Supporting Process Mining by Showing Events at a Glance

Minseok Song and Wil M.P. van der Aalst
Eindhoven University of Technology
P.O.Box 513, NL-5600 MB, Eindhoven, The Netherlands
{ m.s.song, w.m.p.v.d.aalst}@tue.nl

Abstract. *Process mining* has emerged as a way to analyze processes based on the event logs of the systems that support them. Today's information systems (e.g., ERP systems) log all kinds of events. Moreover, also embedded systems (e.g., medical equipment, copiers, and other high-tech systems) start producing detailed event logs. The presence of event logs is an important enabler for process mining. The primary goal of process mining is to extract knowledge from these logs and use it for a detailed analysis of reality. One of the challenging issues in process mining is *process performance analysis*. As a method to analyze process performance and to provide new insights, this paper proposes the dotted chart that shows overall process events at a glance. The chart shows process events in a graphical way such that the analyst gets a "helicopter view" of the process and is able to immediately spot opportunities for process improvement. The approach has been implemented in the context of the ProM framework.

Key words: Process mining, business process management, business performance, workflow, Petri nets

1. Introduction

Process mining has emerged as a way to analyze systems and their actual use based on the event logs they produce. The goal of process mining is to extract information (e.g., process or organizational models) from these logs, i.e., process mining describes a family of a-posteriori analysis techniques exploiting the information recorded in the event logs. One of the challenging topics in process mining is *process performance analysis*. Process performance analysis aims at supporting organizations to measure and analyze the performance of their business processes. The importance of process performance analysis is widely recognized. A lot of research has been performed to measure process performance and there are many commercial solutions provided by software vendors [8,9].

The works on the process performance analysis can be categorized into two approaches. The first approach is the *data driven approach*. It focuses on data elements and shows aggregated views of the data. Commercial BI and BAM tools belong to this category [5]. They typically look at aggregate data seen from an external perspective (frequencies, averages, utilization, service levels, etc.). This approach includes the work on the data visualization tools (e.g. Magna View). These tools show summarized views using various kinds of graphs and nice graphical metaphors (e.g., speedometers) [4,6,11]. The summarized views and proper visualization mechanisms help users to better understand process performance, especially the overall performance of business processes. However, it is difficult to look at the *inside of the process* to identify and locate problems (e.g. delays, bottlenecks).

The second approach is the *process driven approach*. It concentrates on business processes. This approach includes tools such as ARIS PPM, HP Business Process cockpit, ILOG JViews, ProM, etc. [7,10,1] They make it possible to look "inside the process" at different abstraction levels. They normally use a process model as a metaphor for showing process performance. It means that they require "well-formed" process models where the performance information is projected. Thus this approach is not suitable for the analysis of highly complex and/or flexible processes (i.e. do not have a structured and/or stable process model).

To overcome the limitations of both approaches, we come up with the idea of showing process execution as it is. Like existing process mining techniques, we focus on the event logs and have developed the "dotted chart" that shows the event logs in a graphical way. It also calculates various performance metrics. The chart enables users to obtain valuable insights into the phenomena influencing

the performance. Unlike the existing commercial BI and BAM tools, the dotted chart can easily look at the inside of the process. And it is more robust than the process driven approach, since it does not require any underlying process model.

The paper is organized as follows. Before explaining the dotted chart, we briefly introduce process mining in Section 2. In Section 3, the concept of the dotted chart is explained. Section 4 describes the implementation in the ProM framework and the application of the dotted chart to a real process analysis. Finally, Section 5 concludes the paper.

2 Process Mining: A Short Overview

The goal of process mining is to extract information (e.g., process or organizational models) from event logs, i.e., process mining describes a family of a-posteriori analysis techniques exploiting the information recorded in the event logs. Typically, these approaches assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in the process) and is related to a particular case (i.e., a process instance). Furthermore, some mining techniques use additional information such as the performer or originator of the event (i.e., the person / resource executing or initiating the activity), the timestamp of the event, or data elements recorded with the event (e.g., the size of an order).

Process mining addresses the problem that most "process/system owners" have limited information about what is actually happening. In practice, there is often a significant gap between what is prescribed or supposed to happen, and what *actually* happens. Only a concise assessment of reality, which process mining strives to deliver, can help in verifying process models, and ultimately be used in system or process redesign efforts.

The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs. As shown in Figure 1, we consider three basic types of process mining: (1) *discovery*, (2) *conformance*, and (3) *extension*.

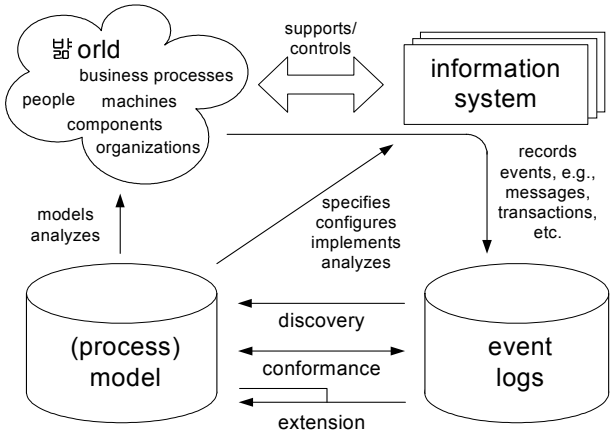

Figure 1. Process mining overview

Traditionally, process mining has been focusing on *discovery*, i.e., deriving information about the original process model, the organizational context, and execution properties from enactment logs. There is no a-priori model, i.e., based on an event log some model is constructed. An example of a technique addressing the control flow perspective is the α-algorithm, which constructs a Petri net model describing the behavior observed in the event [3]. However, process mining is not limited to process models (i.e., control flow) and recent process mining techniques are more and more focusing on other perspectives, e.g., the organizational perspective or the case perspective. For example, there are approaches to extract social networks from event logs and analyze them using social network analysis [2]. This allows organizations to monitor how people, groups, or software/system components are working together. *Conformance* checking compares an a-priori model with the observed behavior as recorded in the log, i.e., reality is compared with e.g. process model or business rule and deviations are detected. The third type of process mining involves *extending* a model with additional information extracted from the log, e.g., information about decisions, timing, resources, frequencies, etc.

The dotted chart analysis can be positioned as a discovery technique. However, unlike the existing techniques it emphasizes the time dimension and puts no requirement on the structure of the process. It provides a useful ``helicopter view'' and can be used to quickly locate performance problems as will be shown in the remainder.

3. Dotted Chart Analysis

3.1. Overview

This section gives an overview of the dotted chart. The dotted chart is a chart similar to a Gannt chart. It shows a spread of events of an event log over time. The basic idea of the dotted chart is to plot dots according to the time. Figure 2 shows an example of the dotted chart. In the chart, a dot on the chart represents a single event in the log. The chart has two orthogonal dimensions: (1) *time* and (2) *component types*. The time is measured along the horizontal axis of the chart. Along the vertical axis, component types such as instance, originator, task, event type, or data elements are shown. Based on the component types, the events are rearranged. For example, Figure 2(a) shows the dotted chart using tasks as component types, i.e., every "row" corresponds to a task and the dots show the times at which event corresponding to this type took place. Figure 2(b) uses process instances as component types, i.e., every row corresponds to a particular case (e.g., patient, order, application, etc.). Component types help users to focus on a particular aspect of the event log. For example, if the instance is used as a component type, the spread of events within each instance can be identified. Then users can easily identify which instance takes longer, which instance has many events, etc. The chart provides several time options and some metrics for performance. The remainder of this section will explain these in detail.

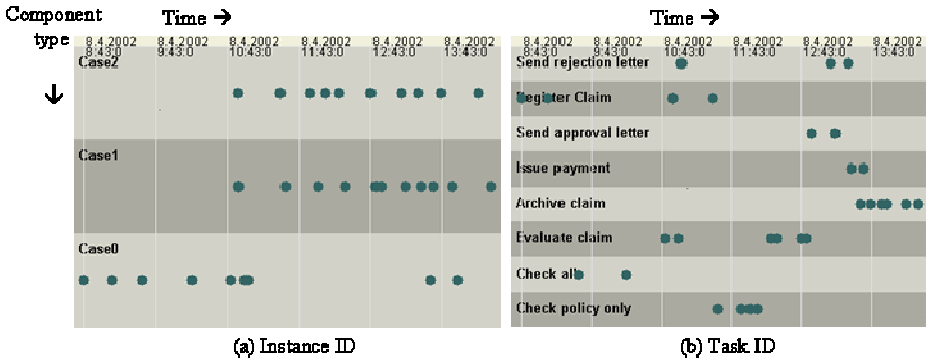

Figure 2. Example of dotted chart

3.2. Multiple Time Options

The dotted chart provides multiple time options. These options determine the position of the event in the horizontal time dimension. The first option is to use actual timestamps, i.e., the time when the event actually happened is used to position the corresponding dot. Figure 2 shows two dotted charts using actual timestamps. However, there are four alternative time options as shown in Figure 3. Figure 3(a) shows an example using the *relative time option*, i.e., the first event for every component type (in this example a process instance) is positioned at time 0. This shows that

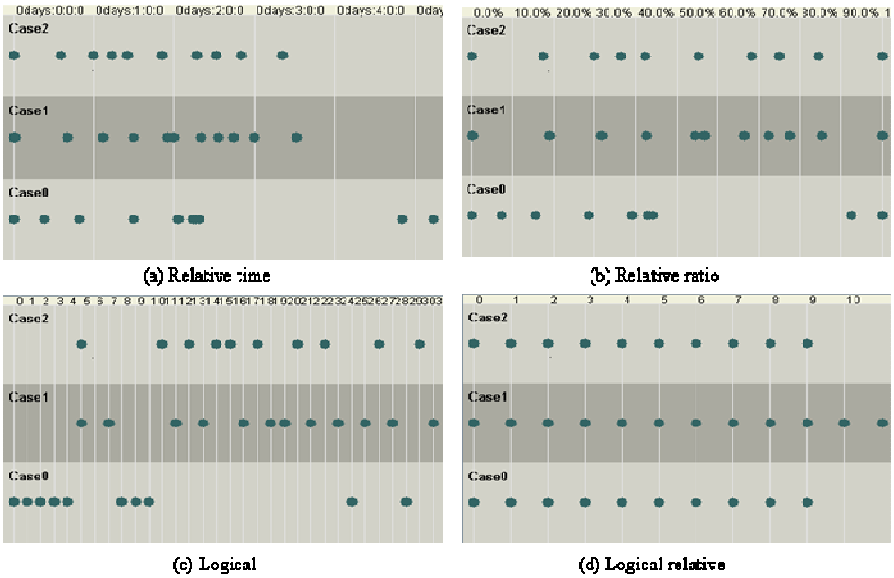

Figure 3. Various time options

process instance "Case0" took the longest although it was finished before the other two instances (compare Figure 2(a) with Figure 3(a)). The *relative ratio option* shown in Figure 3(b) stretches each case to end at the same time. This way one can see the relative distribution of events inside each process instance. Figure 3(c) shows the *logical option*. Here are events sorted based on the timestamps and given a sequence number, i.e., the first event has time 0, the second event has time 1, etc. Figure 3(d) shows the *logical relative option* combining the idea of having sequence numbers rather than timestamps with starting all instances at time 0.

3.3. Performance Metrics

The dotted chart shows the spread of events in an event log. For this reason, it is difficult to show traditional performance values such as waiting times and execution times of tasks. However, it can provide the metrics related to events and their distribution over time (spread). There are two kinds of performance metrics: (1) metrics for the overall event log and (2) metrics for each component. Figure 4 depicts performance metrics in the chart. For the overall event log, (a) the position of the first event in the log, (b) the position of the last event in the log, (c) average spread, (d) minimum spread, and (e) maximum spread can be calculated. For each component type, (1) the position of the first event in a component, (2) the position of the last event in a component, (3) average interval between events, (4) minimum interval between events, and (5) maximum interval between events can be calculated. Based on the component type and time options, the performance values can be interpreted in a various ways. For example, if we use originator as a component type and actual time as a time option, the average spread represents the average length of time that the originators appeared to be involved. If we use instance as a component type and relative time as a time option, the average spread stands for the average execution time of instances.

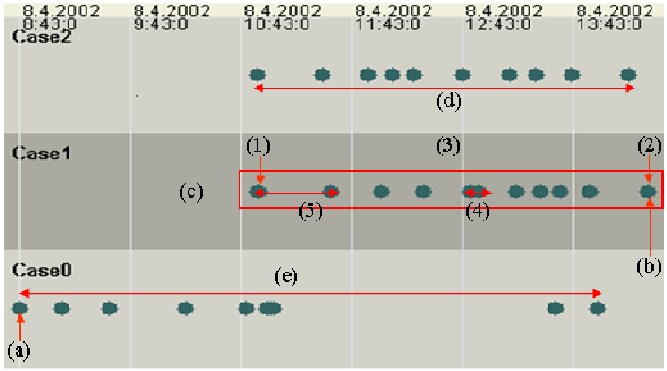

Figure 4. Performance values

4. Implementation and Application

We have implemented the dotted chart as a plug-in for the ProM framework [1]. The ProM framework is open source and uses a plug-able architecture, e.g., people can add new process mining techniques by adding plug-ins without spending any efforts on the loading and filtering of event logs and the visualization of the resulting models. Figure 5 shows a screenshot of the dotted chart plug-in. The plug-in has a menu panel on the left, the chart is shown in the middle, and both the overall view of the diagram and the metrics explained in Section 3.3 are displayed on the right. The menu panel consists of several options for

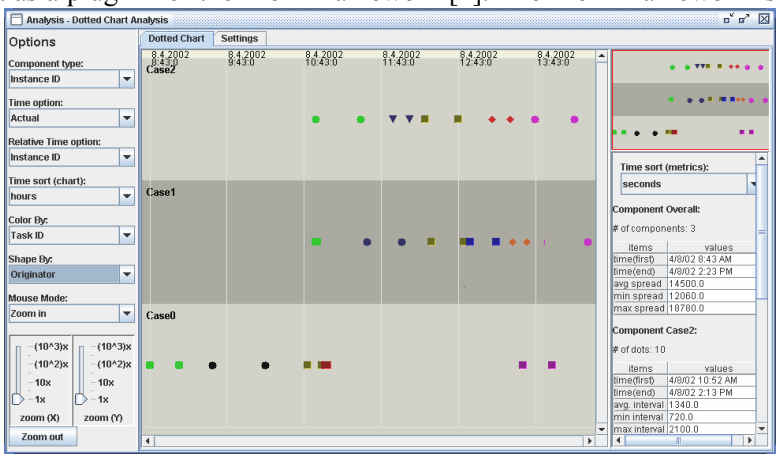

Figure 5. Screenshot of the Dotted Chart

selecting component type, time option, relative time option, etc. One of the interesting features is the zoom in/out function that enables users to focus on the particular area in the diagram (e.g. finding patterns or viewing a complex part). Users use the sliderbars on the menu panel or select a particular area on the diagram to zoom in the area. To display more information, both color and shape based on component type can be used. In the figure, colors refer to tasks, and shapes refer to originators.

We have applied the dotted chart analysis to analyze real event logs from several organizations, including several hospitals, producers of high-tech systems (medical equipment and wafer steppers), municipalities, banks, etc. In this paper, we can only briefly present our analysis of a log from a large hospital in the Netherlands. The event log contains 619 process instances (patients) and 3515 audit trail entries (events). It has 51 district activities and 34 originators (departments) are involved in the process execution.

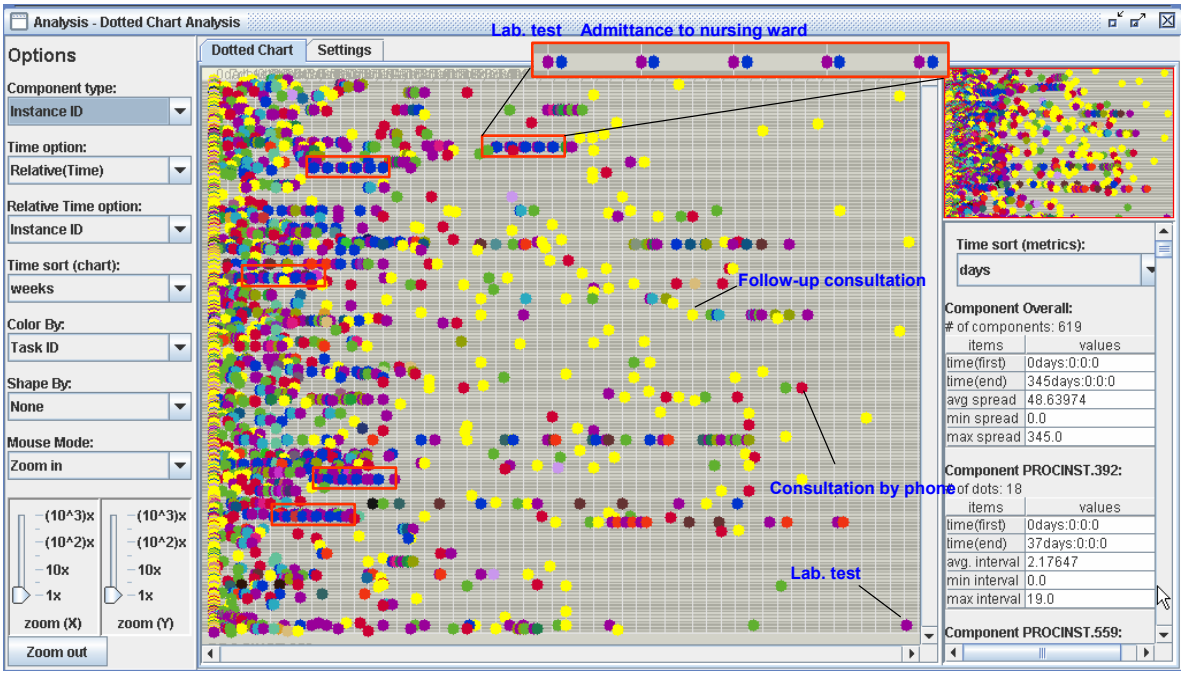

Figure 6. The dotted char for an event log containing 3515 events relating to 619 patients

Figure 6 shows a dotted chart generated for this particular patient log. In the diagram, we use relative time option and the component types are instances (patients), i.e., the first event for every patient takes place at time 0. The results clearly indicate the duration of each instance. The metrics shown on the left-hand side of Figure 6 also indicate the time of the first and of the last events, durations, the number of events in an instance, etc. For example, it is shown that the average case duration is about 49 days and the maximal duration is 345 days.

Users can obtain useful insights from the chart, e.g., it is easy to find interesting patterns by looking at the dotted chart. In Figure 6, the density of events on the left side of the diagram is higher than the density of those on the right side. This shows that initially patients have more diagnosis and treatment events than in the later parts of the process. When we focus on the long duration instances (i.e. the instances having events in the right side of the diagram), it can be observed that they mainly consist of follow-up consultation (yellow dot), consultation by phone (red dot), and lab test (violet dot) activities.
It reflects the situation that patients have regular consultation by visiting or being phoned by the hospital and sometimes have a test after or before the consultation. It is also easy to discover patterns in the occurrences of activities. For example, five instances have the pattern that consists of a lab test and an admittance to the nursing ward activities.

The applicability of the dotted chart has been demonstrated using a real-life event log. It shows that the concept of the dotted chart is relatively simple, but makes it possible for users to gain an overall insight into the underlying process and its performance.

5. Conclusion

This paper presented an innovative way to view events in the context of process mining. Given an event log, the dotted chart shows a powerful overview of the underlying process. It is possible to select different views by changing the component types or time options, or by zooming into parts of the chart. Moreover, based on the dotted chart all kinds of performance metrics are generated and visualized. The dotted chart paradigm has been implemented in the context ProM framework and is made available via www.processmining.org. In the paper, we reported on a case study conducted in a Dutch hospital. This and other practical applications of the dotted chart show its robustness and ability to provide useful "helicopter views" and ideas for process improvement.

Acknowledgement

This research is supported by EIT, NWO-EW, the Technology Foundation STW, and the SUPER project (FP6). Moreover, we would like to thank the many people involved in the development of ProM.

References

[1] W.M.P. van der Aalst, B.F. van Dongen, C.W. Günther, R.S. Mans, A.K. Alves de Medeiros, A. Rozinat, V. Rubin, M. Song, H.M.W. Verbeek, and A.J.M.M. Weijters. ProM 4.0: Comprehensive Support for Real Process Analysis. In J. Kleijn and A. Yakovlev, editors, *Application and Theory of Petri Nets and Other Models of Concurrency (ICATPN 2007)}}*, volume 4546 of *Lecture Notes in Computer Science*, pages 484--494. Springer-Verlag, Berlin, 2007.

[2] W.M.P. van der Aalst, H.A. Reijers, and M. Song. Discovering Social Networks from Event Logs. *Computer Supported Cooperative work*, 14(6):549-593, 2005.

[3] 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.

[4] Chen C.-A, S. Kalvala, and J. Sinclair. A Process-Based Semantics for Message Sequence Charts with Data. In *ASWEC '05: Proceedings of the 2005 Australian conference on Software Engineering*, pages 130-139, Washington, DC, USA, 2005. IEEE Computer Society.

[5] Gartner. Gartner's Application Development and Maintenance Research Note M-16-8153, The BPA Market Catches another Major Updraft. http://www.gartner.com, 2002.

[6] M.-C. Hao, U. Dayal, and F. Casati. Visual mining business service using pixel bar charts. In B.-F. Erbacher, P.-C. Chen, J.-C. Roberts, M.-T. Gröhn, and K. Börner, editors, *Visualization and Data Analysis 2004*, volume 5295 of *SPIE Conference*, pages 117-123, June 2004.

[7] 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.

[8] P.Kueng. Process Performance Measurement System: a tool to support process-based organizations. *Total Quality Management*, 11(1):67-85, 2000.

[9] 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.

[10] 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.

[11] D.P. Tegarden. Business information visualization. *Communications of the AIS*, 1:2-38, 1999.