Chapter 18

Business Process Graphs:
Similarity Search and Matching

Remco Dijkman
Eindhoven University of Technology, The Netherlands

Marlon Dumas
University of Tartu, Estonia

Luciano García-Bañuelos
University of Tartu, Estonia

ABSTRACT

Organizations create collections of hundreds or even thousands of business process models to describe their operations. This chapter explains how graphs can be used as underlying formalism to develop techniques for managing such collections. To this end it defines the business process graph formalism. On this formalism it defines techniques for determining similarity of business process graphs. Such techniques can be used to quickly search through a collection of business process graphs to find the graph that is most relevant to a given query. These techniques can be used by tool builders that develop tools for managing large collections of business process models. The aim of the chapter is to provide an overview of the research area of using graphs to do similarity search and matching of business processes.

INTRODUCTION

Nowadays, most organizations describe their operations in terms of business processes models. As a consequence they have large repositories of business process models. For example, the Reference Model for Dutch Local Government (Documentair Structuurplan, n.d.) contains around 600 business process models and the SAP Reference Model (Curran & Keller, 1999) contains a similar number of business process models.

As the size of these repositories increases, automated techniques for managing them become important. This includes techniques for searching through a repository of business process models in order to retrieve models that are similar to a given one. We call this form of search similarity search. Similarity search can be used for various purposes. For example, when two organizations merge, their business analysts need to understand
the similarity and overlap between the business processes that were previously ran separately, in order to establish the opportunities for operational consolidation. If both organizations have their operations described in terms of business process models, a search through their collections for business process models that are similar would reveal such opportunities. As another example, an organization that maintains a large collection of business process models may want to inspect that collection on a regular basis in order to identify pairs of groups of models with high similarity. High similarity between models may indicate that anti-patterns such as overlap exist in the collection, which should be removed by means of refactoring.

This chapter explains how graph theory can be used as a basis for similarity search. To this end it presents the concept of business process graph and subsequently defines several techniques for graph-based business process similarity search. The chapter aims to present a comprehensive overview of work that we have undertaken in this area (Dijkman, Dumas & García-Bañuelos, 2009; Dijkman, Dumas, García-Bañuelos & Käärik, 2009; Dijkman, van Dongen, Dumas, Käärik & Mendling, 2010; van Dongen, Dijkman & Mendling, 2008).

The remainder of this chapter is structured as follows. First, we explain the general concept of business process similarity, including the work that others have done in this area. Second, we present the concept of business process graph and define the various techniques for measuring business process similarity. Third, we present future research direction on the topic of graph-based business process similarity and finally we present conclusions.

**BACKGROUND**

A business process is a collection of related tasks that can be performed to achieve a certain goal. A business process model represents the tasks in a business process as well as their relations. Typically, the model is enriched with nodes that affect the order in which tasks can be performed. These are called control-nodes or gateways. Figure 1 shows six examples of business process models in the Business Process Modeling Notation (BPMN). Rounded rectangles represent tasks. Circles represent events that occur during a process. An empty circle represents an event that signals the start of the process. Consequently, in the first process, at the start of the process the task ‘buy goods’ is performed. A bulls-eye represents an event that signals the completion of the process. Diamonds represent gateways. A diamond marked with an ‘X’ is called an exclusive gateway and it represents a decision point at which the process can continue along exactly one out of a number of alternative paths. An exclusive gateway can also be used to merge multiple alternative paths. Similarly a diamond marked with an ‘O’ (called an inclusive gateway) represents a point at which one or more alternative paths can be chosen or merged while a diamond marked with a ‘+’ (called a parallel gateway) represents a point where multiple outgoing parallel threads are started, or a point where multiple incoming threads must be synchronized before the execution of the process can continue.

In this Chapter, we will review several approaches to defining business process similarity metrics. For uniformity, we define a business process similarity metric as a function that, given two business process models, returns a value between 0 and 1 (inclusive) that indicates the similarity between these business process models. The similarity metric should correspond to human judgment, such that models that are perceived to be more similar have a higher similarity value.

A business process similarity metric can be used as a query tool in order to perform similarity search queries. A similarity search query returns those models from a collection that have a sufficiently high similarity value when compared to a given model. In that case the models from the collection are called document models and the input