Information Sciences 422 (2018) 558–571

Contents lists available at ScienceDirect

Information Sciences

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

Unsupervised clustering of service performance behaviors

Hamdi Yahyaoui[a,*], Hala S. Own[b]

[a] Computer Science Department, Kuwait University, Kuwait
[b] Department of Solar and Space Research, National Research Institute of Astronomy and Geophysics, Egypt

ARTICLE INFO

Article history:
Received 12 December 2016
Revised 2 July 2017
Accepted 20 August 2017
Available online 12 September 2017

Keywords:
Services
Behaviors
Performance
Unsupervised clustering
Time series

ABSTRACT

We propose in this paper a novel approach for unsupervised clustering of services' behaviors. These behaviors are modeled as multivariate time series that capture the evaluation of several service quality attributes for a period of time. The importance weights of quality attributes are derived based on the Shannon's entropy concept and the service data is flattened in a format that is convenient for clustering. The flattening process spans over a time oriented aggregation transformation, which leverages Haar reduction. The reduction is modeled as a maximization of an objective function. The absence of ground truth is tackled by performing a set of tests to determine the best number of clusters and clustering algorithms. Extensive experiments were conducted to validate the proposed unsupervised clustering approach.

1. Introduction

Nowadays, there is a growing desideratum to use online services for everyday activities such as e-commerce, e-learning, gaming, etc. To handle the wide spectrum of developed services, researchers proposed solutions for clustering services by measuring their semantic similarities. Service semantics is related to its functional aspect. It can be derived from its description and annotations. The main benefit of such clustering is the enhancement of user search operations.

Lately, there is a focus on studying non-functional aspects of services since users seek efficient and reliable services, which can fulfill their requests. A valuable research effort was spent on clustering services based on their quality. Quality attributes such as reliability, response time, and availability are good indicators of service quality. Clustering approaches generally leverage a simple evaluation of each quality attribute and neglect its long term variation. Such short term assessment may not reflect the performance behavior of a service during a certain time period.

We advocate that the long term aspect should be taken into consideration for the sake of devising a more accurate approach in service performance behaviors assessment. An important indicator of such performance is services' quality attributes. Furthermore, we think that clustering services based on their performance behaviors is one of the desirable features that may help users in selecting the best services in terms of performance. For instance, users can be more interested to stable services rather than improving. Others may prefer in the opposite to select improving services. For providers, such clustering is a tool that allows them to do the assessment of their service performance and take adequate corrective measures. For instance, providers can detect degrading services based on clustering so they would investigate the reasons behind such bad performance and fix the related issues. Therefore, a clustering that takes into account the long term variation of the service quality provides users and providers with a better assessment of service performance in terms

* Corresponding author.
E-mail addresses: hamdi@cs.ku.edu.kw (H. Yahyaoui), halaown@gmail.com (H.S. Own).

http://dx.doi.org/10.1016/j.ins.2017.08.065
0020-0255/

of accuracy. The clustering of services faces two main challenges: dimensionality and importance of quality attributes. The dimensionality may hinder the development of efficient clustering techniques whereas the attribute importance is a paramount feature that should be reflected while building the clusters. Another issue that should be considered is the absence of a ground truth, i.e., a prior labeling of the data in which each data item is labeled with the class/cluster to which it belongs. In a supervised/semi-supervised clustering, a part of the input data is labeled while in an unsupervised clustering such labeling is missing. Henceforth, in an unsupervised setting there should be a strategy to determine the best number of clusters in order to obtain good clustering results.

We propose in this paper an unsupervised clustering approach for services based on their long term performance behaviors. Services' behaviors are modeled as multivariate time series. The proposed approach transforms the multivariate data into a flat data based on quality attributes' weights. Furthermore, the transformed data dimensionality is reduced using Haar Wavelet transformation. The reduced data is then clustered based on a two-step clustering algorithm.

The main contributions of this work include:

- The design of a new Time Oriented Aggregation (TOA) transformation for service performance behaviors, which leverages quality attributes' weights.
- The proposal of an unsupervised clustering of service performance behaviors, which relies on Haar Wavelet transformation of the aggregated data.
- The evaluation of the proposed approach using a real dataset of services.

The remainder of the paper is organized as follows. Section 2 is devoted to the related work. In Section 3, we present our approach. Section 4 is dedicated to multidimensional data preprocessing and transformation. In Section 5, we discuss our clustering strategy in the absence of ground truth and the metrics we used to assess its accuracy. Experiments on real world Web services are presented in Section 6. Finally, some concluding remarks and possible extension of this work are provided in Section 7.

2. Related work

With the huge number of deployed services, users are more and more interested in selecting the best services in terms of performance. The main indicators of a service performance are its quality attributes such as response time, availability, reliability, etc. Services quality assessment is one of the extensively explored topics. Most of the research was oriented toward ranking and clustering services using their quality attributes [2,16,24,31–33]. Despite these valuable efforts, most of the proposed approaches tackle the assessment from a short term perspective as we pinpointed in our previous work [29]. A long term assessment of services would lead to more accurate assessment results. Henceforth, services should be rather evaluated based on their performance behaviors for a certain time frame. A natural question arose: which convenient model can be devised to capture such behaviors? Since each service may be assessed based on many quality attributes in different times, we adopt the multivariate time series as a model, which captures the performance behavior of a service during a certain period. Assessment of services' behaviors can be achieved by grouping services, which have similar performance behaviors. We therefore resort to unsupervised clustering of multivariate time series.

Clustering time series was tackled in several previous works. Zhang and Ho [30] proposed an unsupervised feature extraction approach for time series based on Haar transform. The Haar transform can be repetitively applied. The first time it is applied it is considered as applied at level 1. The selected number of levels depends on the desired trade-off between dimensionality and distortion. In fact, the higher is the number of levels the higher is the distortion. The distortion is measured based on the energy of the time series, which is the sum of all its squared values. The authors tackled such trade-off by stopping the reduction process at a level $j-1$, in which the energy of the time series becomes higher than that of level j. The extracted features are then leveraged to do a clustering of the time series. Our solution for the trade-off between the dimensionality and the distortion is to model it as a maximization problem that includes both the dimensionality and distortion. Furthermore, we deal with multivariate time series, which cannot be handled by the aforementioned approach. Several issues arise when dealing with multivariate time series such as attributes' weights computation, dimensionality reduction, and data aggregation. We are providing valuable solutions for all of these issues in this work.

Lin et al. [15] proposed the Symbolic Aggregate Approximation (SAX), which transforms a time series into a sequence of symbols from a pre-defined alphabet. Using a sliding window approach, SAX can reduce the dimensionality. However, there is no discussion regarding the distortion issue. By loosing the time series trend, SAX may cluster time series, which do not have the same trend in the same cluster as pinpointed in [26].

Paparrizos and Gravano [19] proposed a new approach called k-shape that leverages a distance, which uses a normalized version of the cross-correlation measure to consider the shape (trend) of time series while comparing them. They did not consider the dimensionality and the distortion in their framework, which are deeply investigated by our approach. Furthermore, k-shape does not propose a practical solution to the issue of absence of ground truth.

Recently, Barragan et al. [5] proposed a wavelet-based clustering approach of multivariate time series that combines wavelets features, a multiscale PCA similarity technique and fuzzy clustering. The time series are transformed into wavelet coefficients from which similarity values are generated. These values are then used by an extended version of the Fuzzy C-Means algorithm to cluster the time series. The aim of the clustering is to do the recognition of fault patterns. The distortion and the absence of ground truth issues are not discussed in that work.