Contextual Edit Distance for Semantic Trajectories

An enrichment of string metric for semantic contextual comparison

Clément Moreau [1], Thomas Devogele[1], Verónika Peralta[1], Laurent Etienne[1,2]

[1]*LIFAT, University of Tours, Tours – France*

[2]*ISEN, Brest – France*

firstname.lastname@univ-tours.fr

Context

The fine 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**.

Contribution

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 **generic** and can be adapted to different contexts and business needs. It can be used as a building block for defining more complete distances, in particular, combining with well-known spatial and temporal metrics in GIS context.

Motivations

We believe that the edition of sequences should be done in a contextual way, by taking into account the other symbols in the sequences. Thus, we claim these three following properties:

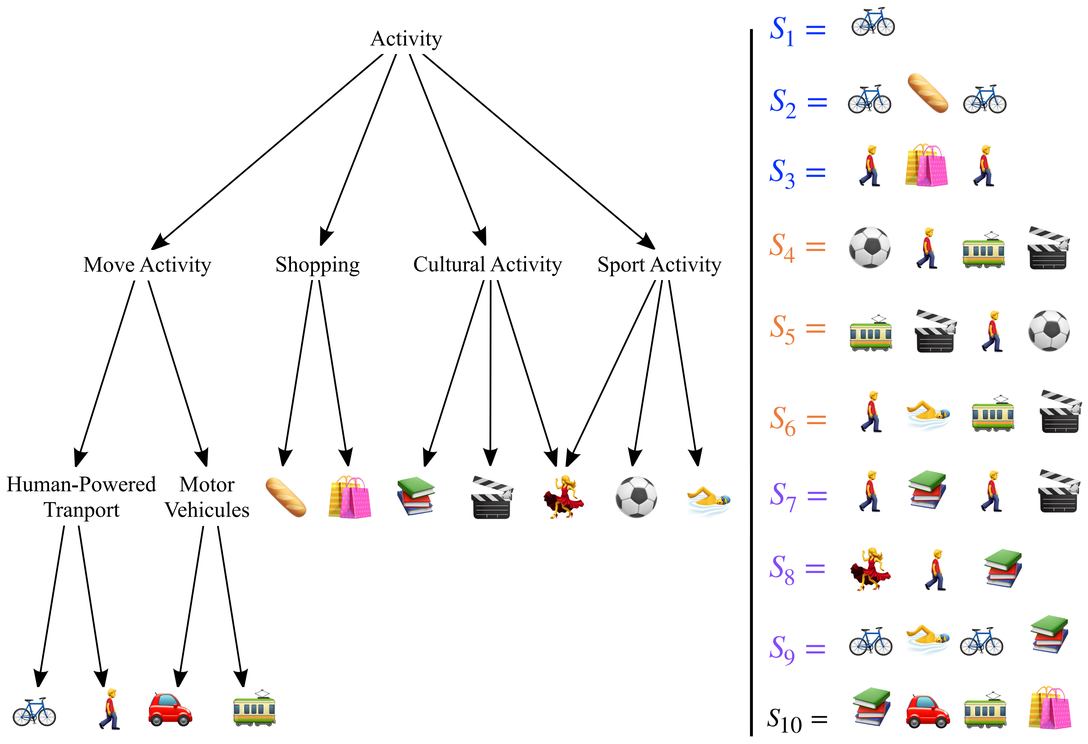

Figure 1: **On the left, symbols of activity organized as an ontology ; On the right, set of work-home semantic trajectories.**

Property 1 (Semantic similarity)**.** *Let Σ be an alphabet. Elements of Σ are represented in a directed acyclic graph to perform a* ***similarity function***

$$sim : \Sigma \times \Sigma \to [0, 1]$$

between symbols and quantify their semantic proximity [Deza and Deza, 2016].

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.***

- Sequences S_7 and S_8 are closed ⇒ Two Cultural activities : 🎬 $\in S_7$ and 🐓 $\in S_8$,
- Sequences S_8 and S_9 are closed ⇒ Two Sport activities 🐓 $\in S_8$ and 🏊 $\in S_9$, and Human-Powered transport 🚶 $\in S_8$ and 🚲 $\in S_9$.
- Sequences S_2 and S_3 are closed ⇒ Human-Powered transport and Shopping activities.

Property 3 (Repetition and permutation)**.** *Given a semantic sequence $S \in \Sigma^n$ and a symbol $a \in \Sigma$ to edit in S.* ***Edition of repeated close symbols has little cost.*** ***Permutations*** *are a corollary of this property and Property 2.*

- Sequences S_1 and S_2 are closed ⇒ 🚲 is repeated.
- Sequences S_4, S_5 and S_6 are closed ⇒ Permutation of closed symbols (Sport activities ⚽ and 🏊.)

Contextual Edit Distance

The **Contextual Edit Distance** is string metric that extended the Edit Distance.
CED deals with set E of three common **contetual edit operations** (`mod`ification, `add`ition, `del`etion) $e : (S, a, k)$ where :

- $S \in \Sigma^n$ is the edited sequence.
- $a \in \Sigma$ a symbol to edit.
- $k \in \mathbb{N}^*$ an index in S.

In order to take into account the notion of contextual proximity, these operations are based on:

- A **contextual function** $f_k : \mathbb{N}^* \to [0, 1]$ which quantifies the relationship between an edited symbol x and another one in S. The greater the value of f_k, the greater the impact of x in S.
- 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 contextual edit operation e.
- A **contextual vector** $v : E \to [0, 1]^n$ which define a proximity coefficient for each symbol by the context edit function such that $v(e) = \langle \varphi_e(i) | i \in [\![1, n]\!] \rangle$. We note $v_i(e) = \varphi_e(i)$.

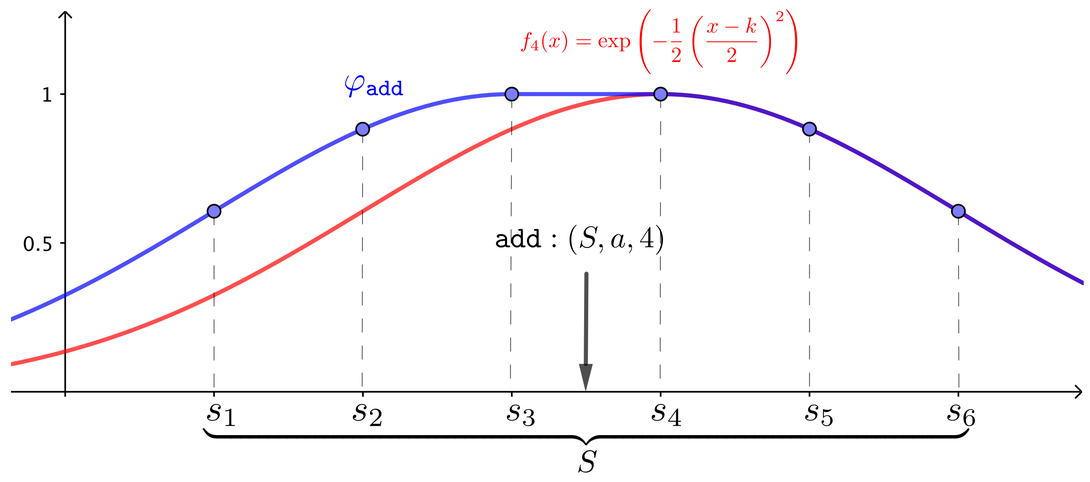

Figure 2: **Add a symbol a at position 4 in the sequence S**

- A **contextual weight** $\alpha \in [0, 1]$ such that if $\alpha \to 0$ the contextual part is maximal ; if $\alpha \to 1$ then CED is equivalent to Edit Distance.

Finally, each Contextual Edit Operation has a **cost** of application $\gamma : E \to [0, 1]$.

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

Thus, the **Contextual Edit Distance** $d_{CED} : \Sigma^n \times \Sigma^p \to \mathbb{R}^+$ is such that:

$$d_{CED}(S_1, S_2) = \max \left\{ \min_{P \in \mathcal{P}(S_1,S_2)} \left\{ \sum_{i=1}^{|P|} \gamma(e_i) \right\}, \min_{P \in \mathcal{P}(S_2,S_1)} \left\{ \sum_{i=1}^{|P|} \gamma(e_i) \right\} \right\}$$

Where $P \in \mathcal{P}(S, S')$ is an edit path $P = (e_1, ..., e_N)$ from S to S'.

The computation of CED is based on Dynamic Programming method [Wagner & Fischer, Journal of the ACM 1974] and guaranteed a **polynomial complexity** in $O(n \times p \times \max(n, p))$.

Experiments

Parameters.

- The following semantic trajectories and ontology detail in Figure 1.
- **Wu-Palmer similarity** [Wu and Palmer, ACL 1994] between symbols.
- A **Contextual function** f_k shows in Figure 2.
- α is set to 0 to take full context into account.
- Hierarchical clustering with Ward aggregation criterion.

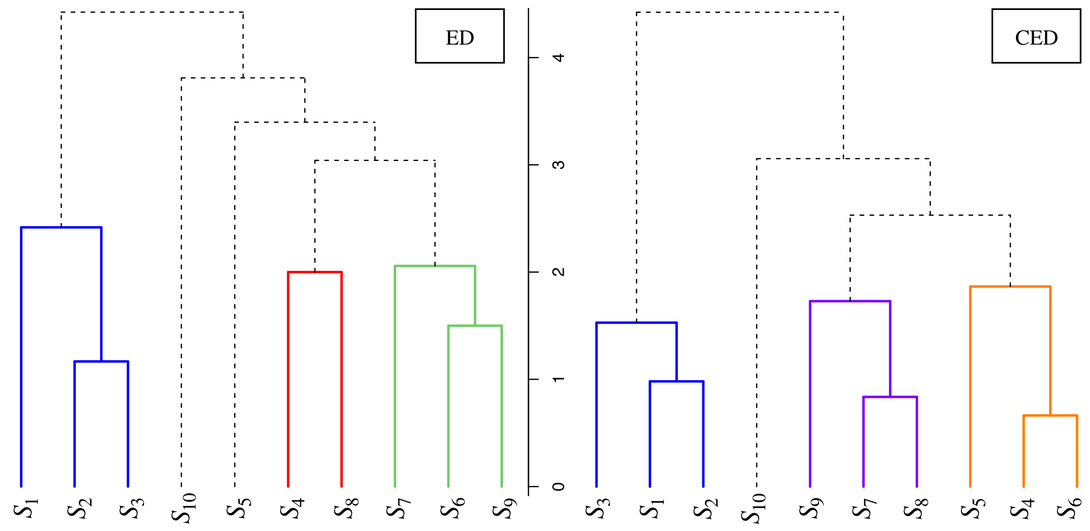

Figure 3: **On the left, HC done with classical Edit distance (CED with $\alpha = 1$); On the right, HC done with CED distance with previous desribed settings. The cut is based on the higher relative loss of inertia criteria.**

Findings.

- CED changes the space topology compared to the Edit Distance and reduces the distance between elements.
- CED implements Properties 1, 2 and 3 and the trajectories are well clustered as described in Motivations section.