Exploring the CSCW Spectrum using Process Mining

Wil M.P. van der Aalst

Department of Technology Management, Eindhoven University of Technology,

P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands.

w.m.p.v.d.aalst@tm.tue.nl

(telephone: +31 40 247.4295/2290, fax: +31 40 243.2612)

Abstract

Process mining techniques allow for extracting information from event logs. For example, the audit trails of a workflow management system or the transaction logs of an enterprise resource planning system can be used to discover models describing processes, organizations, and products. Traditionally, process mining has been applied to structured processes. In this paper, we argue that process mining can also be applied to less structured processes supported by Computer Supported Cooperative Work (CSCW) systems. In addition, the ProM framework is described. Using ProM a wide variety of process mining activities are supported ranging from process discovery and verification to conformance checking and social network analysis.

Keywords: Process Mining, Business Activity Monitoring, Business Process Intelligence, CSCW, Data Mining.

1. Introduction

Buzzwords such as BAM (Business Activity Monitoring), BOM (Business Operations Management), BPI (Business Process Intelligence) illustrate the interest in closing the BPM (Business Process Management) loop [2]. This is illustrated by Figure 1 which shows the level of support in four different years using the BPM lifecycle. The lifecycle identifies four different phases: *process design* (i.e., making a workflow schema), *system configuration* (i.e., getting a system to support the designed process), *process enactment* (i.e., the actual execution of the process using the system), and *diagnosis* (i.e., extracting knowledge from the process as it has been executed). As Figure 1 illustrates, BPM technology (e.g., workflow management systems) started with a focus on getting the system to work (i.e., the system configuration phase) [2]. Since the early nineties BPM technology matured and more emphasis was put on supporting the process design and process enactment phases in a better way. Now many vendors are trying to close the BPM lifecycle by adding diagnosis functionality [4,5]. The buzzwords BAM, BOM, BPI, etc. illustrate these attempts.

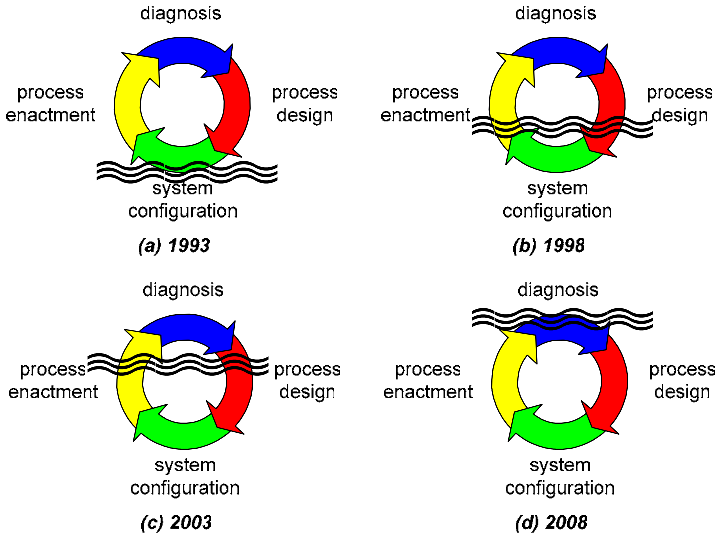


Figure 1: The level of support is rising

The diagnosis phase assumes that data is collected in the enactment phase. Most information systems provide some kind of *event log* (also referred to as transaction log or audit trail). Typically such an event log registers the start and/or completion of activities. Every event refers to a case (i.e., process instance) and an activity, and, in most systems, also a timestamp, a performer, and some additional data.

Process mining techniques [4,5] take an event log as a starting point to extract knowledge, e.g., a model of the organization or the process. For example, the ProM (Process Mining) framework developed at Eindhoven University of Technology provides a wide range of process miming techniques.

This paper discusses process mining techniques, and, in particular, the techniques supported by the ProM framework, in the context of Computer Supported Cooperative Work (CSCW) [11]. The CSCW domain provides a very broad range of systems that support "work" in all its forms. Workflow Management (WFM) systems and BPM systems can be seen as particular CSCW systems aiming at well-structured office processes. In this paper, we explore the application of process mining in the broader CSCW domain. The goal is to trigger new applications of process mining and to define interesting scientific and practical challenges.

The remainder of the paper is organized as follows. First, we discuss the CSCW spectrum of systems. Then we introduce the concept of process mining followed by an introduction to the ProM framework. Then we discuss the application of process mining in several domains of the CSCW spectrum. We use the systems Staffware (Staffware Tibco), InConcert (Tibco), Outlook (Microsoft), SAP R/3 (SAP AG), and FLOWer (Pallas Athena) as concrete examples in the wide range of CSCW systems that can be used as a starting point for process mining. Finally, we discuss related work and conclude the paper.

2. CSCW spectrum

There exist many definitions of the term *Computer Supported Cooperative Work* (CSCW). Some emphasize the support of work processes while others emphasize the fact that people work in groups [11,12,21]. Within the CSCW domain there has been a constant struggle between technological views and sociological views. A nice illustration is the so-called "Winograd-Suchman debate" in the early nineties [17,22,24,25]. Winograd and Flores advocated the use of a system called the "coordinator", a system based on Speech act theory (i.e., the language/action perspective) in-between e-mail and workflow technology [24,25]. People like Suchman and others argued that such systems are undesirable as they "carry an agenda of discipline and control over an organization's members" [22]. Clearly, process mining adds another dimension to this discussion. The goal of process mining is not to control people. However, it can be used to monitor and analyze the behavior of people and organizations. Cleary, such technology triggers ethical questions. However, such questions are beyond the scope of this paper. Instead, we want to focus on the applicability of process mining in the broader context of CSCW. Therefore, we first explore the *CSCW spectrum*.

Many authors provide a classification of CSCW [10,11,12]. The classical paper by Ellis et al. [11] classifies groupware systems using two taxonomies: the space/time taxonomy and the application-level taxonomy. The *space/time taxonomy* classifies interaction into same place/different places and same time/different times. For example, a face-to-face meeting is "same place and same time" interaction while the exchange of e-mails is "different places and different times" interaction. The *application-level taxonomy* classifies systems based on the purpose they serve.

A later classification given by Ellis distinguishes four classes of CSCW systems: (1) Keepers, (2) Coordinators, (3) Communicators, and (4) Team-agents [10].

Keepers support the access to and modification of shared artifacts. Typical issues that are of primary concern to keepers are access control, versioning, backup, recovery, and concurrency control. Examples of keepers include the vault in a Product Data Management (PDM) system, a repository with drawings in a CAD/CAM system, and a multi media database system.

Coordinators are concerned with the ordering and synchronization of individual activities that make up the whole process. Typical issues addressed by coordinators are process design, process enactment, enabling of activities, and progress monitoring. The key functionality of a workflow management system is playing the role of coordinator.

Communicators are concerned with explicit communication between participants in collaborative endeavors. Typical examples are electronic mail systems and video conferencing systems, and basic issues that need to be addressed are message passing (broadcast, multicast, etc.), communication protocols, and conversation management.

Team-agents are specialized domain-specific pieces of functionality. A team agent is typically a system acting on behalf of a specific person or group and executing a specific task. Examples include an electronic agenda and a meeting scheduler.

The classifications described in literature are not very meaningful when considering process mining in the context of CSCW. Moreover, in literature CSCW is typically restricted to a small class of software products named "groupware" while more successful products supporting work are excluded. (Since the "Winograd-Suchman debate" some CSCW researchers consider workflow management software and the like not part of the CSCW spectrum. However, one should realize that widely used software products ranging from ERP to CRM and call-center systems support workflow-like functionality.) Therefore, we propose another classification based on two dimensions as shown in Figure 2. On the one hand we distinguish between *data centric* (i.e., the focus is on the sharing and exchange of data) and *process centric* (i.e., the focus is on the ordering of activities) approaches/systems. On the other hand we distinguish between *structured* (there is a predefined way of dealing with things) and *unstructured* (things are handled in an ad-hoc manner) approaches/systems.

Production workflow systems [2] such as Staffware (Tibco-Staffware), MQ Series Workflow (IBM), etc. are process centric and support structured activities. Note that these systems only support predefined processes and focus on control-flow rather than data-flow. Ad-hoc workflow systems such as InConcert support unstructured activities in a process centric manner, i.e., each process instance has a specific process model that may be modified and extended on-the-fly. Groupware products, including e-mail systems such as Outlook, typically are data centric and support unstructured activities. i.e., they are unaware of some predefined process. Note that here we interpret groupware in a narrower sense, and not as broad as in [10,11,12]. Finally, there is a wide variety of systems that are data centric while focusing on structured processes. A typical example is the ERP system SAP R/3 which can be viewed as a set of applications built on top of a complex database. Parts of SAP R/3 are process-aware (e.g., the workflow module Webflow), but in most cases the presence of data enables certain activities rather than some explicit process model. Case handling systems such as FLOWer (Pallas Athena) support a mixture of structure and unstructured processes using a combination of a data centric and process centric approach [7]. Therefore, they are positioned in the middle of the CSCW spectrum.

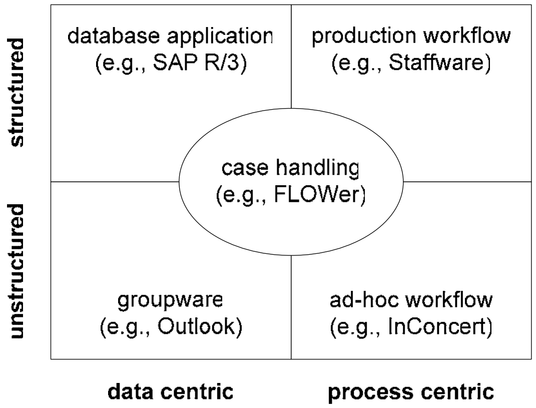


Figure 2: CSCW Spectrum

We will use Figure 2 to discuss the relevance of process mining in the context of CSCW. However, before doing so, we briefly introduce the concept of process mining.

3. Process mining

3.1 Process Mining: An Example

The goal of process mining is to extract information about processes from transaction logs [6]. We assume that it is possible to record events such that (i) each event refers to an *activity* (i.e., a well-defined step in the process), (ii) each event refers to a *case* (i.e., a process instance), (iii) each event can have a *performer* also referred to as *originator* (the person executing or initiating the activity), and (iv) events have a *timestamp* and are totally ordered [4]. In addition events may have associated data (e.g., the outcome of a decision). Events are recorded in a so-called *event log*. To get some idea of the content of an event log consider the fictive log shown in Table 1.

case id	***activity id***	***originator***	***timestamp***
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity D	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity H	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity H	John	10-3-2004:12.34
case 3	activity E	Pete	10-3-2004:12.50
case 3	activity F	Carol	11-3-2004:10.12
case 4	activity D	Pete	11-3-2004:10.14
case 3	activity G	Sue	11-3-2004:10.44
case 3	activity H	Pete	11-3-2004:11.03
case 4	activity F	Sue	11-3-2004:11.18
case 4	activity E	Clare	11-3-2004:12.22
case 4	activity G	Mike	11-3-2004:14.34
case 4	activity H	Clare	11-3-2004:14.38

Table 1: An example of an event log

As we will show later, logs having a structure similar to the one shown in Table 1 are recorded by a wide variety of CSCW systems. This information can be used to extract knowledge. For example, the Alpha algorithm described in [1,6] can be used to derive the process model shown in Figure 3.

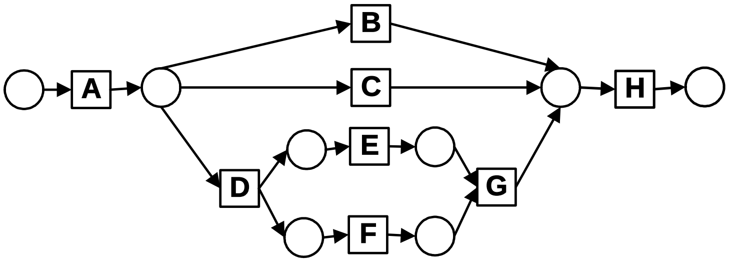


Figure 3: A process model derived from Table 1 and represented in terms of a Petri net

It is important to note that the Alpha algorithm is just one of the many process mining techniques available. For example, it is possible to extract a social network based on an event log. For more details we refer to [3] and Section 6.

3.2 Overview of Process Mining and Related Topics

Figure 4 provides an overview of process mining and the various relations between entities such as the information system, operational process, event logs and process models. Note that although Figure 4 is focusing on process perspective, process mining also includes other perspectives such as the organizational and data perspectives [3]. We will discuss all three perspectives later, but first we focus on the architecture shown in Figure 4.

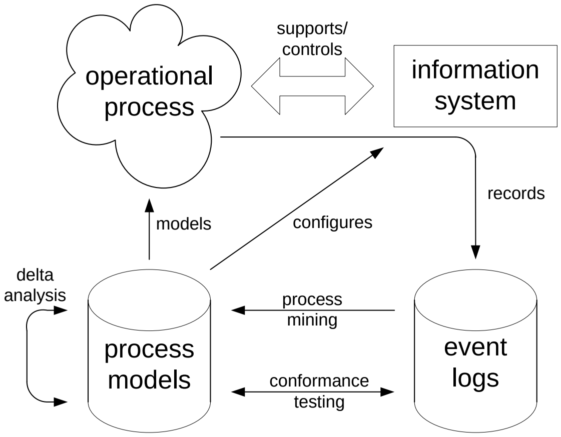


Figure 4: Overview of process mining and related topics

Figure 4 shows the *operational process* (e.g., the flow of patients in a hospital, the handling of insurance claims, the procurement process of a multinational, etc.) that is interacting with some *information system* (e.g., and ERP, CRM, PDM, BPM, or WFM system). Clearly the information system and the operational process exchange information, e.g., the system may support and/or control the process at hand. Related to the information system and processes it supports are *process models* and *event logs*. As discussed before, many systems log events related to some process (cf. the arrow labeled *records* in Figure 4). The role of models is more involved. Clearly, process models can be used to model the operational process for a variety of reasons. Process models can be used to analyze and optimize processes but can also be used for guidelines, training, discussions, etc.   (cf. the arrow labeled *models* in Figure 4). However, increasingly information systems are configured on the basis of models (cf. the arrow labeled *configures* in Figure 4). For example, consider process-aware systems ranging from production workflow systems such as Staffware and COSA to ERP systems like SAP R/3 and BaaN. Models can be *prescriptive* or *descriptive*. If they are used for configuration, they tend to be prescriptive. If they are used for other purposes, they are often descriptive.

Both the models and the event logs can be seen as some abstraction from the operational process. While event logs record the actual events being logged, the process model focuses at the aggregated level (also referred of as "type level"). The goal of process mining is to extract models from event logs (cf. the arrow labeled *process mining* in Figure 4). Based on the observations recorded in the log, some model is derived. Like in classical data mining it is possible to derive relationships, e.g., causality relations, interaction patterns, and dependencies. Pure process mining just focusing on discovery is complemented by delta analysis and conformance testing. Delta analysis is used to compare a predefined model (prescriptive or descriptive) and a discovered model (cf. the arrow labeled *delta analysis* in Figure 4). Conformance testing is concerned with comparing a model and an event log. This can be used to investigate the fitness and appropriateness of a model (cf. the arrow labeled *conformance testing* in Figure 4). For example, it can be used to measure alignment. To illustrate the use of delta analysis and conformance testing, consider the SAP R/3 reference model expressed in terms of Event-driven Process Chains (EPCs). The EPCs describe best practices, but the SAP system does not enforce people to follow these best practices. Using conformance testing, the actual logs can be compared with the EPCs and indicate where organizations/people deviate. Instead of directly comparing the logs and the models, it is also possible to first do process mining and compare the result with the original model using delta analysis.

3.3 Three Mining Perspectives

As indicated before process mining is not restricted to the process perspective (also referred to as control-flow) and also includes other perspectives such as the organizational and data perspectives [3]. In this section, we briefly discuss the three dominant mining perspectives in more detail.

The *process perspective* is concerned with the control-flow, i.e., the causal ordering of activities. Consider again Table 1. For the process perspective only the first two columns are relevant and the goal is to derive a process model, e.g., the Petri net shown in Figure 3. To do this we can first translate the table in an audit trail for each case, i.e., case 1: <A,B,H>, case 2: <A,C,H>, 3: <A,D,E,F,G,H>, and case 4: <A,D,F,E,G,H>. Given these traces we apply Occam's Razor, i.e., "one should not increase, beyond what is necessary, the number of entities required to explain anything". This tells us that the process holds activities A, B, C, D, E, F, G, and H. Every process starts with A and end with H. In-between there is a choice between executing (1) B only, (2) C only, or (3) D, E, F, and G. In the latter case, first D is executed followed by both interleavings of E and F, followed by G. Using Occam's principle we deduce that E and F are in parallel. Using a variety of algorithms (e.g., the Alpha algorithm developed by the author) we can deduce the Petri net shown in Figure 3. It is important to note that process mining *should* not require *all* possible observations to be present in the log. This happens to be the case for Table 1/Figure 3, but in general only fraction of the possible behavior will actually be observed. Consider for example a process with 10 binary choices between two alternative activities. In this case one would need to see 2^{10}=1024 different traces. If 10 activities are in parallel, one would need even 10!=3628800 different traces. In such cases one should not expect to see all possible traces, but simply look for the most likely candidate model. This is the reason we are not only using algorithmic approaches and also use heuristics and genetic mining.

The *organizational perspective* is concerned with the organizational structure and the people within the organizational units involved in the process. The focus of mining this perspective is on discovering organizational structures and social networks. Note that Figure 3 completely ignores the third column in Table 1. Nevertheless this column may be used to derive interesting knowledge. For example, it is possible to discover which people typically work together, which people execute similar activities, etc. This can be used to build social networks, i.e., directed graphs where each node represents a person and weighted arcs connecting these nodes represent some relationship.

The *data perspective* is concerned with case and the data associated to cases. Table 1 does not hold any data. However, in reality case and activities have associated data (e.g., the amount of money involved, the age of a customer, the number of order-lines, etc.). Such information may be combined with the columns shown in Table 1 to answer interesting questions such as: "Do large orders take more time than small orders?", "What is the average flow time of cases where John is involved?", "Does the treatment of male patients differ from the treatment of female patients?".

We have been working on techniques and tools to mine each of the three perspectives mentioned. The next section describes the tool that has been used to mine different work processes in the CSCW spectrum.

4. ProM

After developing a wide variety of mining prototypes (e.g., EMiT, Thumb, MinSon, MiMo, etc.), we merged our mining efforts into a single mining framework: the *ProM framework*. Figure 5 shows a glimpse of the architecture of ProM. It supports different systems, file formats, mining algorithms, and analysis techniques. It is possible to add new (mining) plug-ins without changing the framework.

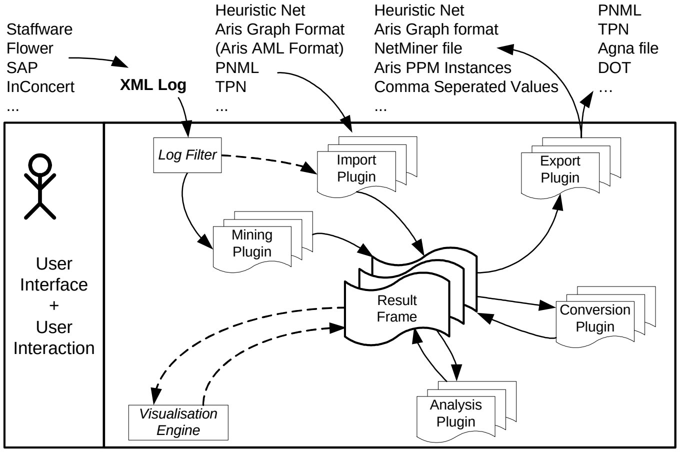


Figure 5: Architecture of ProM

Currently more than 30 plug-ins have been realized to offer a wide variety of process mining capabilities. Instead of elaborating on these plug-ins we show some results based on the log shown in Table 1.

Figure 6 shows the result of applying the Alpha algorithm [6] to the event log shown in Table 1. Note that indeed the process shown in Figure 3 is discovered. Since ProM is multi-format it is also possible to represent processes in terms of an EPC or any other format added to the framework.

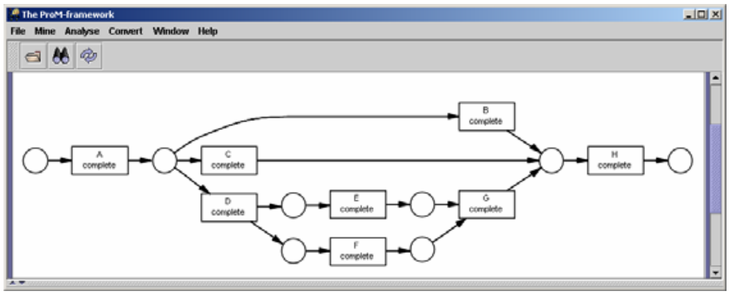


Figure 6: Applying the Alpha plug-in to Table 1

Figure 7 shows a social network [3] based on the event log shown in Table 1. Now nodes represent actors rather than activities.

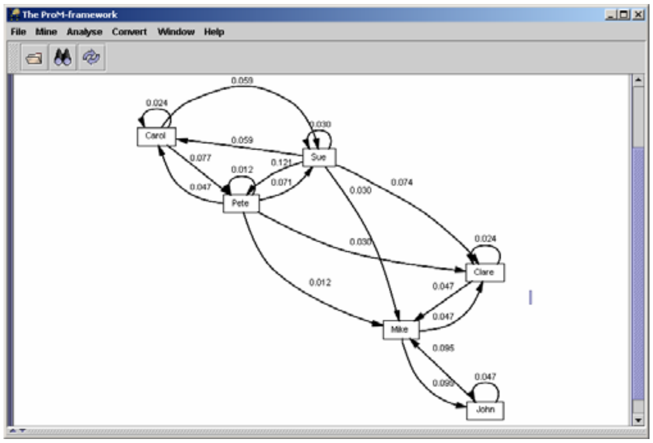


Figure 7: Applying the social network miner plug-in to Table 1

For more information on the ProM framework or to download the toolset we refer to www.processmining.org. In the remainder of this paper we focus on five example systems covering the CSCW spectrum shown in Figure 2.

5. Mining the CSCW spectrum

In Section 2 we introduced a classification of CSCW-like systems, based on two dimensions: (1) data or process centric and (2) structured or unstructured. In this section we give concrete examples and discuss how process mining techniques could be deployed in a meaningful way.

5.1 Example: Staffware

Tibco recently acquired Staffware and its workflow product. Staffware is a classical production workflow system aiming at high-volume highly-repetitive processes. Therefore, it is a typical candidate of the upper-right quadrant in Figure 2 (structure – process centric).

Figure 8 shows the process designer of Staffware. Like most other systems in the upper-right quadrant in Figure 2, Staffware is able to generate audit trails that can be used as input for process mining.

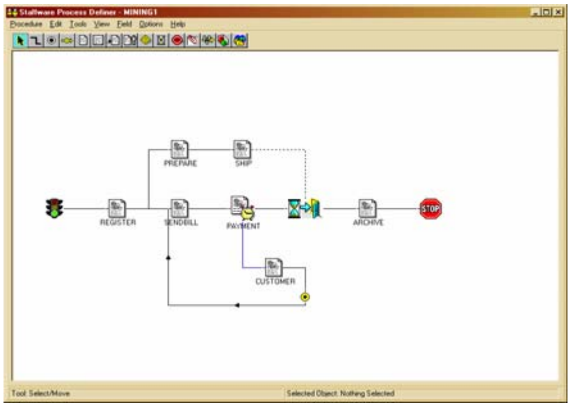
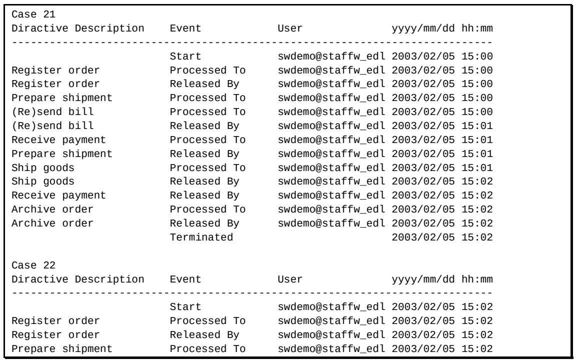

Staffware Process Definer - MINING1
Procedure Edit Tools View Field Options Help
PREPARE
SHIP
REGISTER
SENDBILL
PAYMENT
ARCHIVE
STOP
CUSTOMER
Tool: Select/Move
Selected Object: Nothing Selected

Figure 8: Screenshot of Staffware designer

Figure 9 shows a fragment of a Staffware log. Note that the content of the log is similar to the content of the event log shown in Table 1. Therefore, process mining tools such as ProM have no problems using Staffware logs as input for process mining activities.



```
Case 21
Diractive Description    Event            User              yyyy/mm/dd hh:mm
----------------------------------------------------------------------------
                         Start            swdemo@staffw_edl 2003/02/05 15:00
Register order           Processed To     swdemo@staffw_edl 2003/02/05 15:00
Register order           Released By      swdemo@staffw_edl 2003/02/05 15:00
Prepare shipment         Processed To     swdemo@staffw_edl 2003/02/05 15:00
(Re)send bill            Processed To     swdemo@staffw_edl 2003/02/05 15:00
(Re)send bill            Released By      swdemo@staffw_edl 2003/02/05 15:01
Receive payment          Processed To     swdemo@staffw_edl 2003/02/05 15:01
Prepare shipment         Released By      swdemo@staffw_edl 2003/02/05 15:01
Ship goods               Processed To     swdemo@staffw_edl 2003/02/05 15:01
Ship goods               Released By      swdemo@staffw_edl 2003/02/05 15:02
Receive payment          Released By      swdemo@staffw_edl 2003/02/05 15:02
Archive order            Processed To     swdemo@staffw_edl 2003/02/05 15:02
Archive order            Released By      swdemo@staffw_edl 2003/02/05 15:02
                         Terminated                         2003/02/05 15:02

Case 22
Diractive Description    Event            User              yyyy/mm/dd hh:mm
----------------------------------------------------------------------------
                         Start            swdemo@staffw_edl 2003/02/05 15:02
Register order           Processed To     swdemo@staffw_edl 2003/02/05 15:02
Register order           Released By      swdemo@staffw_edl 2003/02/05 15:02
Prepare shipment         Processed To     swdemo@staffw_edl 2003/02/05 15:02
```

Figure 9: Fragment of a Staffware event log

We have implemented a converter from Staffware logs to the XML format used by the ProM framework. An interesting observation is that Staffware logs the offering of work items to people and the completion of the corresponding activities. However, it does not log the actual start of an activity. As a result, it is not possible to measure service times and the utilization of the workforce.

5.2. Example: InConcert

InConcert is an ad-hoc workflow system that is quite different from production workflow systems like Staffware. It is one of the few tools in the lower-right quadrant in Figure 2 (unstructured – process centric). As such it is an interesting tool with unique capabilities. For example, it is possible to create templates from old cases and use them to process new cases. It is also possible to adapt a single case or to model a process model while executing a case (emerging processes).

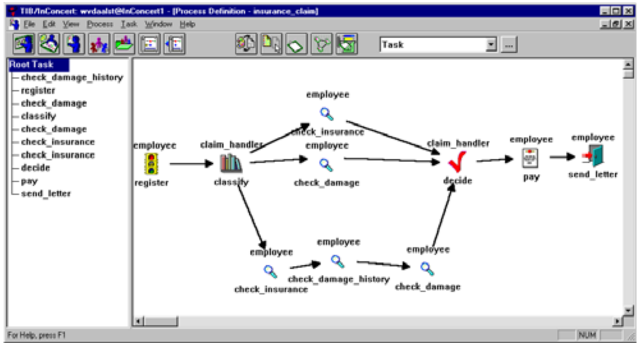


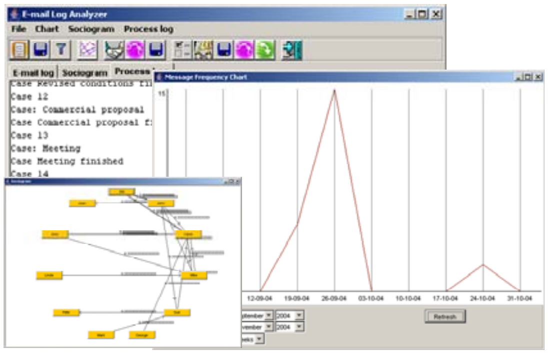
Figure 10: Screenshot of InConcert

Figure 10 shows a screenshot of InConcert. Despite its unique features, the current status of the product is unclear. In 1999 Tibco acquired the tool from Xerox and integrated it into the Tibco BusinessWorks platform. In 2004 Tibco also acquired Staffware making it unclear how Tibco will reconcile the various workflow products.

From a process mining point of view it is interesting that every case has its own process model. In ProM we embedded special mining algorithms ("multi-phase mining") to mine from instance models rather than audit trails. Given the unclear future of InConcert, we did not develop an adaptor for InConcert. Instead the multi-phase mining plug-ins can interface with tools such as ARIS Process Performance Monitor (ARIS PPM, a product of IDS Scheer).

5.3. Example: Outlook

The lower-left quadrant in Figure 2 is more heterogeneous. E-mail programs such as Outlook are probably the most widely used software in this quadrant. Several tools are able to construct social networks from e-mail traffic (e.g., MetaSight, BuddyGraph, etc.). In the context of the ProM framework we have developed a tool to not only generate a social network [3] but also process models.



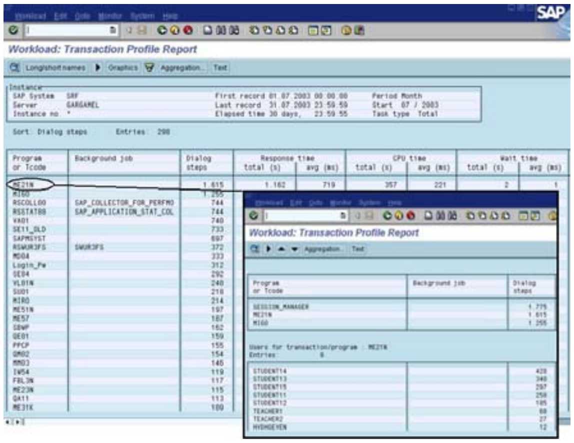
```
Case Revised conditions fi
Case 12
Case: Commercial proposal
Case Commercial proposal f
Case 13
Case: Meeting
Case Meeting finished
Case 14
```

Figure 11: Mining tool to generate event logs from e-mail messages

The challenge of process mining is to identify the case and the task for each event that is recorded. For example, given an e-mail message it is easy to see sender, receiver, timestamp, etc. However, if the e-mail is a step in some process, how to recognize the task and how to link the e-mail message to a specific case. Figure 11 shows the tool we have developed to do such things. Information such as threads, subject information, and special annotations are used to extract meaningful event logs.

5.4. Example: SAP R/3

The upper-left quadrant in Figure 2 is also very heterogeneous. SAP R/3 is probably the most relevant product in this quadrant. In the context of the ProM framework we have applied process mining techniques to the various logs recorded by SAP R/3. At the moment we are also investigating PeopleSoft.



Workload: Transaction Profile Report

```
Instance
SAP System   SRF                     First record 01.07.2003 00:00:00     Period Month
Server       GARGAMEL                Last record  31.07.2003 23:59:59     Start  07 / 2003
Instance no. *                       Elapsed time 30 days,   23:59:55     Task type  Total

Sort: Dialog steps     Entries:  298
```

Program or Tcode	Background job	Dialog steps	Response time total (s)	Response time avg (ms)	CPU time total (s)	CPU time avg (ms)	Wait time total (s)	Wait time avg (ms)
ME21N		1.615	1.162	719	357	221	2	1
MI60		1.255						
RSCOLL00	SAP_COLLECTOR_FOR_PERFMO	744						
RSSTAT88	SAP_APPLICATION_STAT_COL	744						
VA01		740						
SE11_OLD		733						
SAPMSYST		697						
RSWUR3FS	SWUR3FS	372						
MD04		333						
Login_Pw		312						
SE04		292						
VL01N		240						
SU01		218						
MIRO		214						
ME51N		197						
ME57		187						
SBWP		162						
QE01		159						
PPCP		155						
QM02		154						
MM03		146						
IW54		119						
FBL3N		117						
ME23N		115						
QA11		113						
ME31K		109						

Workload: Transaction Profile Report

Program or Tcode	Background job	Dialog steps
SESSION_MANAGER		1.775
ME21N		1.615
MI60		1.255

```
Users for transaction/program : ME21N
Entries:            8
```

STUDENT14		428
STUDENT13		340
STUDENT15		297
STUDENT11		258
STUDENT12		185
TEACHER1		68
TEACHER2		27
HVDHOEVEN		12

Figure 12: Transaction log in SAP R/3 obtained through transaction code ST03

SAP R/3 provides many logs. Unfortunately, the logs are either at a very detailed level or very specific for a given process. For example, using the ST03 Transaction Report shown in Figure 12, we can inspect database transactions. However, these transactions are too fine-grained and do not point to a case and task. SAP R/3 also logs document flows which are more at the business level. Unfortunately, one needs to know the relevant tables and the structure of these tables to use these document flows. Therefore, SAP R/3 can only be mined after considerable efforts. It seems that this is not a limitation of the concept of process mining but a result of the evolutionary growth of SAP R/3 resulting in a wide variety of logs.

5.5 Example: FLOWer

Traditionally, products have been in the four quadrants shown in Figure 2 with the lower-right quadrant being nearly empty. Clearly, real life processes are a mixture of structured/unstructured process/data centric activities. Therefore, some vendors are now aiming at the middle of the CSCW spectrum shown in Figure 2. This is not a trivial pursuit given the trade-offs between the various requirements. For example, it is difficult to develop systems that offer a lot of support without restricting flexibility or requiring a lot of modeling efforts. One of the few tools that is trying to balance between structured and unstructured activities using both a process centric and data centric approach is the case handling system [7] FLOWer of Pallas Athena.

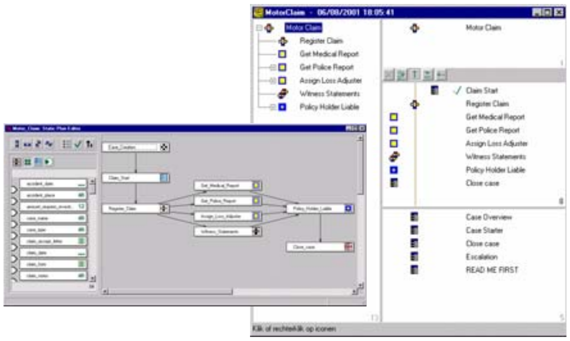


Figure 13: Screenshots of both designer and case guide of FLOWer

Figure 13 shows some screenshots of FLOWer. The basic idea of case handling systems like FLOWer is to allow for implicit routing, i.e., in addition to the predefined routes there are alternative routes that are not modeled explicitly but can only be taken provided proper authorization. Moreover, activities may overlap and are defined in terms of pre- and post-conditions to allow for more flexibility.

We have developed an adaptor for FLOWer in the context of the ProM framework. One of the interesting properties of the adaptor is that it can mine both for process-centric and data-centric events. This allows a more detailed investigation into how people actually work. The adaptor has been applied within several processes of the UWV, a large Dutch organization taking care of work-related regulations (e.g. unemployment).

6. Related work

In Section 2 we already reviewed relevant CSCW literature. In this section we focus on process mining literature. The idea of process mining is not new [4,8,9] but has been mainly aiming at the control-flow perspective. The idea of applying process mining in the context of workflow management was first introduced in [8]. This work is based on workflow graphs, which are inspired by workflow products such as IBM MQSeries Workflow (formerly known as Flowmark). Cook and Wolf have investigated similar issues in the context of software engineering processes. In [9] they describe three methods for process discovery: one using neural networks, one using a purely algorithmic approach, and one Markovian approach. Schimm [20] has developed a mining tool suitable for discovering hierarchically structured workflow processes. Herbst and Karagiannis also address the issue of process mining in the context of workflow management using an inductive approach [15,14]. They use stochastic task graphs as an intermediate representation and generate a workflow model described in the ADONIS modeling language. Most of the approaches have problems dealing with parallelism and noise. Our work in [6] is characterized by the focus on workflow processes with concurrent behavior (rather than adding ad-hoc mechanisms to capture parallelism). In [23] a heuristic approach using rather simple metrics is used to construct so-called "dependency-frequency tables" and "dependency-frequency graphs". These are then used to tackle the problem of noise. The approaches described in [6,23] are based on the Alpha algorithm. Process mining is not limited to the control-flow perspective. As shown in [3], it can also be used to discover the underlying social network. Process mining in a broader sense can be seen as a tool in the context of Business (Process) Intelligence (BPI). In [13,19] a BPI toolset on top of HP's Process Manager is described. The BPI toolset includes a so-called "BPI Process Mining Engine". However, this engine does not provide any techniques as discussed before. Instead it uses generic mining tools such as SAS Enterprise Miner for the generation of decision trees relating attributes of cases to information about execution paths (e.g., duration). In order to do workflow mining it is convenient to have a so-called "process data warehouse" to store audit trails. Such a data warehouse simplifies and speeds up the queries needed to derive causal relations. In [18] Zur Mühlen describes the PISA tool which can be used to extract performance metrics from workflow logs. Similar diagnostics are provided by the ARIS Process Performance Manager (PPM) [16]. The later tool is commercially available and a customized version of PPM is the Staffware Process Monitor (SPM) (www.staffware.com) which is tailored towards mining Staffware logs. Note that none of the latter tools is extracting models, i.e., the results do not include control-flow, organizational or social network related diagnostics. The focus is exclusively on performance metrics. For more information on process mining we refer to a special issue of Computers in Industry on process mining [5] and the survey paper [4].

Note that an earlier version of the paper was presented in a keynote talk at CSCWD 2005 in Coventry [1]. This paper (slightly) extends that paper by structuring the process mining domain in more detail.

7. Conclusion

This paper discussed the application of process mining in the context of the CSCW spectrum. First the spectrum was classified into five domains (cf. Figure 2). Then the topic of process mining was introduced and for each of the five domains an example is given. Although the process mining techniques are maturing and tools such as ProM can easily be applied, there are many open problems and challenges. For example, most of the existing mining techniques have problems dealing with noise and incompleteness. As discussed in this paper we need to apply Occam's Razor to get meaningful results. (Occam's razor is a logical principle attributed to the mediaeval philosopher William of Occam. The principle states that "one should not increase, beyond what is necessary, the number of entities required to explain anything".) One exception should not change the process model completely and should be ignored or marked as such. Moreover, information will always be based on a limited observation period where not all possible combinations of events will occur. Therefore, it does not make sense to assume a "complete" log.

Besides the "discovery aspect" of process mining, complementary approaches such as delta analysis and conformance testing can be utilized. In particular, conformance testing allows for widespread application. In many settings, it is useful to compare some prescriptive or descriptive model with the actual events being logged.

We hope that this paper will inspire researchers and developers to apply process mining in new domains. We also encourage people to use the ProM framework as a platform for such efforts.

Acknowledgements

The author would like to thank Ton Weijters, Boudewijn van Dongen, Ana Karla Alves de Medeiros, Minseok Song, Laura Maruster, Eric Verbeek, Monique Jansen-Vullers, Hajo Reijers, Michael Rosemann, Huub de Beer, Peter van den Brand, Anne Rozinat, Christian Günter, Andriy Nikolov, Wouter Kunst, Martijn van Giessel et al. for their on-going work on process mining techniques. Boudewijn has been the driving force behind ProM, Ana Karla developed the FLOWer adapter, Andriy developed the E-mail mining tool, Wouter and Martijn investigated the application of process mining to FLOWer and SAP R/3. We also thank EIT for supporting the development of the ProM framework, cf. www.processmining.org.

References

[1] W.M.P. van der Aalst. Process Mining in CSCW systems. In W. Shen and A. James et al., editors, Proceedings of the 9th IEEE International Conference on Computer Supported Cooperative Work in Design (CSCWD 2005), pages 1-8, Coventry University/IEEE Computer Society Press, 2005.

[2] W.M.P. van der Aalst and K.M. van Hee. Workflow Management: Models, Methods, and Systems. MIT press, Cambridge, MA, 2002.

[3] W.M.P. van der Aalst and M. Song. Mining Social Networks: Uncovering Interaction Patterns in Business Processes. In J. Desel, B. Pernici, and M. Weske, editors, International Conference on Business Process Management (BPM 2004), volume 3080 of Lecture Notes in Computer Science, pages 244-260. Springer-Verlag, Berlin, 2004.

[4] W.M.P. van der Aalst, B.F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A.J.M.M. Weijters. Workflow Mining: A Survey of Issues and Approaches. Data and Knowledge Engineering, 47(2):237-267, 2003.

[5] W.M.P. van der Aalst and A.J.M.M. Weijters, editors. Process Mining, Special Issue of Computers in Industry, Volume 53, Number 3. Elsevier Science Publishers, Amsterdam, 2004.

[6] W.M.P. van der Aalst, A.J.M.M. Weijters, and L. Maruster. Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, 16(9):1128-1142, 2004.

[7] W.M.P. van der Aalst, M. Weske, and D. Grünbauer. Case Handling: A New Paradigm for Business Process Support. Data and Knowledge Engineering, 53(2):129-162, 2005.

[8] R. Agrawal, D. Gunopulos, and F. Leymann. Mining Process Models from Workflow Logs. In Sixth International Conference on Extending Database Technology, pages 469-483, 1998.

[9] J.E. Cook and A.L. Wolf. Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology, 7(3):215-249, 1998.

[10] C.A. Ellis. An Evaluation Framework for Collaborative Systems. Technical Report, CU-CS-901-00, University of Colorado, Department of Computer Science, Boulder, USA, 2000.

[11] C.A. Ellis, S.J. Gibbs, and G. Rein. Groupware: Some issues and experiences. Communications of the ACM, 34(1):38-58, 1991.

[12] C.A. Ellis and G. Nutt. Workflow: The Process Spectrum. In A. Sheth, editor, Proceedings of the NSF Workshop on Workflow and Process Automation in Information Systems, pages 140-145, Athens, Georgia, May 1996.

[13] D. Grigori, F. Casati, U. Dayal, and M.C. Shan. Improving Business Process Quality through Exception Understanding, Prediction, and Prevention. In P. Apers, P. Atzeni, S. Ceri, S. Paraboschi, K. Ramamohanarao, and R. Snodgrass, editors, Proceedings of 27th International Conference on Very Large Data Bases (VLDB'01), pages 159-168. Morgan Kaufmann, 2001.

[14] J. Herbst. A Machine Learning Approach to Workflow Management. In Proceedings 11th European Conference on Machine Learning, volume 1810 of Lecture Notes in Computer Science, pages 183-194. Springer-Verlag, Berlin, 2000.

[15] J. Herbst. Ein induktiver Ansatz zur Akquisition und Adaption von Workflow-Modellen. PhD thesis, Universität Ulm, November 2001.

[16] IDS Scheer. ARIS Process Performance Manager (ARIS PPM): Measure, Analyze and Optimize Your Business Process Performance (whitepaper). IDS Scheer, Saarbruecken, Gemany, http://www.ids-scheer.com, 2002.

[17] T.W. Malone. Commentary on Suchman article and Winograd response. Computer Supported Cooperative Work, 3(1):37-38, 1995.

[18] M. zur Mühlen and M. Rosemann. Workflow-based Process Monitoring and Controlling - Technical and Organizational Issues. In R. Sprague, editor, Proceedings of the 33rd Hawaii International Conference on System Science (HICSS-33), pages 1-10. IEEE Computer Society Press, Los Alamitos, California, 2000.

[19] M. Sayal, F. Casati, U. Dayal, and M.C. Shan. Business Process Cockpit. In Proceedings of 28th International Conference on Very Large Data Bases (VLDB'02), pages 880-883. Morgan Kaufmann, 2002.

[20] G. Schimm. Generic Linear Business Process Modeling. In S.W. Liddle, H.C. Mayr, and B. Thalheim, editors, Proceedings of the ER 2000 Workshop on Conceptual Approaches for E-Business and The World Wide Web and Conceptual Modeling, volume 1921 of Lecture Notes in Computer Science, pages 31-39. Springer-Verlag, Berlin, 2000.

[21] W. Shen. Special Issue on Collaborative Environments for Design and Manufacturing of International Journal of Advanced Engineering Informatics, 19(2), 2005.

[22] L. Suchman. Do Categories Have Politics? The Language /Action Perspective Reconsidered. Computer Supported Cooperative Work, 2(3):177-190, 1994.

[23] A.J.M.M. Weijters and W.M.P. van der Aalst. Rediscovering Workflow Models from Event-Based Data using Little Thumb. Integrated Computer-Aided Engineering, 10(2):151-162, 2003.

[24] T. Winograd. Categories, Disciplines, and Social Coordination. Computer Supported Cooperative Work, 2(3):191-197, 1994.

[25] T. Winograd and F. Flores. Understanding Computers and Cognition: A New Foundation for Design. Ablex, Norwood, 1986.