Available online at www.sciencedirect.com

ScienceDirect

Information Systems 32 (2007) 685–712

www.elsevier.com/locate/infosys

Mining unconnected patterns in workflows

Gianluigi Greco[a,*], Antonella Guzzo[b,c], Giuseppe Manco[b], Domenico Saccà[b,c]

[a]Department of Mathematics, University of Calabria, Via P.Bucci 30B, 87036 Rende, Italy
[b]ICAR, CNR, Via P.Bucci 41C, 87036 Rende, Italy
[c]DEIS, University of Calabria, Via P.Bucci 41C, 87036 Rende, Italy

Received 21 September 2005; received in revised form 11 March 2006; accepted 22 May 2006
Recommended by M. Weske

Abstract

General patterns of execution that have been frequently scheduled by a workflow management system provide the administrator with previously unknown, and potentially useful information, e.g., about the existence of unexpected causalities between subprocesses of a given workflow. This paper investigates the problem of mining *unconnected* patterns on the basis of some execution traces, i.e., of detecting sets of activities exhibiting no explicit dependency relationships that are frequently executed together. The problem is faced in the paper by proposing and analyzing two algorithms. One algorithm takes into account information about the structure of the control-flow graph only, while the other is a smart refinement where the knowledge of the frequencies of edges and activities in the traces at hand is also accounted for, by means of a sophisticated graphical analysis. Both algorithms have been implemented and integrated into a system prototype, which may profitably support the enactment phase of the workflow. The correctness of the two algorithms is formally proven, and several experiments are reported to evidence the ability of the graphical analysis to significantly improve the performances, by dramatically pruning the search space of candidate patterns.

Keywords: Frequent patterns discovery; Graph mining; Workflow management

1. Introduction

A workflow is a partial or total automation of a business process, in which a collection of *activities* must be executed by humans or machines, according to certain procedural rules. Workflows may be conveniently defined, analyzed and supported by means of *Workflow Management System* (WfMS). These systems represent the most effective technological infrastructure for managing business processes in several application domains (cf. [1–4]), and they are, therefore, more and more utilized into enterprises. In fact, enhancing the functionalities of WfMSs has become a very active research area in recent years, and several efforts have been already spent in order to provide facilities for the human system administrator while designing complex processes as well as to offer an "intelligent" support in the decisions which have to be taken by the enterpriser during the enactment [5–9].

In this paper, we continue on the way of enhancing the functionalities of WfMSs by proposing some data

*Corresponding author. Tel.: +39 0984 496429; fax: +39 0984 839054.
E-mail addresses: ggreco@mat.unical.it (G. Greco), guzzo@icar.cnr.it (A. Guzzo), manco@icar.cnr.it (G. Manco), sacca@icar.cnr.it (D. Saccà).

0306-4379/$ - see front matter
doi:10.1016/j.is.2006.05.001

mining techniques, which are specifically tailored to help the human system administrator to gain some previously unknown, and potentially useful information from the low-level data collected into the *log* files during past enactments of the system. Differently from classical *process mining* techniques, where the focus is on extracting the "hidden" model underlying the low-level data (see, e.g., [10–13] and the discussion in Section 6), in this paper we deal with the enactment phase of the workflow life-cycle: we assume that a workflow model has been already designed and installed over some platforms, and we want to provide an effective support for the decisions to be taken during each execution, based on the history of past executions. To this aim, specific data mining techniques can be used depending on the perspectives workflow specifications are looked from (cf. [14]).

In a simple scenario where one looks at workflows from a *data perspective* only, classical mining techniques, such as the "market basket analysis", can be used to find interesting and potentially useful knowledge about the business data at hand. However, these classical techniques do not fit scenarios where one looks at workflows from a *control-flow perspective*, in which the focus is instead on the causal relations and on the constraints on the occurrence of the tasks. For instance, the administrator may be interested in knowing whether there are subprocesses frequently scheduled in the same enactment, or whether there are correlations (in previous enactments) between the order of execution of a set of activities and the execution of a specific final activity.

In this perspective, important correlations among activities are already hardwired in the workflow model, independently from any execution history and, therefore, the mining may be effectively driven by the workflow structure to detect unexpected and useful information only.

And, in fact, the peculiarities of the domain raise the necessity to deal with the natural graphical representation of control flows, so that specific mining algorithms have to be conceived. An example applicative context is next illustrated by considering the automatization of a "sales ordering" process.

Example 1. Consider the workflow schema depicted in Fig. 1, according to the graphical notation of the *event-driven process chains* (*EPCs*) [15], where the process is represented as chains of *events* (drawn as hexagons) and *functions* (drawn as boxes) connected by means of logical *connectors*. The schema supports a sales order process as follows. As soon as a customer issues a request to purchase a given product, the enterpriser checks the availability of the required items. Three different stocks are available: stocks X and Y, containing items of type A; and stock Z which contains items of type B. If the requested items are not available in the respective stocks, the order is rejected and the corresponding transaction is aborted. Otherwise, the order is accepted and the items are shipped to the customer. In this latter case, two further events may occur: either the trial period elapses without any claim by the customer, or the customer issues a claim for defective item (in which case a substitution has to be accomplished).

In this scenario, it could be crucial to characterize the discriminant factors that will lead to the rejection of the order or to the shipping of a defective item. In fact, specialized mining techniques may reveal important information to be used to diagnose the business process and to identify problems within the supply chain. For instance, by looking at the traces

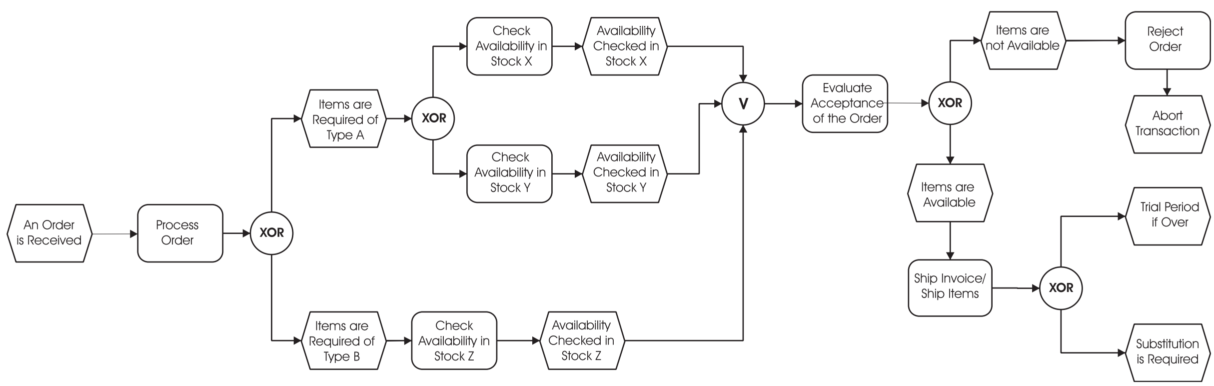

Fig. 1. An example workflow schema.