Log sequence clustering for workflow mining in multi-workflow systems

Xumin Liu, Moayad Alshangiti, Chen Ding, Qi Yu

PII: S0169-023X(16)30135-5

DOI: 10.1016/j.datak.2018.04.002

Reference: DATAK 1641

To appear in: *Data & Knowledge Engineering*

Received Date: 5 August 2016

Revised Date: 23 March 2018

Accepted Date: 3 April 2018

Please cite this article as: X. Liu, M. Alshangiti, C. Ding, Q. Yu, Log sequence clustering for workflow mining in multi-workflow systems, *Data & Knowledge Engineering* (2018), doi: 10.1016/j.datak.2018.04.002.

Log Sequence Clustering for Workflow Mining in Multi-Workflow Systems

Xumin Liu [1], Moayad Alshangiti [1], Chen Ding [2], Qi Yu [1]

[1] *Golisano College of Computing and Information Science, Rochester Institute of Technology, USA*

[2] *Department of Computer Science, Ryerson University, Canada*

Abstract

Current workflow mining efforts aim to discover process knowledge from user-system interaction logs and represent it as high-level workflow models. They assume there is one single workflow model in a system, or rely on the information that can explicitly link each log sequence to the underlying workflow model. Such assumptions may not be applicable to multi-workflow systems where the instances of different workflow models are mixed together without being differentiated. To address this issue, this paper proposes to apply sequence clustering methods to group similar log sequences together. Each sequence cluster corresponds to a workflow model and the log sequences in the cluster are the corresponding instances. This paper investigates different similarity measures, including structure-based and user-based, as well as different clustering algorithms, including one-side clustering and co-clustering. In order to incorporate user factors into sequence clustering, which is novel to the current sequence clustering methods, this paper proposes to model User Behavior Patterns (UBPs) as probabilistic distributions over sequences and learn it from the event log. We represent a UBP as a Probabilistic Suffix Tree and use it to measure sequence similarity. The co-clustering method leverages the dyad relationship between UBPs and log sequences to improve the clustering accuracy. An experimental study has been conducted and the result indicates that user-based methods outperform structure-based methods in terms of accuracy and they are more effective on dealing with noises in the log and the increase of log size. The UBP-sequence co-clustering method achieves the best performance which indicates the effectiveness of incorporating user factors and applying co-clustering.

Keywords: Workflow Mining, Sequence Clustering, User Behavior Pattern, Probabilistic Suffix Tree, Non-negative Matrix Factorization