Multidimensional Process Mining mit dem Process Cube Explorer

Jannik Arndt, Thomas Meents, Bernd Nottbeck
Universität Oldenburg
{jannik.arndt, thomas.meents, bernd.nottbeck}@uni-oldenburg.de

Abstract: In vielen Anwendungsfällen werden Arbeits- und Prozessabläufe in Log-Dateien protokolliert. Das Process Mining findet in diesen Eventlogs das zugrundeliegende Prozessmodell, welches dann zur Kontrolle und Optimierung der Abläufe genutzt werden kann. Der Process Cube Explorer bietet ein Framework für Multidimensional Process Mining, mit dem Eventlogs aus einem Data Warehouse extrahiert und in Prozessmodellen dargestellt werden. Der multidimensionale Ansatz erlaubt es dem Anwender, die Datenbasis anhand der Dimensionen einzugrenzen und so die Auswirkung verschiedener Eigenschaften auf das Prozessmodell zu entdecken.

1 Einleitung

Beim Process Mining handelt es sich um eine Mischung aus Machine Learning, Data Mining und automatisierter Prozessmodellierung. Ziel ist es Wissen über Prozesse aus Eventlogs zu gewinnen [vdA11]. Dabei unterteilt sich das Process Mining in die drei Anwendungsbereiche *Process Discovery* (Finden neuer Prozessmodelle), *Conformance Checking* (Überprüfung von existierenden Prozessmodellen) und *Model Enhancement* (Erweiterung der bestehenden Modelle).

Das in dieser Arbeit vorgestellte Programm *Process Cube Explorer* ist primär im Bereich *Discovery* einzuordnen, in dem aus einer Menge von automatisch oder manuell aufgezeichneten Events das zugrundeliegende Prozessmodell erkannt wird. Hierfür sind die einzelnen Events in chronologischer Abfolge in einem *Eventlog* gespeichert, auf das verschiedene Mining-Algorithmen angewandt werden können. Der multidimensionale Analyseansatz der dieser Arbeit zur Grunde liegt, erleichtert die Analyse großer Datenmengen.

Seit dem Frühjahr 2013 entwickelt eine Projektgruppe aus elf Studenten im Rahmen des Dissertationsvorhabens von Thomas Vogelgesang an der Universität Oldenburg ein Forschungsframework für Multidimensional Process Mining.

2 Herausforderungen

Process Discovery hat in realen Anwendungen zwei große Herausforderungen, die beide aus dem Big-Data-Umfeld stammen: Erstens ist die Menge an automatisch generierten Events

in Eventlogs in der Regel nur sehr schwierig zu überschauen und in ihrer Gesamtheit ohne Vorauswahl oft nicht aussagekräftig. Zweitens steigt der Rechenaufwand der Algorithmen mit der Menge der Events die betrachtet werden deutlich. Es ist also essenziell, dass der Benutzer dieser Anwendung die Möglichkeit hat, seine Daten auf eine sinnvolle Teilmenge zu begrenzen.

An diesem Punkt setzt das Multidimensional Process Mining an: Joel Ribeiro schlägt die Verwaltung vorberechneter Teilergebnisse in einer multidimensionalen Struktur, dem sogenannten "Event Cube", vor [RW11]. Hierzu wird bereits beim ETL-Prozess eine Vielzahl von Aggregationen und weiterer Funktionen ausgeführt. Dieser Ansatz benötigt also eine umfangreiche Vorberechnung aller Daten. Der Ansatz von Thomas Vogelgesang [Vog13] speichert bereits die Eventlogs multidimensional und führt das Process Mining dann auf einer durch den Benutzer eingegrenzten Teilmenge durch. Dafür wird das Konzept des Data Warehousings (DWH) genutzt und die Eigenschaften aus dem Eventlog als Dimensionen dargestellt, anhand derer gezielt zusammengehörige Eventdaten ausgewählt werden können.

Im Programm werden die Abläufe weitestgehend automatisiert und der Benutzer gezielt durch den Prozess geführt. So können alle Schritte im selben Programm vollzogen werden, vom Laden der Daten über die Auswahl durch das bekannte multidimensionale Modell ("Cube"), die Auswahl zwischen verschiedenen Mining-Algorithmen, dem eigentlichen Mining bis hin zur übersichtlichen Darstellung der Daten und weiteren Vergleichs- und Analysemöglichkeiten wie z. B. *Comparing Footprints*, ein Vergleich zwischen dem generierten Prozessmodell und dem zugrundeliegendem Eventlog.

3 Herangehensweise

Die Datenanalyse ist mit allen Eventlogs möglich, die in einem ETL-Prozess in ein DWH überführt werden. Dieses muss sich grob an das vereinfachte Schema in Abbildung 1 halten. Zur Überprüfung der Miningergebnisse wurden Beispieldatensätze erstellt und reale aus der Physionet bereitgestellten *MIMIC II*-Datenbank[1] verwendet. Die *MIMIC II* enthält eine große Menge anonymisierter Krankenhausdaten zu Behandlungsabläufen und verwandten Ereignissen. Den Kern bildet eine Faktentabelle *Fact* die auf mehrere Dimensionstabellen verweist. In diesen sind die Eigenschaften der Patienten wie Alter und Geschlecht, der Krankheit und der Diagnosen gespeichert. Die Zellen der Faktentabelle fassen Fälle (*Cases*) zusammen, bei denen alle Dimensionsausprägungen identisch sind. Ein *Case* wiederum fasst alle *Events* eines Prozessablaufs zusammen. Dies entspricht z. B. der Behandlung eines Patienten. Zusätzlich können auch Events Dimensionen besitzen, z. B. der verantwortliche Arzt für einen Behandlungsschritt. Hierdurch werden feinere Analyseeinstellungen möglich.

An das Framework können verschiedene Datenbankentypen generisch angebunden werden, zur Zeit werden u. a. Oracle, MySQL und PostgreSQL unterstützt. Die geladenen Daten können dann im Programm anhand der Dimensionswerte, also der Eigenschaften der Fälle,

[1] `http://mimic.physionet.org`

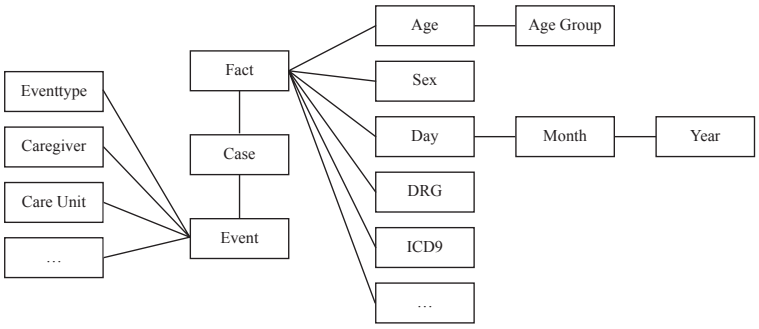

Abbildung 1: Die Datenstruktur des Data Warehouse, aus dem die Eventdaten geladen werden.

in verschiedenen Aggregationsstufen sowie durch weitere Filter ausgewählt werden.

Im nächsten Schritt wählt der Benutzer einen Mining-Algorithmus aus. Bereits implementiert sind drei Varianten des bekannten Alpha-Algorithmus [dMvDvdA04], eine erweiterte Version des HeuristicMiners [WvdAdM06] und eine Implementierung des kürzlich veröffentlichten "Infrequent Miner - inductive" [LFvdA13]. Weitere Algorithmen können leicht über eine Schnittstelle in das Framework integriert werden.

Für jede Dimensionsausprägung, also jede gewählte Zelle des Würfels, erstellen die Mining-Algorithmen ein Prozessmodell, welches als Petrinetz dargestellt wird (siehe Abbildung 2). Diese Modelle können exportiert, gedruckt und miteinander verglichen werden. Die Software bietet nun die Möglichkeit, die Qualität der Modelle zu beurteilen. Mittels *Conformance Checking* kann überprüft werden, auf wie viele der Cases ein generiertes Prozessmodell zutrifft. Außerdem werden dem Benutzer für jedes Prozessmodell weitere Qualitätskennzahlen, z. B. der Anteil der berücksichtigten Events, angezeigt.

4 Ergebnisse

In der Zeit vom April 2013 bis März 2014 wurde ein Framework entwickelt, das zum einen den Benutzer beim kompletten *Process Discovery*-Prozess auf multidimensionalen Daten unterstützt, und zum anderen sehr einfach zu erweitern ist. Der Quellcode des Projekts wird zum Ende der Laufzeit im März 2014 veröffentlicht, sodass insbesondere die Analysemöglichkeiten weiter ausgebaut und weitere Algorithmen integriert werden können. Mit diesem Tool können Forscher und Wissenschaftler, bspw. im Health Care-Bereich, die Software einsetzen um bestehende Prozessmodelle unter Berücksichtigung der individuellen Eigenschaften der Patienten darzustellen, zu analysieren und zu verbessern.

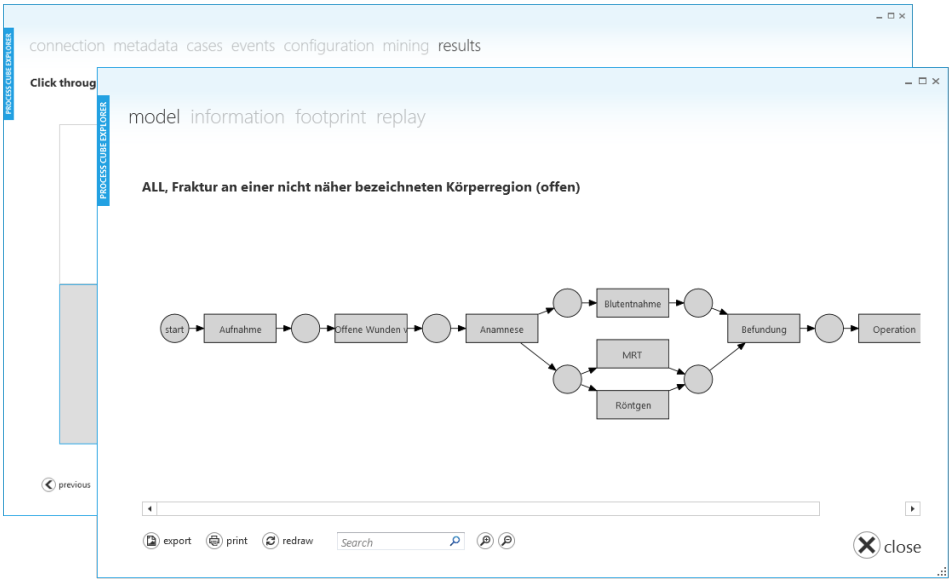

Abbildung 2: Die generierten Prozessmodelle werden dem Nutzer übersichtlich angezeigt, außerdem kann er Modelle vergleichen und einen *Comparing Footprint* erstellen.

Literatur

[dMvDvdA04] A.K. Alves de Medeiros, B.F. van Dongen und Wil M. P. van der Aalst. Process Mining : Extending the α-algorithm to Mine Short Loops. *BETA Working Paper Series, Eindhoven University of Technology*, 2004.

[LFvdA13] Sander J. J. Leemans, Dirk Fahland und Wil M. P. van der Aalst. Discovering Block-Structured Process Models From Event Logs Containing Infrequent Behaviour. 2013.

[RW11] J.T.S. Ribeiro und A.J.M.M. Weijters. Event Cube: Another Perspective on Business Processes. In Meersman et al., Hrsg., *On the Move to Meaningful Internet Systems: OTM 2011*, Jgg. 7044 of *Lecture Notes in Computer Science*, Seiten 274–283. Springer Berlin Heidelberg, 2011.

[vdA11] Wil M. P. van der Aalst. *Process Mining - Discovery, Conformance and Enhancement of Business Processes.* Springer, 2011.

[Vog13] Thomas Vogelgesang. Multidimensional Process Mining A flexible analysis approach for health services research Categories and Subject Descriptors. In *EDBT/ICDT 2013*, Genoa, Italy, 2013.

[WvdAdM06] A.J.M.M. Weijters, W.M.P. van der Aalst und A.K. Alves de Medeiros. Process mining with the heuristics miner-algorithm. 2006.