AN NOVEL SIMILARITY MEASURE FOR TRACE CLUSTERING BASED ON NORMALIZED GOOGLE DISTANCE

Hong Nhung BUI*, Quang-Thuy HA, Tri-Thanh NGUYEN

VNU-University of Engineering and Technology, Hanoi, Vietnam

Abstract

In trace clustering, a problem of process mining, traditional distance measures only focus on the local relationship between trace pairs. In this paper, we propose a new method to measure the global relationship of the traces based on the Normalized Google Distance. Experimental results show that our method not only outperforms alternatives but also helps to speed up the trace clustering.

I. Introduction

Process discovery, i.e., construct the business process model from an event log, is one of the most important tasks of process mining. Since the output of this task is used in several other tasks, the performance of these tasks depends heavily on the quality of the generated process model. Many studies have shown that trace clustering is one of the most effective solution to improve the quality of the generated process model when the input event log is complex, large and heterogeneous [2, 3, 4]. Concretely, instead of discovering the process directly from the whole event log, the event log is divided into a set of clusters containing similar events. A process discovery algorithm is applied on each cluster to figure out sub-process models. Consequently, in this approach, the clustering step plays in important role in this task.

Common algorithms used for trace clustering depends greatly on the distance measure between trace pairs. Traditional distance measures focus

* Corresponding Author, email: nhungbth@hvnh.edu.vn
Keywords and phrases: process mining, process discovery, trace clustering, normalized Google distance, similarity measure.

on calculating the local relationship between two traces, i.e., only the features of the two vectors are used for calculation [7]. To overcome this issue, we propose a new approach to consider global relationship of traces in a whole event log based on the Normalized Google Distance [6] in distance calculation.

II. The Normalized Google Distance

Normalized Google Distance (NGD) is a relative semantic distance that was suggested by Cilibrasi and Vitányi [1, 6]. It can calculate the similarity between two terms in a natural language based on their context of use on the world-wide-web by using the Google search engine.

The distance between the terms x and y (called the normalized Google distance) is defined as

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (1)$$

where *f(x)* and *f(x,y)* denote the number of web pages containing *x* and both *x* and *y*, correspondingly; *N* is the total number of web pages indexed by the Google search engine. NGD obviously takes the whole indexed webpages (i.e., global semantic of *x* and *y*) into account.

III. The normalized trace weight (NTW) based on NGD

An event log is a collection of events each of which is composed of a set of properties (e.g., *event id*, *case id*, *activity*, *timestamp*, etc.). A sequence of events with the same *case id* is a process instance, and is referred to as a *trace*. When we consider on activities, then a trace can be represented by sequence of activities, then the event log is a collection of traces. We propose to map *activity* and *trace* to a *term* and a *web page,* correspondingly to exploit the advantage of NGD. In traditional approach, a trace can be represented as a vector of k-grams for clustering, i.e., distance is calculated between every two vectors. We propose to estimate the weight of each k-gram and a trace (in global context) as follows:

Definition 1. The normalize weight of a 1-gram activity *x* is

$$NW(x) = \log(f(x))/(\log(N) - \log(f(x)) \quad (2)$$

Definition 2. The normalize weight of 2-gram activities *xy* is

$$NW(xy) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (3)$$

Definition 3. The normalize weight of 3-gram activities *xyz* is

$$NW(xyz) = \frac{\max\{\log f(x), \log f(y), \log f(z)\} - \log f(x,y,z)}{\log N - \min\{\log f(x), \log f(y), \log f(z)\}} \quad (4)$$

where *f(x), f(x,y),* and *f(x,y,z)* denote the number of traces containing *x*; (*x,y*), and (*x, y*, *z),* correspondingly; *N* is the total number of traces.

Definition 4. The normalize trace weight of *t* is

$$NTW(t) = \sum_{kgram_i \in t} NW(kgram_i) / |kgram_i \in t| \quad (5)$$

In this approach, each trace is represented by a real number not a vector, hence, the complexity of clustering algorithms is reduced. The calculation of *NW*(.) can be effectively implemented using FP-tree [9], which is fast.

Let *L= {abdeh, agbch, acdcfdbc, dcdeg, acdgfdcgf, acgefbgeg}* be an event log. The first trace *t=abdeh* is represented in 1-gram, 2-gram, and 3-gram as {*a,b,d,e,h*}, {*ab,bd,de,eh*}, and {*abd, bde, deh*}, respectively. Applying the corresponding formula pairs (2, 5), (3, 5), and (4, 5) for each k-gram set, the *NTW*(*t*) is *3.58*, *1.24,* and *1.84*, respectively.

VI. Experimental results

In this paper, the three-phase framework, i.e., "Trace representation and Clustering", "Process discovery", and "Model evaluation" for process discovery was used [5] for experiments. Binary vector space with Euclidean distance was also implemented as the baseline for evaluation. Instead of evaluating the clustering, we evaluate the generated processes using the *fitness* and *precision* measures [8]. Three different event logs (Lfull [1], prAm6 [2], prHm6 [3]), and K-means clustering algorithm were

[1]www.processmining.org/event_logs_and_models_used_in_book/Chapter7.zip
[2]http://data.3tu.nl/repository/uuid:44c32783-15d0-4dbd-af8a-78b97be3de49

adopted for experiments. Table 1 describes some characteristics of these event logs. Three trace representation methods, i.e., 1-gram, 2-gram and 3-gram, were used.

Table 1. The characteristics of three event logs

Event log	#cases	#events	Characteristics
Lfull	1391	7539	Duplicated traces, repeated activities in a trace
prAm6	1200	49792	Few duplicated traces, no-repeated activities
prHm6	1155	1720	No-duplicated traces, no-repeated activities

Table 2. Results of Euclidean and NTW measures

Event log / Measure	**Lfull**		**prAm6**		**prHm6**	
Euclid	*Fitness*	*Precision*	*Fitness*	*Precision*	*Fitness*	*Precision*
1-gram	0.991	0.754	0.968	**0.809**	0.902	0.66
2-gram	0.951	**0.958**	0.968	0.809	0.902	0.66
3-gram	0.955	**0.962**	0.968	0.809	0.902	0.66
NTW	*Fitness*	*Precision*	*Fitness*	*Precision*	*Fitness*	*Precision*
1-gram	**0.994**	**0.806**	**0.97**	0.606	**0.919**	**0.795**
2-gram	**0.995**	0.913	**0.972**	**0.995**	**0.921**	**0.809**
3-gram	**0.9999**	0.93	**0.973**	**0.933**	**0.911**	**0.861**

The experimental results in Table 2 indicate that NTW has better performance. The fitness of NTW based trace representation is always higher than that of Euclid. When the duplicated traces in the event log are high, the NTW exposes its effectiveness on the fitness. The precision is influenced by the characteristic of duplicated traces and repeated activities of event logs. In the case of prHm6, an event log without duplicated traces and repeated activities, the precision of NTW based is higher than that of Euclid based trace representation.

[3]http://data.3tu.nl/repository/uuid:44c32783-15d0-4dbd-af8a-78b97be3de49

V. Conclusions and future work

This paper proposed a normalized trace weight measure, which takes global context into account for clustering traces in event logs. The experimental results prove that the global context helps to improve the performance of process discovery task. The results are comparable to state-of-the-art approaches. With the ability of transforming a trace vector into a single value, the complexity of clustering algorithm is reduced.

This NTW is just the preliminary version, further refined version should be studied to make it better. Another direction, the order of activities in a trace should be incorporated in the next version of NTW.

References

[1] A. R. Cohen, and P. M. B. Vitányi, Normalized Google Distance of Multisets with Applications, IEEE Trans. Pattern Anal. Mach. Intell (2015), 1602-1614.

[2] I. Fischer, and J. Poland, New Methods for Spectral Clustering. In Proc. ISDIA (2004).

[3] J. Poland, and T. Zeugmann, Clustering the Google Distance with Eigenvectors and Semidefinite Programming, Knowledge Media Technologies (2006), pp. 61–69.

[4] P. D. Koninck, et al. Explaining clusterings of process instances, Data Min. Knowl. Discov. 31(2017), pp. 774-808.

[5] Q. T. Ha, et al., A trace clustering solution based on using the distance graph model, In proceedings of ICCCI (2016), pp. 313-322.

[6] Rudi L. Cilibrasi, and Paul M.B. Vitanyi, The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering (2007), pp. 370-383.

[7] R. P. J. C. Bose and WMP V. Aalst, Context Aware Trace Clustering: Towards Improving Process Mining Results, SDM (2009), pp. 401-412.

[8] WMP V. Aalst. Process Mining: Data Science in Action (2nd edition). Springer, (2016).

[9] J. Han, et al. Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD'00, Dallas, TX). ACM Press, (2000).