Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 61–71
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original article

Exploiting structural similarity of log files in fault diagnosis for Web service composition

Xu Han [a,*], Binyang Li [b], Kam-Fai Wong [c], Zhongzhi Shi [d]

[a] *Beijing Key Laboratory of Electronic System Reliability Technology, Capital Normal University, Beijing, China*
[b] *University of International Relations, Beijing, China*
[c] *Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, Hong Kong, China*
[d] *The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*

Available online 7 June 2016

Abstract

With increasing deployment of Web services, the research on the dependability and availability of Web service composition becomes more and more active. Since unexpected faults of Web service composition may occur in different levels at runtime, log analysis as a typical data-driven approach for fault diagnosis is more applicable and scalable in various architectures. Considering the trend that more and more service logs are represented using XML or JSON format which has good flexibility and interoperability, fault classification problem of semi-structured logs is considered as a challenging issue in this area. However, most existing approaches focus on the log content analysis but ignore the structural information and lead to poor performance. To improve the accuracy of fault classification, we exploit structural similarity of log files and propose a similarity based Bayesian learning approach for semi-structured logs in this paper. Our solution estimates degrees of similarity among structural elements from heterogeneous log data, constructs combined Bayesian network (CBN), uses similarity based learning algorithm to compute probabilities in CBN, and classifies test log data into most probable fault categories based on the generated CBN. Experimental results show that our approach outperforms other learning approaches on structural log datasets.

Keywords: Web services composition; Fault diagnosis; Combined Bayesian network (CBN); Similarity; Probability

1. Introduction

In recent years, as a promising computing paradigm, service-oriented computing (SOC) [1] has changed the way of design, delivery and consumption of software applications. Web services technology aiming at implementing service oriented architecture has been widely applied to different areas for research or business purposes. Accordingly, more and more business functions are published as Web services (WS) by various organizations and companies. There are two main approaches of developing Web services, including SOAP-based Web services [2] and RESTful Web services [3,4]. Numerous atomic Web services can be regarded as access points for applications without relying on other Web services. When a user request cannot be fulfilled by atomic Web services, composite Web service plays an important role in providing complex collaboration and interaction between multiple Web services.

During the past ten years, a large number of existing standards [5] for Web service composition have been defined. Web service orchestration language is one of these standards for describing executable business processes, which are composed of Web services. Based on OASIS standards, Web Services Business Process Execution Language (WSBPEL or BPEL) [6] is an executable orchestration language for modeling business processes with Web services. Depending on Web service composition, business processes can be executed by the support of BPEL engine from a third party. Some recent

* Corresponding author.
E-mail addresses: hanxu@cnu.edu.cn (X. Han), byli@uir.cn (B. Li), kfwong@se.cuhk.edu.hk (K.-F. Wong), shizz@ics.ict.ac.cn (Z. Shi).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.03.006

research work has focused on extending BPEL engine to enable composition of SOAP-based, RESTful and OSGi services [7]. Since there are more and more Web services (esp. RESTful Web services) used in practice, the complexity of Web service composition is becoming higher than before. For this reason, how to ensure dependability and availability is essential for Web service composition. To enhance the dependability of service flow execution, fault tolerant Web service orchestration [8] was proposed. As a critical aspect in fault tolerance framework, fault diagnosis aims at identifying or locating the causes with high probability to explain process exceptions and failures based on runtime information. Since most detailed running information in processes execution is recorded in log files, fault diagnosis by learning from log data is becoming an important issue in this area.

Currently, semi-structured data formats, including XML and JSON, are used as the standards of information representation on the Internet. Due to its good flexibility and interoperability, more and more log files of software running information are represented using the XML/JSON format, especially for Web services. Thus, it becomes a key topic of fault diagnosis research which focuses on analyzing semi-structured and XML/JSON like log. Generally, semi-structured documents have much richer structural information than flat ones, which has potential influence on classification accuracy. Taking this into account, the main task of learning from this kind of documents will have more challenges than before. However, for most classification methods of log analysis, IR-based methods are commonly used ignoring a significant amount of structural information, which leads to low classification accuracy. Therefore, how to learn the structural information from the log has great impact on the accuracy of fault classification.

In this paper, we propose a similarity-based Bayesian learning approach for fault classification of semi-structured logs. Our method is to first estimate similarity degrees of structural elements from heterogeneous log data. Then the basic structure of combined Bayesian network (CBN) is constructed, and the similarity-based learning algorithm is used to compute probabilities in CBN. Finally, test log data can be classified into most probable fault categories based on the generated CBN.

The rest of this paper is organized as follows. Section 2 introduces the related work concerning fault diagnosis of Web service composition and existing classification methods for (semi-)structured documents. Section 3 provides an overview of the similarity based structural classification approach and the CBN model. The details of CBN generation, including how to compute probabilities of CBN using similarity-based learning algorithm, are presented in Section 4. Experimental results of this approach compared to other learning approaches are shown in Section 5. Finally, Section 6 draws the conclusion.

2. Related work

Over the past years, some research work in Web services area has concentrated on how to enhance dependability and availability of Web services. Fault tolerant Web services orchestration [8] is supported by fine grained identification of exception and fault causes and the consequent execution of effective exception and fault handlers [9,10]. As an important step of fault tolerance, fault diagnosis has attracted wide attention of academic community increasingly. From the methodology perspective, the existing work on fault diagnosis in this area can be divided into two main categories, including model-based diagnosis and data-driven diagnosis.

With respect to model-based approaches, the basic idea is to model the behavior and inner logic of the diagnosed service, and then discover runtime faults based on its model. The on-going work has been described in some published papers. WS-DIAMOND [11] is a European research project which eight research agencies have participated in. In this project, model-based diagnosis is adopted as the principal approach. Yan et al. [12,13] presented a model-based approach for diagnosing orchestrated Web service processes. In their approach, Web services with faults can be deduced from the variable dependency on execution trajectory, which is represented by the generated automata of BPEL description. With the assumption that behavioral descriptions of individual activities may not be totally given, Mayer et al. [14] presented an approach of isolating minimal sets of faulty activity executions based on the process structure. Considering composite service adaption to the dynamic execution environment, Dai et al. [15] analyzed the error propagation relation between any two services and gave uncertain casual relation between exceptions and services by computing error propagation degree. In addition, some research groups proposed testing frameworks [16] and hybrid models [17] for fault diagnosis of Web service composition.

For data-driven approaches, the diagnosis problem is usually transformed into the classification problem. Then it can be solved by using data mining and machine learning algorithms on log file data. As we know, many reported research efforts have focused on mining log files of computing systems. And there are some common places in log mining methods for regular computing systems and Web services. For this reason, we can make reference to those existing methods in the related area. Considering the differences between two basic data types — plain text data and semistructured data, mining approaches are often designed and implemented according to the type of training and test data. For plain-text log data, Li et al. used Bayes method [18,19] to categorize text messages in log files, and utilize the temporal information to improve classification performance. In Ref. [20], Bayes classifier, semi-supervised learning, and decision trees, are used to automatically recognize symptoms of recurrent faults. As for semi-structured log data, there are also some corresponding classification approaches. In Ref. [21], a database of failure signatures against which undiagnosed failure data can be matched, is constructed from monitoring data. And then anomaly-based clustering method is proposed to generate right clusters for diag- nosing failures with low-confidence match. Denoyer and Gallinari [22] provided a generative model for classification task based on Bayesian networks, which can handle both structure and content. Zaki and Aggarwal [23] presented an effective rule based classifier for XML data using frequent discriminatory