Discovery and Analysis of Process Models: A Case Study

Skroch O, Hombrecher M

openrheinmain.org
13-Sep-2019

Event log

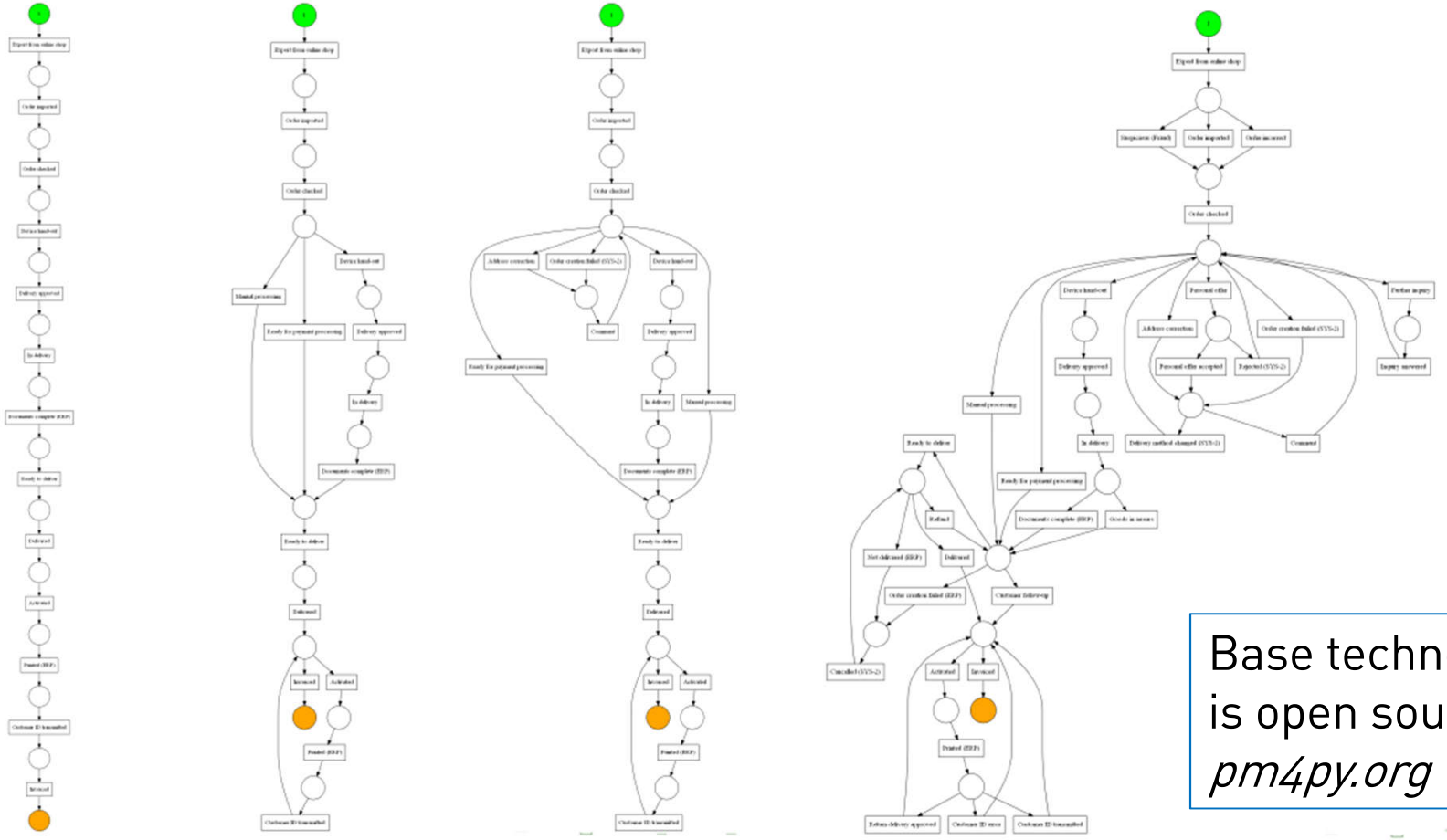
Starting point is a 3-month order processing event log with approx. 3 million entries.

```
Sequence-ID;Text;Timestamp

...

10427186;Webshop export;20.10.2013 13:43:05
10427186;10 Order imported;20.10.2013 13:58:04
10427186;15 Order checked;20.10.2013 13:58:13
10427186;23 Device handout;20.10.2013 14:01:43
10427186;30 Ready for delivery;20.10.2013 14:01:43
10427186;50 In delivery;20.10.2013 14:01:43

...
```

Pre-processing and data clean-up
in a sub-project using Python and C++

Discovery

A modified version of the *α–algorithm** generates a fully formal workflow net from the event log.



Base technology is open source: *pm4py.org*

* van der Aalst W, Weijters A, Maruster L (2004), "Workflow Mining: Discovering Process Models from Event Logs". *IEEE Transactions on Knowledge and Data Engineering*, (16) 9: 1128-1142.

Analysis

Investigation into questions with business relevance for the client, for example:
can the success of a trace be predicted?

- **Preparation**
- Classifying each trace as „success“ or „no success“
  - Semantic analysis based on the final event of the trace
- “Charging“ the pure process traces
  - With additional feature data made available from the client
    E.g. device, brand, channel
  - With basic computations on the data itself
    E.g. day of week, hour of day, time since predecessing event, number of devices per trace
  - We gained 224 attribute values (220 binary, 4 numeric) all together

Analysis

Investigation into questions with business relevance for the client, for example: *can the success of a trace be predicted?*

- **Application of a self-calibrating** random decision forest*
- Self-calibration with a *grid-search* using 900 parameter combinations for a single tree and 300 parameter combinations for the whole forest

Base technology is open source: *scikit-learn.org*

- Validation with unknown data (a random subset of 20% of the data that was not used for the supervised learning phase)
- Disappointing validation result:

  Only 78,3% accuracy, while 78,5% of the traces from the event log were classified as "success"

* Ho TK (1995), "Random Decision Forests". In: Kavanaugh M, Storms P (eds), *Proceedings of the Third International Conference on Document Analysis and Recognition: Volume 1*: 278-282. IEEE Computer Society Press, Los Alamitos, California, USA et al.

Analysis

Investigation into questions with business relevance for the client, for example: *can the success of a trace be predicted?*

- **Application of a self-calibrating** random decision forest*
- That means: the best random forest predicted slightly worse than a pure guess on the empirical probability distribution…
- In hindsight, the reason for the poor prediction quality seems intuitively clear to a human analyst: *the logged traces are (neccessarily) highly „redundant“*
- However, further anaysis of the discriminating attributes in the *random forest* provided valuable insight…

DEVICE_TYPE_135	0.066454
Onlineshop	0.121465
Telemarketing	0.113806
dayOfWeek	0.089961
hour	0.204486

* Ho TK (1995), “Random Decision Forests”. In: Kavanaugh M, Storms P (eds), *Proceedings of the Third International Conference on Document Analysis and Recognition: Volume 1*: 278-282. IEEE Computer Society Press, Los Alamitos, California, USA et al.

Analysis

Investigation into questions with business relevance for the client, for example: *can the success of a trace be predicted?*

Base technology is open source: *keras.io*

- **Application of a recurrent artificial neural network***
  - Long short-term memory on 2 blocks with 32 cells each.
- Approach:
  - Predict the final outcome of a trace with its last n events missing.
  - Decreasing prediction accuracy with increasing n:

    $n = 1$ 98% of LSTM preditions correct
    $n = 6$ 97%
    $n = 14$ 92%
- A suspicion at second glance, open for further investigation:
  - Median length of „success“ traces is 12 events and 0,75 quantile is 13 events, median length of „no success“ traces is 14 events and 0,75 quantile is 15 events
  - With the approach „reversed“ – using only the first n events – the LSTM was better than the 78,5% benchmark only with $n \geq 12$.
  - This might call for an analysis if the LSTM has „recognized“ only the length of the trace…

* Hochreiter S, Schmidhuber J (1997), “Long Short-Term Memory”. *Neural Computation, (9) 8 : 1735-1780.*

Summary and outlook

- The “mined” information represents valuable input for business decisions.
- Interpretation makes sense only within the actual context we examined.
- Therefore, it can hardly be generalized.
- Further work has already started, including
  - a possible formalization of specific traits and characteristics of process models to improve the statistical handling of process schemes
    - “out of scope” in the standard data analysis approaches
  - the application of further advanced mining methods to retrieve unique business relevant facts which are possibly missed by other approaches.

Your questions?