Overview of Workflow Mining Technology*

Chunqin GU
School of Information Science and Technology, Sun Yat-sen University
chunqingu@163.com

Hui-you Chang
School of Information Science and Technology, Sun Yat-sen University
isschy@mail.sysu.edu.cn

Yang YI
School of Information Science and Technology, Sun Yat-sen University
issyy@mail.sysu.edu.cn

Abstract

The purpose of workflow mining is to build proper model based on event-based logs and to support the analysis and design of workflow models. The workflow mining technology had become a hot research field among computer application from the late 1990's. Firstly, the definition and research significance of the workflow mining are presented. Then, we made an overview of the current workflow mining technologies including the origin of workflow mining, the mining of the workflow models, the mining of the workflow performance and the improvement in workflow models based on the workflow mining technology. Finally, the facing challenge and the development direction of workflow mining are presented.

1. Introduction

In 1993, the standardization organization of workflow technology – Workflow Management Coalition (WFMC) was built. The definition of workflow in WFMC is as follows: workflow is concerned with the automation of procedures where documents, information or tasks are passed between participants according to a defined set of rules to achieve, or contribute to an overall business goal [1]. So, all kinds of activities of enterprises are organized and corresponded by business processes modeling and defined business logic relation.

Workflow mining means the knowledge discovery of workflow system, which can be induced by the definition of data mining. The ultimate target of workflow mining is to mine the transactional logs of workflow system, and to discover the knowledge of workflow including workflow models mining, workflow performance mining and workflow models improving.

The life circle of workflow system includes four periods[2]: (1) workflow modeling; (2) workflow configuration; (3) workflow enactment (4) workflow diagnosis. In the past, the research on workflow technology mainly focused on the workflow modeling and configuration, but less on workflow enactment and least on workflow diagnosis. The workflow mining technology research focuses on the period of workflow enactment and workflow diagnosis. The workflow mining technology can reversely support the workflow modeling and analysis by mining the logs generated during the period of enactment.

2. Workflow mining research status

2.1. Origin of workflow mining

Cook [3] firstly discovered models of software processes in software engineering from event-based data, and presented three software processes mining methods: one using neural networks, one using a pure algorithmic approach, and one Markovian approach. The authors consider the latter two the most promising approaches. The pure algorithmic approach builds a finite state machine (FSM) that can't model concurrent activities. Agrawal [4] firstly introduced the concept of process mining to workflow system and applied an algorithm to process logs obtained from an IBM's workflow management platform – MQSeries[5]. After that research on workflow mining deeply developed.

2. 2. Workflow modes mining

There exists some disadvantage in the process of a traditional workflow modeling, such as

- Modeling needs much experience
- Modeling needs much time
- Modeling is very expensive
- Modeling do not reflect the real process
- Models change quickly

* Supported by the National Natural Science Foundation of China under Grant Nos. 60573159.

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.86

IEEE computer society

The disadvantage of traditional workflow modeling can be overcome by mining workflow models. So, process models can be obtained by mining not by designing them. Figure 1. demonstrates the process of workflow models mining.

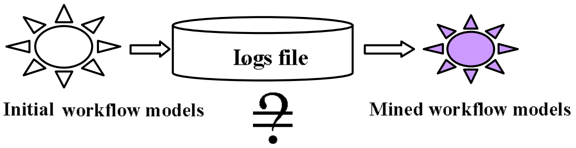

Figure 1. Process of workflow models mining

[6] is an extension and optimization of [4]. [4]'s research was based on the atomicity of activities. That is to say those activities are not time-consuming. [6] is different from [4] that viewed the execution of an activity as a time interval and define the time interval from ready status to close status as a span of an activity. Because of the introduction of activities' lifespan, the concurrent activities are ascertained from the following two aspects: (1) two activities appearing in the two kinds of sequence; (2) two activities time interleaving of active status. So long as one of two conditions is met, the two activities can be judged as concurrent activities. On the condition of incomplete logs, if only condition 1 is used to judge the concurrent activities, the error of modeling is prone to be brought on. So, the judge of concurrent activities is more comprehensive, and real situation of workflow can be reasonable reflected. [6]'s graphs are more faithful in the sense that the number of excess and absent edges is consistently smaller and it depends on the size and quality of the log.

In [7], their approach is considered a special data mining that is based on a model representation and a special mining procedure. Its model representation is defined in the form of a block-oriented workflow meta-model. It covers the standard workflow control constructs and combines the construction of sound workflow models with an enlarged expressive power.

2.2.1. Models mining based on Petri net. Aalst [8] presents a new algorithm designated as α-algorithm to extract a process model from workflow logs that meets the definition of integrity and represents it in terms of a Place/Transition (P/T) net based on a Petri net. The α-algorithm can be applied to mining all structured workflow net (SWF-net), but it is not possible to discover duplicate tasks and non-free choice structure. In addition, it can't discover short cycle such as single cycle and double single. And the α-algorithm can't be applied in the condition of incomplete logs or logs with noise.

[9] identifies single-cycle and double-cycle logs and deletes the circular logs by preprocessing the noisy logs. After finding out the short cycle, α-algorithm can be applied to mining logs with short cycle, so the problem of short cycle can be solved by the step of preprocessing.

2.2.2. Non-structure workflow models mining. Except for structured workflow processes, there are many processes that are highly dynamic that we call ad-hoc process. [14] is one of the few process-aware collaboration systems allowing for ad-hoc processes. Unlike in classical workflow systems, the users are no longer restricted by the system. Dustdar [15] presents a newly developed extraction tool named Teamlog that can be applied into the mining of ad-hoc workflow processes.

2.2.3. Noise processing in the models mining. [10,11] researches on the noise of logs and presents three steps of solving the noise and modeling: (1)constructing dependence frequency table(D/T-table); (2)mining activities relation table(R-table) by D/F-table; (3)reconstructing the workflow net(WF-net) by R-table. The threshold is used in the period of processing noise, but the setting of the threshold depends on the experience data. It is not suitable to people who has no experience.

Maruster [12] presents a method of gaining optimal threshold. The optimal threshold is gained by machine learning. The method improves the past ones of delete noise. In [13] a Bayesian method is completely applied to mining logs with noise.

2.3. Workflow models mining tools

Table 1.[2] lists some classic workflow modes mining tools such as EmiT, Little Thumb, InWoLvE and Process Miner. The aspects such as data structure, time and concurrent are focused.

2.4. Workflow performance mining

Prophase workflow mining works have concentrated their efforts on process behavioral aspects. Although powerful, those proposals are found lacking in functionalities and performance when used to discover transactional workflow that can't be seen at the level of behavioral aspects of workflow. Their limitations mainly come from their incapacity to discover the transactional dependencies between process activities, or activities transactional properties.

Gaaloul [20] describes mining techniques, which can discover workflow models, and improve its transactional behavior from event logs. Handling workflow transactional behavior remains a main problem to ensure a correct and reliable execution. It is obvious that the discovery and the explanation of this behavior would better understand and control workflow recovery. Gaaloul [21] firstly describes mining techniques, which are able to discover a workflow model, and to improve its transactional behavior from event based logs. In [21], first an algorithm was proposed to discover workflow patterns, and then proposes techniques to discover activities transactional dependencies that allow to mine workflow recovery techniques. Finally, based on this mining step, the paper uses a set of rules to improve workflow design.

Table 1. Comparing EmiT, Little Thumb, InWoLvE and Process Miner

	EmiT [16]	Little Thumb [17]	InWoLvE [18]	Process Miner [19]
Structure	Graph	Graph	Graph	Block
Time	√	×	×	×
Basic parallelism	√	√	√	√
Non-free choice	×	×	×	×
Basic loops	√	√	√	√
Arbitrary loops	√	√	×	×
Hidden tasks	×	×	×	×
Duplicate tasks	×	×	√	×
Noise	×	√	√	×

Zhang [22] views workflow as a queue system, and models the traffic volume by analyzing the process's arrival time. The paper mines the factors affecting service time by analyzing the process's service time, and models users' waiting time limitation by analyzing the waiting time. Zhao [23] brings out a role identification methods based on techniques of workflow logs mining and process decomposition. Role identification was analyzed from several layers, and carried from bottom to top through workflow mining. In the paper, relevancy parameter was used to define the role granularity so as to form the leaf roles. Then, the concept of process manager is introduced to identify the topper roles and define hierarchical division of roles. Finally, the fruit of workflow model mining is borrowed to depict the activity dependences of relevant roles, meanwhile to complete interactive roles identification.

2.5. Improving workflow models

Since the control flow specifications of workflow are manually designed, they entail assumptions and errors, leading to inaccurate workflow models. Decision points, the XOR nodes in a workflow graph model, determine the path chosen toward completion of any process invocation. [24] presents that positioning the decision points at their earliest points can improve process efficiency by decreasing their uncertainties and identifying redundant activities, presents novel techniques to discover the earliest positions by analyzing workflow logs and to transform the model graph, presents that the transformed model is more efficient with respect to its average execution time and uncertainty, when compared to the original model.

Decision mining, also referred to as decision point analysis, aims at the detection of data dependencies that affect the routing of a case. [25] describes how machine learning techniques can be leveraged for decision mining, and presents a Decision Miner implemented within the ProM framework.

3. Facing challenge and direction of workflow mining

3.1. Current facing mainly problems

There are problems that are not solved or not solved best at the aspect of control flow. Such problems include short cycle, noise, non-integrity data, duplicated tasks, non-free choice and hided tasks problems. The researchers present many ideas and methods involving the problems of short cycle, duplicated tasks and hided tasks. There are still many problem needed solving, such as the problem of non-free choice. There exists a common problem: When one or some aspects are considered, the other aspects are easy to be neglected. There is no a good method that can consider all those problems.

Restricted by the way of workflow modeling and the workflow mining, some algorithms can deal with some problems but can't deal with other problems. Sometimes these algorithms are restricted by so many conditions, and not universal.

3.2. Direction of workflow mining

Facing current main problems, later there are more researchers who will do some works focusing the following aspects:

(1) Making deeper researches on some complicated structure such as hided tasks, duplicated tasks, non-free choice and short cycle.

(2) Improving the current algorithms and building the proper workflow models formalization description in order to solve all kinds of complicated mining tasks.

(3) Making researchers from other perspective, such as organization perspective, information perspective

and application perspective, which is different from control flow aspect. Workflow mining should not restrict to model and analyze control flow, but should analyze workflow performance, data flow, transaction recovery and resource usability etc. in order to support the enterprise process re-engineering.

(4) The problems of privacy preservation in workflow mining will be considered, so that the workflow algorithms can construct correct workflow models without destroying the logs' privacy.

4 Conclusion

It's just a start to research into workflow mining. Although there are some research progresses, there are some problems that are worth of making researches.

References:

[1] Workflow Management Coalition. Workflow management coalition terminology and glossary. Technical Report, WfMCTC-1011, Brussels: Workflow Management Coalition, 1996

[2] W.M.P. van der Aalst, B.F. van Dongen, J. Herbst, L. Maruster. Workflow mining: A survey of issues and approaches[J]. Data & Knowledge Engineering, 2003, 47: 237-267

[3] E. Cook, A. L. Wolf. Discovering models of software processes from event-based data. ACM Transactions on Software Engineering and Methodology, 7(3), 215 ~ 249, 1998.

[4] R. Agrawal, D. Gunopulos, F. Leymann. Mining Process Models from Workflow Logs. Lecture Notes In Computer Science, vol. 1377, 469~483, 1998.

[5] HB. Luo, YY. Fan, C. WU. Overview of Workflow Technology. Journal of Software, vol. 11(7), 899-907, 2000

[6] S.S. Pinter, M. Golani. Discovering workflow models from activities' lifespans. Computers in Industry, 53(3), 283~296, 2004.

[7] G. Schimm. Mining exact models of concurrent workflows. Computers in Industry, 53(3), 265 ~281, 2004.

[8] W.M.P. van der Aalst, T. Weijters, L. Maruster. Workflow mining: discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering, 16(9), 1128~1142, 2004.

[9] A.K.A. de Medeiros, W.M.P. van der Aalst, A.J.M.M. Weijters. Workflow mining: current status and future directions. Lecture Notes in Computer Science, vol. 2888, 389~406 ,2003.

[10] A.J.M.M. Weijters, W.M.P. van der Aalst, Process mining: discovering workflow models from event-based data, in: B. Kroose, M. de Rijke, G. Schreiber, M. van Someren (eds.), Proceedings of the 13th Belgium–Netherlands Conference on Artificial Intelligence (BNAIC 2001), 2001, pp. 283–290.

[11] A.J.M.M. Weijters, W.M.P. van der Aalst, Workflow mining: discovering workflow models from event-based data, in: C. Dousson, F. Hooppner, R.Quiniou (Eds.), Proceedings of the ECAI Workshop on Knowledge Discovery and Spatial Data, 2002, pp. 78–84.

[12] L. Maruster, A.J.M.M. Weijters, W.M.P. van der Aalst, A. van den Bosch, Process mining: discovering direct successors in process logs, in: Proceedings of the 5th International Conference on Discovery Science (Discovery Science 2002), Lecture Notes in Artificial Intelligence, vol. 2534, Springer-Verlag, Berlin, 2002, pages 364–373.

[13] Colombo, E. Damiani, G. Gianini. Discovering the software process by means of stochastic workflow analysis. Journal of Systems Architecture, vol. 52, 684~692, 2006.

[14] S. Dustdar. Caramba – A Process-Aware Collaboration System Supporting Ad hoc and Collaborative Processes in Virtual Teams. Distributed and Parallel Databases, vol.15, 45-56,2004.

[15] S. Dustdar, T. Hoffmann, W.M.P.van der Aalst. Mining of ad-hoc business processes with TeamLog. Data & Knowledge Engineering, vol.55, 129-158, 2005.

[16] B.F.van Dongen, W.M.P.van der Aalst. Emit: A Process Mining Tool. ICATPN 2004, LNCS 3099, pp. 454-463, 2004.

[17] A.J.M.M. Weijters, W.M.P van der Aalst. Rediscovering workflow models form event-based data using little thumb. Integrated Computer-Aided Engineering, vol. 10, 151-162, 2003

[18] J. Herbst, D. Karagiannis. Workflow mining with InWoLvE. Couputers in Industry, vol. 53, 245-264, 2004.

[19] G. Schimm. Process Miner – A Tool for Mining Process Schemes from Event-Based Data. Lecture Notes in Computer Science, vol.2424, 525-528, 2002.

[20] W. Gaaloul, S. Bhiri, C. Godart. Discovering workflow transactional behavior from event-based log. Lecture Notes In Computer Science, vol. 3290, 3~18, 2004.

[21] W. Gaaloul, C. Godart. Mining workflow recovery from event-based logs. Lecture Notes In Computer Science, vol. 3649, 169~185, 2005.

[22] P. Zhang, N. Serban. Discovery, visualization and performance analysis of enterprise workflow. Computational Statistics & Data Analysis, vol.51, 2670`2687, 2007.

[23] J. Zhao, WD. Zhao. Process roles identification based on workflow logs mining. Computer Integrated Manufacturing System. vol. 12, 1916-1920,2006.

[24] S. Subramaniam, V. Kalogeraki, D. Gunopulos. Improving process models by discovering decision points. Information Systems, 2006

[25] Rozinat, W.M.P. van der Aalst. Decision Mining in ProM. LNCS 4102, pp. 420~425, 2006