Crowd-Based Mining of Reusable Process Model Patterns

Carlos Rodríguez, Florian Daniel, and Fabio Casati
University of Trento, Via Sommarive 9, I-38123, Povo (TN), Italy
{crodriguez,daniel,casati}@disi.unitn.it

Abstract. Process mining is a domain where computers undoubtedly outperform humans. It is a mathematically complex and computationally demanding problem, and event logs are at too low a level of abstraction to be intelligible in large scale to humans. We demonstrate that if instead the data to mine from are models (not logs), datasets are small (in the order of dozens rather than thousands or millions), and the knowledge to be discovered is complex (reusable model patterns), humans outperform computers. We design, implement, run, and test a crowd-based pattern mining approach and demonstrate its viability compared to automated mining. We specifically mine mashup model patterns (we use them to provide interactive recommendations inside a mashup tool) and explain the analogies with mining business process models. The problem is relevant in that reusable model patterns encode valuable modeling and domain knowledge, such as best practices or organizational conventions, from which modelers can learn and benefit when designing own models.

Keywords: Model patterns, Pattern mining, Crowdsourcing, Mashups.

1 Introduction

Designing good business processes, i.e., modeling processes, is a non-trivial task. It typically requires not only fluency in the chosen modeling language, but also intimate knowledge of the target domain and of the common practices, conventions and procedures followed by the various actors operating in the given domain. These requirements do not apply to business processes only. We find them over and over again in all those contexts that leverage on model-driven formalisms for the implementation of process-oriented systems. This is, for instance, the case of data mashups, which are commonly based on graphical data flow paradigms, such as the one proposed by Yahoo! Pipes (http://pipes.yahoo.com).

In order to ease the modeling of this kind of data mashups (so-called pipes), in a previous work, we developed an extension of Pipes that interactively recommends mashup model patterns while the developer is modeling a pipe. A click on a recommended pattern automatically weaves the pattern into the partial pipe in the modeling canvas. Patterns are mined from a repository of freely accessible pipes models [11]. We mined patterns from a dataset of 997 pipes...
Fig. 1. A Yahoo! Pipes data mash-up model pattern for plotting news on a map. The mashup logic is expressed as data flow diagram.

taken from the “most popular” category, assuming that popular pipes are more likely to be functioning and useful. Before their use, patterns were checked by a mashup expert assuring their meaningfulness and reusability (e.g., see Figure 1 for an example of a good pattern). The extension is called Baya [13], and our user studies demonstrate that recommending model patterns has the potential to significantly lower development times in model-driven environments [12].

The approach however suffers from problems that are common to pattern mining algorithms: identifying support threshold values, managing large numbers of produced patterns, coping with noise (useless patterns), giving meaning to patterns, and the cold start problem. Inspired by the recent advent of crowdsourcing [6], the intuition emerged that it might be possible to attack these problems with the help of the crowd, that is, by involving humans in the mining process. The intuition stems from the observation that pure statistical support does not always imply interestingness [2], and that human experts are anyway the ultimate responsible for deciding about the suitability or not of patterns.

In this paper, we report on the results of this investigation and demonstrate that crowd-based pattern mining can indeed be successfully used to identify meaningful model patterns. We describe our crowd-based mining algorithm, the adopted software/crowd stack, and demonstrate the effectiveness of the approach by comparing its performance with that of the algorithm adopted in Baya. We also show how our results and lessons learned are applicable to and impact the mining of model patterns from business process models.

2 Background and Problem Statement

2.1 Reference Process Models: Data Mashups

Mashups are composite web applications that are developed by integrating data, application logic, and pieces of user interfaces [1]. Data mashups are a special type of mashups that specifically focuses on the integration and processing of data sources available on the Web. Typical data sources are RSS or Atom