A Contextual Edit Distance for Semantic Trajectories

Clement Moreau, Thomas Devogele, Veronika Peralta and Laurent Etienne
University of Tours
Blois, France
firstname.lastname@univ-tours.fr

ABSTRACT

The understanding of daily human activity is an active research topic. Thanks to GPS and smartphones, human movements can be monitored and analyzed. In addition, by exploiting Linked Open Data and user personal data, semantic labels and annotations can be added to movements. Thus, semantic trajectories can be considered as sequences of timestamped activities where each activity is described by a semantic label. In this context, a major challenge is the comparison of such semantic trajectories, looking to extract and learning similar human mobility behaviors. We propose *CED (Contextual Edit Distance)*, a generic similarity measure for semantic sequences comparison which improve the Edit Distance to take into account the context similarity between elements in the sequence. CED is configurable to any sequence data and business needs.

CCS CONCEPTS

• **Information systems** → **Similarity measures**; *Geographic information systems.*

KEYWORDS

edit distance, semantic trajectory, human behavior analysis, similarity measure, clustering

ACM Reference Format:
Clement Moreau, Thomas Devogele, Veronika Peralta and Laurent Etienne. 2020. A Contextual Edit Distance for Semantic Trajectories. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3341105.3374125

1 INTRODUCTION

Semantic trajectories enrich the classical trajectories (sequences of space-time positions [7]) with semantic labels representing activities [11], POI [5] or sets of aspects [1, 3, 8]. They allow to model richer aspects of human behavior that were not captured by traditional trajectories. In particular, they allow the semantic comparison of trajectories, in order to capture frequent behavior, detect atypical moves, and more generally, understand human activity. For example, from a sociological point of view, two trajectories representing two people' moves "home-football" and "home-basketball" may represent similar behavior (trips to play sport), even if they live at different neighbourhoods and move at different times. Therefore, *similarity measures*, which are at the core of several mobility data mining methods, should take advantage of semantic labels in the comparison of semantic trajectories. There exists many measures to compare semantic trajectories like [3, 4, 6]. However, They generally use trivial distances (e.g binary, euclidean) between symbols of the sequences. Moreover, they do not take into account the notion of context to compare the sequences. We understand the term *context* as the semantic content, or a portion, of the sequence (i.e. trajectory).

In this paper we present a new distance for comparing semantic trajectories. Thus, we modify the Edit Distance to build a new similarity measure, a Contextual Edit Distance (CED) between semantic trajectories. The running example (figure 1) illustrate CED results.

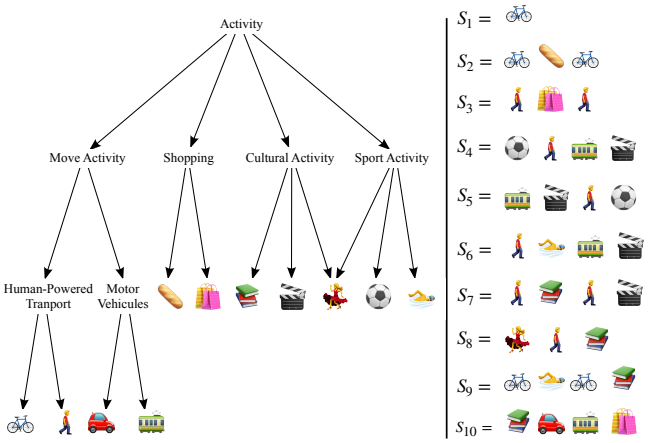

Figure 1: Ten semantic trajectories and the associated concept hierarchy based on an simple ontology

In the rest of this paper, we note Σ a set of symbols e.g 🏊 or 🎬. $S \in \Sigma^n$ is a finite sequence of symbols, $|S| = n$ denotes the length of sequence S. The main interest of CED is the three following properties:

Property 1 (Semantic similarity). *We consider there exists a similarity metric* $sim : \Sigma \times \Sigma \rightarrow [0, 1]$ *between symbols and quantify their semantic proximity. This sim measure can be obtained by building a directed acyclic graph (e.g. a concept hierarchy based on an ontology) on* Σ*'s elements [2].*

Property 2 (Contextual semantic edition). *Given a semantic sequence* $S \in \Sigma^n$ *and a symbol* $a \in \Sigma$ *to edit in* S*. The edition cost of* a *depends on the similarity of nearby symbols in* S*.*

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03.
https://doi.org/10.1145/3341105.3374125

For example, similarity between S_7 and S_8 is high. Indeed, the main difference is the presence of 🎬 $\in S_7$ and 🐓 $\in S_8$. Nevertheless, theses two cultural activities are closed.

Property 3 (Repetition and permutation). *Given a semantic sequence $S \in \Sigma^n$ and a symbol $a \in \Sigma$ to edit in S. Edition or permutations of repeated close symbols has little cost.*

For example, similarity between S_1 and S_2 is good enough. Indeed, 🚲 is repeated. The adding cost repeated activities is low. Although S_4, S_5 and S_6 are made up of the same or close symbols but swapped, so we can say they are similar.

Contribution. In this paper we present a new similarity measure inspired by the Edit Distance, the Contextual Edit Distance (CED), for semantic sequences apply on semantic trajectories. This measure focus on the semantic dimension and takes into account the similarity between symbols of the sequence and the context. The remainder of this article is organized as follows. Section 2 mathematically describes the CED while Section 3 presents a simple clustering process to illustrate the advantages of this measure. Section 4 concludes the paper.

2 CONTEXTUAL EDIT DISTANCE

Let Σ be a finite alphabet of symbols. $S \in \Sigma^n$ is a finite sequence of symbols, $|S| = n$ denotes the length of sequence S. Furthermore, we consider s_i for $i \in [\![1, n]\!]$ denotes the i^{th} symbol of sequence $S = \langle s_1, s_2, ..., s_n \rangle$.

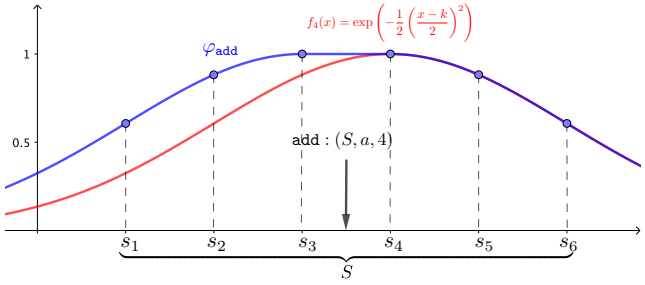

Figure 2: Add a symbol a at position 4 in the sequence S

Definition 1 (Contextual edit operation). *A contextual edit operation e is a function $e : (\Sigma^* \times \Sigma \cup \{\varepsilon\} \times \mathbb{N}) \to \Sigma^*$ whose arguments are a sequence, a new symbol to be included in the sequence (or none) and the index (position) in the sequence where the edition takes place.*

We consider the following set $O = \{\texttt{mod}, \texttt{add}, \texttt{del}\}$ of contextual edit operations:

- $\texttt{mod} : (S, a, k) \mapsto \langle s_1, ..., s_{k-1}, a, s_{k+1}, ..., s_n \rangle$
  Replace the symbol at index k by the new symbol a.
- $\texttt{add} : (S, a, k) \mapsto \langle s_1, ..., s_{k-1}, a, s_k, ..., s_n \rangle$
  Insert symbol a at index k. The symbols at and after index k are shifted forward.
- $\texttt{del} : (e, \varepsilon, k) \mapsto \langle s_1, ..., s_{k-1}, s_{k+1}, ..., s_n \rangle$
  Delete the symbol at position k. The symbols after position k are shifted backward.

In order to take into account duplication, contextual information and permutation, we introduce the notion of context vector and context function that control the level of temporal proximity of two activities. Thus, thanks to a context vector, relationship factor is computed.

Definition 2 (Context function). *Consider a contextual edit operation $e(S, a, k)$ and a function $f_k : \mathbb{R} \to [0, 1]$ which holds the following properties:*

1. *$f_k(k) = 1$.*
2. *f_k is a monotonically increasing function on $]-\infty, k]$.*
3. *f_k is symmetrical center on k.*

The third property guarantees to take with the same importance previous and future symbols located at equal distance from s_k.

We denote f_k as context function. A context edit function $\varphi_e : \mathbb{N}^* \to [0, 1]$ is a transformation of f_k stretching the function according to the type of operation e. We distinguish three cases for e in $\{\texttt{mod}, \texttt{add}, \texttt{del}\}$:

- $\varphi_{\texttt{mod}}(x) = f_k(x)$
- $\varphi_{\texttt{add}}(x) = \begin{cases} f_k(x+1) & \text{if } x \leq k-1 \\ f_k(x) & \text{if } x \geq k \end{cases}$
- $\varphi_{\texttt{del}}(x) = \begin{cases} f_k(x+1) & \text{if } x \leq k-1 \\ 0 & \text{if } x = k \\ f_k(x-1) & \text{if } x \geq k+1 \end{cases}$

Finally, a context vector $v : E \to [0, 1]^n$ is defined as

$$v(e) = \langle v_1, \ldots, v_n \rangle$$

where $v_i = \varphi_e(i)$.

Intuitively, the context vector quantifies the relationship between symbol s_k and another symbol s_i. Bigger $|k-i|$ is, lesser the symbols s_i has an impact on s_k. Thus, the relationship factor weighted the similarity values between two symbols.

The cost function γ of CED generalizes the classical definition of Edit Distance and takes into account the local context of each query in the exploration.

Definition 1 (Cost function γ). Given an operator $e(S, a, k)$ a cost function $\gamma : E \to [0, 1]$ for the contextual edit operations is defined as:

$$\gamma(e) = \alpha \times \delta(e) + (1 - \alpha)\left(1 - \max_{i \in [\![1,n]\!]} \{sim(s_i, a) \times v_i(e)\}\right) \quad (1)$$

where:

- $\alpha \in [0, 1]$ is a contextual parameter.
  If $\alpha \to 0$ the contextual part is maximal and therefore the distance between two queries will be strongly evaluated according to the content of the sequence being edited ; if $\alpha \to 1$ then cost of edition is fixed and CED is equivalent to Edit Distance.
- $\delta(e) = \begin{cases} 1 - sim(s_k, x) & if\ e = \texttt{mod} \\ 1 & else \end{cases}$
  is the local cost of the Edit Distance.

Definition 2 (Edit path). Given two sequences $S, S' \in \Sigma^*$, an edit path $P = \langle e_1, e_2, ..., e_m \rangle$ from S to S' is a sequence of operations that transform S in S'. We note $\mathcal{P}(S, S')$ the set of all edit paths to transform S in S'.

An important point to be mentioned here is that, for some reasons of complexity and computation time, the calculation of context vector is based on the start sequence S i.e $S^{(0)}$. Moreover, there is an asymmetric relation between elements due to the fact that add operator is not the reverse operator of del and vice versa.
Thus, by re-using the classic definition of the Edit Distance [9], we can define the one-side distance from S_1 to S_2.

Definition 3 (One-sided Contextual Edit Distance). *Let* $\tilde{d}_{CED} : \Sigma^n \times \Sigma^p \rightarrow \mathbb{R}^+$ *be the Contextual Edit Distance from* S_1 *to* S_2 *such that:*

$$\tilde{d}_{CED}(S_1, S_2) = \min_{P \in \mathcal{P}(S_1, S_2)} \left\{ \sum_{i=1}^{|P|} \gamma(e_i) \right\} \quad (2)$$

To save the symmetry, we use the same trick as Hausdorff distance and we apply the max operator on each one-sided contextual edit distance.

Definition 4 (Contextual Edit Distance). *Let* $d_{CED} : \Sigma^n \times \Sigma^p \rightarrow \mathbb{R}^+$ *be the Contextual Edit Distance such that:*

$$d_{CED}(S_1, S_2) = \max \left\{ \tilde{d}_{CED}(S_1, S_2), \tilde{d}_{CED}(S_2, S_1) \right\} \quad (3)$$

About time complexity, we re-use Wagner and Fisher algorithm and dynamic programming approach [9]. This guarantees CED a polynomial complexity in $O(n \times p \times \max(n, p))$.

3 RUNNING EXAMPLE

To compute the similarity between concepts for this running example, we choose the Wu-Palmer similarity [10]. The main goal of our experiments is to compare CED and Edit Distance results for trajectories having several degrees of similarities and confirm properties in section 2. The dendogram produced with Edit Distance is a baseline for understanding the interest and properties of CED. Figure 3 shows the resulting dendrograms with Ward agglomeration criteria performed in Python. To obtain this dendrogram, we set up CED with $\alpha = 0$ and the same context function f_k show in Figure 2. We highlight several interesting results of the analysis of CED values between some sets of sequences of activities . First, the contextual information is effectively used and has an impact in the computation of CED. For example, S_7, S_8 have a similar context : Two cultural activities and walking. And, indeed $d_{CED}(S_7, S_8) < d_{ED}(S_7, S_8)$. We assume that this couple is correctly grouped together with CED. In addition, if an activity is split by a new activity, only the cost of adding this new activity is added. We can notice this property on S_6, S_7 but also on the cluster $\{S_1, S_2, S_3\}$ with S_1 and S_2. Finally, the cluster $\{S_4, S_5, S_6\}$ is composed by sequences with same (or close e.g. foot, swim) elements but with different order. This set confirms the permutation property. Globally, CED changes the topology of the space by bringing together sequences with similar content. To summarize, using CED as similarity measure significantly improves the clustering process of theses semantic trajectories according to properties claim in introduction. In the same way, CED improves outlier detection processes.

4 CONCLUSION

This paper addressed the problem of comparing semantic trajectories, which is a hot topic for the understanding of human behavior. Concretely, it defines a new measure, CED, to compute the distance between semantic sequences. This measure extends the Edit Distance and its operations by taking context into account. It uses ontologies to describe symbols and to calculate the similarity between them. Thus, new aspects are taken into account as semantic similarity, local context of each activity and permutations. Moreover, the polynomial complexity of our algorithm (based on Dynamic Programming) gives hope to process large data sets.

In near future, CED will be tested on two large data sets concerning two real application domains: tourism circuits for recommendation and children daily mobility for clustering behaviour. We would also consider time aspect of symbol in order to express, for example, the duration of an activity. Furthermore, we will investigate other domains of application of CED, for example to measure the contextual similarity between log files.

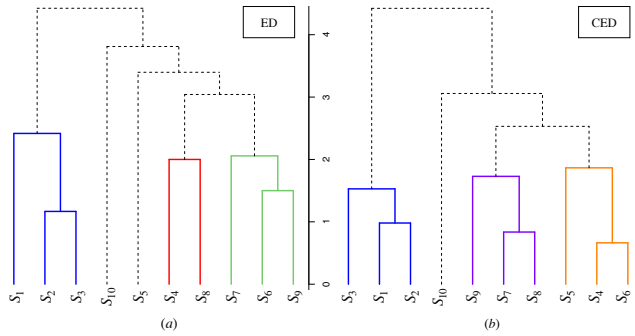

Figure 3: Hierarchical clustering based on (*a*) Edit Distance and (*b*) CED. Cut based on the higher relative loss of inertia.

ACKNOWLEDGMENTS

This project is funded by the ANR Mobi'kids and the CVL region (FRANCE).

REFERENCES

[1] V. Bogorny, C. Renso, A. Ribeiro de Aquino, F. de Lucca Siqueira, and LO. Alvares. 2014. CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. *Transactions in GIS* 18 (2014), 66–88.
[2] M. M. Deza and E. Deza. 2016. *Encyclopedia of Distances.* Springer, Chapter 22 - Distances in Internet and Similar Networks.
[3] R. dos Santos Mello, V. Bogorny, LO. Alvares, L.HZ Santana, CA. Ferrero, AA. Frozza, GA. Schreiner, and C. Renso. 2019. MASTER: A multiple aspect view on trajectories. *Transactions in GIS* (2019), 1–20.
[4] AS. Furtado, D. Kopanaki, LO. Alvares, and V. Bogorny. 2016. Multidimensional Similarity Measuring for Semantic Trajectories. *Transactions in GIS* 20 (2016), 280–298.
[5] F. Gianotti, M. Nanni, D. Pedreschi, and F. Pinelli. 2007. Trajectory Pattern Mining. *ACM SIGKDD* (2007), 330–339.
[6] A.L. Lehmann, L.O. Alvares, and V. Bogorny. 2019. SMSM: a similarity measure for trajectory stops and moves. *IJGIS* 33, 9 (2019), 1847–1872.
[7] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, V. Bogorny, ML. Damiani, A. Gkoulalas-Divanis, JA. de Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. 2013. Semantic trajectories modeling and analysis. *Comput. Surveys* 45 (2013), 1–32.
[8] LM. Petry, CA. Ferrero, LO. Alvares, C. Renso, and V. Bogorny. 2019. Towards Semantic-Aware Multiple-Aspect Trajectory Similarity Measuring. *Transactions in GIS* (2019).
[9] R. Wagner and M. Fisher. 1974. The String-to-String Correction Problem. *J. ACM* 21 (1974), 168–173.
[10] Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* 133–138.
[11] C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta. 2014. Splitter: Mining Fine-Grained Sequential Patterns in Semantic Trajectories. *VLDB Endowment* 7 (2014), 769–780.