Graph Matching Algorithms for Business Process Model Similarity Search

Remco Dijkman[1], Marlon Dumas[2], and Luciano García-Bañuelos[2,3]

[1] Eindhoven University of Technology, The Netherlands
r.m.dijkman@tue.nl
[2] University of Tartu, Estonia
marlon.dumas@ut.ee
[3] Universidad Autonoma de Tlaxcala, Mexico
lgbanuelos@gmail.com

Abstract. We investigate the problem of ranking all process models in a repository according to their similarity with respect to a given process model. We focus specifically on the application of graph matching algorithms to this similarity search problem. Since the corresponding graph matching problem is NP-complete, we seek to find a compromise between computational complexity and quality of the computed ranking. Using a repository of 100 process models, we evaluate four graph matching algorithms, ranging from a greedy one to a relatively exhaustive one. The results show that the mean average precision obtained by a fast greedy algorithm is close to that obtained with the most exhaustive algorithm.

1 Introduction

As organizations reach higher levels of Business Process Management (BPM) maturity, repositories with hundreds of business process models become increasingly common [18]. For example, the SAP reference model contains over 600 business process models. A similar number of process models can be found in the reference model for Dutch Local Governments [6]. On a larger scale, tool vendors distribute reference model repositories (e.g. the IT Infrastructure Library – ITIL) with over a thousand process models each[1]. These models are used, for example, to document and to communicate internal procedures or to enable the re-design and automation of business processes. In order to effectively fulfil these tasks, tool support is needed to retrieve relevant models from such repositories.

In this paper, we focus on the problem of similarity search in process model repositories: Given a process model or fragment thereof (the *search model*), find those process models in the repository that most closely resemble the search model. The need for similarity search arises in multiple scenarios. For example, when adding a new process model into a repository, similarity search allows one to detect duplication or overlap between the new and the existing process models. Meanwhile, in the context of reference process model repositories, such

[1] See for example CaseWise's ITIL repository (http://www.casewise.com/Gateway/)

U. Dayal et al. (Eds.): BPM 2009, LNCS 5701, pp. 48–63, 2009.

as ITIL, similarity search allows one to retrieve reference models that overlap with an existing "as is" process model.

Answering a similarity search query involves determining the degree of similarity between the search model and each model in the repository. In this context, similarity can be defined from several perspectives, including the following.

- Text similarity: based on a comparison of the labels that appear in the process models (task labels, event labels, etc.), using either syntactic or semantic similarity metrics, or a combination of both.
- Structural similarity: based on the topology of the process models seen as graphs, possibly taking into account text similarity as well.
- Behavioural similarity: based on the execution semantics of process models.

In previous work, we evaluated several similarity metrics across all three perspectives [5,19]. We found that a structural similarity metric based on graph matching achieved the highest retrieval quality (precision and recall). However, the operationalization of this metric is hindered by the fact that the underlying graph matching problem, namely the graph-edit distance problem, is NP-complete [14]. This is not only a theoretical limitation, but a practical one: our experiments show that for real-life process models with more than 20 nodes, exhaustive graph matching algorithms lead to combinatorial explosion. Therefore, heuristics are needed that strike a tradeoff between computational complexity and precision. This paper presents and compares four heuristic algorithms for calculating the similarity of business process models based on graph matching.

The rest of the paper is structured as follows. Section 2 formulates the problem and introduces the structural similarity metric studied in the paper. Section 3 presents four algorithms that provide alternative operationalizations of the structural similarity metric. Section 4 presents an experimental evaluation of these algorithms. Section 5 discusses related work and Section 6 concludes.

2 Preliminaries

This section defines the notion of business process used in this paper and formulates the structural similarity metric used for comparing pairs of process models.

2.1 Business Process

A business process is a collection of related tasks that lead to a specified goal. Many modeling notations are available to capture business processes, including Event-driven Process Chains (EPC), UML Activity Diagrams and the Business Process Modeling Notation (BPMN) [20]. In this paper, we seek to abstract as much as possible from the specific notation used to represent process models, to allow for measuring similarity of business processes modeled using different notations. Accordingly, we adopt an abstract view in which a process model is a directed attributed graph, as captured in the following definition.