Visual Interactive Process Monitoring

Sebastian Maier[1(✉)], Hannes Kühnel[2], Thorsten May[1], and Arjan Kuijper[2]

[1] Fraunhofer Institute for Computer Graphics Research,
Fraunhoferstr. 5, 64283 Darmstadt, Germany
{Sebastian.Maier,Thorsten.May}@igd.fraunhofer.de
[2] TU Darmstadt, Darmstadt, Germany

Abstract. Sensor data has been coined the oil of the 21st century. We present a technique for the visual analysis of multivariate sensor event log data. This technique tackles two challenges: Firstly, in a complex process the relation of causes and effects is often masked by indirections. Secondly, the metrics to measure success might be different from the measures that identify causes. Thus, our approach does not require that all sensor data is equal. Our techniques combines automated and interactive grouping to identify candidate sets sharing properties relevant for cause and effect analysis. Interactive visual probes offer immediate information on the statistical relevance of an identified connection.

1 Introduction

Large quantities of data are created nowadays by sensors embedded in larger systems. Sensor data is created in virtually any company, in an increasing number of devices, and even in our private lives. In an industrial setting, sensor data are often part of a complex production or service process. A single instance of such a process typically traverses a number of components. Many components are equipped with various sensors each creating additional data about this process instance. During process execution, these sensors create events containing a timestamp and further specific data. For example sensors are used to monitor input or output quality, operational parameters of the process component, like control parameters, throughput, resource consumption or wear. In our scenario, we assume that processes are executed multiple times. Thus, in summary they produce a large amount of data which can be used to evaluate process performance and cost.

Applying concepts from visual analytics to present and analyze this data helps to unveil those nuggets of information which have potential to improve the process and also the product. A key requirement is the ability to identify the root cause for a specific subset of process outcomes. For example, an operator might be interested in the cost-drivers for a specific industrial or service process. This can be further complicated as the effects might be expressed in one unit of measurement like revenue while the cause might be in another unit of measurement like execution time.

Given a process model and a reasonably large set of sensor data, it is difficult to find those interesting nuggets of information within such a sea of data.

S. Yamamoto (Ed.): HIMI 2015, Part I, LNCS 9172, pp. 74–85, 2015.
DOI: 10.1007/978-3-319-20612-7_8

Manyika et al. [11] have estimated the potential value for the US health care system to $ 300 billion. According to them this value can be created in several ways, two of them being creating transparency and by enabling experimentation to improve performance. Both ways require human access to the data.

In the following section we will give an overview of *related work* from the process mining and process monitoring community as well as previous works on visualizing temporal event sequences. In the *method* section we will discuss our approach in detail. Describing the visual representation as well as the analytical components. The *results* section will provide feedback from the user evaluation on specific visual design decisions and the overall application.

In summary, our contribution is a technique for the visual analysis of multivariate event logs data. This technique tackles two challenges: Firstly, in a complex process the relation of causes and effects is often masked by indirections. Secondly, the metrics to measure success might be different from the measures that identify causes. Thus, our approach does not require that all sensor data is equal. Our techniques combines automated and interactive grouping to identify candidate sets sharing properties relevant for cause and effect analysis. Interactive probes offer immediate information on the statistical relevance of an identified connection.

2 Related Work

As we consider this work to be at the corner of *process mining* and *visual analytics* we reviewed publications which are relevant to understand the field of process mining and those applying visual analytics techniques to process data.

2.1 Process Mining

Process mining is a combination of data mining, process modeling and analysis. As opposed to process design and operation, process mining is considered a "bottom-up" approach. This means, processes are modeled, evaluated and compared based on empirical measurement. The idea is to discover, monitor and improve real processes through extraction of information from event logs. The basis for process mining is an event log, with every event representing a sensor in a process step. Event typically include a timestamp and potentially multivariate information, depending on the sensor.

Van der Aalst et al. [3] distinguishes three high-level tasks for process mining - discovery, conformance and enhancement. *Discovery* is the task of building a process model based upon the patterns identified in the event logs. *Conformance* basically combines process mining and process design. Event data is mapped onto a specific process model. The result of a conformance test defines if and where the model matches the measurement. *Enhancement* aims at extending a process model to adapt to new measurements or newly emerging patterns.

Van der Aalst et al. defined six guiding principles for process mining in the Process Mining Manifesto [3], we will name those two we see as most important for our application:

- Log extraction should be driven by questions, because they give better understanding and meaning for event data.
- Models should be treated as purposeful abstractions of reality, because it is helpful to have multiple views on data because there can be multiple views.

In the same Manifesto they defined eleven challenges, we will focus on the challenges eight to eleven, which we do aim for with this paper:

- Provide operational support because process mining is not only an offline analysis but can also be used for online operational support, where three activities can be found: detect, predict, and recommend.
- Process Mining also should be combined with other types of analysis like e.g. visual analytics, simulation and data mining to get more information from event data.
- Usability for non-experts should be improved in form of user-friendly interfaces in front of the process mining and algorithms.
- Understandability for non-experts also has to be improved because there are problems by understanding the results of the analysis, so they should be clearly shown.

2.2 Process Visualization and Analysis

Process visualization techniques are a subclass of time-oriented data visualization techniques (see Tominski and Aigner for a survey [5]). In this section, we focus on techniques primarily aimed at *discrete* time-oriented data, such as event logs. With this focus, we may distinguish techniques that show explicit connections between discrete events (i.e. they show the process) and those that do not. Furthermore, we distinguish whether a technique shows a single instance or multiple instances. By "instance" we refer to either a process instance or an event series, in case a process is not defined. In these terms, our own technique is a process visualization for multiple process instances, each of which containing heterogeneous multivariate data. Visualization of process models actually have been first proposed outside the visualization community. In fact, visual business process design is part of a number of commercial tools. In virtually all cases, the visualization is a node-link diagram. Albrecht et al. [6] present an approach to layout models defined according to BPEL standard. However, the visualization of the model only, does not serve a direct purpose for a "bottom-up" approach. For visual process mining, the event data has to be included in the visualization.

Rind et al. [13] compare fourteen visualization techniques for electronic health records, which mostly consist of discrete event data. Most of the techniques surveyed do not show explicit connections between events. Events, like, diagnoses, therapies and outcomes are often arranged along a time axis. *Lifelines2* [18] is an example for such a visualization and temporal comparison of multiple patient records. Lifelines2 also features so called *temporal summaries*; these summaries present the distribution of durations or time between events over all patients in a specific group. This approach is similar to our probes, except that our probes are

moved freely over the visualization to define events for analysis. In comparison, the *Lifeflow* [20] approach emphasizes the comparison of event sequences instead of time or durations. Such a sequence visualization offers a "middle ground" between anchoring events on a time axis and connecting events to a process model. Finally the Outflow [19] visualization anchors the patient data to the visualization of the model. Like our own approach, it combines properties from Sankey Diagrams [12] and techniques for layouting directed acyclic graphs [8]. Sankey Diagrams have originally been invented to visualize import and export volumes traded across the world. Riehmann et al. [12] present an interactive version to show an energy distribution network. Sankey Diagrams are directed graphs, whose edge thickness is defined by the volume of flow. With outflow and our approach, the edge thickness is defined by the number of instances. Directed acyclic graphs are easier to layout than general graphs. Thus, specific layout techniques have been proposed, and many layout techniques preserve the sequence of events, which we adopted for our technique. In comparison to Lifelines2 or similar approaches, the relative time is not preserved in the spatial layout. We capture time and duration in the nodes. *Sessionviewer* [10] is a visualization for web session logs, offering multiple levels of granularity. It relates views on different levels of aggregation. A model view, which is a state transition graph, an aggregate view over time and detailed information for all sessions. Most notably it is aimed at distinguishing sessions by different usage patterns.

Aside from the visualization itself, a number of approaches have focused on analytical challenges. One of these challenges is sequence aggregation and identification of patterns. For multivariate process data, patterns may be defined in terms of the process structure, in terms of node attributes or any combinations thereof. Fails et al. present a query interface for the comprehensive investigation of event patterns. Queries are edited in visual building blocks. Thus, the user must establish the correspondence between query and result on his own. Another approach to identify patterns between processes has been presented by Wongsuphasawat and Lin [21]. This approach analyses event logs, to identify different patterns of web client use for customer analysis. Specifically their visualization highlights changes of usage patterns for dynamic data. Usage patterns are distinguished by criteria with a predefined ordering. Our approach differs from Fails' approach that a query specification is done by interacting with probes in the process visualization. Our approach differs from Wongsuphasawat's and Lin's approach that patterns are defined as free combinations of node attributes and a specific process structure.

The *parallel sets* approach [9] has been proposed for multivariate categorical data instead of process data. However, parts of its design and behavior are similar to our approach. Most notably, parallel axes are connected by groups of data items, that can in turn be highlighed for detailed inspection. Furthermore the axes offer information about potential correlations between adjacent axes. We provide this function by means of virtual probes.

The idea of an interactive virtual probe has been mainly inspired from similar approaches in scientific visualization [16]. Probes are used to avoid clutter in

complex three-dimensional views. In essence, they provide a pivot area (points, planes or volumes) of a domain, where detailed information is mapped onto. Typically, multiple probes of different types may be moved around freely, which is also supported by our approach.

3 Method

Following Shneidermans well known InfoVis mantra "Overview first, Filter and zoom, Details on demand" [15] we start with providing an overview of the process structure including the information about the event flow to the user. This process overview is provided using a customized Sankey diagram [7,14]. Each node in this diagram represents one of the sensors, see Fig. 1.

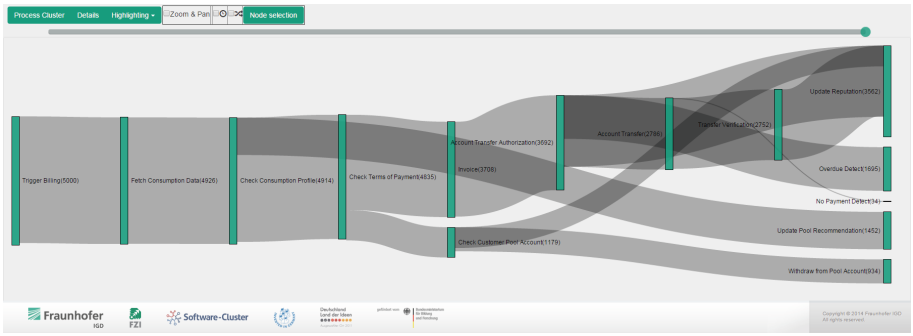


Fig. 1. Initial overview of the full process. The bars (in green) represent steps in the business process. We can immediately see that the process has different endpoints and various junctions (Color figure online).

This is the starting point for all analytical paths the user can choose to explore. We provide two complementing ways to explore the process data. One makes use of the identified event clusters while the other relies on the concept of data probes. A data probe is a small overlay visualization providing detailed information on the attribute(s) of a single sensor. Both ways enable the user to filter the data flow. We can filter based on cluster affiliation and based on the value at one or multiple sensors. For clusters we provide a means to compare two clusters based on their properties. Data probes allow us to perform what-if analysis to explore dependencies between different sensors and sensor attributes.

We will detail each of the described analytical paths in the following subsections after a quick discussion of the data preprocessing and the overview visualization.

3.1 Data Preprocessing

Most, if not all data analysis tools require some sort of data preprocessing to enable a reasonable (visual) presentation for the user.

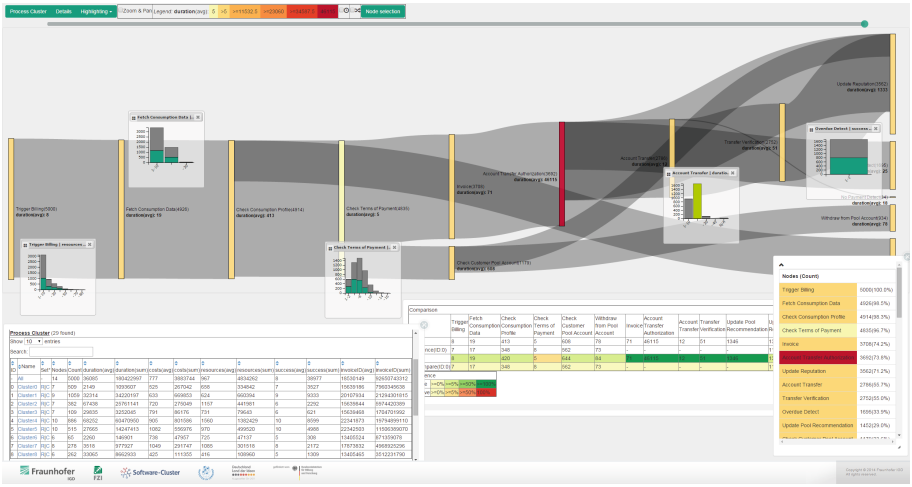


Fig. 2. Visualization of the entire process model with active filters and multiple overlays showing detailed information. Five probes have been places over the process visualization. They show the distribution of performance and cost indicators for a subset of event sessions.

The data consists of a process model, a directed acyclic graph, and a vector of readings for each sensor and session. Each reading is associated with a session id, enabling us to follow the flow of any session through the process model. The process model provides not only information about the relationship between sensors but also information about the split join semantics. Process and business model notations, as one instance of process models, usually provide two different split and join modes: *AND* and *XOR*. See Vergidis et al. for an overview of different business model notations and their supported patterns [17].

To collect sensor properties, like average duration, costs, etc. we attach the event data to the process model and calculate those metrics for each sensor. The two clustering steps are described in more detail in Subsect. 3.3. The overview visualization will be described in the following section (Fig. 2).

3.2 Overview

Providing users with an initial overview is considered good practice in the InfoVis community. Van der Aalst considers a good visual representation of a process model as one of the challenges of process mining [1,2]. We decided to use a *Sankey* diagram for the visualization of the events flowing through the process model. We decided against a standard node-link diagram to emphasize the amount of sessions following a certain path in the model. With a standard node-link diagram this would not be so explicit. As the maximum width of the links is restricted by the size of the nodes - it is not reasonable to draw the links larger

then the attached nodes. Another option would have been a matrix visualization of the underlying graph. However, a matrix visualization masks the temporal ordering of the process. We assume, that preserving the ordering is helpful for cause and effect analysis.

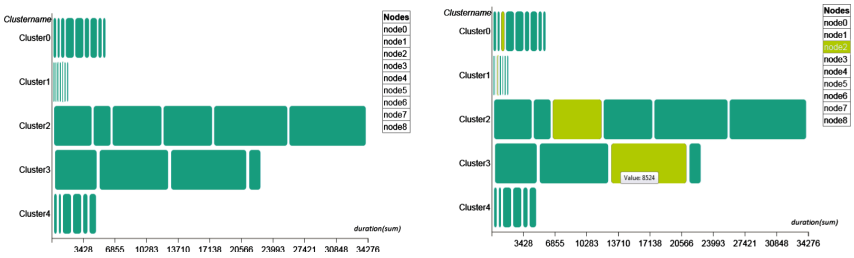
The overview visualization provides zooming and panning for the user to adjust it to her needs. Additional overlays can be added to the overview to support the detailed analysis. They will be describes in the following subsections.

3.3 Clustering

As useful as it is to work on the event sessions, sometimes one is interested on a higher level of abstraction. Our approach offers a method to cluster the event series by their sequence. This is done with an modified *apriori* algorithm. The classical example for this is the identification of *items* which are bought together in the same *transaction* [4]. We interpret sensors as *items* and event sessions as *transactions*. This leads to a clustering of those event sessions that traverse the same set of sensors, independent of the sequence of traversal. In addition, it is possible to define custom clusters by selecting a set of sensors such that all event sessions traversing these sensors will be part of the custom cluster.

For each cluster a number of data attributes are calculated based on the event sequences. This includes information like the number of sequences in the cluster and the number of sensors traversed by this cluster. Also we aggregate each data attribute available for the events. This information is presented to the user in form of a data table.

An additional visualization is available to compare all clusters at once according to quantitative attributes, which has been calculated in the cluster creation phase (see Fig. 3).



(a) Visual comparison of clusters

(b) Highlighting a single sensor in each cluster

Fig. 3. The views are used to compare a quantitative attribute (here: duration) between different clusters (vertical axis). Every cluster defines a set of sensors, represented by green rectangles. The attribute is mapped on the width of a rectangle. Measurements on the same sensor, but different clusters can be compared as well (right) (Color figure online).

Cluster definitions can further be used to filter the full event dataset to only those events comprising a single cluster. This filtering is applied directly to the overview visualization, highlighting all sensors and event sessions which are part of the selected cluster (Fig. 4).

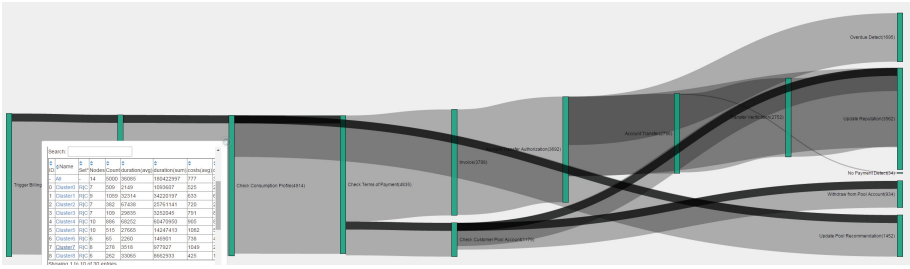
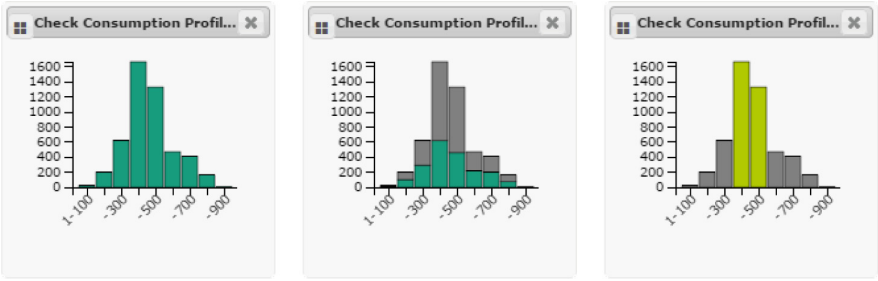


Fig. 4. Highlighting a single cluster



(a) No filter has been defined

(b) A filter has been defined on another probe

(c) A filter has been defined on this probe

Fig. 5. Data probe for a single sensor (here: consumption) in three different states. A value range can be chosen by selecting the corresponding bar of the histogram (Color figure online).

3.4 Data Probes

In contrast to the clustering, our concept of data probes provides a bottom up analytical access. Our user can start his analysis at any sensor, selecting any data attribute that is available for this sensor see Fig. 5a. A data probe will visualize the distribution of the selected attributes values for a single sensor.

There is no technical limitation - except available screen space - to the number of probes a user can attach to the process model. It is even possible to attach multiple probes for different attributes to the same sensor. As space is limited, we render the probes at about 200 square pixels, we perform an automated bucketing of the attributes values to reduce space requirements. To facilitate comparison the bucketing for any attribute is the same for all sensors.

Probes can either be linked to a sensor, in that case we can move them freely in the overview visualization and they will still show the linked sensors readings. Alternatively we can set them into *quick show* mode, in this case, they will always show the readings for the sensor currently touched with the mouse.

Data probes also allow the interactive definition of filters on sensor attributes. Uses can choose to select a single or multiple values by simply clicking on one or multiple of the columns in the visualization see Fig. 5c. This will be immediately reflected on all other open probes, and those to be opened later, by showing the original column values in grey as well as the filtered column values in green see Fig. 5b.

4 Results and Evaluation

We will give a quick overview of the results we obtained so far with this system and also give a summary of the evaluation we conducted with information visualization experts.

4.1 Results

The presented system allows us to analyze large sets of process instances. For our prototype we used different sets with up to five thousand artificially created process instances. Event attribute values have been generated using different stochastic distributions to model real world behavior. The generation of an custom event set has allowed us to work with a data set containing known correlations

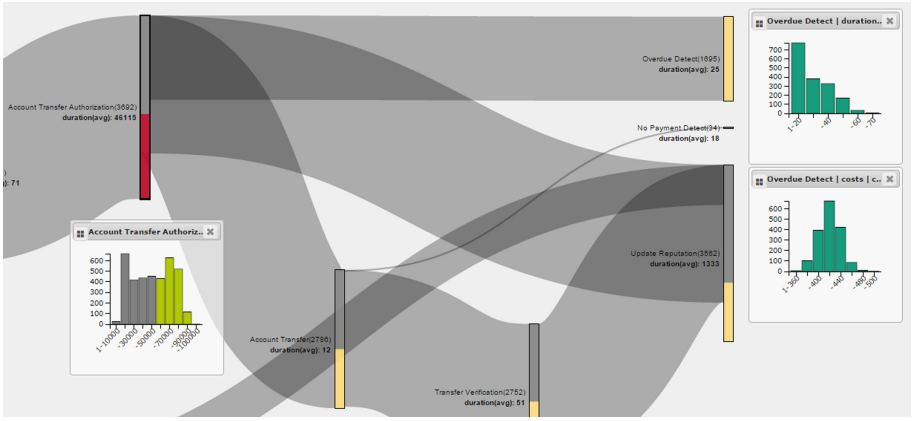


Fig. 6. Using our data probes to analyze correlations in the event data. We have mapped the average duration of events to the color of the sensors. For the leftmost sensor we selected those event sequences with the highest duration. We can see that the top right sensor is only visited by those long running sequences (Color figure online).

between attributes and sensors. The main data set shown in the included figures represents a billing process. The data set contained four dimensions for every event: duration, cost, resources and an invoice id. Parts of the user evaluation are based on smaller artificial sets to focus on specific aspects.

Figure 6 shows a dependency analysis highlighting the artificial correlation between high durations at one sensor and their final process step being *Overdue Detect*. The combination of two different techniques, clustering and data probes, allows to access the data from different perspectives, possibly answering different questions.

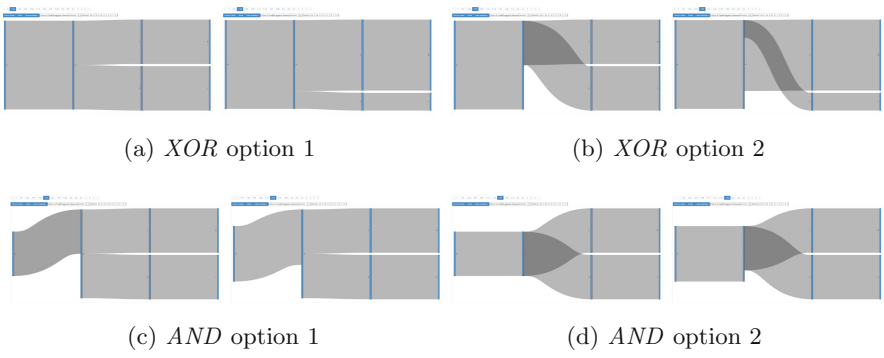
4.2 Evaluation

We conducted a two day evaluation with ten information visualization experts, all with an academic background. Most of them had previous experiences with graph visualizations but only a minority had been working with business processes in the past.

During design phase we identified that for a sankey diagram there exist multiple variants to split flows with *AND* or *XOR* semantic. As the split-join semantic is a central information for any process model we evaluated two different visualizations for *AND* and *XOR* splits, see Fig. 7.

All our participants of the evaluation identified 7a as *XOR* splits. And most of them identified 7d as *AND* splits. We subsequently used those representations for the *AND* and *XOR* split.

We also compared the usefulness of a visual mapping of a single sensor attribute to the color of the sensor with labels containing the exact attributes value. Hardly surprising the preference for one or the other depended largely on the task we asked the users to perform.



(a) *XOR* option 1 (b) *XOR* option 2

(c) *AND* option 1 (d) *AND* option 2

Fig. 7. Evaluation options for the visualization of a split (Color figure online)

5 Conclusion

We present a web-based system to enable human centered visual analysis of process based sensor readings. The application supports exploratory analysis on the aggregated process level as well as on the single sensor level. It is possible to explore the relation between cause and effect using our data probes. A top down approach using precalculated event clusters enables us to filter on process instances with a similar behavior. We evaluated different visual representations for process splits, providing an indication of understandability of those.

We have shown that it is possible to perform data intensive analysis of process log data within a browser. Using different visualization techniques for two distinct analytical paths. Using a sankey diagram for the process visualization has proven to be possible and understandable by experienced users.

Although we used a manually created process model, it would be possible to perform a process mining step on event data to create the process model needed for our approach.

Acknowledgements. This work was partially funded by the German Federal Ministry of Education and Research (BMBF) in the INDINET project under grant number 01IC10S04I.

References

1. van der Aalst, W.: Cartography and Navigation. In: Process Mining, pp. 321–335. Springer, Heidelberg (2011)
2. van der Aalst, W.: Epilogue. In: Process Mining, pp. 337–340. Springer, Heidelberg (2011)
3. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM Workshops 2011, Part I. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012)
4. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco (1994)
5. Aigner, W., Miksch, S., Schumann, H., Tominski, C.: Visualization of Time-Oriented Data. Human-Computer Interaction Series. Springer, London (2011)
6. Albrecht, B., Effinger, P., Held, M., Kaufmann, M.: An automatic layout algorithm for BPEL processes. In: Proceedings of the 5th International Symposium on Software Visualization, SOFTVIS 2010, pp. 173–182. ACM, New York (2010)
7. Bostock, M.: Sankey Diagram (2012). http://bost.ocks.org/mike/sankey/
8. Gansner, E.R., Koutsofios, E., North, S.C., Vo, K.P.: A technique for drawing directed graphs. IEEE Trans. Softw. Eng. **19**(3), 214–230 (1993)
9. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: interactive exploration and visual analysis of categorical data. IEEE Trans. Vis. Comput. Graph. **12**(4), 558–568 (2006)
10. Lam, H., Russell, D., Tang, D., Munzner, T.: Session viewer: visual exploratory analysis of web session logs. In: 2007 IEEE Symposium on Visual Analytics Science and Technology, VAST 2007, pp. 147–154, October 2007

11. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: the next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, June 2011
12. Riehmann, P., Hanfler, M., Froehlich, B.: Interactive sankey diagrams. In: 2005 IEEE Symposium on Information Visualization, INFOVIS 2005, pp. 233–240 (2005)
13. Rind, A., Wang, T.D., Aigner, W., Miksch, S., Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Interactive information visualization to explore and query electronic health records. Found. Trends Hum. Comput. Interact. **5**(3), 207–298 (2013)
14. Schmidt, M.: Der Einsatz von Sankey-Diagrammen im Stoffstrommanagement. Technical report 124, Hochschule Pforzheim (2006)
15. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: 1996 Proceedings. of IEEE Symposium on Visual Languages, pp. 336–343, September 1996
16. Speray, D., Kennon, S.: Volume probes: interactive data exploration on arbitrary grids. In: Proceedings of the 1990 Workshop on Volume Visualization, VVS 1990, pp. 5–12. ACM, New York (1990)
17. Vergidis, K., Tiwari, A., Majeed, B.: Business process analysis and optimization: beyond reengineering. IEEE Trans. Syst. Man Cybern. C Appl. Rev. **38**(1), 69–82 (2008)
18. Wang, T.D., Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Extracting insights from electronic health records: case studies, a visual analytics process model, and design recommendations. J. Med. Syst. **35**(5), 1135–1152 (2011)
19. Wongsuphasawat, K., Gotz, D.: Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE Trans. Vis. Comput. Graph. **18**(12), 2659–2668 (2012)
20. Wongsuphasawat, K., Guerra Gómez, J.A., Plaisant, C., Wang, T.D., Taieb-Maimon, M., Shneiderman, B.: LifeFlow: visualizing an overview of event sequences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 1747. ACM Press (2011)
21. Wongsuphasawat, K., Lin, J.: Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at Twitter. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 113–122, October 2014