Computers in Industry 61 (2010) 463–471

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

A workflow net similarity measure based on transition adjacency relations

Haiping Zha [a,b,e,1], Jianmin Wang [b,c,d,*], Lijie Wen [b,c,d,2], Chaokun Wang [b,c,d,3], Jiaguang Sun [b,c,d]

[a] Department of Computer Science, Tsinghua University, Beijing, China
[b] School of Software, Tsinghua University, Beijing, China
[c] Key Laboratory for Information System Security, Ministry of Education, China
[d] Tsinghua National Laboratory for Information Science and Technology, Beijing, China
[e] Institute of Specifications and Standards, Shanghai 200235, China

ARTICLE INFO

Article history:
Received 12 March 2009
Accepted 8 January 2010
Available online 12 March 2010

Process similarity
Process distance
Transition adjacency relation
Workflow net
Model reduction

ABSTRACT

Many activities in business process management, such as process retrieval, process mining, and process integration, need to determine the similarity or the distance between two processes. Although several approaches have recently been proposed to measure the similarity between business processes, neither the definitions of the similarity notion between processes nor the measure methods have gained wide recognition. In this paper, we define the similarity and the distance based on firing sequences in the context of workflow nets (WF-nets) as the unified reference concepts. However, to many WF-nets, either the number of full firing sequences or the length of a single firing sequence is infinite. Since transition adjacency relations (TARs) can be seen as the genes of the firing sequences which describe transition orders appearing in all possible firing sequences, we propose a practical similarity definition based on the TAR sets of two processes. It is formally shown that the corresponding distance measure between processes is a metric. An algorithm using model reduction techniques for the efficient computation of the measure is also presented. Experimental results involving comparison of different measures on artificial processes and evaluations on clustering real-life processes validate our approach.

1. Introduction

Nowadays, process-aware information systems, such as WFM (Workflow Management system), SCM (Supply Chain Management), PDM (Product Data Management system) and ERP (Enterprise Resource Planning) have been widely adopted in industry to offer generic modeling and enactment capabilities for structured business processes [12]. Business processes accumulated in various information systems have become important intellectual assets which represent the real-life business handling procedures of the organizations. A deep insight into these business processes and their mutual relationship is necessary to business management activities. There are many applications in business process management that require measuring the similarity between business processes, such as process retrieval, process mining, and process integration. For example, the task items of a workflow process in Teamcenter PDM [25] is established before the control-flow structure. Therefore, different workflow designers may construct different processes with the same set of task items. We need to determine whether the behaviors of these models are equivalent or not, and how different their behaviors are in case of inequivalence. Another example is China CNR Corporation Limited [10] which is a newly regrouped company which has more than 20 subsidiary companies. Before the corporation was established, most of these subsidiary companies independently deployed ERP systems with a total of more than 200,000 process models. Now, CNR needs to integrate these ERP systems. How to group or cluster these processes is a big problem. Measuring the similarity between process models automatically will be helpful in tackling these challenges.

Researchers working on formal methods have proposed a variety of equivalence notions to compare the behaviors between processes, such as trace equivalence, bisimulation, and branching bisimulation [21,15,19]. However, these equivalence notions can only tell a binary answer, i.e., equivalence or inequivalence. If we measure the similarity between processes based on equivalence notions, e.g., mapping equivalence to 1 and inequivalence to 0, such a measure is not very useful because it cannot distinguish between complete difference and slight difference of process behaviors. As to Fig. 1, we investigate process behaviors in the context of trace semantics. The firing sequences of Processes N_1,

* Corresponding author at: Institute of Information System & Engineering, School of Software, Tsinghua University, Room 819, Main Building, Tsinghua University, Beijing 100084, China 100084. Tel.: +86 10 62781776; fax: +86 10 62781776.
E-mail addresses: chp04@mails.tsinghua.edu.cn (H. Zha), jimwang@tsinghua.edu.cn (J. Wang), wenlj00@tsinghua.edu.cn (L. Wen), chaokun@tsinghua.edu.cn (C. Wang), sunjg@tsinghua.edu.c (J. Sun).
[1] Room 1402B, Building Zijing No 14, Tsinghua University, Beijing 100084, China. Tel.: +86 10 51537905; fax: +86 10 62773417.
[2] Tel: +86 10 62773417; fax: +86 10 62773417.
[3] Tel.: +86 10 62795393; fax: +86 10 62795393.

0166-3615/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.compind.2010.01.001

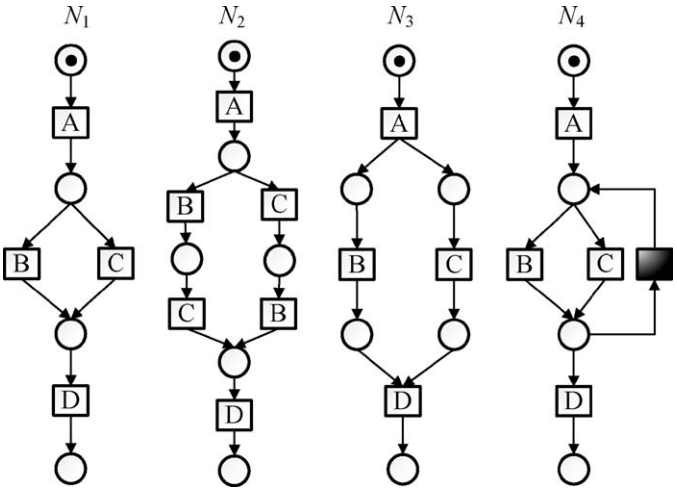

Fig. 1. Sample process models in WF-nets (Note that a black block in process N_4 represents a τ task).

N_2, and N_3 are $\{ABD, ACD\}$, $\{ABCD, ACBD\}$, and $\{ABCD, ACBD\}$, respectively. The firing sequences of Process N_4 cannot be enumerated one by one (i.e., infinite) because there is a loop structure in Process N_4. Intuitively, these models are similar in some sense. However, with the trace equivalence notion, we cannot tell the degree of similarity between them except that Processes N_2 and N_3 are equivalent.

Recent literature has proposed different approaches to quantifying the similarity between business processes [2,18,11]. Although firing sequences are a good representation of the behavior of a process, they are neither possible (e.g., with loop structures) nor practical due to the high complexity of exploring state space of concurrent models [14]. Therefore, the proposed approaches for measuring similarity are all based on substitute representations, such as on causal footprints [11] or observed behavior [18] in the context of different definitions of the similarity notion. However, the results of these similarity measures cannot be compared with each other because they lack a unified definition of the similarity between processes.

In this paper, we define the similarity and the distance between two business processes on full firing sequences of the processes as the unified reference concepts. However, the full firing sequences of a process are not always available. For example, to a WF-net with loop structures, either the number of full firing sequences or the length of a single firing sequence is infinite. Considering that transition adjacency relations (TARs) can be seen as the genes of the firing sequences because the TAR set can specify the exact behavior of a large category of processes (e.g., SWF-nets) [6], we define a computable similarity based on the TAR set. The TAR set describes transition orders that appear in all possible firing sequences (one directly follows the other). For example, the TAR sets of Processes N_1, N_2, N_3 and N_4 are $\{AB, AC, BD, CD\}$, $\{AB, AC, BC, CB, BD, CD\}$, $\{AB, AC, BC, CB, BD, CD\}$ and $\{AB, AC, BC, CB, BB, CC, BD, CD\}$, respectively, as shown in Fig. 2. We can see that the TAR set of Process N_4 is finite despite of its infinite firing sequences. The TAR set has three features. Firstly, the TAR set represents the essence of the firing sequences of a process. Secondly, for any processes with a finite number of tasks, the cardinality of the TAR set is finite. Thirdly, the TAR set of a WF-net can be generated in less time than that of a full firing sequence generation. At the same time, the corresponding distance measure between business processes is also defined. The distance measure is proved to be a metric which can be used in a wide range of applications.

This paper makes three contributions: firstly, it presents a definition of the similarity (and the distance) between processes based on firing sequences as a unified reference concept; secondly, it proposes a similarity measure (and distance measure) based on transition adjacency relation set; thirdly, both distance measures are proved to be metrics.

The remainder of this paper is organized as follows. Section 2 introduces some basic concepts used throughout the paper. Section 3 describes the details of our approach to measuring similarity between business processes. Section 4 proposes an algorithm intended to decrease the time required for the TAR set generation. Section 5 shows the experimental results. Section 6 discusses the related work, and Section 7 concludes the paper and outlines future work.

2. Preliminaries

In this section, we introduce some basic concepts used throughout the paper, such as Petri nets and WF-nets. For a more detailed introduction, we refer the readers to [20,1].

2.1. Petri nets

Petri net is a formal language which can be used to specify processes. Since the language has formal and executable semantics, processes modelled in Petri nets can be executed by an information system. A Petri net consists of three modeling elements: *Transition*, which typically corresponds to either an activity (task, process step) which needs to be executed, or a silent step (i.e., τ task) which is used for routing purposes; *Place*, which is used to define the preconditions and postconditions of transitions. A transition can be fired (executed) if the precondition is satisfied. The result of such a firing will be that the postcondition holds; transitions and places are connected through directed *arcs* in such a way that (i) places and transitions have at least one incident edge and (ii) in every path, transitions and places alternate (no place is connected to another place and no transition is connected to another transition). A formal definition of conventional Petri nets is presented as follows.

Definition 1. (Petri nets) A Petri net is a triple (P, T, F), where:

- P is a finite set of places,
- T is a finite set of transitions, $P \cup T \neq \varnothing$ and $P \cap T = \varnothing$,
- $F \subseteq (P \times T) \cup (T \times P)$ is a set of directed arcs representing flow relations, joining places and transitions together.

To denote the state of process execution, the concept of a *token* is used. A *token* in a place shows that a certain condition holds. Each place can arbitrarily contain many of such tokens. If a transition

N_1	A	B	C	D
A		1	1	
B				1
C				1
D				

N_2	A	B	C	D
A		1	1	
B			1	1
C		1		1
D				

N_3	A	B	C	D
A		1	1	
B			1	1
C		1		1
D				

N_4	A	B	C	D
A		1	1	
B		1	1	1
C		1	1	1
D				

Fig. 2. The TAR sets of processes shown in Fig. 1 where 1 means there is a TAR.