Data Improvement to Enable Process Mining on Integrated Non-log Data Sources

Reinhold Dunkl

University of Vienna, Austria
Faculty of Computer Science
reinhold.dunkl@univie.ac.at

Abstract. Process models derived using Process Mining (PM) are often very complex due to Data Quality Issues (DQIs). Some of those DQIs arise from integration of different data sources or the transformation of non-process oriented data, hence are structural and can be abstracted from the domain. Activity Sequencing and Activity Hierarchy are two concepts for improving certain DQIs in order to improve PM outcomes. The approaches are evaluated by showing the improvement of derived process models using a simplified real world scenario with simulated data.

Keywords: Data Enrichment, Data Quality Improvement, Data Integration, Process Mining.

1 Introduction

Process Mining (PM) – or more specific Process Discovery – aims at analyzing data in order to derive process models [1]. ProM is a PM framework that offers diverse mining algorithms to discover such models, e.g: apriori, heuristic or genetic algorithms. For any process discovery activity ProM expects as input process oriented data (event data) in a log file format. Normally the execution log files of information systems are used to derive the underlying process model that led to these log files. We will use the ProM framework and the heuristic miner to show the virtues of our approaches to improve the data quality and therefore the mined process models.

Data Quality issues (DQIs) – related to process oriented data – are manifold and arise from diverse real world situations like integration of diverse data sources or preparatory data transformations to generate process oriented data out of diverse structured data. Being confronted with such situation in the EBMC^2 project [3,6] we identified different DQIs related to process oriented data. We are facing diverse DQIs at once which heavily impairs PM outcomes. The combination of different DQIs makes analyzing of causation of one DQI to the outcome of PM intransparent. In order to generate meaningful process models we need to separate single DQIs and develop concepts to tackle each of them.

Section 2 motivates our research and connects it to related work followed by Section 3 which confines and abstracts DQIs within our setting. Concepts to

R. Moreno-Díaz et al. (Eds.): EUROCAST 2013, Part I, LNCS 8111, pp. 491–498, 2013.

improve data quality in order to overcome the identified DQIs are presented in Section 4 whereas Section 5 evaluates how mined models improve by rectified log data based on our proposed concepts. Section 6 summarizes the paper and points out future and follow up work and research.

2 Motivation and Related Work

DQIs have different causes from obvious simple operational causes, like data input errors, to more structural causes, like data model designs. In this paper we want to deal with DQIs that can be abstracted and solved by adding knowledge in order to improve existing data. Therefore we do not deal with errors arising from operation but with errors based on structural differences.

One way to improve DQIs is by purging log data based on constraint violation [7]. This way whole cases are purged which is not always intended. If we want to preserve deviant cases we need to tackle DQIs in another way.

As mentioned before, we want to use PM to discover process models which makes process oriented data necessary, expected in log file format. We know from the EBMC^2 project on skin cancer treatment – as well as stated in the Process Mining Manifesto [2] – that such process oriented data is not always at hand when it comes to realistic data sources. Such data sources are designed for a certain application without a particular process structure in mind and therefore lacking in – for PM purpose – necessary details e.g.: detailed temporal information.

The Process Mining Manifesto [2] categorizes the quality of data sources in terms of PM from one to five starts (* - *****) where PM results from * and ** data sources are not trustworthy. Following this categorization we are dealing with data sources that are categorized as * or **. In order to make those data sources usable – deriving meaningful process models using PM – we aim to raise the quality to at least ***. These categories are already aggregated and therefore hard to use in order to identify single DQIs. Bose et al. [4] collected and categorized a multitude of DQIs in this area, also covering the ones we identified, but offers no solutions or concepts to improve data relating to theses DQIs.

Any event stream is limited in the knowledge that is included – and therefore also what can be derived from it – as it contains instantiations of the underlying process which might not represent the whole process model at all. In order to overcome DQIs that impede PM outcomes we need to extract more knowledge and prepare or pre-process the log data. Or in other words, by adding some knowledge we intend to raise the possibilities to derive further knowledge.

A PM project following a process model like the L* life-cycle model described in the Process Mining Manifesto [2] or the PMMF approach [5] helps in avoiding certain DQIs but do not guarantee to solve all of them, hence this process models can be enhanced by adding data improvement based on additional knowledge. E.g. the L* life-cycle model consists of four stages describing a PM project where the first stage deals with extraction by understanding the available data

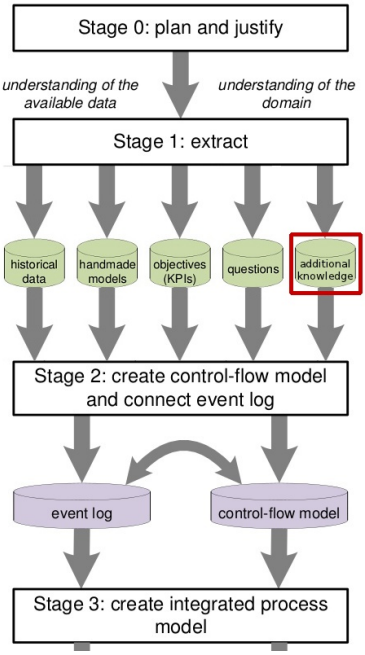

Fig. 1. Extended extract of the L* life-cycle model [2]

and the domain. Here we need to attach the extraction of additional knowledge for the concepts (cf. Subsection 4.1 and 4.2) we suggest to solve certain identified DQIs. The proposed concepts correspond to the second stage ("*create control-flow model and connect event logs*") that are dealing with filtering and adaption of the event log in order to improve results. Figure 1 shows an extended extract of the L* life-cycle model with the mentioned stages where our approach is embedded.

3 Abstractions from the Problem Setting

One aim of the EBMC^2 project is to derive patient treatment models for skin cancer treatment [3,6]. For that purpose we identified and integrated diverse databases and used them as a basis to generate process oriented data. After first attempts using this data for PM we were able to identify several DQIs that made the mining of meaningful process models impossible, especially because of combinations of different DQIs.

In order to provide PM with data we developed the Data Integration Layer for Process Mining (DIL/PM) – partially described in [3] – that eases integration and transformation steps. Figure 2 shows the central extract of the underlying data model. The entities "*case*", "*event*", "*attribute*" and "*attributeValue*" represent all necessary information to generate log files for PM. Meta information is stored

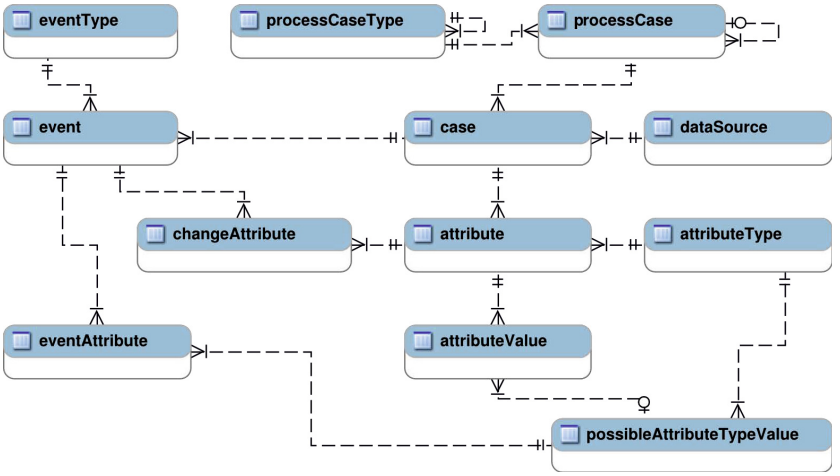

Fig. 2. Data model from the Data Integration Layer for Process Mining (DIL/PM)

in the corresponding type entities, hence allowing us to extend the model for holding additional information to solve DQIs. A transformation and/or filtering step can make use of this additional data to improve DQIs while generated required log files.

3.1 The Problem Environment and Setting

Some DQIs arise from the integration of different data sources, e.g.: Activities from one data source representing other activities from another data source or existing different granularity levels of activities within different data sources. Certain DQIs arise from the non-process orientation of data that makes transformations necessary, e.g.: Missing temporal information leading to incomplete process models or rough temporal granularity. Every data source holds attributes that are set at a certain point when an event happens. The information which events happened and the order they happened in is often missing in the data source.

3.2 DQI Confinement and Abstraction

Being confronted with actual diverse and non-log data sources raises actual problems [3,6]. Some of them are domain specific and need to be solved individually, others can be abstracted, solutions conceptualized and used for improving by pre-precessing or enriching data. In this paper we concentrate on two different domain independent DQIs. First the already mentioned rough temporal granularity that leads to parallelisms in the process model and second different granularity levels of activities leading to unnecessary overloaded process models.

To improve these DQIs we need additional knowledge that is not found within the data itself. The persons knowing the process itself or conducting the activities within the process are possessing the missing knowledge. In order to improve the data quality or enrich the data it is necessary to make use of this knowledge by identifying, extracting and storing it in a structured way so it can be used for processing. To overcome rough temporal granularity, meaning several activities have the same date and time even though they are conducted at different times, we need to know simple sequencing information on activities that allows us to correct time information. For solving different granularity levels of activities, meaning multiple different activities having a common denominator and can be represented by one activity, we need this hierarchical links between the activities. A filter using this information can generate log files with less activities leading to simpler and easier understandable process models.

4 Concept Design to Improve Data Quality

The last section presented some DQIs we identified. We now suggest to introduce parts of the wanted process model it self to improve these DQIs. This section presents two concepts to improve two of those DQIs: Activity Sequencing to improve rough temporal granularity and Activity Hierarchy to improve different granularity levels of activities. Both concepts are integrated into the DIL/PM model, cf. Figure 3.

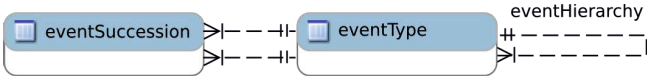

Fig. 3. Extension of the DIL/PM to hold information on Activity Sequencing and Activity Hierarchy

4.1 Activity Sequencing

Activity Sequencing aims for collecting basic information on the antecedent and subsequent activities (e.g.: wake up, breakfast, lunch, dinner, sleep) to resolve parallelisms caused by imprecise temporal information. Figure 3 (left side) shows how eventTypes (activities) can be extended with an additional entity eventSuccession that way allowing multiple succession eventTypes to be stored to one eventType. In a pre-processing step events with the same time stamp can be corrected by searching for a succession relation in eventSuccession. The time correction should be minimal just to allow PM to resolve parallelisms without changing the rest of the model. This very simple algorithm can be extended to find transitive eventTypes, in which case loops within the sequencing have to be recognized. Hence the usage of e.g. reachability algorithms will be necessary, especially if we want to resolve the DQI of missing temporal information. For this paper and the DQI of rough temporal granularity we stick to the simple algorithm as we want to show how the derived process models improve.

4.2 Activity Hierarchy

Activity Hierarchy aims for collecting subtype information on activities (e.g.: swimming with subtypes front crawl, back crawl, breaststroke) that way resolving overloaded process models caused by different granularity levels of activities. Figure 3 (right side) shows how eventTypes (activities) can be extended with an recursive relationship that way allowing to store a rooted tree hierarchy of eventTypes. In a filtering (choosing the level of granularity) and a pre-processing step the event name can be unified to parent eventType names. If we want to use more than one level of this hierarchy – transitive activity name unification – we need a more complex algorithm. For this paper and to be able to show how derived process models improve we stick to the simple case of one level.

5 Evaluation

For evaluation we take a closer look at a part of the skin cancer treatment process which is reflecting the problems we identified. The process starts with the first visit by the patient, hence with the activity "*medical history*". After that a "*skin check*" is performed followed by the "*surgical excision of primary tumor*". Some days later the "*histology of primary tumor*" is finished that decides about the further examination activities (some processes already end here, if no further follow-up examinations are necessary). The examination part that follows is a loop and the severeness of the illness defines which examinations will be performed in one iteration. First a "*clinical examination*" is performed where it is decided if a blood test ("*draw blood sample*" and "*check blood sample*") and/or imaging examinations ("*sonography*", "*x-ray*", "*MRT*", "*CT*", "*PET*" and/or "*PET-CT*") will be conducted in order to find distant metastases.

5.1 Simulation

Real world data we collected within the EBMC^2 project is to small in quantity and has diverse other DQIs. Therefore we use the developed simulation tool *iMine*^{Sim} to generate a log file containing entries with imprecise temporal information as well as the six defined imaging examination activities. The activities *medical history*", "*skin check*" and "*surgical excision of primary tumor*" happen in this order but on the same day. The generated time stamps granularity is on a day basis, therefore all three activities have the same time stamp. Figure 4 (left side) shows the derived process model using the heuristic miner from the ProM 5.2 framework. As we can see the first visit activities that happened on the same day are assumed to be parallel. The found particular control flow is based on the order of the entries in the log file, which was randomized. Further we can see how examinations are conducted parallel which is represented with the four branches leaving "*histology of primary tumor*" to "*sonography*" which is the primary imaging examination that is used, "*draw blood sample*", "*check blood sample*" and "*clinical examination*". We can also see the loop of examinations with the backwards directed branches.

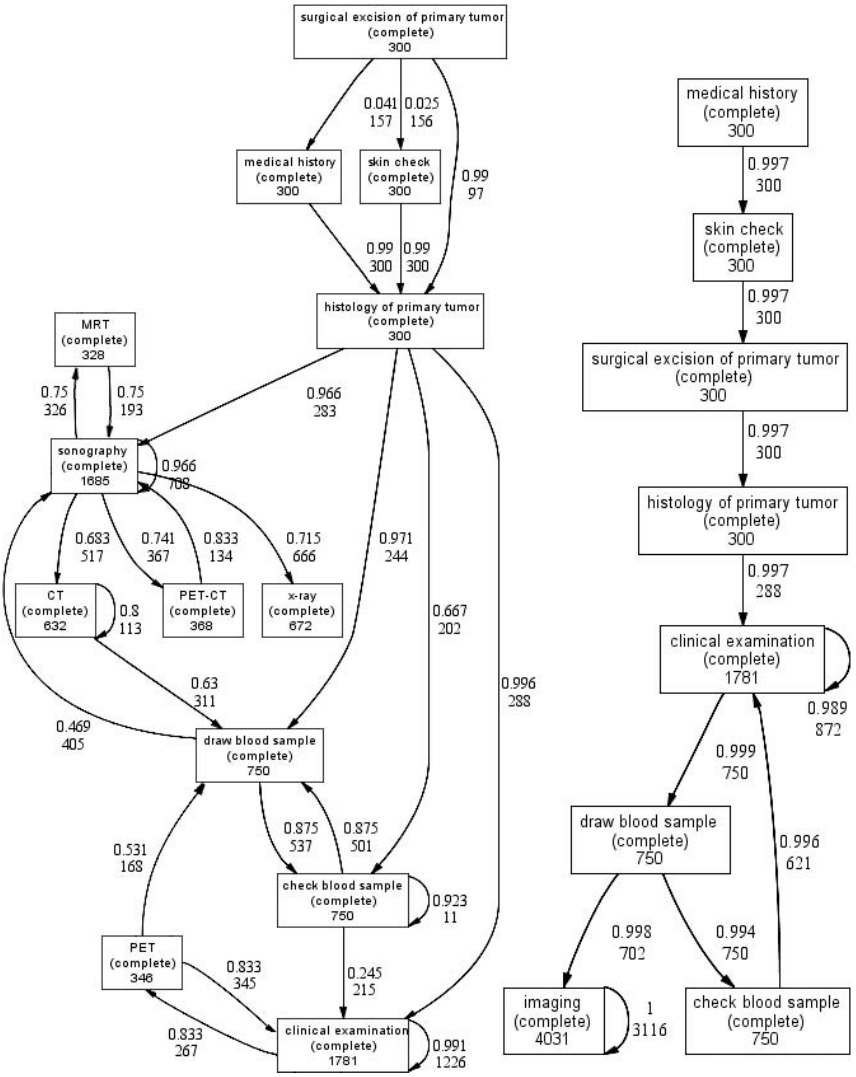

Fig. 4. Derived process models before (left) and after (right) DQIs improvement

5.2 Process Model Improvement

After specifying times stamps using Activity Sequencing and reducing the activities for imaging examinations to one activity "*imaging*" using Activity Hierarchy we can see the improvements in Figure 4 (right side). Parallelism at the first three activities have been resolved as well at "*draw blood sample*" and "*clinical examination*". The diverse imaging examinations – that bloated the model – have been reduced to one activity making the model much more easy to understand without changing the data source. Different filter options allow single imaging activities to be added easily.

6 Summary and Outlook

We showed on a simplified real world example the virtues of the two concepts Activity Sequencing and Activity Hierarchy for improving certain Data Quality Issues (DQIs) and therefore with PM derived process models. The presented data model from the Data Integration Layer for Process Mining (DIL/PM) was extended to hold the additional knowledge for preparing log files for PM.

The algorithms of two concepts Activity Sequencing and Activity Hierarchy will be extended to solve additional DQIs and further concepts will be developed to improve other DQIs. Evaluations with future real data from the EBMC^2 project [3,6] as well as a data set on higher eduction system will be conducted.

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., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011 Workshops, Part I. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012)
3. Binder, M., et al.: On analyzing process compliance in skin cancer treatment: An experience report from the evidence-based medical compliance cluster (ebmc2). In: Ralyté, et al. (eds.) [8], pp. 398–413
4. Bose, J.C., Mans, R., van der Aalst, W.M.P.: Wanna improve process mining results? Tech. rep., BPM Center Report (2013)
5. De Weerdt, J., Schupp, A., Vanderloock, A., Baesens, B.: Process mining for the multi-faceted analysis of business processes: A case study in a financial services organization. Computers in Industry 64(1), 57–67 (2013)
6. Dunkl, R., Fröschl, K.A., Grossmann, W., Rinderle-Ma, S.: Assessing medical treatment compliance based on formal process modeling. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 533–546. Springer, Heidelberg (2011)
7. Ly, L.T., Indiono, C., Mangler, J., Rinderle-Ma, S.: Data transformation and semantic log purging for process mining. In: Ralyté, et al. (eds.) [8], pp. 238–253
8. Ralyté, J., Franch, X., Brinkkemper, S., Wrycza, S. (eds.): CAiSE 2012. LNCS, vol. 7328. Springer, Heidelberg (2012)