Fast Business Process Similarity Search with Feature-Based Similarity Estimation

Zhiqiang Yan, Remco Dijkman, and Paul Grefen

Eindhoven University of Technology
P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands
{z.yan,r.m.dijkman,p.w.p.j.grefen}@tue.nl

Abstract. Nowadays, business process management plays an important role in the management of organizations. More and more organizations describe their operations as business processes, and the intra- and inter-organizational interactions between operations as services. It is common for organizations to have collections of hundreds or even thousands of business processes. Consequently, techniques are required to quickly find relevant business process models in such a collection. Currently, techniques exist that can rank all business process models in a collection based on their similarity to a query business process model. However, those techniques compare the query model with each model in the collection in terms of graph structure, which is inefficient and computationally complex. Therefore, this paper presents a technique to make this more efficient. The technique selects small characteristic model fragments, called features, which are used to efficiently estimate model similarities and classify them as *relevant, irrelevant or potentially relevant* to a query model. Only *potentially relevant* models must be compared using the existing techniques. Experiments show that this helps to retrieve similar models at least 3.5 times faster without impacting the quality of the results; and 5.5 times faster if a quality reduction of 1% is acceptable.

1 Introduction

Nowadays, service and business process management technologies develop quickly in both academic and industrial fields. To increase the flexibility and controllability of the management of organizations, business processes are used to describe functions organization provides and services are used to describe the both intra- and inter-organizational cooperations. As a result, it is common to see collections of hundreds or even thousands of business process models. For example, the SAP reference model consists of more than 600 business process models [16], and the reference model for Dutch Local Governments contains a similar number of models [9]. As business process model collections increase in size, tools and techniques are required to manage them. This includes tools and techniques for quickly searching a collection for business process models that meet certain criteria. These criteria can be specified by means of a search query [2,5], but also by means of (a part of) a business process model for which similar models must be retrieved [6,7].

R. Meersman et al. (Eds.): OTM 2010, Part I, LNCS 6426, pp. 60–77, 2010.

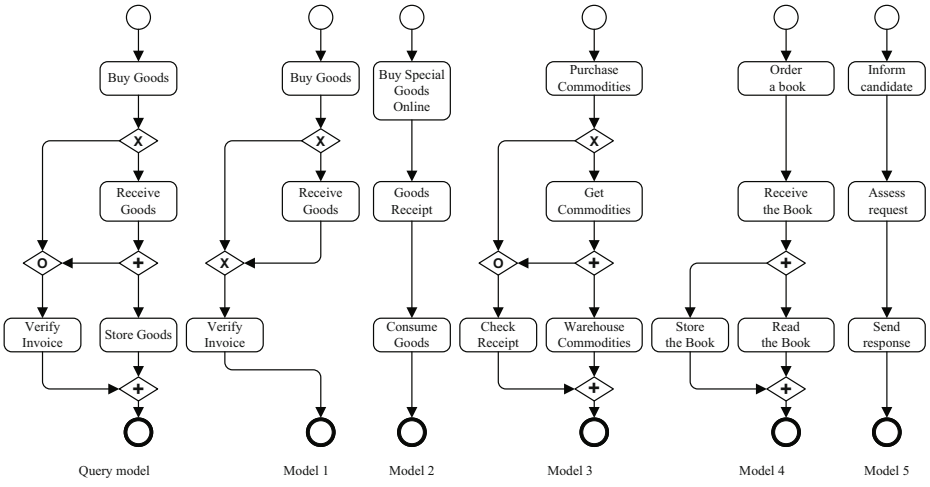

Fig. 1. Searching a collection of business process models

This paper focuses on the second class of search techniques, which are also called *similarity search* techniques. Similarity search can, for example, be applied to search a collection of reference process models for the model that best matches a process model from a specific organization; or in case of merger between two organizations to search which business process model from one organization matches which business process model of the other. Figure 1 shows an example of a business process similarity search. It shows one *query model* and five *process models* in the BPMN notation. Given a search query model, a similarity search technique should only return those process models that are similar to the search query model and it should return those similar process models in order of their similarity to the search query model. In the example, the technique could return models 1, 2 and 3.

There currently exist similarity search techniques [6,7]. However, these techniques focus on defining a metric to compute the similarity between two process models. To rank the business process models in a collection, the similarity of each of the process models to the query model must be computed. Subsequently, the process models are ordered according to their similarity. This is time consuming and can cause a similarity search operation to take multiple seconds or even minutes, depending on the metric and algorithm that is used, while a search query should be performed within milliseconds by a search engine, e.g., Google.

Therefore, the goal of this paper is to develop a similarity search technique that is both accurate *and* fast. The technique works by quickly classifying process models as 'relevant', 'irrelevant' or 'potentially relevant' to a search query model, based on an estimation of their similarity to the search model. Existing similarity search techniques then only have to be used to rank the process models in the 'potentially relevant' category, which typically contains much fewer models than the collection as a whole (in our evaluation set only 10% of the number of models in the collection), therewith significantly reducing the search time.