Multidimensional Process Mining: Questions, Requirements, and Limitations

Thomas Vogelgesang[1,3], Georg Kaes[2], Stefanie Rinderle-Ma[2], and H.-Jürgen Appelrath[1]

[1] Department of Computer Science, University of Oldenburg, Germany
{thomas.vogelgesang | appelrath}@uni-oldenburg.de
[2] Faculty Of Computer Science, University of Vienna, Austria
{georg.kaes | stefanie.rinderle-ma}@univie.ac.at
[3] Work conducted during stay at University of Vienna

Abstract. Multidimensional process mining is an emerging approach that adopts the concept of data cubes to analyze processes from multiple views. This enables analysts to split event logs into a set of homogenous sublogs according to the case and event attributes. Each sublog is independently analyzed using process mining techniques resulting in an individual process model for each sublog. These models can be compared to identify group-related differences between the process variants. In this paper, we derive a number of general research questions addressed for multidimensional process mining by a literature review. We analyze the requirements for its application and point out its limitations and challenges. We conduct two case studies applying multidimensional process mining in two different use cases to evaluate our findings.

Keywords: Multidimensional Process Mining, Process Cubes

1 Introduction

Process mining [1] is a set of techniques that allow for the automatic analysis of processes. These techniques are typically based on so-called event logs which are collections of events recorded during process execution. Each event is related to a particular process instance (case) and represents the execution of an activity. Typically, the events of a case are chronologically ordered forming the trace of the case. Event logs may also contain arbitrary attributes describing the properties of cases and events. Process discovery is a subset of process mining techniques that analyzes the event log and automatically creates a descriptive process model from it. Other kinds of process mining are conformance checking which measures how good the process model reflects the event log and process enhancement which adds new perspectives (e.g., waiting times) to the process model.

Usually, process mining techniques only consider one event log resulting in a single process model. However, analysts are often interested in particular parts of the process or particular cases (e.g., not meeting a given quality requirement).

In: S. España, M. Ivanović, M. Savić (eds.): Proceedings of the CAiSE'16 Forum at the 28th International Conference on Advanced Information Systems Engineering, Ljubljana, Slovenia, 13-17.6.2016, published at http://ceur-ws.org

Although it is possible to filter the event log to derive the desired view on the data, this is kind of laborious and time-consuming if done manually, especially if multiple process models should be compared. The notion of multidimensional process mining (MPM) is an emerging approach that tackles this problem by adopting the concept of data cubes from the data warehouse domain.

The basic idea of MPM is to interpret the attributes from the event log as dimensions which form the edges of a multidimensional data cube. Each unique combination of dimension values refers to a cell of the cube that typically contains event data (sublog). Using OLAP (Online analytical processing) [7] operators like roll-up and drill-down, it is possible to change the granularity of the data cube. Slice and dice operators can be used to filter the data in order to restrict it to a subset. Each cell of the resulting data cube is independently mined by process discovery techniques, resulting in an individual process model for each cell. Finally, these models can be compared to identify differences between different cells (e.g., representing groups of patients, customers or products).

The objective of this paper is to analyze, for which kind of non-domain-specific questions MPM is suited for. Based on a literature review, we specify a classification of dimensions that characterizes a dimension with respect to the underlying process. Furthermore, we analyze the general requirements that have to be considered to successfully apply MPM. Additionally, we identify limitations and challenges of MPM in general that are not specific to a particular approach and which should be tackled by future research. We validate our findings in two case studies applying MPM in different scenarios.

2 Analysis of existing approaches

We reviewed of MPM-related literature to identify typical questions, dimension classes, data requirements, limitations, and challenges. Table 1 summarizes its results. Aspects that are considered in or that are relevant for the publication are marked with ✓, while irrelevant or unconsidered aspects are marked with ·. We discuss the results in the following subsections in more detail.

We have to point out that this literature analysis underlies a number of limitations. First, our analysis is based on a relatively low number of publications because MPM is an emerging part of process mining research and has not attracted that much attention so far. However, the discussed publications clearly show the upcoming demand for MPM in the process mining community. Second, the reviewed publications mainly focus on conceptual work mostly giving only little information about the questions driving their case studies, the available dimensions and the identified limitations.

2.1 MPM-typical Questions

In contrast to traditional process mining, MPM does not discover a process model from a single event log, but multiple models from a set of sublogs. Therefore, the main question of MPM is not only (1) *how does the process look like?*, but also

	Vogelgesang et al. [15,14]	van der Aalst [2]	Mamaliga [10]	Bolt, van der Aalst [5]	van der Aalst [3]	Ribeiro, Weijters [13]	Ribeiro [12]	van Eck et al. [6]	Mans et al. [11]	van der Aalst et al. [4]	Case study 1 (Sec. 3.1)	Case study 2 (Sec. 3.2)
MPM-related questions												
Question (1)	✓	✓	✓	✓	✓	✓	·	✓	·	·	✓	✓
Question (2)	✓	✓	✓	✓	✓	✓	·	✓	·	·	✓	✓
Question (3)	✓	·	✓	✓	✓	✓	·	✓	·	·	✓	✓
Question (4)	✓	✓	✓	✓	✓	·	·	·	·	·	✓	✓
Dimension classes												
Customer-related	✓	✓	✓	✓	✓	✓	✓	✓	✓	·	✓	✓
Product-related	·	·	✓	·	·	✓	✓	✓	✓	·	✓	·
Provider-related	✓	✓	·	✓	✓	·	·	·	✓	·	✓	✓
Execution-related	✓	·	✓	✓	·	·	✓	✓	·	✓	✓	✓
Quality-related	✓	·	✓	·	✓	✓	·	✓	✓	✓	✓	✓
Time-related	✓	✓	·	✓	✓	✓	✓	✓	·	✓	✓	✓
Data requirements												
Integrate all attributes	✓	·	·	·	·	·	✓	✓	·	·	✓	·
Representativeness	✓	·	✓	·	·	·	·	✓	✓	✓	✓	✓
Limitations and challenges												
Comparison of cells	✓	✓	✓	✓	·	·	✓	✓	✓	·	✓	✓
High effort for data integration	✓	·	·	·	·	·	·	·	·	·	✓	·
Performance optimizations	✓	✓	✓	·	·	·	·	✓	·	·	·	✓
Interactivity	✓	·	·	·	·	·	·	·	·	·	✓	✓
Handling of concept drifts	·	✓	·	✓	·	·	·	·	·	·	·	·

Table 1. Overview of results (✓: considered/relevant / ·: not considered/irrelevant)

(2) *what are the differences between several processes or process variants?* and (3) *how are these differences determined by the dimensions representing instance or event related features?* Another typical question for MPM is (4) *How does the process evolve over time?* Of course, the specific questions are domain-related and depend on the process. However, there is a number of dimensions that are supposed to be of a strong interest throughout several domains. Based on the literature review, we identified six classes of dimensions.

Customer-related dimensions describe the features of the customer and are typically used to group similar customers in MPM. Typical examples are age and gender.

Product-related dimensions describe features of the product (or some other subject that can be considered as a "product" in the widest sense of meaning) related to the process. Typical examples are product types and variants.

Provider-related dimensions characterize the environment of process execution, namely the executing organization. Typical examples are the executing department, the location of a branch office, or the executing organization itself if multiple organizations are compared to each other.

Execution-related dimensions describe features related to the execution of a process instance. In contrast to the previous dimensions, they do not char-

acterize external entities but the process execution itself. Therefore, these features are determined during the execution of the process. Examples are the urgency of treatment, the type of surgical intervention, and the diagnose.

Quality-related dimensions This kind of dimension characterizes the quality of the process result. In contrast to execution-related dimensions, their values are not captured throughout the execution of the process, but at the end of the execution or even shortly before it if the result is already foreseeable. Examples are the outcome of the treatment and the reason for discharge.

Time-related dimensions describe the temporal classification of cases and events. Even though there is only one notion of time, it may be expressed by multiple dimensions (e.g., start or end time of the process execution).

Time-related dimensions are special as they may be used to define chronological orders of process instances. This is necessary to answer the question of process evolution (4) which is analyzed by creating and mining sublogs according to chronologically ordered time intervals, e.g. by the year of the process execution. Comparing these process "snapshots" can reveal changes of the process over time indicating concept drifts.

2.2 Data Requirements

Generally, the more dimensions are available in the data cube, the more multifaceted questions arise during the analysis. Therefore, **all available attributes should be integrated** into the data cube to avoid restricting the possible analysis space of the OLAP queries, even though the contribution of a dimension to the analysis might not be clear in advance. This enables the analyst to define unforeseen ad-hoc queries.

Another important data requirement in the context of MPM is the **representativeness** of data. To discover a meaningful process model, the event log must contain a representative number of cases. MPM partitions the event data into a set of multiple sublogs, each containing only a fraction of the original data. Therefore, MPM usually requires a multiple number of cases and events to ensure representativeness. Additionally, the data should be to some extent equally spread across the multidimensional space. Otherwise, the data distribution limits the number of meaningful models. As the distribution of data is given by the data attributes, it is not possible to enforce representativeness for all combinations of dimensions. Therefore it is mandatory, that the analysts carefully check the number of cases and events to avoid misinterpretations. However, we recommend to consider this problem during data integration by defining categorial values for the dimensions. Non-categorial values should be mapped to meaningful classes of values to avoid sparsity and improve the representativeness. For categorial attributes with many values, we also recommend to define a classification hierarchy (e.g., introducing additional 5-year and 10-year age classes) to allow for aggregation which typically improves the representativeness of cells.

2.3 Limitations of MPM

The **comparison of cells** seems to be the major challenge of MPM. Even with advanced concepts like difference views and process model consolidation [15] it is still difficult to handle dozens of models during analysis. A further limitation is the **high effort for data integration** (e.g., for defining the dimension hierarchies) which should be tackled by a better tool support. Other challenges are outstanding **performance optimizations** (e.g. due to materialization concepts), a lack of **interactivity** and the missing **handling of concept drifts**.

3 Evaluation

We conducted two case studies to evaluate the findings presented in Section 2.

3.1 Higher Education Processes (HEP)

The HEP data contains events related to university courses in Computer Science collected based on a service-oriented teaching platform [9]. It contains 354 cases (students participating in a particular course) and 28,129 events. The processes start with a kick-off meeting and typically end with a final marking. Between these endpoints, the students iteratively work on exercises (e.g., uploading solutions) while the lecturers give feedback. Throughout the whole process, students and lecturers can create forum posts e.g., to discuss the exercises. We integrated the data into the database of PMCube Explorer [14], defining 12 dimensions (e.g., course, lecturer, student, semester) and 12 non-dimensional attributes (e.g., number of compiler errors, points scored for a particular exercise).

The case study revealed changes in the structure of a course held in two consecutive years. For example, the number of exercises was reduced while more feedback was given by the lecturers (cf. Fig. 1). Comparing the process variants of a particular course for the different final marks also showed interesting similarities and differences. E.g., the first steps of the process were exactly the same for all variants, starting to diverge with the evaluation activity for the first project phase. The process variants also clearly show that students with worse marks tend to struggle earlier during the course and that students who failed the course gave up working on the final exercises. This is also reflected by the average number of events per case, clearly showing that good students are significantly more active than students who failed.

In this case study, all questions (1) to (4) were addressed (see Table 1). The considered dimensions cover all dimension classes, also the product-related dimensions (e.g. course dimension, considering the knowledge transferred to the students to be the universities "product"). Moreover, the case study confirmed that all available attributes should be considered to build up the data cube. E.g., we used the number of compiler errors and warnings recorded with the compile events in several ad-hoc queries. These attributes were integrated into the data cube, though they were initially considered to be useless for the analysis.

The case study also confirmed that the representativeness is mandatory for MPM. We were not able to pose several queries due to missing data resulting in sparsity. Furthermore, the case study showed that MPM in general still has to face a number of limitations and challenges. The main limitation was the difficulty to compare the processes, even though we applied advanced concepts like the automatic creation of difference models. The integration of data also took a significant higher effort (e.g. for defining dimensions) compared to the creation of a normal event log. This supports the identified limitation and emphasizes the need for a better tool support for integrating the data into the cube structure. Finally, the lack of interactivity was confirmed as filtering and similar operations cannot be triggered directly from the process model, which makes it sometimes quite cumbersome to change the view on the process. However, performance issues could not be confirmed during this case study due to the relatively small set of data. The challenge of concept drifts was not confirmed as well, because changes of the process during run-time were not expected (all process instances were performed in parallel during the semester).

3.2 Multidimensional Change Mining

This case study analyzes the data from [15] in a novel manner: we combine change mining [8] with MPM which has not been considered so far. Due to the lack of adequate change logs, we apply change mining to an execution log interpreting the variations in the process as ad-hoc changes.

We analyzed process logs from a department in emergency cases. The change analysis was well suited to examine exceptional situations which should not occur in an usual setting. For example in some cases, the termination of anesthesia was initiated before the clearance of the surgeon. We also found some process instances where additional radiological examinations occurred after the surgery – this may be a hint that something went wrong and additional tests have been required. In some cases, we found admission events after the discharge. This can happen when data acquired at the patients admission had to be corrected or complemented, e.g. due to missing health insurance information. Finally, we found some cases containing steps which hinted at another upcoming surgery.

Change mining provided interesting results, especially for detecting exceptional situations. Typically, the change trees grow in width very fast, which can be partially avoided by filtering non-relevant events which e.g. only exist to add missing data to a patient's profile. Also the aggregation of self loops of certain events reduced the trees' depth and improved the analysis (cf. Fig. 2).

During the case study, we considered all general MPM-related questions (see Table 1). Aspects related to data integration were not considered, because we reused an existing data cube. For the same reason, the available dimension classes were similar to [15]. The representativeness turned out to be very challenging, because some of the OLAP queries resulted in multiple cells each containing only a few cases. Therefore, the results cannot be generalized. The comparison of results turned out to be the major limitation for the application of change mining in MPM, because there are currently no supporting concepts (e.g., difference

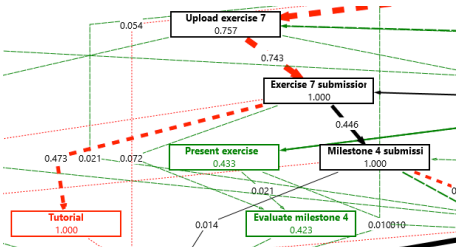

Fig. 1. Comparing two semesters of the same university course (excerpt)

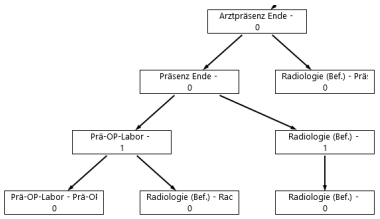

Fig. 2. Change Tree showing exceptional situations in the medical domain (excerpt)

views) available for change trees. Due to a low bandwidth connection to the database, we also had to face performance issues. However, we were able to achieve acceptable loading times by omitting all unneeded attributes. Beyond the analysis of process evolution by comparing process variants for different months, the challenge of concept drift was not considered.

4 Discussion and Recommendations

The results of the case studies generally confirm the findings of the literature analysis. Note that the issue of concept drift (beyond the general question of process evolution) was not relevant in both case studies. Due to the little number of publications, more case studies are required to demonstrate the potential of MPM and identify its limitations, providing a clear direction for future research.

Nonetheless, we were able to identify requirements, that need to be met for the successful application of MPM. An extensive data base is crucial for MPM. On the one hand, MPM requires a multiple number of cases and events compared to traditional process mining. On the other hand, it requires many (preferably all available) additional attributes that can be exploited to aggregate and filter the data. To ensure representativeness of results, the different dimension values should – at least to some extent – be equally distributed across the cases and events. Dimension values should not be related to only a few cases and events, which is a stricter requirement than only avoiding sparsity.

The biggest challenge remains the comparison of cells, even though there are initial approaches to tackle this problem. Another major challenge is the performance, because long processing times disrupt the explorative process analysis. The problem of missing interactivity of the process mining results (in the sense of direct interactions with the process models) has not been considered, so far. However, we believe that *interactive process mining* will draw more attention in the future, also in the context of traditional process mining. Minor limitations are the high initial effort for data integration and the handling of concept drifts.

Recommendations: We recommend to use MPM if the analysis aims to compare processes or process variants or a multitude of filtering is required (even without comparing different models). It may also contribute to the analysis

of process evolution. Furthermore, we recommend to use MPM if the process-specific research questions are not clear in advance and should be refined during the analysis. On the contrary, there are also situations that are not suitable for MPM. We recommend to use traditional process mining approaches if only a very limited number of cases and events or only a very few attributes are available in the event data. Traditional process mining can also be the better choice if the additional effort for creating the dimensions etc. exceeds its utility. This may be the case if only a few and simple filtering of the event log are expected or if only a (fast) overview of the overall process is needed.

References

1. van der Aalst, W.M.P.: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
2. van der Aalst, W.M.P.: Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. In: Song, M., et al. (eds.) Asia Pacific Business Process Management. LNBIP, vol. 159, pp. 1–22. Springer (2013)
3. van der Aalst, W.M.P.: How People Really (Like To) Work - Comparative Process Mining to Unravel Human Behavior. In: Human-Centered Software Engineering. pp. 317–321 (2014)
4. van der Aalst, W.M.P., et al.: Comparative process mining in education: An approach based on process cubes. In: Ceravolo, P., et al. (eds.) Data-Driven Process Discovery and Analysis. LNBIP, vol. 203, pp. 110–134. Springer (2013)
5. Bolt, A., van der Aalst, W.M.P.: Multidimensional Process Mining Using Process Cubes. In: Gaaloul, K., et al. (eds.) Enterprise, Business-Process and Information Systems Modeling. LNBIP, vol. 214, pp. 102–116. Springer (2015)
6. van Eck, M.L., et al.: PM^2 : A Process Mining Project Methodology. In: Zdravkovic, J., Kirikova, M., Johannesson, P. (eds.) Advanced Information Systems Engineering. LNCS, vol. 9097, pp. 297–313. Springer (2015)
7. Golfarelli, M., Rizzi, S.: Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, Inc., New York, NY, USA, 1 edn. (2009)
8. Kaes, G., Rinderle-Ma, S.: Mining and querying process change information based on change trees. In: Int'l Conf. on Service-Oriented Computing (2015)
9. Ly, L.T., Indiono, C., Mangler, J., Rinderle-Ma, S.: Data transformation and semantic log purging for process mining. In: Advanced Information Systems Engineering - 24th International Conference. pp. 238–253 (2012)
10. Mamaliga, T.: Realizing a Process Cube Allowing for the Comparison of Event Data. Master's thesis, TU Eindhoven (Aug 2013), letzter Aufruf: 21.08.2015
11. Mans, R., et al.: Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes. Springer Briefs in BPM, Springer (2015)
12. Ribeiro, J.T.S.: Multidimensional Process Discovery. Ph.D. thesis, TU Eindhoven (2013)
13. Ribeiro, J.T.S., Weijters, A.J.M.M.: Event Cube: Another Perspective on Business Processes. In: Robert Meersman et al. (ed.) On the Move to Meaningful Internet Systems: OTM 2011. LNCS, vol. 7044, pp. 274–283. Springer (2011)
14. Vogelgesang, T., et al.: Multidimensional process mining with pmcube explorer. In: Proceedings of the BPM Demo Session 2015. pp. 90–94. ceur-ws.org (2015)
15. Vogelgesang, T., et al.: PMCube: A Data-Warehouse-based Approach for Multidimensional Process Mining. In: Business Process Management Workshops (2015)