Expert Systems with Applications 42 (2015) 4207–4215

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A knowledge-intensive approach to process similarity calculation

Stefania Montani [a], Giorgio Leonardi [a,*], Silvana Quaglini [b], Anna Cavallini [c], Giuseppe Micieli [c]
on behalf of the Stroke Unit Network (SUN) collaborating centers

[a] DISIT, Computer Science Institute, Università del Piemonte Orientale, Viale T. Michel 11, 15121 Alessandria, Italy
[b] Department of Electrical, Computer and Biomedical Engineering, Università di Pavia, Via Ferrata 1, I-27100 Pavia, Italy
[c] Istituto di Ricovero e Cura a Carattere Scientifico Fondazione "C. Mondino", Via Mondino 2, I-27100 Pavia, Italy

ARTICLE INFO

Article history:
Available online 24 January 2015

Keywords:
Process comparison
Graph edit distance
Process mining
Stroke management

ABSTRACT

Process model comparison and similar processes retrieval are key issues to be addressed in many real world situations, and particularly relevant ones in some applications (e.g., in medicine), where similarity quantification can be exploited in a quality assessment perspective.

Most of the process comparison techniques described in the literature suffer from *two main limitations*: (1) they adopt a purely syntactic (vs. semantic) approach in process activity comparison, and/or (2) they ignore complex control flow information (i.e., other than sequence). These limitations oversimplify the problem, and make the results of similarity-based process retrieval less reliable, especially when domain knowledge is available, and can be adopted to quantify activity or control flow construct differences.

In this paper, we aim at *overcoming both limitations*, by introducing a framework which allows to extract the actual process model from the available process execution traces, through process mining techniques, and then to compare (mined) process models, by relying on a **novel distance measure**.

The novel distance measure, which represents the main contribution of this paper, is able to address issues (1) and (2) above, since: (1) it provides a **semantic, knowledge-intensive approach** to process activity comparison, by making use of domain knowledge; (2) it explicitly takes into account **complex control flow constructs** (such as AND and XOR splits/joins), thus fully considering the different semantic meaning of control flow connections in a reliable way.

The positive impact of the framework in practice has been tested in stroke management, where our approach has outperformed a state-of-the art literature metric on a real world event log, providing results that were closer to those of a human expert. Experiments in other domains are foreseen in the future.

1. Introduction

Process model comparison is a key issue to be addressed in many real world situations. For example, when two companies are merged, process engineers need to compare processes originating from the two companies, in order to analyze their possible overlaps, and to identify areas for consolidation. Moreover, large companies build over time huge process model repositories, which serve as a knowledge base for their ongoing process management/enhancement efforts. Before adding a new process model to the repository, process engineers have to check that a similar model does not already exist, in order to prevent duplication.

Particularly interesting is the case of medical process model comparison, where similarity quantification can be exploited in a quality assessment perspective. Indeed, the process model actually implemented at a given healthcare organization can be compared to the existing reference clinical guideline, e.g., to check conformance, or to understand the level of adaptation to local constraints that may have been required. As a matter of fact, the existence of local resource constraints may lead to differences between the models implemented at different hospitals, even when referring to the treatment of the same disease (and to the same guideline). A quantification of these differences (and maybe a ranking of the hospitals derived from it) can be exploited for several purposes, like, e.g., auditing purposes, performance evaluation and funding distribution.

* Corresponding author. Tel.: +39 0131 360340.
E-mail addresses: stefania.montani@unipmn.it (S. Montani), giorgio.leonardi@unipmn.it (G. Leonardi), silvana.quaglini@unipv.it (S. Quaglini), anna.cavallini@mondino.it (A. Cavallini), giuseppe.micieli@mondino.it (G. Micieli).

http://dx.doi.org/10.1016/j.eswa.2015.01.027
0957-4174/

Various process model comparison techniques are described in the literature (see Section 4). However, most of them suffer from two main limitations:

1. they adopt a purely syntactic approach in process activity comparison, ignoring the semantics of the activities being compared, often referring just to their names: activities with a different name are considered as not matching, while they could share very similar characteristics (e.g., have the same goal);
2. they ignore complex control flow information (other than sequence): in this way, a construct with, e.g., two parallel activities, can be matched to a construct involving the same activities, but in mutual exclusion.

Issues (1) and (2) above correspond to a strong simplification of the process model semantic meaning, and may lead to unreliable results in process comparison. This can be really unacceptable in many real world domains, like the already mentioned medical ones, where physicians and hospital managers need to guarantee the highest quality of service to patients.

In this paper, we aim at *overcoming the limitations* outlined above, by introducing a framework which allows to mine the actual process model from the available process execution traces, and then to compare (mined) process models.

While the framework, in its current version, relies on already published process mining techniques to extract the process model from traces, process comparison exploits a **novel distance measure**, which represents the main contribution of the paper.

Our distance measure is very innovative with respect to available literature approaches (see detailed discussion in Section 4). Indeed, it is able to address issues (1) and (2) above, since:

1. it provides a **semantic approach** to process activity comparison, by making use of domain knowledge. Indeed, it rates two activities as very similar, if they are connected through semantic (i.e., ontological) relations. Specifically, the metric can be properly adapted to operate with different knowledge representation formalisms (e.g., taxonomy vs. semantic network with different characteristics). Very interestingly, it also exploits all the information that can be extracted through process mining (e.g., temporal information), always in a **semantic and knowledge-intensive** perspective;
2. it explicitly takes into account **complex control flow constructs** (such as AND and XOR splits/joins – also called *gateway nodes* henceforth), thus considering the different **semantic meaning** of control flow connections in a reliable way.

Fully exploiting the semantics of process models in comparison and similarity quantification along the lines illustrated above represents a major development with respect to the literature in the field, as extensively discussed in Section 4. Such a development is likely to provide a significant impact in supporting the expert's work in quality assessment, particularly in those applications where domain knowledge is rich and well consolidated, as is often the case in medicine (Basu, Archer, & Mukherjee, 2012).

Indeed, the positive impact of the framework in practice has already been tested in stroke management (see Section 3), where our approach has outperformed a state-of-the-art metric (La Rosa, Dumas, Uba, & Dijkman, 2013) on a real world event log, providing results that were closer to those of a human expert.

The paper is organized as follows. Section 2 provides the details of our methodological approach. Section 3 showcases experimental results. Section 4 compares our contribution to related works. Section 5 illustrates our conclusions and future research directions.

2. Methods

As stated in the Introduction, our framework first extracts the actual process model from the execution traces, and then performs process model comparison by means of a novel metric. The methodological techniques supporting the first step (process mining) are briefly presented in subSection 2.1, while subSection 2.2 is devoted to the detailed description of our metric, which represents the main contribution of this paper.

2.1. Mining process models

Process mining describes a family of a posteriori analysis techniques (Van der Aalst et al., 2003) exploiting the information recorded in process execution trace repositories (also called *event logs*), to extract process related information (e.g., process models). Typically, these approaches assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well defined step in the process) and is related to a particular process instance. Furthermore, some mining techniques use additional information such as the timestamp of the event, or data elements recorded with the event.

Traditionally, process mining has been focusing on discovery, i.e., deriving process models and execution properties from event logs. It is important to mention that, in discovery, there is no a priori model, but, based on logs, some model, e.g., a Petri Net, is constructed. However, process mining is not limited to process models (i.e., control flow), and recent process mining techniques are more and more focusing on other perspectives, e.g., the organisational perspective, the performance perspective or the data perspective. Moreover, as well stated in the Process Mining Manifesto (IEEE Taskforce on Process Mining, 2011), process mining also supports conformance analysis and process enhancement. In this paper, however, we only deal with the process perspective.

In our work, we are currently relying on mining algorithms available within ProM (Van Dongen, Alves De Medeiros, Verbeek, Weijters, & Van der Aalst, 2005), an open source tool which supports a wide variety of process mining and data mining techniques.

In particular, we have mainly exploited ProM's *heuristic miner* (Weijters, Van der Aalst, & de Medeiros, 2006) for mining the process models. Heuristic miner takes in input the event log, and considers the order of the events within every single process instance execution. The time stamp of an activity is used to calculate this ordering. Heuristics miner can be used to express the main behavior registered in a log. Some abstract information, such as the presence of composite tasks (i.e., tasks semantically related to their constituent activities by means of the ''part-of'' relation), cannot be derived by heuristic miner, that will only build a model including ground (i.e., not further decomposable) activities. On the other hand, it can mine the presence of short distance and long distance dependencies (i.e., direct or indirect sequence of activities), and information about parallelism, with a certain reliability degree (see also Section 2.2). The output of the mining process is provided as a graph, also called ''dependency graph'', where nodes represent activities, and arcs represent control flow information.

We have chosen to rely on heuristic miner because it is known to be tolerant to noise, a problem that may affect many real world event logs (e.g., in medicine sometimes the logging may be incomplete). Moreover, heuristic miner labels the output graph edges with several mined information, that we are explicitly considering in process comparison (such as reliability, see Section 2.2). The output of heuristic miner can also be automatically converted into a