonstrated that the ML approach GP is superior to the other TS and ML approaches in terms of forecasting accuracy; however, GP raises a tradeoff because it takes more time to train and evolve a predictor (i.e., increased predictor generation time, PGT). As indicated in Yang et al. [4], to reach a compromise between forecasting accuracy and time cost, it is worth considering using the two most widely used TS approaches, namely, ARIMA models and ES methods. Finally, most surveyed works follow a single-approach strategy to address the problem (e.g., using GP or ARIMA models to predict future QoS values regardless of the characteristics of historical QoS data). However, there are also a few exceptions that use a cross-approach selection strategy (Ye et al. [7] and Amin et al. [12]) or a model combination strategy (Amin et al. (a). [11]), with experimental results that suggest these strategies can further improve forecasting accuracy.

For evaluation, forecasting accuracy is the most common and important concern in this research area. The existing research considers a number of time-series measures of accuracy, including mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). Based on these measures, to quantitatively calculate the improvement of a proposed approach over the others in terms of forecasting accuracy, most studies adopt relative prediction accuracy improvement (RPAI) as a numerical indication, including Fanjiang et al. [2], Nourikhah et al. [5], Amin et al. (a). [11], and Amin et al. (b). [12]. Basically, MAE, MAPE, and RMSE (and the other measures of accuracy in time series research, as listed and introduced in [19]) calculate the overall accuracy of a measured approach, integrating and expressing it as a single number for quick comprehension and comparison. To assess the forecasting stability of an approach, some studies, such as Leitner et al. [9] and Cavallo et al. [6], also calculate the standard deviations (SDs) of their original forecasting point errors. Finally, except for accuracy, as shown in Table 1, the cost of time is also an important concern in many of the reviewed studies. Mostly, they consider the time that is required to train, fit, and produce a predictor (i.e., predictor generation time, PGT) and to obtain a set of forecast values through the predictor (forecast production time, FPT).

For time-aware dynamic QoS forecasting, most studies used real-world QoS data to perform their empirical evaluation and experiments. Thus far, the QoS time series dataset collected and provided by Cavallo et al. [6] is most widely used, with application in the early research Cavallo et al. [6] and Amin et al. [11] along with more recent work in Nourikhah et al. [5], Fanjiang et al. [2], and Yang et al. [4]. The other authors performed experiments on original dynamic QoS datasets. These self-collected QoS datasets show no consistency in terms of properties, such as the number of observed services and dynamic QoS attributes or the length of gathered QoS time series. In fact, no widely acceptable standard or consensus exists for the features of a valid time-aware dynamic QoS dataset, and we this may cause problems for research validity. In addition, these self-collected datasets are typically not publicly available, making it difficult or impossible to reproduce the experimental results for comparison with newly proposed forecasting approaches. This lack of a large-scale, widely-acceptable QoS dataset (benchmark) is further discussed in the next section.

Finally, as indicated in Table 1, some research exclusively focuses on forecasting problems, while the others seek to integrate an QoS-aware application with the developed/applied forecasting approaches. A number of the latter works only theoretically describe or discuss the combining of both types of QoS-based approaches, without providing any practical implementation or evaluation, while the others have integrated and tested both types of approaches and reported their performance to demonstrate the usage and influences of time-aware dynamic QoS forecasting approaches. Among the combined QoS-aware applications, most are approaches to provide proactive detection (prediction) for SLA violations and automated service selection (composition). We consider these applications as demonstrating the usefulness and value of time-aware dynamic QoS forecasting approaches.

3 Future Challenges and Research Directions

Despite nearly a decade's work on the time-aware (time series-based) dynamic QoS forecasting problem, there are still a number of challenges and issues that must be addressed:

3.1 Time-Aware Dynamic QoS Datasets

Aside from various forecasting approaches, another indispensable component for any prediction research is the use of a valid and referable dataset (benchmark) for performance evaluation and comparison. However, we consider the QoS datasets used in the current work to be somewhat defective in several different aspects. First of all, as mentioned in the previous section, many of the time-aware QoS datasets are not retrievable, and thus do not allow for the reproduction of experimental results. Second, Cavallo et al.'s QoS dataset [6] and other self-collected datasets are outdated and most likely to not represent current QoS status. For example, a paper published in 2010 uses the Cavallo et al. dataset [6] was recorded and gathered in 2006, thus the underlying data is over a decade old. Most datasets, such as those used in Cavallo et al. [6], Amin et al. [12], Senivongse et al. [13], and Yilei et al. [14], were produced by gathering QoS values from a set of SOAP/WSDL-based Web services. However, currently the RESTful style [20] is the most popular and prevalent form of cloud/Web services. These two types of services have many intrinsic differences; thus, the QoS values obtained from SOAP/WSDL-based Web services may not be representative of today's services. Finally, the properties of certain QoS datasets are not suitable for large-scale experiments. For example, in the dataset used in Ye et al. [7], for each recorded service, its QoS history is 6 months old and has only 28 time slots (i.e., observations, thus the length of a QoS time series is 28). Thus it has too few observations and too long a time interval (6 months/28). Another issue in many datasets is that the number of their QoS time series samples (i.e., the number of services observed) are poor. For example, in the dataset provided in Cavallo et al. [6], only the data from ten real-world services were recorded, which may raise doubts about this dataset's validity because currently there are hundreds of thousands of services available in the real world. Experimental results based on poor quality datasets may not be reliable (i.e., their generality could be questioned).

A time-aware (time series-based) dynamic QoS dataset without the above issues must be considered and developed to offer a robust basis for studies and experiments in this research area. When gathering QoS data for the production of such a dataset, its collection time and duration, the number of observed services and dynamic QoS attributes, and the type of services must be carefully considered in terms of the dataset's validity; otherwise, the dataset must be easy to retrieve or publicly available. A referable time-aware QoS dataset is the third dataset reported and studied in [20] (namely, the dataset used in Yilei et al. [14]). However, this dataset suffers from several shortcomings including its relative age (the data were collected in March 2011), the type of services used (SOAP/WSDL-based Web services), and the length of each QoS time series (only 64 observations). In addition to forecasting performance, we believe that a large-scale, modern QoS time series dataset would also be very useful for understanding the current status of cloud/Web services and networks, as discussed in the next section.

3.2 Statistical Analysis to QoS Time Series

According to our survey, most current research presents empirical studies, without applying formal, statistical analytics to the targeted QoS data processing, modeling, and prediction. For the most part, the forecasting approaches and QoS data were simply tested for performance comparison to other approaches. In other words, the studied QoS time series were block boxes being used as a benchmark to test various approaches, and their intrinsic natures and characteristics remain unclear. A few studies briefly explored the processed QoS data; for example, in the approach proposed in [12], the authors used the Hanset test to check the linearity of the QoS data: the Engle test to verify volatility in [11], and the Hurst exponent to test the existence of long memory in [5]. However, this type of one-sided analysis mainly serves to support or verify of an authors' assumptions regarding the QoS time series and are thus insufficient to provide comprehensive insight into the QoS data.

The characteristics of time-aware QoS data differ from many other types of time series (e.g., financial time series). For example, the response-time time series of real-world Web services feature sudden peaks (i.e., very long response times), potentially caused by short-term network congestion or server problems. But such phenomenon are very rare in financial time series, such as the overall index of a stock market. In the time series of stock market indexes, major fluctuations (i.e., exceeding ten percent of the previous value) are extremely rare, but such oscillations are common in response-time QoS time series due to service and network instability.

An insightful and comprehensive analysis is required of QoS time series focusing on identifying their unique characteristics and specialties (the differences from the other types of time series). We believe that such knowledge would be very helpful in developing dedicated modeling and forecasting approaches for QoS time series, taking advantage of the presented findings (for example, in developing a sophisticated criterion for the cross-approach model selection strategy, as discussed in the next section), and in answering various questions, such as why the machine learning approach GP can outperform conventional time series methods in this kind of forecasting task [2, 4]. Analyzing and exploring a large-scale, modern QoS dataset might also help indicate the current status of services and networks. For example, observing the QoS variations of the services in a given region or from a single provider may indicate stability and usability (i.e., service and/or network conditions).

3.3 Forecasting Approaches and Strategies

According to the relatively recent work by Yang et al. [4] and Fanjiang et al. [2], the ML approach GP has superior forecasting accuracy than the other methods. Another common ML approach, ANNs, is also widely considered in the existing research. We believe that this demonstrates the potential and superiority of ML approaches for such forecasting problems. However, several ML approaches applicable to time series modeling and forecasting have yet to be considered, including support vector machines (SVMs), Markova chains, and some more complicated ANNs developed and used in deep learning (e.g., recurrent neural networks, RNNs). A complete performance evaluation and comparison including all applicable approaches would be informative for this research area.

As noted in Table 1, most studies follow the single-approach strategy (applying only an individual forecasting method to address the problem). Two exceptions are Ye et al. [7] and Amin et al. [12]; the former selects between ARIMA models and the Holt-Winters method based on their modeling/fitting accuracy, and the later chooses between ARIMA models and SETARMA models according to the linearity of QoS data. However, they suffer defects in terms of the poor number of candidate approaches and the non-robust selection criterion. In time series research, each forecasting method has its own rationale and most suitable data type. When selecting the most suitable approach, one would ideally take more approaches into account (as shown in [4], many statistical and ML approaches can perform time series forecasting). On the other hand, thus far, a well-developed, sophisticated selection criteria have not been identified for QoS time series forecasting; the two above-mentioned examples only choose based on a single process data property and lack comprehensive and insightful consideration of the data. A more complex and robust selection criterion, such as a rule-based guidance, would be useful.

In addition to the single-approach and cross-approach selection strategies discussed above, another common strategy is to combine the forecasting results of different approaches (i.e., the cross-approach combination strategy). For example, the authors of [21] reported that a combination of the forecasting results of their top three individual approaches can effectively increase forecasting accuracy for cloud workload. However, our survey finds no instance of this strategy being applied to this kind of forecasting task. Overall, in terms of the three different strategies for time series forecasting (namely, the single-approach, the cross-approach selection, and the cross-approach combination strategies), the existing studies are somehow insufficient.

3.4 Integration

Without a QoS-aware (QoS-based) application to be integrated and used with the developed dynamic QoS forecasting approach, the forecasting approach is actually useless (and even meaningless). Despite several instances of viable integration, a more efficient way to combine both types of approaches is still needed. For example, most surveyed forecasting approaches use single-predictor mode, thus a trained and fitted predictor is used for prediction for long periods (e.g., the next 100 future time points) without any update. However, in the real world, as time passes, new QoS values can be observed and recorded, and these new QoS values should be incrementally and efficiently applied to maximize predictor performance because more recent values are more valuable and referable for the performance of the training and fitting task of a predictor). A simple way to ensure the most recent data is used is to re-train a new predictor each time a newer QoS value becomes available. However, this entails repetitive costs for training and fitting. For example, if we consider using the most suitable approach for time-aware QoS forecasting (i.e., GP), each time it must wait for GP to evolve and then search for a new predictor from scratch. GP's biggest problem is its very long training time, as demonstrated in [4]. To overcome this problem, we would suggest using the dynamic GP approach demonstrated in [22], which is specifically developed for the performance of continuous time series forecasting with lower evolving costs and higher forecasting accuracy. In [22], the authors proposed several techniques for the progressive evolution and searching of a set of expression-based predictors for prediction at consecutive future time points. Using this approach, it is no longer necessary to start the evolution and searching from scratch each time the prediction moves in time, and it can produce a set of predictors adapted to different data generation processes. We consider that the integration, development, or searching of such a relatively efficient approach would be useful in addressing the problem.

4 Conclusion

Because the values of time-aware dynamic QoS attributes vary over time, considerable research attention has focused on predicting such attributes. The present paper provide a comprehensive survey, comparison, and discussion of relevant studies over the past decade, and proposes challenges and research directions that should be carefully considered for future research efforts.

Six general criteria were used for comparison, including year of publication, the problem specification addressed in each work, the approaches proposed and/or considered for comparison, method of performance evaluation, the QoS dataset adopted for empirical experiments, and finally the type of application for which a forecasting approach was developed. Insufficiencies in the existing research were categorized into four groups: improper current time-aware QoS datasets, lack of an insightful and comprehensive analysis to QoS time series data (i.e., to the QoS dataset), incompleteness in the approaches and strategies developed, and finally a more efficient way to integrate a QoS forecasting approach and its application. Examples and possible solutions are proposed for each.

Acknowledgement. This research is partially sponsored by the Ministry of Science and Technology (Taiwan) under the Grant MOST106-2811-E-001-003 and MOST106-2221-E-001-007-MY2.

References

1. Kritikos, K., Plexousakis, D.: Requirements for QoS-based web service description and discovery. IEEE Trans. Serv. Comput. **2**(4), 320–337 (2009)
2. Fanjiang, Y.-Y., Syu, Y., Kuo, J.-Y.: Search based approach to forecasting QoS attributes of web services using genetic programming. Inf. Softw. Technol. **80**, 158–174 (2016)
3. Zheng, Z., Lyu, M.R.: Personalized reliability prediction of web services. ACM Trans. Softw. Eng. Methodol. **22**(2), 1–25 (2013)
4. Syu, Y., Kuo, J.-Y., Fanjiang, Y.-Y.: Time series forecasting for dynamic quality of web services: an empirical study. J. Syst. Softw. **134**, 279–303 (2017)
5. Nourikhah, H., Akbari, M.K., Kalantari, M.: Modeling and predicting measured response time of cloud-based web services using long-memory time series. J. Supercomput. **71**(2), 673–696 (2015)
6. Cavallo, B., Penta, M.D., Canfora, G.: An empirical comparison of methods to support QoS-aware service selection. In: Proceedings of the 2nd International Workshop on Principles of Engineering Service-Oriented SystemsCape Town, South Africa, pp. 64–70. ACM (2010)
7. Ye, Z., Mistry, S., Bouguettaya, A., Dong, H.: Long-Term QoS-aware cloud service composition using multivariate time series analysis. IEEE Trans. Serv. Comput. **9**(3), 382–393 (2016)
8. Rahman, Z.U., Hussain, O.K., Hussain, F.K.: Time series QoS forecasting for management of cloud services. Presented at the Proceedings of the 2014 Ninth International Conference on Broadband and Wireless Computing, Communication and Applications (2014)
9. Leitner, P., Ferner, J., Hummer, W., Dustdar, S.: Data-driven and automated prediction of service level agreement violations in service compositions. Distrib. Parallel Databases **31**(3), 447–470 (2013)
10. Xia, Y., Ding, J., Luo, X., Zhu, Q.: Dependability prediction of WS-BPEL service compositions using petri net and time series models. In: 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), pp. 192–202. IEEE, Redwood City (2013)
11. Amin, A., Colman, A., Grunske, L.: An approach to forecasting QoS attributes of web services based on ARIMA and GARCH models. In: 2012 IEEE 19th International Conference on Web Services (ICWS), Honolulu, HI, pp. 74–81. IEEE (2012)
12. Amin, A., Grunske, L., Colman, A.: An automated approach to forecasting QoS attributes based on linear and non-linear time series modeling. In: Proceedings of the 27th IEEE/ACM International Conference on Automated Software EngineeringEssen, Germany, pp. 130–139. ACM (2012)
13. Senivongse, T., Wongsawangpanich, N.: Composing services of different granularity and varying QoS using genetic algorithm. In: Proceedings of the World Congress on Engineering and Computer Science 2011. Lecture Notes in Engineering and Computer Science, San Francisco, CA, USA, pp. 388–393 (2011)
14. Yilei, Z., Zibin, Z., Lyu, M.R.: WSPred: a time-aware personalized QoS prediction framework for web services. In: 2011 IEEE 22nd International Symposium on Software Reliability Engineering (ISSRE), Hiroshima, pp. 210–219. IEEE (2011)