Lecture Notes in Business Information Processing 412

More information about this series at http://www.springer.com/series/7911

Gert Janssenswillen

Unearthing the Real Process Behind the Event Data

The Case for Increased Process Realism

Springer

Author
Gert Janssenswillen
Research Group Business Informatics
Hasselt University
Diepenbeek, Belgium

ISSN 1865-1348 ISSN 1865-1356 (electronic)
Lecture Notes in Business Information Processing
ISBN 978-3-030-70732-3 ISBN 978-3-030-70733-0 (eBook)
https://doi.org/10.1007/978-3-030-70733-0

This book is a revised version of the PhD dissertation written by the author at Hasselt University in Belgium.
The original PhD dissertation is accessible at: http://hdl.handle.net/1942/28268.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

In loving remembrance of Joannis JANSSEN († 1695)
and Adriana SWILDEN († 1712)

In appreciation of their legacy,
and for a lifetime of spelling, correcting,
and explaining my surname.

Preface

Companies in the 21st century possess a large amount of data about their products, customers and transactions. The prominent role of business processes in the modern organisation in recent decades has led to a remarkable increase in the amount of event data that is available. Event logs are *logbooks* that contain information about everything that happens in a company on a daily basis. A customer who places an order, an employee who logs in to the customer management system to handle the order, a supplier who delivers a quotation for the products, a production line that is started, etc. The digitisation of all these events enables us to analyse business processes at a level that was previously unthinkable.

The increase in available event data gave rise to process mining, a discipline that focuses on extracting insights about processes from event logs. However, correctly displaying business processes is not a trivial task. Due to the high complexity of most processes, event logs contain only a limited sample of all the possible ways and combinations in which business processes can be performed. Errors and inconsistencies in the available data create additional difficulties. In response to these challenges, process discovery algorithms were developed - algorithms that discover process models based on event logs. However, the crucial question is: how good are these discovered models? Are they able to correctly represent business operations?

The concept of process realism is introduced in this dissertation. To optimise processes, evidence-based decision making is needed. Consequently, it is essential to map these processes in a realistic way. Blindly relying on both partial and/or inconsistent data and on algorithms can lead to wrong actions being taken.

Process realism is approached from two perspectives in this dissertation. First, quality dimensions and measures for process discovery are analysed on a large scale and compared with each other on the basis of empirical experiments. Which measures are best suited to assess the quality of a discovered process model? What are their weaknesses and strengths? And what challenges still need to be overcome in order for it to evolve into a reliable quality measurement?

The experiments in this thesis show that there are important differences between the different quality measures in terms of feasibility, validity and sensitivity. Moreover, the role and meaning of the generalisation dimension is unclear. Existing generalisation measures do not succeed in adequately assessing the fit between process models and the underlying process. Fitness and precision measures also do not constitute unbiased estimators of the quality of the model as a representation of the underlying process. Furthermore, with regard to experimental set-up, various challenges have been identified that are necessary for measures to evolve towards a correct quality measurement.

In addition to the focus on process models, process realism is also approached from a data point of view. By developing a transparent and extensible toolset, a framework is offered to analyse process data from different perspectives. Exploratory and descriptive analysis of process data and testing of hypotheses again leads to increased process realism.

The developed framework is applied in this dissertation to two case studies. First, how can we use process data to better understand students' study trajectories and to better guide students? Secondly, how can applying process analysis in a railway context map out the use of the rail infrastructure and analyse deviations between the timetable and implementation in order to achieve a smoother service for passengers?

Both case studies show that the framework has clear added value, and that the answers to the questions asked can help to improve the processes under consideration. At the same time, however, unresolved challenges within process mining are also emphasised, such as the analysis of processes at the right level of granularity, and the assumption that process instances are independent of each other.

From both perspectives, process model and process data, recommendations are made for future research, and a call is made to give the *process realism* mindset a central place within process mining analyses.

Acknowledgements

> Many great actions are committed in small struggles.
>
> Victor Hugo

There are so many who have—each in their own way—left their fingerprints on this dissertation, and thereby also on my heart. Like so many projects, this has not been a solitary journey, and I am indebted to all those who have travelled with me along the way. All those who have shared their expertise, passion, support, and laughter. Growing as a person is an incremental process influenced by a near-infinite sequence of experiences and influences. It would be impossible to mention all those who have contributed to this growth, but allow me to spend some words on those whose influence was paramount.

First and foremost, my genuine gratitude goes to my supervisor, Benoît Depaire. Over the years, you have not only been my supervisor, but also have become a true mentor and friend. You have been my strength when I felt weak. My guiding light when I felt lost. You have continually challenged the limits of what I thought I could do. Teaching together with you has taught me to take all details into account, and to never assume that things are happening by themselves. You have given me the freedom and encouragement to explore other topics and take roads less travelled, which made this dissertation what it is now, and me the person I am today.

On the first day of my PhD, your advice to me was to learn R programming. Honestly, I was sceptical at first—I could already program in Python, why would I need R? But you also set me a goal: *try to recreate this graph in Disco using R.* About six months later, there was a first R package for process analysis which started to look like something useful. Another eight months later, the first version of `edeaR` was published on CRAN. Today, I can confidently say `bupaR` can do everything which Disco can, and much more. Thank you for giving me that initial spark which has ignited a fiRe within me. And thank you for the opportunity to pass along that spark to my students each year—I hope that one day they will appreciate its true value.

A sincere thank you also goes to Mieke Jans. Through your diverse experience from both industry and academia, you have always provided my with an alternative perspective on my work and valuable feedback that few others would be able to give. Your distinctive angle on things has influenced and improved my dissertation in more ways than you would guess. Teaching together with you has shown me how an effective, well-oiled process should run. On a personal level, you have taught me to be proud of my achievements and challenged me to leave my comfort zone—which in my case is not at all straightforward.

Furthermore, my appreciation also goes to Koen Vanhoof. Thank you for all the things you do for our research group. Together with Mieke and Benoît, you have made the business informatics group into what it is today—an excellent place to work (and study). An achievement you should all three be genuinely proud of.

I would like to profoundly thank Benoît, Mieke, Koen and all the members of the jury for reading my thesis and for providing me with useful remarks and suggestions, which have markedly improved the quality of my dissertation.

In particular, I would like to thank Sabine Verboven. While you have not followed my PhD from up-close from the very start, you were there before it started. At Infrabel, you have guided me in my very first process mining adventures, an experience and opportunity I am grateful for and will never forget. Together we also learned the importance of a welcoming culture and executive-level support for process mining and data science in general. I am hopeful we will continue to be partners in this never-ending endeavour.

Moreover, I would like to especially thank Jorge Munoz-Gama and his colleagues, particularly Marcos Sepúlveda, Wai Lam Jonathan Lee, and Juan Pablo Salazar. Not only have you all influenced or improved my dissertation in several ways, you have also provided me with a warm welcome in your group during my visit, of which I hope many more will follow.

You can never go wrong when you have a great team to work with. I am grateful for every member of our research group, for creating the extraordinary atmosphere in which we work. Especially, I would like to thank my office mates Frank Vanhoenshoven, Mathijs Creemers and Mehrnush Hosseinpour, for welcoming me in their office during the last months of my PhD. Thank you for allowing me to occasionally disturb your work with weird facts, anecdotes, and frustrations.

Particular heartfelt thanks goes to the best colleagues one can possibly imagine, my *breakfast besties*, Hanne Pollaris and Marijke Swennen. Thank you both for being my personal stylists. Hanne, even on the saddest and gloomiest days, you are that one person who can put a smile on my face, just by entering the room. Your positivity in life is an inspiration for us all. Thank you for always believing in me and supporting me.

Marijke, I would honestly not know where I would be without you. Whenever anything happens, the first thing I think of is telling you, and asking your advice. You have stayed with me through so many ups and downs. I am forever in your debt, and count myself as one of the luckiest persons on earth to have you as a friend.

Together with you, Jeroen Corstjens, Niels Martin and Stef Moons, we have been on countless trips, enjoyed innumerable meals and, most of all, shared an everlasting amount of laughter, happiness, and joy. But despite having spent all these moments together, I think I may have often forgot to thank you all for being terrific friends!

Finally, and most importantly, I would like to deeply thank my family for their ongoing support.

I am grateful for my siblings, for providing me with endless distractions from my, at times tedious, work. Thank you also for occasionally reminding me what a *real* job looks like. Thank you for reminding me not to be too hard on my students, because *they are all doing their best.* But most of all, thank you for being the best siblings one can imagine! Words will never be enough to show my appreciation for you.

But the two persons who deserve the most praise are my parents. Dear mom and dad, I have let so many years pass without thanking you both. But you haven't let a single second pass without loving me unconditionally. Thank you for who I am, and thank you for all the things I'm not. Forgive me for the words unsaid. Thank you for the wings you have given me, for having taught me how to soar up into the sky and expand my horizons. It may take a lifetime, but I'll do everything to repay for what you have done for me. Thank you for being there, even when I'm stupid enough to think I'd rather be alone.

So much of me,
is made of what I learned from you.
You'll be with me forever,
like a handprint on my heart.

December 2020 Gert Janssenswillen

Contents

Part I Introduction

1 Process Realism . . . 3
1.1 Introduction to Process Mining . . . 4
1.1.1 Business Process Management . . . 6
1.1.2 The Emergence of Process Mining . . . 8
1.1.3 Perspectives . . . 10
1.1.4 Tools . . . 12
1.1.5 Towards Evidence-Based Business Process Management . . . 13
1.2 The Case for Process Realism . . . 14
1.2.1 Motivation . . . 14
1.2.2 Research Objective . . . 15
1.3 Methodology and Outline . . . 17
1.3.1 Process Model Quality . . . 17
1.3.2 Process Analytics . . . 19

Part II Process Model Quality

2 Introduction to Conformance Checking . . . 23
2.1 Introduction to Process Mining . . . 23
2.1.1 Preliminaries . . . 25
2.1.2 Process . . . 26
2.1.3 Event Log . . . 26
2.1.4 Model . . . 28
2.2 Quality Dimensions . . . 29
2.2.1 Fitness . . . 31
2.2.2 Precision . . . 32
2.2.3 Generalization . . . 33
2.2.4 Simplicity . . . 34
2.3 Quality Measures . . . 34
2.3.1 Fitness . . . 35
2.3.2 Precision . . . 39
2.3.3 Generalization . . . 43
2.4 Conclusion . . . 45
2.5 Further Reading . . . 46

3 Calculating the Number of Distinct Paths in a Block-Structured Model . . . 49
3.1 Introduction . . . 49
3.2 Formal Algorithm . . . 50

3.2.1 Assumptions and Used Notations 50
3.2.2 Generic Approach 51
3.2.3 Block Functions 52
3.2.4 Limitations. 56
3.3 Implementation. 57
3.3.1 Preliminaries 58
3.3.2 Algorithm 59
3.3.3 Extended Block Functions 60
3.3.4 Silent Transitions and Duplicate Tasks 64
3.4 Performance 67
3.5 Conclusion and Future Work 69
3.6 Further Reading 70

4 Comparative Study of Quality Measures. 71
4.1 Introduction 71
4.2 Problem Statement 72
4.3 Methodology 73
4.3.1 Generate Systems 74
4.3.2 Calculate the Number of Paths 78
4.3.3 Simulate Logs 78
4.3.4 Discover Models. 82
4.3.5 Measure Quality 83
4.3.6 Statistical Analysis 84
4.4 Results 86
4.4.1 Feasibility 86
4.4.2 Validity 90
4.4.3 Sensitivity 99
4.5 Discussion 107
4.6 Conclusion. 108
4.7 Further Reading 110

5 Reassessing the Quality Framework. 111
5.1 Introduction 111
5.2 Exploratory Versus Confirmatory Process Discovery 112
5.2.1 Problem Statement 114
5.3 Methodology 115
5.3.1 Generate Systems 115
5.3.2 Simulate Logs 115
5.3.3 Discover Models. 116
5.3.4 Measure Log-Quality. 117
5.3.5 Measure System-Quality 117
5.3.6 Statistical Analysis 118
5.4 Results 119
5.4.1 Log Versus System-Perspective 119
5.4.2 Generalization 123

5.5 Discussion 125
5.6 Conclusion. 125
5.7 Further Reading 127

6 Towards Mature Conformance Checking 129
6.1 Synthesis 129
6.1.1 Fitness. 129
6.1.2 Precision 131
6.1.3 Generalization 132
6.2 Future Research 132
6.2.1 System-Fitness and System-Precision 132
6.2.2 Improving the Experimental Setup 135

Part III Process Analytics

7 Reproducible Process Analytics. 139
7.1 Introduction 139
7.2 Problem Statement 140
7.3 Requirements Definition 142
7.3.1 Functionality Requirements 142
7.3.2 Design Requirements. 144
7.4 Design and Development of Artefact 145
7.4.1 Core Packages 146
7.4.2 Supplementary Packages 147
7.5 Demonstration of Artefact 149
7.5.1 Event Data Extraction 149
7.5.2 Data Processing 154
7.5.3 Mining and Analysis. 162
7.6 Discussion 172
7.7 Conclusion. 174
7.8 Further Reading 174

8 Student Trajectories in Higher Education 177
8.1 Learning Analytics and Process Mining 178
8.2 Data Understanding. 178
8.3 Followed Versus Prescribed Trajectories 179
8.3.1 Root Causes. 181
8.3.2 Impact. 183
8.4 Failure Patterns. 184
8.4.1 Bags 184
8.4.2 High-Level Analysis 186
8.4.3 Low-Level Analysis 188
8.5 Understanding Trajectory Decisions 189
8.6 Discussion 192
8.7 Conclusion. 193
8.8 Further Reading 193

9 Process-Oriented Analytics in Railway Systems 195
9.1 Introduction 195
9.2 Problem Statement and Related Work 196
9.3 Methodology 198
9.3.1 Rerouting Severity 201
9.3.2 Rerouting Diversity 203
9.3.3 Discovering Patterns 207
9.4 Results 209
9.4.1 Rerouting Severity 215
9.4.2 Rerouting Diversity 216
9.5 Discussion 220
9.6 Conclusions 222
9.7 Further Reading 223

Part IV Conclusions

10 Conclusions and Recommendations for Future Research 227
10.1 Process Model Quality 227
10.1.1 Lessons Learned 228
10.1.2 Recommendations for Future Research 229
10.2 Process Analytics 233
10.2.1 Lessons Learned 233
10.2.2 Recommendations for Future Research 235

Afterword 239

A Additional Figures and Tables Chapter 4 245

B Function Index `bupaR` packages 257
B.1 `bupaR` 257
B.2 `edeaR` 259
B.3 `evendataR` 260
B.4 `xesreadR` 261
B.5 `processmapR` 261
B.6 `processmonitR` 261
B.7 `petrinetR` 262
B.8 `ptR` 263
B.9 `discoveR` 263

C Scripts Chapter 8 265

D Scripts Chapter 9 273

References 277