A model based slicing technique for process mining healthcare information *

Fazle Rabbi[1,2], Yngve Lamo[2], and Wendy MacCaull[3]

[1]Department of Information Science and Media Studies, University of Bergen, Norway
[2]Department of Computer Technology, Electrical Engineering and Science, Western Norway University of Applied Sciences, Norway
[3]Department of Computer Science, St. Francis Xavier University, Nova Scotia, Canada
Fazle.Rabbi@uib.no, Yngve.Lamo@hvl.no, wmaccaul@stfx.ca

Abstract. Process mining is a powerful technique which uses an organization's event data to extract and analyse process flow information and develop useful process models. However, it is difficult to apply process mining techniques to healthcare information due to factors relating to the complexity inherent in the healthcare domain and associated information systems. There are also challenges in understanding and meaningfully presenting results of process mining and problems relating to technical issues among the users. We propose a model based slicing approach based on dimensional modeling and ontological hierarchies that can be used to raise the level of abstraction during process mining, thereby more effectively dealing with the complexity and other issues. We also present a structural property of the proposed slicing technique for process mining.

Keywords: healthcare systems, ontology, process mining, slicing, abstraction

1 Introduction

Today's vast amount of healthcare related information needs to be accessed easily and integrated intelligently to support better healthcare delivery. Systematic analysis of healthcare data can help to detect patterns so that healthcare providers can optimize their resource allocation and clinicians can optimize treatment plans for individuals leading to better health outcomes. To improve the quality of health services delivery, healthcare professionals are particularly interested to know what are the common pathways for patients, how can a process model be improved, and to what extent do existing systems follow clinical guidelines. However it is not easy to answer these questions as healthcare processes are highly dynamic, complex, sometimes ad-hoc, increasingly multidisciplinary [15], and involve numerous points of care and a variety of clinicians and treatment plans.

Data analysis techniques such as process mining present the opportunity to analyse healthcare information, from numerous viewpoints (i.e., contexts)

* Partially funded by Intromat. (www.intromat.no)

such patient populations with specific diseases, ages, gender, incidences of comorbidity, or type of healthcare service setting (e.g., clinic, hospital, nursing home), home location (urban or rural), or procedures used, etc., in order to learn from this information. Process mining [4] techniques hold great potential to improve health services delivery.

van der Aalst presented four different analysis perspectives in [4] which include control-flow perspective, organizational perspective, case perspective, and time perspective. These perspectives are useful to derive useful insight with respect to ordering of activities, the roles of resources, the attributes related to a particular case, and the frequency and timing of events. However, this approach lacks an abstraction mechanism allowing health professionals to both mine relevant information from highly discipline specific data sources and also to process event data from often highly individualistic patient pathways in order to discover common pathways.

In this paper, we address some of the challenges of process mining and learning from the large amount and variety of healthcare information. We present a model based slicing technique for process mining which utilizes dimensional modeling and ontological representations of healthcare information. Ontologies are increasingly being used to standardize terminologies in healthcare and other areas. For example, the ICD-10 ontology [16] provides diagnostic codes for classifying diseases, including a wide variety of signs, symptoms, abnormal findings, etc., while SNOMED-CT [3] provides a comprehensive terminology for clinical health. Our proposed slicing technique proposes that we use graphical representations incorporating an easy-to-use graphical interface allowing health professionals to apply process mining on a variety of abstraction levels.

de Medeiros and van der Aalst, [11], discussed the necessity of relating elements in event logs with their semantic concepts, allowing them to perform concept-based reasoning and analysis. They illustrated their ideas with an example process model to repair telephones in a company which included three different ontologies and implemented their technique in ProM. Our approach is similar; however, we illustrate the value and potential of using a graphical query language for specifying slicing requirements of process mining. The graphical query language uses dimensional modelling and ontology which makes it suitable for applying the slicing technique in the very complex healthcare domain.

Mans et al., [10], discussed the application of process mining in healthcare and provided an overview of frequently asked questions by medical professionals in process mining projects. The questions reflect the medical professionals' interest both in learning common pathways of different patient groups, to determine their compliance with internal and external clinical guidelines, and also in gathering information about the throughput times for treating patients. The authors pointed out the need for accumulating data from different data sources claiming this to be a major challenge in healthcare. and urging the exploitation of ontology-based process mining approaches in the healthcare domain.

Bistarelli et al. presented a prototype tool called PrOnto in [6] which can discover business processes from event logs and classify them with respect to a business ontology. The tool takes an event log file as input and produces an UML based activity diagram in XML format. The aim is to raise the level of abstraction in process mining by utilizing business ontologies. They proposed an ontology representing the hierarchy of resources to define which level of abstraction will be used [6]. Here we use both dimensional models and ontologies to classify event logs allowing us to be more specific in one portion of the process model while being more generic in another portion of it. Our mining technique uses several pre-processing steps, allowing us to specify both context and level of abstraction.

The rest of this paper is organized as follows. Section 2 discusses the characteristics of and challenges of process mining in the healthcare domain. Section 3 presents data abstraction methods involving dimensional modelling, ontological hierarchies and graphical representations resulting in novel approaches to process mining. We study the structural property of our slicing technique in section 4. Section 5 concludes the paper and gives directions for our future work.

2 Characteristics of and challenges in careflow analysis

In this section, we briefly discuss several characteristics of healthcare data which make process mining so difficult. We argue that the current practice of process mining needs to be advanced by means of a rich information model in order to accommodate the requirements of the many and various stakeholders.

In many developed countries, the majority of the citizens use public healthcare services provided by a variety of service providers using a large number of software applications. For instance, Helse Vest IKT (HV-IKT) [2] an IT company in Norway that supplies equipment and services within the ICT area to specialist health services in the western part of Norway, has more than 1000 software systems to support the regional healthcare. As a result, healthcare data are often in silos and different standards are followed by different health facilities to code diagnosis, lab test results, medical procedures and drugs. This presents a major problem as patients frequently need to visit various health facilities so for effective process mining, event logs representing activities of various systems must be considered. In healthcare, the data preparation task is very critical as healthcare data are very sensitive and therefore semantics of the data must be preserved [14]. Lack of standards for data definitions and reporting mechanisms in health data across various disciplines makes it very difficult to analyse the large spectrum of health profiles. Event logs from a variety of systems using various data definitions and formats must therefore be harmonized before they can be analysed by any process mining algorithm.

Identifying common pathways for patient flow in healthcare systems is complicated by the large variety of patient conditions, points of care and diagnoses. Consider Table 1 which shows a portion of such an event log. The resulting process model is usually too large to provide meaningful information. Even after filtering to keep, for example, only the patient cases which are admitted to the radiology department, the model is still too large. By pre-processing event logs so that we exploit the hierarchical information structure in healthcare such as organizational structure of hospitals and clinics, or ontologies of healthcare terminologies, we can support various levels of abstraction and reduce the size of the process model. We need a mechanism to specify how to exploit ontological hierarchies to change abstraction levels.

Table 1. Portion of a healthcare event log (Sample data)

Id	Event time	Event name	Resource
1	2017-03-20 13:30	Surgical Clinic	Kristi Salazar
1	2017-03-20 13:30	(N39_9) Disorder of urinary system	Darla Ramirez
2	2017-03-07 14:00	Radiology Department	Ricky Alvarado
2	2017-03-07 15:00	(N63) Unspecified lump in breast	Deborah Tyler
2	2017-03-07 15:15	(N64_5) Other symptoms in breast	Johanna Buchanan
3	2017-04-06 08:30	Division of Mental Health Protection	Henrietta King
3	2017-04-06 08:30	(F321) Depressive episode	Beatrice French

While analyzing common pathways for patients, different contexts are required to allow clinicians to focus on different groups of patients and to visualize their careflows. For instance, for patients with mental disorders, psychologists may be interested in patient visits to various service points while researchers may be interested in investigating the efficiency of new medications or procedures. The psychologists need to select a patient group based on diagnosis but display information about patient visits to different service points, while researchers require the health assessment related information. We need efficient tool support where one can filter the data and specify the context for data visualization. Existing process mining techniques may be used to visualize flow but event logs must be prepared in different ways to support a wider variety of queries. Existing process mining tools such as Fluxicon Disco [1] support filtering over event logs; however, since the events are not categorized, they do not support grouping of similar events from an abstraction level (e.g., disease group from ICD-10) to extract an abstract process model over another type of events (e.g., visit to health service points).

3 Model based pre-processing step for process mining

In this section we propose a model-based slicing technique for handling the complexity of process mining in the healthcare domain. The proposed slicing technique is centered on the use of dimensional modeling and ontologies and it consists of pre-processing steps for process mining. The slicing technique includes filtering and grouping of activities where we use ontological perspectives to identify semantically related cases and provide a higher level of abstraction in the data, which at the same time gives meaningful information to healthcare professionals. Usually these professionals focus on a specific domain (their area of specialization) but often they are interested in also getting an abstract view of aspects of their patients' use of services from other disciplines. An interdisciplinary view of process models along with zoom-in features are essential to allow healthcare professionals to understand the flow of patients and/or their symptoms and diagnoses. We address this issue by incorporating dimensional models where we use ontologies along various dimensions. We allow healthcare professionals to specify on which area they wish to focus and for which area they want an abstract view. This requirement of mixing abstraction levels has not been addressed in the state-of-art process mining literature. In the following subsections we present a slicing technique for process mining by means of a graphical language based on ontologies and dimensional modeling.

3.1 Role of ontologies in process mining

Ontologies can play two roles in process mining. First, they provide vocabularies to integrate data from multiple healthcare information systems. We propose using the SNOMED-CT and ICD-10 ontologies to group activities in event logs across disciplines enabling process mining over multiple data sources. We also propose using hospital- or regional health authority-specific ontologies to standardize information of an organizational or administrative nature. Second, ontologies provide suitable levels of abstraction for (i) specifying filters to select particular patient groups and for (ii) grouping information to capture and/or visualize the careflow process at a high (and more useful) level of abstraction.

3.2 Dimensional modeling for contextual analysis of careflow

The concept of dimensional modeling originated from data warehousing and business intelligence (DW/BI). The DW/BI systems emphasize collecting and processing raw data quickly and turning them into useful information while preserving the consistency of the data. Dimensional models package the data in a simple format so that information may be displayed in a manner easily understood by business users, and support efficient data analytic tools in terms of query performance. We present a new application area of dimensional modeling, showing that it can be used for specifying context-related requirements for process mining.

Figure 1 shows how a dimensional model can be used in a healthcare setting to capture process mining requirements. The purpose of this dimensional model is to provide a visualization allowing the user to investigate care flow from different contexts. Typically dimensional models are used to represent detailed atomic information captured by a business process. However, we propose to use ontological hierarchies to provide hierarchical representation of healthcare information along each dimension of a dimensional model for healthcare.

Suppose we have a dimensional model for Healthcare as indicated in Figure 1 with the dimensions: Episode of Care, Clinical Finding, Procedure, Observation

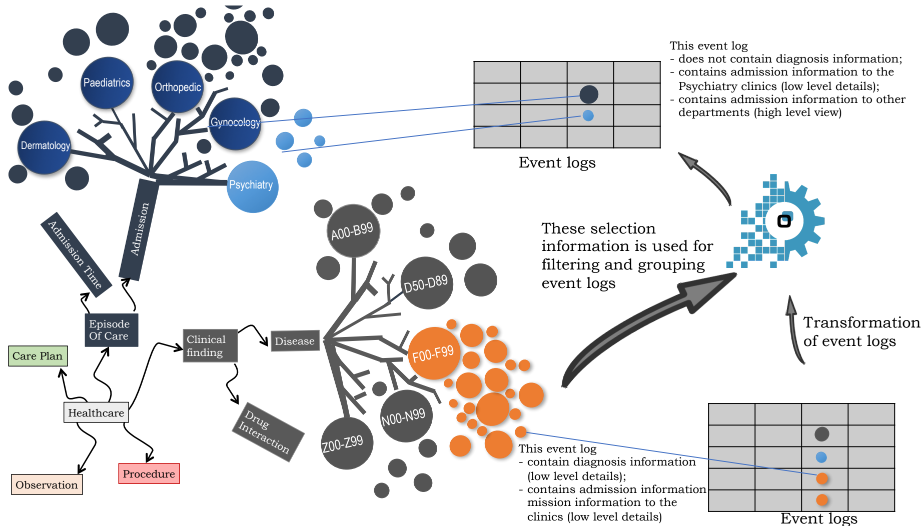

Fig. 1. Use of dimensional model for specifying process mining requirements

and Care Plan. Figure 1 shows a fragment of the ICD-10 ontology representing several diseases linked to the Clinical Finding dimension. We now show how dimensional modeling can be used to support analysis of healthcare processes, by supporting both filtering and selecting the level of abstraction (i.e., grouping) for visualizing the process mining output.

Suppose an analyst is interested in investigating the flow of patients with mental and behavioral disorders admitted to the various departments in a hospital. The analyst is interested in the movement of the patient in psychiatry clinics but the specifics of other clinics in not needed, only the information at the department level. We will assume that the department hierarchy of the hospital is used for the 'Episode of Care - Admission' dimension. We illustrate the situation in Figure 1 to help visualize how the dimensional model and the hierarchical representation of data can be utilized to obtain the information. In Figure 1, 'F00-F99' is the ICD-10 code for 'mental and behavioral disorder' diseases. Selecting 'F00-F99' for filtering means to filter based on the sub-diseases under 'F00-F99' (depicted as small orange circles in the figure), extracting event logs for patients with one of the sub-disease codes of 'F00-F99'. From these event logs we now extract events related to admission. An event log with patient's admission information contains information about the patient's visits to different clinics. We are interested in clinic information only for psychiatry but we need to display patient admission to other departments. We use the department hierarchical information (Ontology) to group the results displaying a higher level of abstraction for departments (i.e., department names, e.g., Gynocology,

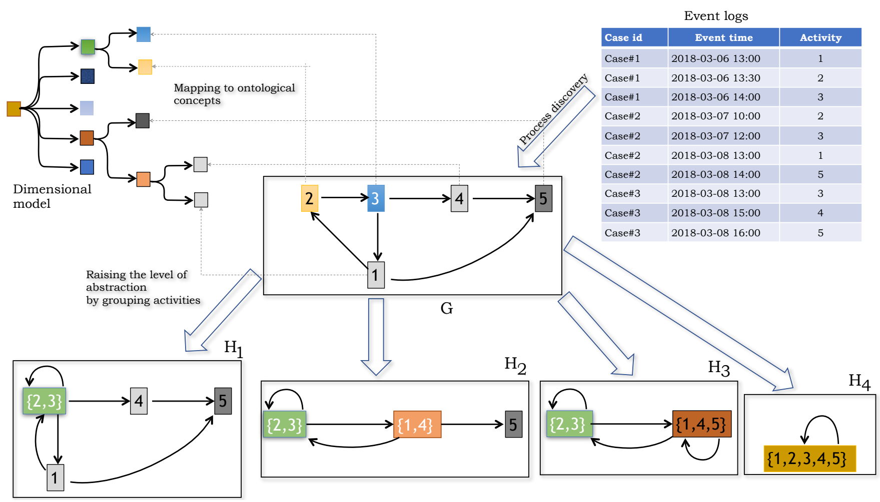

Fig. 2. A process model and its different abstractions

Orthopedics, etc.) displaying only department name rather than clinic name for visits to clinics other than psychiatry clinics.

4 Structural property

There are several process mining tools which produce process models of various kinds. For example, ProM [7], an open-source process mining tool, produces Petri nets [12] as output to describe the process models extracted from event logs. Disco by Fluxicon [1], a commercial process mining tool, produces attributed graphs to present discovered process models. RapidProM [5] consists of state-of-the-art process mining algorithms and produces process models with various formalisms such as BPMN Models [13], Petri nets [12], colored Petri nets [8], process trees [9], etc. Since these process models are (essentially) graph-based, we illustrate the abstraction in graph-based process model and discuss a structural property which is needed to ensure that abstract process model correctly reflects the abstraction in the ontology.

Suppose we apply process mining technique over a set of traces T representing event logs of a system and obtain a directed graph G representing the discovered process model (see Figure 2). We refer to this graph G as the base process model. The use of a dimensional model and ontological hierarchies allows us to group some of the nodes from the base process model. On the upper left of Figure 2 we see a dimensional model with five dimensions, two of which have associated ontological hierarchies. The ontology-based slicing approach allows us to group the activities according to the ontological hierarchies. While grouping activities based on the ontological hierarchies, one can choose the level of abstraction and this choice allows us to transform the original event logs into an abstract version. The transformed event log (which is now at a more abstract level) is used for discovering a more abstract process model. The remainder of Figure 2 illustrates the effects of some groupings of activities on the base process model. The colours of the five tasks in G reflect the position the associated task in the dimensional model. These colours show that activities 2 and 3 belong to the ontological hierarchy associated to the top dimension. Using this higher level of abstraction we group activities 2 and 3, producing a single node representing the set of activities $\{2, 3\}$. See H_1 at the lower left hand side. It is possible to also group nodes based on their ontological hierarchy from other dimensions as shown in process models H_2, H_3 and H_4. (Two levels of the second ontological hierarchy are used in succession to get H_2 and H_3.) These groupings of activities provide even more compressed process models. It is desirable to obtain an abstract process model H such that there exists a graph homomorphism between the base process model G and H. The graph homomorphism ensures that the abstract process model reflects the abstraction in the ontology.

Theorem 1 (Correctness of abstraction). *Given a dimensional model and a set of traces T representing event logs, if the mapping of activities to the ontological concepts relates each element from its domain to a maximum of one leaf node of the dimensional model then combining the activities based on the dimensional model compresses the base process model G to a graph H such that there exists a graph homomorphism between G and H.*

Proof hint: The theorem can be proved by showing an equivalence class relationship from the nodes and edges of the process model from G to the abstract process model H following the level of abstraction in the ontological models.

5 Conclusion

To analyze the vast amount of healthcare information, better visualization techniques are needed to get an abstract and transparent view of what processes have been executed. In this paper we propose a model based approach based on dimensional modelling and associated ontological hierarchies that will allow analysts to specify process mining requirements such as the context and abstraction level. The idea of using a combination of dimensional modeling with ontologies is novel in this paper. We envision a healthcare information system that provides access to information from various healthcare providers. We are currently developing a careflow analysis tool which offers a diagrammatic language for slicing event logs.

This diagrammatic representation of the dimensional model gives an easy-to-use interface allowing the user to collect process mining requirements. After the needed preprocessing, the tool uses an existing process mining tool called *Disco* [1] to show the results of the process mining. The tool provides a visualization to aid healthcare workers in specifying their process mining requirements. In future an expert user group from the Helse Vest IKT will be involved in evaluating the tool.

References

1. Fluxicon Disco, https://fluxicon.com/disco/
2. Helse Vest IKT, https://helse-vest-ikt.no/seksjon-engelsk
3. SNOMED CT, www.snomed.org/snomed-ct
4. van der Aalst, W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer Publishing Company, Incorporated, 1st edn. (2011)
5. van der Aalst, W.M.P., Bolt, A., van Zelst, S.J.: Rapidprom: Mine your processes and not just your data. CoRR **abs/1703.03740** (2017), http://arxiv.org/abs/1703.03740
6. Bistarelli, S., Noia, T.D., Mongiello, M., Nocera, F.: Pronto: an ontology driven business process mining tool. Procedia Computer Science **112**, 306 – 315 (2017). https://doi.org/https://doi.org/10.1016/j.procs.2017.08.002, http://www.sciencedirect.com/science/article/pii/S1877050917313418
7. van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: The prom framework: A new era in process mining tool support. In: Ciardo, G., Darondeau, P. (eds.) Applications and Theory of Petri Nets 2005. pp. 444–454. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
8. Jensen, K., Kristensen, L.M.: Coloured Petri Nets: Modelling and Validation of Concurrent Systems. Springer Publishing Company, Incorporated, 1st edn. (2009)
9. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Scalable process discovery with guarantees. In: Gaaloul, K., Schmidt, R., Nurcan, S., Guerreiro, S., Ma, Q. (eds.) Enterprise, Business-Process and Information Systems Modeling. pp. 85–101. Springer International Publishing, Cham (2015)
10. Mans, R.S., van der Aalst, W.M.P., Vanwersch, R.J.B., Moleman, A.J.: Process mining in healthcare: Data challenges when answering frequently posed questions. In: Lenz, R., Miksch, S., Peleg, M., Reichert, M., Riaño, D., ten Teije, A. (eds.) Process Support and Knowledge Representation in Health Care. pp. 140–153. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
11. de Medeiros, A.K.A., van der Aalst, W.M.P.: Process mining towards semantics. In: Advances in Web Semantics I, Lecture Notes in Computer Science, vol. 4891, pp. 35–80. Springer (2009)
12. Murata, T.: Petri nets: Properties, analysis and applications. Proceedings of the IEEE **77**(4), 541–580 (Apr 1989). https://doi.org/10.1109/5.24143
13. OMG: Business Process Model and Notation (BPMN) Version 2.0 (January 2011), http://www.omg.org/spec/BPMN/2.0/
14. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edn. (1999)
15. Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Process mining in healthcare: A literature review. Journal of Biomedical Informatics **61**, 224 – 236 (2016). https://doi.org/https://doi.org/10.1016/j.jbi.2016.04.007, http://www.sciencedirect.com/science/article/pii/S1532046416300296
16. WHO: International classification of diseases (version 10), icd-10, www.who.int/classifications/icd/en/