Expert Systems with Applications 41 (2014) 1377–1388

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A graph distance based metric for data oriented workflow retrieval with variable time constraints

Yinglong Ma [a,b,*], Xiaolan Zhang [a], Ke Lu [c]

[a] School of Control and Computer Science, North China Electric Power University, Beijing 102206, China
[b] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Business process management
Data oriented workflow
Similarity computation
Process mining
Process retrieval

ABSTRACT

There are many applications in business process management that require measuring the similarity between business processes, such as workflow retrieval and process mining, etc. However, most existing approaches and models cannot represent variable constraints and achieve data oriented workflow retrieval of considering different QoS requirements, and also fail to allow users to express arbitrary constraints based on graph structures of workflows. These problems will impede the customization and reuse of workflows, especially for data oriented workflows. In this paper, we will be towards workflow retrieval with variable time constraints. We propose a graph distance based approach for measuring the similarity between data oriented workflows with variable time constraints. First, a formal structure called Time Dependency Graph (TDG) is proposed and further used as representation model of workflows. Similarity comparison between two workflows can be reduced to computing the similarity between their TDGs. Second, we detect whether two TDGs of workflows for similarity comparison are compatible. A distance based measure is proposed for computing their similarity by their normalization matrices established based on their TDGs. We theoretically proof that the proposed measure satisfies the all the properties of distance. In addition, some exemplar processes are studied to illustrate the effectiveness of our approach of similarity comparison for workflows.

1. Introduction

Business process management is an established area that aims at the automation of a business process (Cook & Wolf, 1998), and has been widely applied in many fields such as e-Science (Taylor, Deelman, & Gannon, 2007), medical healthcare (Lyng, Hildebrandt, & Mukkamala, 2009; Maximini & Schaaf, 2003), search (Frefimann, 2006) and information integration (Hung & Chiu, 2004; Lee, Ho, Ho, & Lau, 2011), etc. In the recent years, the use of business processes has significantly expanded from the original domain of business processes towards new areas in scientific data processing, such as data oriented workflows, etc. Data oriented workflows (Ikeda, Park, & Widom, 2011) have been widely applied in many scientific areas with the large amount of data and complex computation tasks.

Data-oriented workflows can be modeled as graphs where nodes denote tasks/services for data computation and data manipulation, and edges denote the flow of data input to and output from the tasks/services. Scientific workflows (McPhillips & et al., 2009) are a kind of typical data oriented workflows that are applied in many fields such as bioinformatics, astronomy, ecology, earth science, etc. A variety of scientific workflow systems such as Kepler (Ludascher & et al., 2005) and Taverna (Oinn & et al., 2005) were developed, which accelerate the pace of scientific progress in these scientific areas.

Process similarity measure is often used in process retrieval (Bergmann, 2011; Leake & Kendall-Morwick, 2008; Madhusudan, Zhao, & Marshall, 2004), process mining (Greco, Guzzo, Manco, & Saccia, 2005; Huang, Wang, Zhao, Zhang, & Yuan, 2006; Lim, Lee, & Raman, 2012; van der Aalst et al., 2003; Wen, van der Aalst, Wang, & Sun, 2007), process scheduling (R-Moreno, Borrajo, Cesta, & Oddi, 2007) and process integration (von Berg, Schmidt, & Wendler, 2001; Wang, Shen, & Hao, 2006; Yan, Maamar, & Shen, 2001; Zha, Wangb, Wen, Wang, & Sun, 2010). This is especially true for data oriented workflows retrieval. What users are indeed concerned about is how to apply advanced methods on new data for discovering new facts. However, the users in real life usually do not know how to develop a workflow for applying these advanced methods, and they even do not know how to program. Fortunately, data oriented workflows are often developed with the aim of scientific experimentation and can be "repeatedly" executed with

* Corresponding author at: School of Control and Computer Science, North China Electric Power University, Beijing 102206, China. Tel./fax: +86 10 61772643.
E-mail addresses: yinglongma@gmail.com, m_y_long@otcaix.iscas.ac.cn (Y. Ma), zhangxiaolan0504@126.com (X. Zhang), luk@ucas.ac.cn (K. Lu).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.035

different data or different parameters (Ellisman, Fahringer, Fox, Gannon, et al., 2007). The methods of using data oriented workflows for data analysis and processing are often highly similar, so users always concentrate on producing large amounts of high-quality analysis data instead of deeply understanding these methods. In the situation, when analysis data are available, users would be highly interested in searching a repository of workflows. They can select from the repository the most suitable workflow through which the best possible methods can analyze these data. Generally, data oriented workflow retrieval is a promising solution that can help users find a suitable workflow satisfying their requirements to a given problem by matching the expressed constraints and ranking them according to some criteria instead of developing one themselves. Some approaches for workflow retrieval have demonstrated potential and initial success in business workflow retrieval (Awad & Sakr, 2010; Beeri, Eyal, Kamenkovich, & Milo, 2008).

However, some important open issues for data oriented workflow retrieval remain to be resolved. On the one hand, most existing researches in workflows retrieval do not involve quality of services. A complex workflow for data analysis and processing is composed of dozens of distributed tasks/services which are often provided by external service providers. Although some of services are free to access, the availability and quality of services (QoS) can be guaranteed only by paying for these services. Most important, the services provided also possibly have different levels of QoS. Quality of services possibly includes many aspects such as execution time, response speed, service cost, etc. The price of a service can be determined by its levels of QoS. Service providers can charge higher prices for higher levels of QoS. Users may not always need that workflows can be completed in a higher level of QoS than they require. They sometimes may prefer to use cheaper services with a lower QoS that is sufficient to meet their requirements. However, most existing approaches and models cannot represent variable constraints and achieve data oriented workflow retrieval of considering different QoS requirements of users.

On the other hand, graph based representation of workflows is an intuitive and effective approach describing data/control flows of workflows. It is desirable for users making data analysis and processing to use the graph based constraints for expressing their QoS requirements of workflows they require. However, most existing approaches fail to allow users to express arbitrary constraints based on graph structures of workflows. Although some workflow query languages such as BPQL (Beeri et al., 2008) and BPMN-Q (Awad & Sakr, 2010), were proposed for workflow query, users are often not experts in any query language. There is also a lack of tools to be able to express their retrieval requests as easily as possible and support a rich set of graph edit operations (e.g., adding/removing/replacing of a flow or a task), assignments of QoS constraints, and automatic similarity computation of data oriented workflows. All these problems will impede the customization and reuse of scientific workflows.

We confine our work to the time constraints of data oriented workflows in this paper. Services developed by external providers have variable execution speeds which correspond to different levels of QoS. A higher execution speed means a more execution time. If user hopes that tasks in workflows can be completed in shorter execution time, these tasks would be completed in higher levels of QoS, and therefore users need to pay more for executing these tasks with higher levels of QoS. However, different users have different requirements of execution time. They may not always need to complete workflows earlier than they require. This paper will be towards workflow retrieval with variable time constraints.

In this paper, we propose a graph distance based approach for measuring the similarity between oriented workflows with variable time constraints. We define a graph based distance metric by which we can measure and compare the similarity between two workflows based on variable time constraints. First, a formal structure called Time Dependency Graph (TDG) is proposed and further used as representation model of workflows. Similarity comparison between two workflows can be reduced to computing the similarity between their TDGs. Second, we detect whether two TDGs of processes for similarity comparison are compatible from the perspective of functionality of workflows. Then, a distance based measure is proposed for computing their similarity by the normalization matrices based on their TDGs. We theoretically proof that the proposed measure satisfies the all the properties of distance. In addition, a case example is studied to illustrate our approach of similarity comparison for workflow retrieval.

This paper is organized as follows. In Section 2, we review the related work. Section 3 is to propose a graph based representation of workflows with variable time constraints, which is called Time Dependency Graph (TDG). In Section 4, we detect the compatibility of two workflows by the δ-compatibility. In Section 5, we compare the similarity between two workflows based on their TDGs. Normalization matrices are proposed to represent business processes. In Section 6, we further propose a distance based measure d for measuring the similarity between business processes. Most important, we theoretically proof that the measure satisfies all the properties of distance. Section 7 is to illustrate our approach by using a case of process retrieval. Section 8 is to introduce the developed prototype system. Section 9 is the conclusion.

2. Related work

Reuse of data oriented workflows is an active research area in recent years (Bowers, Ludascher, Ngu, & Critchlow, 2006; Simmhan, Plale, & Gannon, 2008; Slominski, 2007), and workflow retrieval provides a feasible solution for their reuse. Users would like to simply choose one of the matching workflows from workflow repository, provide its input data, and obtain the output results. Recently, most existing approaches describe their data and intended analysis by using a set of keywords (Liu, Shao, & Chen, 2010; van der Aalst et al., 2003). Keywords based workflow retrieval does not consider the structures of workflows, and therefore cannot really satisfy users' retrieval requirements. Some approaches with more expressivity were developed for workflow query, such as BPQL (Beeri et al., 2008) and BPMN-Q (Awad & Sakr, 2010). However, users are often not the experts in process query languages, and the use of these languages is not easy to them.

Workflows including data oriented workflows are often modeled as directed acyclic graphs (DAGs), where nodes denote tasks and services, and arcs denote scheduling dependencies between computation tasks/services (Deelman, Singh, et al., 2004; Thain, Tannenbaum, & Livny, 2005). Several graph algorithms have been proposed for similarity assessment such as sub-graph isomorphism, maximal common sub-graphs, or edit-distance measures (Bunke, Foggia, Guidobaldi, Sansone, & Vento, 2002; Bunke & Shearer, 1998; Vijayalakshmi, Nadarajan, Roddick, Thilaga, & Nirmala, 2011). The similarity between business processes is reduced to the distance between graphs (Bunke & Shearer, 1998; Zhang, Wang, & Shasha, 1996). Reference Goderis, Li, and Goble (2006) attempted to build a gold standard for workflow ranking. via a workflow discovery tool, a mechanism for ranking workflow fragments was provided based on graph sub-isomorphism matching. Reference Bergmann and Gil (2011) reported an interesting work that adopts semantic based approach to retrieve workflows, but it failed to consider the cost and QoS constraints. Reference Jin (2006) proposed an approach for similarity of Weighted Directed Acyclic Graphs (wDAGs). Reference Kiani, Bhavsar, and Boley (2013) also proposed the structure-weight graph similarity and explored its application in E-Health. However, the graph similarity that these