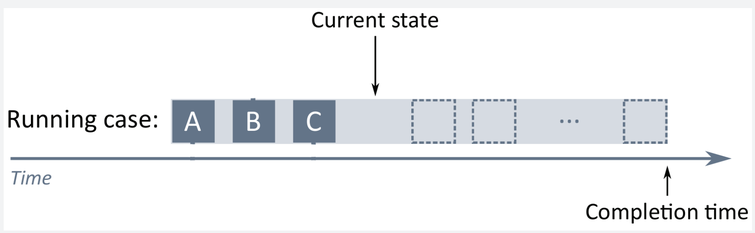
Sequence completion prediction

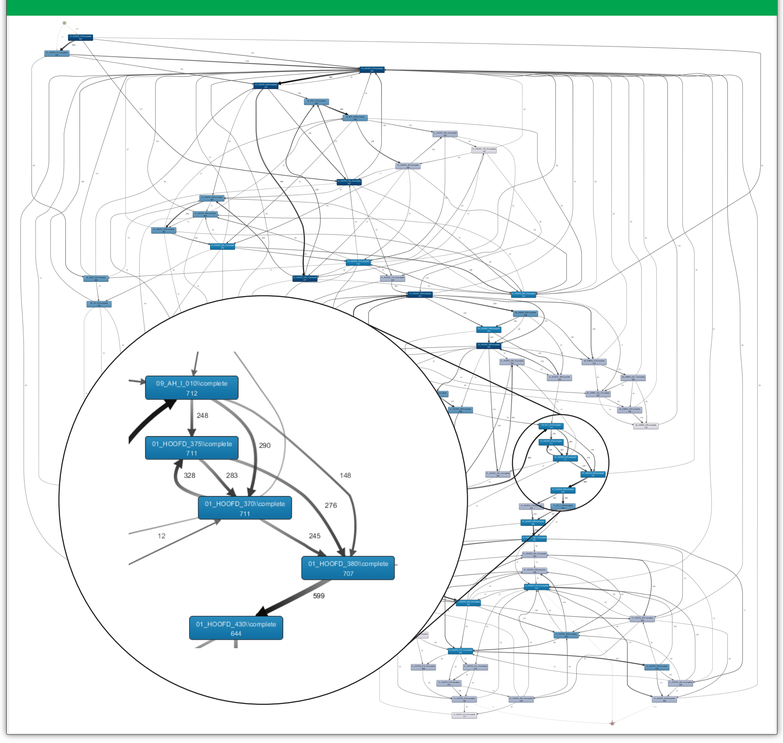
Andrii Rozumnyi[1] · Dmytro Chasovskyi[2]
Institute of Computer Science, University of Tartu
Curriculum: [1]Computer Science, [2]Software Engineering

1 Abstract

We develop a framework for predicting how an on-going instance of a business process will unfold up to its completion. Specifically, for a given point during the process case execution, we extract all possible completion scenarios how this case may unfold up to its completion and estimate probabilities of various scenarios. In this project we report only the most likely scenario.



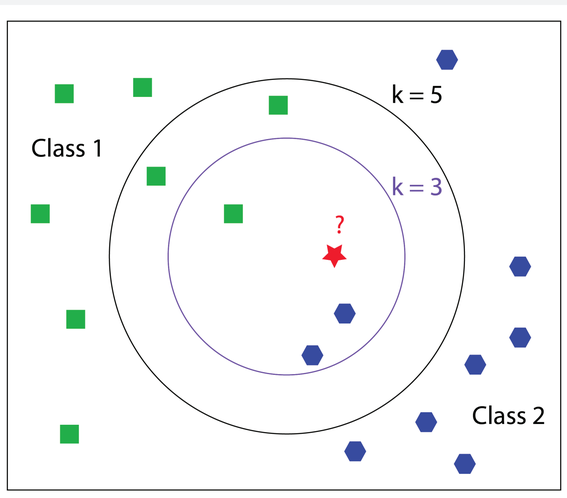
Example of a Process Model



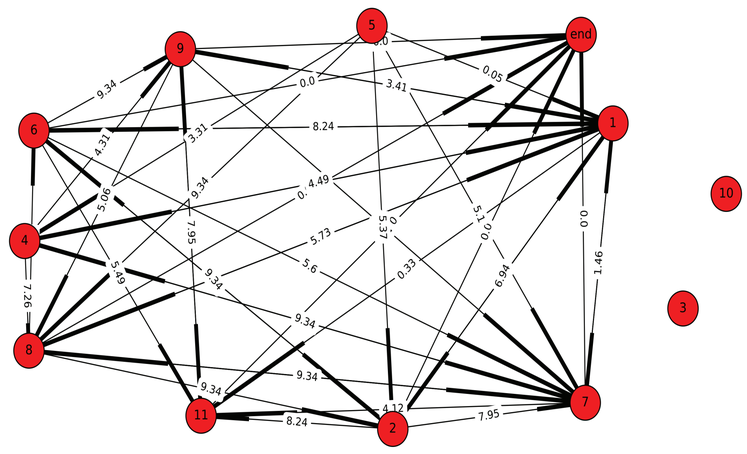
Example of a Log

Case Id	Timestamp	Resource	Activity	Category	Amount
65923	20-02-2002:11.11	Jack	A	-	1000
65923	20-02-2002:13.31	Jack	B	Gold	1000
65923	21-02-2002:08.40	John	C	Gold	900
65923	22-02-2002:15.51	Joe	F	Gold	900
65924	19-02-2002:09.10	Jack	A	-	200
65924	19-02-2002:13.22	John	B	Standard	200
65924	20-02-2002:17.17	John	D	Standard	200
65924	21-02-2002:10.38	Joe	F	Standard	200

2 Approach 1: K-Nearest Neighbors



3 Approach 2: Annotated transition system (ATS)



-A transition system is a triplet TS = (S, A, T), where S is the set of states, $A \subseteq A$ is the set of activities and $T \subseteq S \times A \times S$ is the set of transitions. $S^{start} \subseteq S$ is the set of initial states, and $S^{end} \subseteq S$ is the set of final (accepting) states.

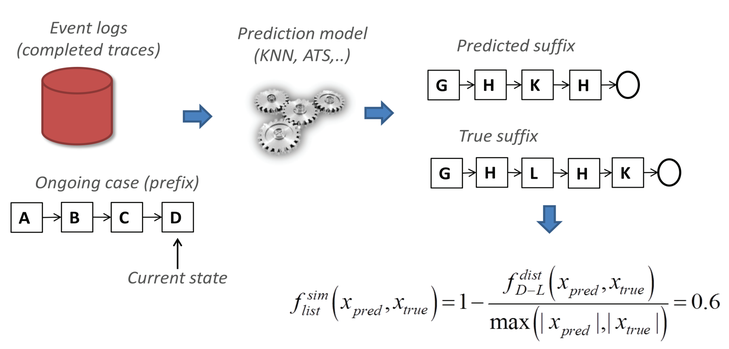
- We build transition system from the adjacency matrix (based on the trainin set)

- For each test sample, we annotate edges based on the transition probabilities learned from the training samples

4 Datasets

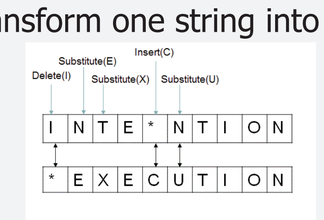
	Domain	# of cases	Mean case length	Median case length	# of events	# of event classes (distinct events)
Business Process Intelligence Challenge 2011	Healthcare	915	131	44	120,045	583
Business Process Intelligence Challenge 2012	Healthcare	225	132	93	29,694	316
Environmental permit	Municipality	937	41.6	43	38,944	381
Road traffic fines	Government	28,530	3.59	3	102,396	11

5 Solution Framework



Edit Distance

Edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.



Damerau-Levenshtein similarity

Damerau-Levenshtein distance is a distance (string metric) between two strings, i.e., finite sequence of symbols, given by counting the minimum number of operations needed to transform one string into the other, where an operation is defined as an insertion, deletion, or substitution of a single character, or a transposition of two adjacent characters.

$$f^{sim}_{list}(x_1, x_2) = 1 - \frac{f^{dist}_{D\text{-}L}(x_1, x_2)}{\max(|x_1|, |x_2|)}$$

6 Results

- KNN outperforms ATS- based approach, but the latter is more stable
- Both are better than random work.
- Performance plateaus with longer suffixes

7 Acknowledgement

The project was supervised by Fabrizio Maggi. The studies of Andrii Rozumnyi and Dmytro Chasovskyi are supported by the Estonian Foreign Ministry's Development Cooperation and Humanitarian Aid funds.

8 References

[1] Polato, Mirko, et al. "Time and Activity Sequence Prediction of Business Process Instances." arXiv preprint arXiv:1602.07566 (2016).

[2] Van der Aalst, Wil MP, M. Helen Schonenberg, and Minseok Song. "Time prediction based on process mining." Information Systems 36.2 (2011): 450-475.

[3] SPiCe - Sequence Prediction ChallengE http://spice.lif.univ-mrs.fr/

Repo: https://github.com/JaakTree/predict_sequence_completion