Behavior Pattern Mining: Apply Process Mining Technology to Common Event Logs of Information Systems

Jinliang Song, Tiejian Luo and Su Chen

Abstract—**Although information systems support a wide range of recreational and social activities, the pattern of users' behavior in such systems is not clear. We extend the process mining technology to work on common event logs that have no workflow cases reference, and name the new technology as behavior pattern mining. This paper gives out a brief survey on issues, challenges, approaches and related tools in behavior pattern mining area, and compares behavior pattern mining with workflow mining technology, which is the other sub field of process mining. To reply the problems proposed by behavior pattern mining, the process mining area gets its new motivation.**

I. Introduction

As information technology increasingly support a wide range of recreational and social activities, more and more people are able to connect online and benefit from the flexible virtual spaces. The computer and network supported systems enable organizations and individuals to communicate, to cooperate, and to compete in a more flat environment. However, these systems have changed the behavior customs heavily as such systems helped users' daily work. To release the behavior pattern of computer supported work, is one of the key points to understand the modern working method, and produces recommendation to improve information systems, as well as indicates how to raise the work efficiency of people.

A large group of scientists analyze behavior patterns through psychological, social, and organizational perspectives. From the year 2001 to 2004, NSF (National Science Foundation) funded SOC project (Science of Collaboratories) [1] to survey more than a hundred of collaborative projects in multiple fields, and aimed to define, abstract, and codify the broad underlying technical and social elements that lead to successful collaboratories. In detail, SOC project analyzed these projects from money, information and communication flow for institutions, IT resources and users. We notice that the money, information and communication condition in a project is the phenomenon and results caused by collaborators' behavior, and the analysis through such three kinds of flow is an indirect method to discover behavior pattern. In this paper, we emphasize the direct way to do research on behavior pattern. Our plan is to mine the event logs that recorded by information systems, and to reveal the hidden relation of behavior in logs. This automated analyzing approach can accelerate the discovery of behavior pattern efficiently.

The automatic pattern discovery technology is cross-disciplinary among KDD (Knowledge Discovery in Data) and machine learning. The most well-known application area to discover pattern in logs is web usage mining, which is defined as applying data mining techniques to log interactions between users and a website [2]. Web usage mining tries to find out the relationship between independent operations and the successful clicking sequence for specific task, and uses the mined rule to classify particular customers and improve personalization of systems. Another hot pattern discovery technology is process mining, it tries to extract an explicit process model from event logs, i.e., to create a process model given a log with events such that the model is consistent with the observed dynamic behavior [3]. Compare to web usage mining, process mining support major workflow patterns in target model, which is efficient to express comprehensive relations for multi users who cooperate to accomplish a task. In addition, with the workflow pattern as its symbols, process mining supports the rediscovery of workflow model from workflow logs naturally, which is known as workflow mining. Further more, many researchers in process mining area use process mining as the same term of workflow mining. Although workflow mining doesn't assume the presence of a workflow management system, it assumes that workflow logs with event data can be collected [4]. Such workflow logs are usually known as workflow traces.

In recent years, workflow mining technology developed a range of effective algorithms and approaches, and starts to apply on real world. This paper discusses the problem to immigrate workflow mining technology on common information system event logs but workflow traces, which will extend the application area of automatic pattern discovering method. We name the behavior pattern discovering on common system event logs as the term behavior pattern mining, and consider it as another sub field of process mining besides workflow mining. We discuss the requirements, challenges, structure of data source, possible approaches, and related tools of it basing on the achievement of workflow mining. In organization, section 2 introduces

This work was supported by Ministry of Science and Technology of the People's Republic of China under Grant 2005DKA64100 and 2005DKA10201.

Jinliang Song is Ph.D candidate with Graduate University of Chinese Academy of Sciences, No.19A Yuquan Road,Beijing,P.R.C, Beijing, 100049, China (phone: 86-10-88256663-40; fax: 86-10-88256668; e-mail: songjl@mails.gucas.ac.cn).

Tiejian Luo is professor with Graduate University of Chinese Academy of Sciences, No.19A Yuquan Road,Beijing,P.R.C, Beijing, 100049, China (e-mail: tjluo@gucas.ac.cn).

Su Chen is Ph.D candidate with Graduate University of Chinese Academy of Sciences, No.19A Yuquan Road,Beijing,P.R.C, Beijing, 100049, China (e-mail: chensu@mails.gucas.ac.cn)..

what is behavior pattern mining, including the definition, the input data format, the targets and the challenges of it. Section 3 talks about the major approaches transferred from workflow mining area. Section 4 gives out some useful tools for test-bed. At last section 5 discusses and concludes the paper.

II. What is Behavior Pattern Mining

Behavior pattern mining is to discover the law of relationships and restricts between users' behavior connected to a common goal from event logs, and targets to represent how people accomplish their works in multi users' interaction environment. Behavior pattern mining is a sub field of process mining, which targets the automatic discovery of information from an event log and the discovered information can be used to deploy new systems that support the execution of business processes or as a feedback tool that helps in auditing, analyzing and improving already enacted business processes [5]. Workflow mining is another sub field of process mining, and workflow mining is the reverse engineering of workflow modeling basing on workflow logs.

Behavior pattern mining and workflow mining have some common points, such as both of them use workflow patterns as the symbols to present the mined results, and some mining approaches can be applied on both the two. On the contrary, the two are different in data source format and their target. Workflow mining grounds on workflow trace logs, whereas behavior pattern mining grounds on common event logs that are not partitioned. Behavior pattern mining focuses on how to structure and present users' interactions, whereas workflow mining mainly focuses on workflow model improvement and verification. The knowledge from workflow mining area is a valuable base to discuss behavior pattern mining.

A. The Data Source of Behavior Pattern Mining

Users' behavior in information systems can be recorded as system logs. When information systems coordinate with workflow management engine, the recorded information can be workflow logs. Table I is an example of workflow event log came from [6] but sorted by case column. When we partition the workflow event log by different case instances, we get the well-known workflow traces data. In this table, we can get the six kinds of typical information in workflow event log:

- Case: represents a process instance.
- Activity: identifies an activity instance for process.
- Event: specifies the event type of an activity instance (i.e. START or COMPLETE).
- Repository: specifies the data repositories an activity instance writes to or reads from.
- Participant: specifies who performs an activity instance.
- Time: specifies the time an event occurred.

Most of workflow mining approaches treat workflow traces as the format of source data. However, to point out which case instance does an event belong to is not obvious when we have no reference information came from workflow management engine. It means that logs without case instance reference are more common structure of recorded data by information systems. We drop the case value in Table I and sort events by time, and then we get Table II as an example of common event logs to be the source data of behavior pattern mining.

In brief, the input data of behavior pattern mining is a stream of logged events with the five kinds of information as activity, event, repository, participant, and time.

Usually information systems keep their own format of log file storage. In order to render the design of process mining techniques and tools independent of the target system implementation, Aalst introduced the MXML event log format that was designed to meet the requirements of process mining tools in the best possible way. Further more, Aalst's team provided ProM Import Framework to do the conversion from custom log formats to MXML. [7]

TABLE I
WORKFLOW EVENT LOG EXAMPLE

CASE	ACTIVITY	EVENT	REPOSITORY	PARTICIPANT	TIME
1	A	START	CUST	T	12
1	A	COMPLETE	APF	T	24
1	B	START	APF	N,Z	45
1	B	COMPLETE	APF	N,Z	100
1	C	START	APF	T	120
1	D	START	APF	U	121
1	D	COMPLETE	MAF	U	200
1	C	COMPLETE	EAF	T	205
1	F	START	APF	X	230
1	F	COMPLETE	PAF	X	234
2	A	START	CUST	M	23
2	A	COMPLETE	APF	M	30
2	B	START	APF	U,Z	125
2	B	COMPLETE	RJF	U,Z	175
2	E	START	RJF	X	214
2	E	COMPLETE	MRF	X	221
2	F	START	APF	M	290
2	F	COMPLETE	PAF	M	299

TABLE II
COMMON EVENT LOG EXAMPLE

ACTIVITY	EVENT	REPOSITORY	PARTICIPANT	TIME
A	START	CUST	T	12
A	START	CUST	M	23
A	COMPLETE	APF	T	24
A	COMPLETE	APF	M	30
B	START	APF	N,Z	45
B	COMPLETE	APF	N,Z	100
C	START	APF	T	120
D	START	APF	U	121
B	START	APF	U,Z	125
B	COMPLETE	RJF	U,Z	175
D	COMPLETE	MAF	U	200
C	COMPLETE	EAF	T	205
E	START	RJF	X	214
E	COMPLETE	MRF	X	221
F	START	APF	X	230
F	COMPLETE	PAF	X	234
F	START	APF	M	290
F	COMPLETE	PAF	M	299

B. *The Targets of Behavior Pattern Mining*

Document [8] introduces questions that managers usually have about processes in organizations, and some of them can be answered by workflow mining technology. When we go deep into these questions and reconsider the questions under a view of more common information systems, we get the motivation of behavior pattern mining. Further more, we add some particular questions caused by the characteristics of behavior pattern mining.

1) What is the most frequent path for users' operation?

2) Does users' operation fit the expecting model? How frequent is the mismatch? Where are the problems?

3) What is the average/minimum/maximum throughput time for users' major steps of operation?

4) Which paths take too much time on average? How frequent are users follow these patterns? What are the critical steps for these paths?

5) What is the average service time for specific task?

6) How much time was spent between any two tasks of a mission?

7) How do users actually use the system?

8) What are the business rules in the system?

9) Are the rules indeed being obeyed?

10) How many people are involved in a particular work?

11) What is the communication structure and dependencies among people?

12) How many transfers happen from one role to another role?

13) Who are important people in a particular work?

14) Who contribute useful information to whom?

15) Who work under same pattern?

16) How the patterns of users' operation change during the time passes?

C. *The Challenges of Behavior Pattern Mining*

Both behavior pattern mining and workflow mining are sub fields of process mining, which means to extract the process model from event logs. As a result, behavior pattern mining and workflow mining share some technology basement, and face up some shared challenges. We review the challenges in [3] and get the following common ones:

1) Mining hidden tasks;

2) Mining duplicate tasks;

3) Mining non-free-choice constructs;

4) Mining loops;

5) Using time and repository information;

6) Mining different perspectives;

7) Dealing with noise;

8) Dealing with incompleteness;

9) Gathering data from heterogeneous sources;

10) Visualizing results;

Besides, behavior pattern mining uses common event log as its data source, whereas workflow mining uses workflow traces as its data source, so that behavior pattern mining encounters some new challenges that workflow mining does not have.

11) Lacking process reference data: Behavior pattern mining has no knowledge about the expecting process model or which process does an event belong to. An automatic data discovering problem with no future knowledge is always challenging.

12) Recognizing the start and end point of a process: Without the case instance information in event logs, where does a mission start and where does it finish are not obvious. To avoid mined model too huge to handle, it requires a mechanism to divide the whole process model to be suitable sub processes.

13) Event data stream mining: Common event log is a stream of recorded tasks, which is not supported by most approaches in process mining area, and the approaches usually require partitioning the data stream. Further more, behavior pattern mining is influenced by concept drift problem came from data stream mining area, because of the variation nature of the data source during the time passes.

14) Incremental mining: When new events are logged, the existing process models require update to represent the latest relations of users' operation. However, the mined models may be destroyed by the coming events that have not been covered by existing models.

To reply each challenges of behavior pattern mining, and to apply them on every kinds of information systems, the algorithms and approaches for process mining can be improved. In addition, the related KDD and machine learning technology will benefit from the behavior pattern mining scenario.

III. ALGORITHMS AND APPROACHES

The process model determines the form of mined results, because no algorithm could provides an output that beyond its representing model. Further more, the process model affects the design of mining algorithms and related approaches. However, when we use behavior pattern mining technology

to analyze data came from real world, we have some common problems to consider. Such as: how to collect source data? How to get rid of noise? How to design the mining approach? And how to verify the results? In this section, we describe the common problems as typical stages of behavior pattern mining first, and then discuss detailed features of related algorithms.

A. *The Typical Stages of Behavior Pattern Mining*

We present the major stages of behavior pattern mining as Fig 1, most of the process mining algorithms and approaches can be mapped into the four stages. In addition, the four stages support iterative and incremental usage to improve the mining performance. Such stages were first introduced in paper [9].

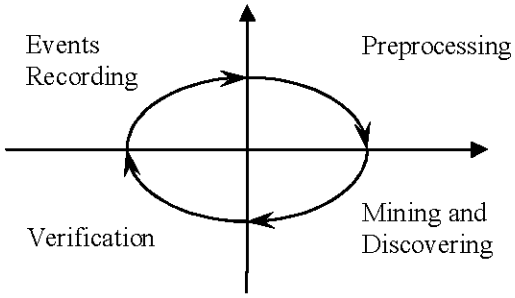

Fig. 1. The Stages of Behavior Pattern Mining

1) Events recording: This stage prepares the original data for behavior pattern mining. Custom information systems may record their own classes of information in specific log format. The basic types of information include activity, event, repository, participant, and time. The logged information may be extended as particular requirement.

2) Preprocessing: Traditionally, this stage transfers the log format to fit mining method, and deals with noise and incomplete record. Depending on different mining algorithm, preprocessing does statistics, possibility calculating and so on, to filter information and reduce the computing loads of mining. As a characteristic of behavior pattern mining, common event logs may be transferred to be workflow trace data to utilize existing workflow mining methods.

3) Mining and discovering: Performs mining algorithm to discover behavior pattern. Algorithms consider previous stage data as input and output the results of discovering. Thus, this stage is the main phase of behavior pattern discovering.

4) Verification: Applies the discovered results on system improvement and research, and check out if the results fit the experience of real world. Thus, this stage is the end of a discovering cycle, and is the start point of the next one.

B. *Algorithms in Behavior Pattern Mining Context*

The major difficulties of behavior pattern mining come from the characteristics of input data. The raw data of common event logs hides a large part of knowledge instead of showing them. For example, we have no knowledge about which events relate to the same behavior pattern in expecting model, and events came from the huge number of users in entire information system have great chance to disturb the relationships of events came from users keeping closer connection. Such specialty of input data leads to new challenges advanced by behavior pattern mining (i.e., the No. 11 to 14 challenges in section II-C).

As a result, a straightforward idea is to bridge the knowledge gap between common event logs and workflow traces, and then we can benefit from algorithms of workflow mining again. According to paper [9], we consider the gap bridging is a new part of preprocessing stage. In other words, we try to transfer common event logs to be workflow traces, and then we get the same format of data source for mining. Nevertheless, such transformation is not easy to accomplish, and the effect of transformation depends on the understanding of the inner cohesion in properties of separate events.

In most lucky situation, we could get a transferred data source as similar as traces. We should notice that the data transformation in preprocessing is not to do whatever we want. In fact, such transformation is a kind of dividing and resorting, and it influences the meaning of mined model seriously. Therefore, paper [9] provides the concept of perspectives to standardize the transfer operation, and discusses related meaning of mined model under each kind of perspectives. Considering most process mining algorithms suppose relative events keep a high possibility to appear one by one, so a useful strategy to produce perspectives is to keep perspectives as small as possible to avoid the destroy to the connection between relative events. One of such minimal perspective not mentioned in paper [9] is the task perspective. We divide the business logic of information system into small pieces targeting minor tasks so that each minor task can be accomplished under a little number of steps. If some tasks are still too big, we could try to further divide such tasks into subtasks. In e-business system, a task can be a successful purchase. In e-collaboration system, a task can be a small mission for a virtual group. The task perspective outputs finite number of time ordered events relating to same tasks. This perspective divides event logs into a structure very similar to traces, and the mining methods in next stage will classify tasks and find out patterns in each class of tasks.

In other situation, perspectives still output a long continuous event logs containing comprehensive structure of patterns. An example of this situation is the user perspective in paper [9]. Then the algorithms came from workflow mining need improvement to solve the new challenges advanced by behavior pattern mining. The bad news is existent workflow mining algorithms seldom have the ability to discover all major kinds of pattern structure. Thesis [5] gives out an overview of major approaches in process mining, and concludes the result as Table III, and the genetic mining algorithm described in [5] is a kind of method could discover all major kinds of pattern structure but shows poor performance in computation time. The good news is most of workflow mining algorithms have chance to be extended to fit behavior pattern mining context. A large number of mining algorithms focus on local dimension which is common for long continuous event logs and workflow traces, but

algorithms define some operator binding to traces. For example, the genetic mining in [5] describes its fitness measurement function according to the number of specific traces, and we have to redefine such function under behavior pattern mining context to accomplish algorithm extension.

The discovery of loop structure plays a key role when we analyze none trace data source, because a successful expecting mined model should generalize the characteristics of original event logs, and finding the existence of loop structure is one of the most effective way to improve generality. For mining methods can not discover loop structure perfectly, one of the potential compromise is to cut off logs in every positions related to loop structure, then we get traces without any loop structure, but the mined behavior pattern will be frequently appeared fragments of process model without the coverage of whole process.

IV. Tools and Simulation

ProM [20] is an extensible framework and a full functional open source test-bed for process mining research and application. ProM covers a wide range of process mining technology. It supports log format transformation, events filtering, model mining, events statistics and verification. Nearly all major process mining approaches have implementation as plug-ins in ProM, including Alpha algorithm, Heuristics miner, and Genetic algorithm. New algorithm for behavior pattern mining can be implemented grounding on ProM's api, and improvement for existing algorithm can be tested and compared on ProM.

To simulate the data source of process mining, some researchers use commercial workflow management engine, such as Staffware to generate expecting event logs. Some others choose CPN Tools [21] to generate logs. First, the target model should be transferred to Colored Petri Net, and use CPN Tools to export the transferred model as XML file. Second, use ProM Import [22] to load the target model from CPN Tools file, and generate related event log in MXML format. At last, we merge multi copies of such log like crossover to simulate an event log that comes from multi users in an information system.

Collecting event logs from real world systems is a good way to get needed data as well. Paper [23] introduces a framework for collaboration supporting systems, and such framework has already included log and event recording function as one of its core services.

V. Conclusion

Workflow mining is the largest and most successful sub field of process mining area till now. It treats workflow traces as its input data. To extend process mining technology on

TABLE III
Overview of Major Process Mining Approaches

	Cook et al. [10]	Agrawal et al. [11]	Pinter et al. [12]	Herbst et al. [13]	Schimm [14]	Grecco et al. [15]	Van der Aalst et al. [16]	Weijters et al. [17]	Dongen et al. [18]	Wen et al. [19]
Event log:										
- Atomic tasks	√	√		√		√	√	√	√	√
- Non-atomic tasks			√		√					
Mined model:										
- Dependencies	√	√	√	√	√	√	√	√	√	√
- Nature split/join	√			√	√	√	√	√	√	√
- Whole model		√	√	√	√	√	√	√	√	√
Mining approach:										
- Single step	√	√	√				√	√		
- Multiple steps				√	√	√			√	√
Sequence:	√	√	√	√	√	√	√	√	√	√
Choice:	√	√	√	√	√	√	√	√	√	√
Parallelism:	√	√	√	√	√	√	√	√	√	√
Loops:										
- Structured	√	√	√	√	√		√	√	√	√
- Arbitrary		√	√				√	√	√	√
Non-free-choice:										
- Local	√	√			√	√	√	√	√	√
- Non-local										√
Invisible Tasks:										
- Skip	√	√	√	√	√	√		√	√	
- Split/join	√			√						
Duplicate Tasks:										
- Sequence				√						
- Parallel				√						
Noise:										
- Dependency pruning	√	√	√	√						
- Other								√		

common event logs that have no workflow cases reference, which technology is named as behavior pattern mining, will provide new motivation for the development of process mining technology, and brings the process mining technology into wider range of systems to help understand users' behavior pattern in information systems through an automatic way.

This paper gives out the definition of behavior pattern mining, and briefly surveys such area, including the motivation, challenges, data format, major technology and related tools. The key point of behavior pattern mining is how to combine the power of four typical stages, including Events recording, Preprocessing, Mining and discovering, Verification, to get rid of the specialty that the source data lacks future knowledge about target model. Following the proposed issues, the future research direction of behavior pattern gets a basic outline. However, limited by the length of writing and the short history of behavior pattern mining, many works need to go deeper, especially to test existent approaches and algorithms through real world data.

ACKNOWLEDGMENT

This work is funded by Ministry of Science and Technology, PRC. The project grant contract numbers are 2005DKA64100 and 2005DKA10201.

The idea of five typical collaborative scenarios, which is advanced by Professor Zhanqiu Dong of Graduate University of Chinese Academy of Sciences, attracts our interest to discover the rules in collaborative environment, and guides us to put forward the concept of behavior pattern mining.

REFERENCES

[1] The Science of Collaboratories (SOC) project is supported by the National Science Foundation under Grant No. IIS 0085951. The address is: http://www.scienceofcollaboratories.org/

[2] Cooley, R., P.-N. Tan, and J. Srivastava, "Discovery of intersting usage patterns from Web data," presented at WEBKDD, 1999.

[3] W.M.P. van der Aalst and A.J.M.M. Weijters, " Process Mining: A Research Agenda", Computers in Industry, 53(3):231-244, 2004.

[4] W.M.P. van der Aalst a, B.F. van Dongen a, J. Herbst b, L. Maruster a, G. Schimm c, A.J.M.M. Weijters a, "Workflow mining: A survey of issues and approaches", Data & Knowledge Engineering 47 (2003) 237–267.

[5] Alves de Medeiros, Ana Karla, "Genetic Process Mining", CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN, printed in 2006.

[6] Clarence, A. E. Kwang-Hoon, K. Aubrey, J. R. Workflow Mining: Definitions, Techniques, and Future Directions. Workflow Handbook 2006, Springer Press, 213-228.

[7] Christian W. G¨unther and Wil M.P. van der Aalst, "A Generic Import Framework For Process Event Logs", Workshop on Business Process Intelligence (BPI), Vienna, Austria, 5-7 September, 2006.

[8] Ana Karla Alves de Medeiros and A.J.M.M. (Ton) Weijters, " ProM Framework Tutorial", November 2006, http://is.tm.tue.nl/~cgunther/dev/prom/PromTutorial.pdf

[9] Jinliang, S. Tiejian, L. Su, C. Feng G. The Data Preprocessing of Behavior Pattern Discovering in Collaboration Environment. The 2007 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2007), Silicon Valley, USA, will be published in Nov 2007.

[10] J.E. Cook, Z. Du, C. Liu, and A.L.Wolf. Discovering Models of Behavior for Concurrent Workflows. Computers in Industry, 53(3):297-319, 2004.

[11] R. Agrawal, D. Gunopulos, and F. Leymann. Mining Process Models from Workflow Logs. In I. Ramos G. Alonso H.-J. Schek, F. Saltor, editor, Advances in Database Technology - EDBT'98: Sixth International Conference on Extending Database Technology, volume 1377 of Lecture Notes in Computer Science, pages 469-483, 1998.

[12] S.S. Pinter and M. Golani. Discovering Workflow Models from Activities Lifespans. Computers in Industry, 53(3):283-296, 2004.

[13] J. Herbst and D. Karagiannis. Workflow Mining with InWoLvE. Computers in Industry, 53(3):245-264, 2004.

[14] G. Schimm. Mining Exact Models of Concurrent Workflows. Computers in Industry, 53(3):265-281, 2004.

[15] G. Greco, A. Guzzo, L. Pontieri, and D. Sacca. Mining Expressive Process Models by Clustering Workflow Traces. In H. Dai, R. Srikant, and BIBLIOGRAPHY 365 C. Zhang, editors, PAKDD, volume 3056 of Lecture Notes in Computer Science, pages 52-62. Springer, 2004.

[16] 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.

[17] 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.

[18] B.F. van Dongen and W.M.P. van der Aalst. Multi-phase Process mining: Aggregating Instance Graphs into EPCs and Petri Nets. In Proceedings of the Second International Workshop on Applications of Petri Nets to Coordination, Workflow and Business Process Management (PNCWB), 2005.

[19] L. Wen, J. Wang, and J. Sun. Detecting Implicit Dependencies Between Tasks from Event Logs. In Xiaofang Zhou, Jianzhong Li, Heng Tao Shen, Masaru Kitsuregawa, and Yanchun Zhang, editors, APWeb, volume 3841 of Lecture Notes in Computer Science, pages 591-603. Springer, 2006.

[20] The homepage of ProM: http://is.tm.tue.nl/~cgunther/dev/prom/

[21] CPN Tools: http://wiki.daimi.au.dk/cpntools/cpntools.wiki

[22] ProM Import: http://is.tm.tue.nl/~cgunther/dev/promimport/

[23] T. Luo, J. Song, and S. Chen, etc. A Services Oriented Framework for Integrated and Customizable Collaborative Environment. In the 2007 IEEE International Conference on Information Reuse and Integration, August 2007.