Clustering sequences of multi-dimensional sets of semantic elements

Clément Moreau, Alexandre Chanson, Veronika Peralta, Thomas Devogele, Cyril de Runz

Clustering Sequences of Multi-dimensional Sets of Semantic Elements

Clement Moreau, Alexandre Chanson, Verónika Peralta, Thomas Devogele and Cyril de Runz
University of Tours
Blois, France
firstname.lastname@univ-tours.fr

ABSTRACT

The study of semantic aspects of human behavior is an hot topic. Most of the time, semantic sequences describe these complex behaviors. Indeed, sequences include several information as type of human activities or places. To study these complex data, we need to define new similarity measures and select appropriate clustering processes. This article proposes a semantic similarity measure, based on ontologies, which manages complex semantic elements with different levels of detail and incertitude. An application of this approach from the domain of touristic mobility shows the interest of this process.

CCS CONCEPTS

• **Information systems** → **Clustering**; **Similarity measures**; • **Computing methodologies** → *Ontology engineering*;

KEYWORDS

Clustering, Data mining, Edit distance, Human behavior, Semantic sequences, Similarity measure, UMAP

ACM Reference Format:
Clement Moreau, Alexandre Chanson, Verónika Peralta, Thomas Devogele and Cyril de Runz. 2021. Clustering Sequences of Multi-dimensional Sets of Semantic Elements. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3412841.3441920

1 INTRODUCTION

The analysis and clustering of semantic sequences, representing sequences of human activities, is a hot topic receiving increasing interest in many research communities (e.g. machine learning, information systems, but also social sciences and psychology). Indeed, semantic sequences are widely used for representing sequences of varied types of elements, as semantic trajectories [24], music playlists [11], database exploration logs [32], among others, and are at the kernel of many techniques for recommendation [8, 16], prediction [41] or outlier detection [9]. Remark that to keep general, the *semantic elements* in the sequence may represent many types of human activities but also other types of entities, like listened songs or visited points of interest.

Yet, the main interest of these studies on semantic sequences, in particular of semantic mobility sequences, is the clustering of the dataset in order to extract coherent and understanding behaviors or patterns over the sequences. [17, 21, 31, 35]. Moreover, a difficult problem in this context is the comparison of semantic sequences, generally concerning complex semantic elements.

As a motivating example, consider the two following scenarios:

Alice is on vacation in the Loire Valley region. She leaves her Relais & Chateaux hotel to visit a museum on religion in the Middle Ages during the morning, then she has a picnic in a park. In the afternoon, she visits a castle and a church. In the evening she has dinner at a pub / restaurant before going to a sound and light show at a renaissance castle. She finally returns to her hotel.

Bob is also on vacation in the Loire Valley region. Bob is staying in a campsite. In the morning he visits a castle with flower gardens. He has lunch in a bistro. In the afternoon he goes to a baroque concert at a gothic church and then takes a guided tour of the local town by bicycle. In the evening he has dinner again in a bistro before returning to his campsite.

How similar are these two trips ? The comparison of sequences of semantic elements is a hard problem. Indeed, the fine-grained comparison of semantic elements, like a *castle* and a *church* is already a challenge and many techniques for the semantic comparison have been proposed [48]. This difficulty is increased when the elements to be compared are no longer one-dimensional but multi-dimensional, i.e. they are defined by means of several *concepts* expressing diverse semantics. For example, how to compare a "*Sound and light show at a Renaissance castle*" and a "*Baroque concert at a Gothic church*" ? Should the comparison consider the type of event (show vs. concert), the place (castle vs. church) or the architectural style (renaissance vs. gothic) ? Furthermore, some elements may concern a set of concepts for a same dimension, for example a *castle with flower gardens*.

Thanks to domain ontologies capturing specific business needs and the explosion of Semantic Web and Linked Open Data (LOD), the ability to accurately compare complex semantic elements seems more than ever possible. Nevertheless, to the best of our knowledge, the issue of comparing sets of semantic elements that are multi-dimensional, ambiguous, and domain-specific is an open problem [20]. The question remains even more barbed when we are dealing with the comparison of sequences of multi-dimensional sets of semantic elements.

In this paper, we propose a new approach to cluster sequences of multi-dimensional semantic elements. To this end, we extend

P

and combine state of the art similarity metrics, exploiting multiple ontologies for representing multiple semantics, and conforming to typical characteristics of human behavior (ex. redundancy, repetition and cyclicity of activities [40, 41]). We also reuse and combine off the shelf clustering algorithms. Nevertheless, the clustering process is non-trivial. Indeed, the methods for comparing complex sequences form abstract spaces difficult to apprehend. Then, it is relevant to look at several types of algorithms in order to determine the approach best adapted for processing this type of data.

Our contributions are the following:

- A new approach for clustering sequences of multi-dimensional sets of semantic elements.
- A similarity measure for comparing sequences of multi-dimensional sets of semantic elements, which manages elements with different levels of detail and incertitude.
- An application of the approach in the domain of touristic mobility, illustrating the extraction of human behavior.

The remaining of the paper is organized as follows. Section 2 describes related work. Section 3 formulates the clustering problem and proposes similarity measures suited for the comparison of multi-dimensional sets of semantic elements and semantic sequences. Section 4 introduces the case study in tourism domain, describes our experimental protocol and discusses the obtained results. Finally, Section 5 concludes and discusses open challenges.

2 RELATED WORK

In this section, we describe similarity measures proposed for both, comparing semantic concepts and sequences of semantic elements, and review clustering algorithms and techniques commonly used for sequences. Please, remark that, through misuse of language and for the sake of generality, we will use the term *similarity measure* to refer to similarity, dissimilarity, metric or any method to compare entities, regardless of the mathematical properties.

2.1 Similarity measures for semantic concepts

This subsection describes several approaches proposed for comparing semantic concepts and mention the similarity measures more commonly used. All along this subsection, x and y refer to semantic concepts.

Approaches based on **Search engines** compute the similarity between semantic concepts based on the results of search engines. Proposals are essentially based on the work of Cilibrasi et al. [10] and the Normalized Google Distance (NGD) which computes the distance between a given pair of concepts from the number of hits returned by the Google Search engine. This approach has shown interesting experimental results [7]. However, it is very difficult to verify its theoretical properties such as similarity axioms (separability, identity of indiscernible and triangular inequality). Moreover, this approach is not contextual i.e. it is not adapted for a specific business domain.

On the other hand, **Ontology-based** approaches use knowledge graphs such Wikidata[1], Wordnet[2] or other specialized business ontologies to compare concepts. These methods are based on the computation of the shortest path between concepts x and y in the knowledge graph (e.g. Leacock-Chodorow similarity [23]) and the lowest common ancestor of x and y (e.g. Wu-Palmer [46] and Li et al. [26] similarities). Thus, more a concept c is detailed (i.e. its distance $d(c)$ from root node is high) more similarity scores are accurate. Consequently, an inconvenient of this approach is the requirement to have depth graphs. This approach gathers many methods and careful readers can find a survey in [48].

Feature-based approaches are mainly build on the Ratio Model of Jaccard. Considering the set of features related to a concept, these similarities compute the ratio of common features of concepts x and y. The Tversky Ratio Model [42], is a generalization of the Jaccard and Dice models, which also considers the distinctive characteristics of each concept (i.e. the features of one concept which are not part of the other). Feature-based models are applicable in contexts in which entities are or can be represented as sets of features, making them very flexible approaches. However, the semantics of features is not taken into account.

Finally, approaches using **Information Content** are based on the use of corpora and the definition of Information Content (IC) given by [39]. For a concept c and a $corpus$, $IC_{corpus}(c) = -\log p(c)$ where $p(c)$ denotes the probability of encountering subsumers of c in the corpus. More IC is high, more the concept is specific and informative. Resnik similarity [39] computes IC of lowest common ancestor of concepts x and y, the intuition being that concepts sharing a more specific parent concept, share more information and thus are more similar. Lin similarity [27] extends Resnik similarity by computing the ratio among IC of lowest common ancestor and IC of concepts. This approach is complex to implement because it requires large corpora to be efficient. Moreover, these similarities are not normalized in [0,1] and do not respect similarity axioms.

Among these methods, the one best suited for semantic comparison of domain-dependent concepts is the Ontology-based approach. In our proposal, we reuse this method for concept comparison.

2.2 Similarity measures for semantic sequences

Many methods have been proposed for the comparison of categorical sequences. Most of approaches are based on Optimal Maching (OM) Methods [2] and typical measures are LCSS [19], DTW [5] and those of the Edit Distance family [25, 32, 44].

Let remark that although OM methods were not initially proposed for sequences of semantic nor multi-dimensional elements, they can be easily extended to handle them whether providing a similarity measure for comparing complex elements. In this sense, several works propose the use of Euclidean distance combined with LCSS or DTW to compare multi-dimensional elements in time-series [43], Cosine similarity to compare vectors [32] and Hausdorff or Halkidi [18] distances to compare subsets of elements (see [12] for a survey).

The increasing interest of the mobility community in the design of trajectories with semantic information [3, 6, 24, 36] opens new challenges. Indeed, the enrichment of trajectories can be done thanks to applications as SeMiTri [47] or specific domain ontologies like the framework Baquara[2] [14]. Examples of semantic information related to the stops can be, for instance, the Place Of Interest

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page
[2] https://wordnet.princeton.edu/

(POI) category (e.g. Hotel, Museum, Restaurant), the event or activity (e.g. visit, work, leisure), or other information on the POI (e.g. architecture, price of entrance).

In this context, new similarity measures, specialized for semantic trajectories, have been developed [15, 24]. MSM [15] and SMSM [24] measures consider elements composed of many dimensions, namely spatial, temporal and semantic. Authors propose a framework for the comparison of multi-dimensional sequences where distances are defined separately for each dimension, and then aggregated as a weighted sum. Although these measures provide a richer representation of mobility sequences w.r.t. previous spatio-temporal approaches, they do not support the comparison of complex semantic elements like sets of concepts. Indeed, they compare elements using the discrete metric, i.e. a simple Boolean comparison between elements. In addition, they require many parameters and thresholds to initialize and to tune them. Finally, they do not exploit mobility properties such that repetitions of activities or potential permutations.

Alternatively, Moreau et al. propose the CED similarity measure [31], which extends Edit Distance measures adapting cost computation to typical mobility characteristics, in particular the redundancy of certain elements, repetition [41] and a certain form of cyclicity [40]. CED measure answers the following requirements: (i) edition cost depends on the similarity of nearby elements (the more similar and closer the elements, the lower the cost of operations), (ii) edition of repeated close elements has low cost, and (iii) similar and close elements can be exchanged with a low cost. In addition, CED measure can be paired with any similarity measure even if authors propose the use of ontologies to compare elements. However, as other initial OM-based methods, it was designed for categorical elements and should be extended to deal with multi-dimensional semantic elements.

To the best of our knowledge, there is no similarity measure able to compare sequences of multi-dimensional sets of semantic elements. In this paper we propose to reuse the CED [31] measure, pairing it with an extended similarity measure adapted to multi-dimensional sets of semantic elements.

2.3 Clustering methods

The extraction of behavior from a dataset is a process usually performed thanks to unsupervised machine learning. Indeed, clustering methods are based on similarity measures like the ones described in previous subsections and are widely used for the discovery of human behavior, in particular in sequences of mobility [21, 31, 35]. However, the topology created by similarity measures for semantic sequences is hard to apprehend. In particular, for OM methods, spaces are not euclidean nor metric.

A pairwise comparison of semantic sequences results in a distance matrix that is the input of the clustering process. To the best of our knowledge, the clustering algorithms able to deal with arbitrary distances (not necessarily metrics) are PAM [37] (or K-medoid), hierarchical clustering [22], density clustering (DBSCAN [13], OPTICS [4]) and spectral clustering [33], each one making different hypothesis about cluster topology.

According to the similarity measure and the representation of the sequences, dimensionality reduction methods can be used in order to extract primary dimensions [21]. However, commonly used methods like PCA can only be used for Euclidean spaces in practice. Alternatively, methods like UMAP [30], allow the reduction of a complex topology defined by an arbitrary metric into a low Euclidean space, which facilitates the visualisation of clustering results and enable the usage of other clustering methods, in particular, those requiring an Euclidean space like K-means [29]. In addition, UMAP offers a better preservation of the data global structure, fewer hyperparameters to tune and better speed that previous techniques like t-SNE [28].

Previous proposals for clustering of mobility sequences have been based on K-means [21, 45], DBSCAN [34] or hierarchical clustering [32]. Therefore, in this paper we test several clustering approaches, in order to empirically find the most adapted to sequences of multi-dimensional sets of semantic elements.

3 SIMILARITY MEASURES FOR CLUSTERING OF SEMANTIC SEQUENCES

In this section we formalize the comparison of sequences. We start by describing the comparison of concepts, based on a knowledge graph, then the comparison of complex elements (concerning multiple concepts and multiple ontologies), and finally the extension of the CED measure for dealing with complex elements. This describe our proposal for clustering sequences of complex elements. This step allows to subsequently apply a clustering process defined in the next section in order to extract groups of similar sequences representing coherent behaviors.

3.1 Comparison of concepts

Let O be a set of concepts. In order to compare concepts in O, we introduce a partial order in the set, expressing the semantic containment among concepts. The following definition of knowledge graph embeds such order.

Definition 3.1 (Knowledge graph). Let O be a set of concepts including a special concept denoted *all*. A *knowledge graph* resulting from O is a connected directed acyclic graph $G_O = (O, E)$ with $E \subset O \times O$ where $(x, y) \in E$ iff the concept x (meronym) CONTAINS semantically the concept y (holonym) and $\forall (x, y) \in E, y \neq all$.

Remark that such a knowledge graph is commonly called *meronymy*. Moreover, for any two concepts $x, y \in O$, we denote $LCA(x, y)$ the last common ancestor of x and y and $d(x)$, the depth of x, i.e. its minimal distance from the all node.

We use the Wu-Palmer similarity measure [46], $sim_{WP} : O \times O \rightarrow [0, 1]$, to compare concepts. It is a well-established state of the art measure that takes into account both the depth of the concepts in the knowledge graph and their closest ancestor, being normalized in $[0, 1]$.

$$sim_{wup}(x, y) = \frac{2 \times d(LCA(x, y))}{d(x) + d(y)} \quad (1)$$

This measure is the reference used for comparing concepts in knowledge graphs. Nevertheless, we remark that our proposal (described in next subsections) is independent of the concepts similarity measure and can be easily adapted to other measures.

3.2 Comparison of multi-dimensional elements

Now, consider a set $\mathcal{O} = \{O_1, ..., O_q\}$ of q sets of concepts structured as knowledge graphs as defined in Def. 3.1 and such that for $i \neq j, O_i \cap O_j = \{all\}$.

Each knowledge graph describes a family of properties of elements, thus, their concepts are disjoints, with the exception of the *all* concept, which for facility, is common to all knowledge graphs.

Then, we define an element as a q-dimensional vector, where each dimension is a subset of concepts of a knowledge graph.

Definition 3.2 (Multi-dimensional semantic element). Let $\Sigma = \bigtimes_{k=1}^{q} \mathcal{P}(O_k)$ where $\times$ denotes the cartesian product and $\mathcal{P}$ the power set. A *multi-dimensional semantic element* $\sigma \in \Sigma$ is a q-uplet where the k-th component (noted $\pi_k(\sigma)$) is a subset of O_k.

Example 3.3. In next examples, we revisit our Alice and Bob motivating example in a more formal fashion, using concepts of the DATAtourisme ontology, sketched in Figure 1 and later described in Section 4.1.

Consider one of Bob's activities: *In the morning he visits a castle with flower gardens (Chaumont castle).* Consider three dimensions describing touristic activities: place of interest, type of event and architectural style, taking concepts from three domain ontologies. Bob's activity can be formalized as:

$$\sigma_1 = \langle \{Castle, ParkAndGarden\}, \emptyset, \{Renaissance\} \rangle$$

stating the activity takes place in both a castle and a park/garden, with no information about a specific type of event, the place being of Renaissance style.

Now, let us formalize one of Alice's activities: *Going to a sound and light show at a renaissance castle (Blois castle).* This activity can be formalized as:

$$\sigma_2 = \langle \{Castle\}, \{VisualArtEvent\}, \{Renaissance\} \rangle$$

□

In order to compare elements (i.e. vectors), we propose to separately compare each dimension (i.e. subsets of concepts) and aggregate the obtained similarity scores. Notably, our method deals with incomplete vectors, which is a very frequent situation in touristic activities and other types of trajectories.

Formally, we need a similarity measure $\zeta : \mathcal{P}(O) \times \mathcal{P}(O) \rightarrow [0, 1]$ between two subsets of concepts. We use the Halkidi measure [18], an extension of Wu-Palmer measure in the context of ontologies, defined such that:

$$\zeta(X, Y) = \frac{1}{2}\left(\frac{1}{|X|}\sum_{x \in X}\max_{y \in Y}(sim(x, y)) + \frac{1}{|Y|}\sum_{y \in Y}\max_{x \in X}(sim(x, y))\right) \quad (2)$$

Based on means, this measure is tolerant of outlier similarity scores. Furthermore, results for sets of concepts comparison are empirically in accordance with intuition.

Finally, the similarity between two elements $sim_\Sigma : \Sigma \rightarrow [0, 1]$ is computed as the aggregation of the similarity scores for each dimension, as follows:

$$sim_\Sigma(\sigma, \sigma') = Agg_{k=1}^{q}\left(\zeta(\pi_k(\sigma), \pi_k(\sigma'))\right) \quad (3)$$

where $Agg : [0, 1]^q \rightarrow [0, 1]$ denotes any aggregation function.

Frequently, we have to deal with incomplete elements, i.e. having missing values for some dimensions. For example, in Example 3.3, the second component of σ_1, noted $\pi_2(\sigma_1)$, is missing. Missing values are tricky because they can indicate unknown, irrelevant or non existent values. To tackle this problem, we propose the use of the *average_if* aggregation function which computes an average but ignoring the dimensions where one of the elements has missing values (noted $\emptyset$). In other words only dimensions k such $\pi_k(\sigma) \neq \emptyset \wedge \pi_k(\sigma') \neq \emptyset$ are considered. Weighs are equitably distributed on the remaining dimensions.

Example 3.4. Let instances σ_1 and σ_2 defined in Example 3.3. As the type of event of σ_1 is empty, the computation of $sim_\Sigma(\sigma_1, \sigma_2)$ is such that:

$$sim_\Sigma(\sigma_1, \sigma_2) = \frac{1}{2}\zeta(\{Castle, ParkAndGarden\}, \{Castle\}) + \frac{1}{2}\zeta(\{Renaissance\}, \{Renaissance\})$$

For the last part we have: $\zeta(\{Renaissance\}, \{Renaissance\}) = 1$. For the first part, we have to compute the Wu-Palmer similarity between some concepts. We recall that, $\forall x \in O, sim_{wup}(x, x) = 1$. Thus, we only need to compute $sim_{wup}(Castle, ParkAndGarden) = \frac{1}{2}$, according to the Equation 1 and the ontology of Figure 1. Then, according to Equation 2:

$$\zeta(\{Castle, ParkAndGarden\}, \{Castle\}) = 0.875$$

Thus, $sim_\Sigma(\sigma_1, \sigma_2) = \frac{1}{2} \times 0.875 + \frac{1}{2} \times 1 = 0.9375$ □

3.3 Comparison of sequences

Thus, thanks to previous definitions, we define semantic sequences as an ordered sequence of multi-dimensional semantic elements.

Definition 3.5 (Semantic sequence). A *semantic sequence* $S \in \Sigma^n$ is an ordered sequence of elements $\langle \sigma_1, \sigma_2, ..., \sigma_n \rangle$ such that $\forall k \in [\![1, n]\!], \sigma_k \in \Sigma$ and for $i < j, \sigma_i$ precedes σ_j.

Intuitively, such a sequence indicates that σ_1 is done firstly then σ_2, ..., finally σ_n.

Example 3.6. Returning to Alice and Bob motivating example. Let's revisit our Alice's day in a more formal fashion.

$$\begin{aligned} S_{Alice} = &\langle \langle \{Hotel\}, \emptyset, \{Renaissance\} \rangle, \\ &\langle \{ReligiousSite, Museum\}, \emptyset, \{Medieval\} \rangle, \\ &\langle \{ParkAndGarden\}, \emptyset, \emptyset \rangle, \\ &\langle \{Castle\}, \emptyset, \{Renaissance\} \rangle, \\ &\langle \{Church\}, \emptyset, \{Roman\} \rangle, \\ &\langle \{PubAndBar, Restaurant\}, \emptyset, \emptyset \rangle, \\ &\langle \{Castle\}, \{VisualArtEvent\}, \{Renaissance\} \rangle, \\ &\langle \{Hotel), \emptyset, \{Renaissance\} \rangle \rangle \end{aligned}$$

□

In order to compare semantic sequences, we pair the CED measure [31] with the similaritiy measure among elements defined in Equation 3.

The CED measure is a generalization of the Edit Distance to deal with common characteristics of semantic mobility sequences, which makes it particularly appropriate for mobility analysis. Indeed, the fact that repetition and edition of similar elements in the sequence has a low cost, just like permutations, tends to group elements with same semantics while taking into account a flexible timeframe.

Equations 4 to 6, indicate the CED computation as described in [31]. Firstly, CED includes a modification of the cost operation function γ which generalizes the classical definition of Edit Distance and takes into account the local context of each element in the sequence.

Consider contextual edit operations of the form $e = (o, S, \tau, k)$, denoting the operation $o \in \{\texttt{add}, \texttt{mod}, \texttt{del}\}$ on sequence S at index k by element $\tau \in \Sigma$. Let E be the set of all possible contextual edit operations, the cost function $\gamma : E \longrightarrow [0, 1]$ for the contextual edit operations is defined as:

$$\gamma(e) = \alpha \times \ell(e) + (1 - \alpha) \left(1 - \max_{i \in [\![1,n]\!]} \{sim(\sigma_i, \tau) \times v_i(e)\} \right) \quad (4)$$

where:

- $\alpha \in [0, 1]$ is a contextual coefficient.
  If $\alpha \longrightarrow 0$ the cost will be strongly evaluated according to the near content at index k in the sequence being edited; if $\alpha \longrightarrow 1$ then cost is fixed and CED tends toward the Levenshtein Distance with substitution cost.
- $\ell(e) = \begin{cases} 1 - sim(\sigma_k, \tau) & if\ o = \texttt{mod} \\ 1 & else \end{cases}$ is the cost function of Levenshtein Distance with substitution cost.
- $sim : \Sigma \times \Sigma \longrightarrow [0, 1]$ is a similarity measure between two elements.
- $v(e) \in [0, 1]^n$ is a contextual vector which quantifies the notion of proximity between elements. Usually, the larger $|i - k|$ is, the smaller $v_i(e)$ is.

Let $\mathcal{P}(S_1, S_2)$, all the edit paths to transform a sequence S_1 into S_2, the one-sided contextual edit distance from S_1 to S_2 noted $\tilde{d}_{CED} : \Sigma^n \times \Sigma^p \longrightarrow \mathbb{R}^+$ is defined such that:

$$\tilde{d}_{CED}(S_1, S_2) = \min_{P \in \mathcal{P}(S_1, S_2)} \left\{ \sum_{i=1}^{|P|} \gamma(e_i) \right\} \quad (5)$$

where $P = (e_1, ..., e_q) \in E^q$ is a vector of contextual edit operations.

Computation of Equation 5 is done using dynamic programming and Wagner-Fisher algorithm [44]. Finally, $d_{CED} : \Sigma^n \times \Sigma^p \longrightarrow \mathbb{R}^+$ is computed using the following equation:

$$d_{CED}(S_1, S_2) = \max \left\{ \tilde{d}_{CED}(S_1, S_2), \tilde{d}_{CED}(S_2, S_1) \right\} \quad (6)$$

4 EXPERIMENTS

In this section, we describe our experiments to validate our approach. We firstly introduce a case study concerning tourist mobility, framed by the Smart Loire project[3] (in Subsections 4.1 and 4.2), then, we present our experimental protocol and discuss the obtained results.

[3]in french: https://intelligencedespatrimoines.fr/smart-loire-apr-ir-2017/

4.1 Case study description

The Smart Loire project is part of a French regional initiative in order to help users determine customized touristic itineraries in the Loire Valley region [8]. In particular, the identification of clusters of coherent behaviors with similar visiting patterns is essential for guiding tourism actors on user profiles but also for designing better recommendation tools based on the extracted knowledge.

We represent touristic trips as sequences of activities, where activities are multi-dimensional sets of semantic elements. Indeed, each activity is described by three dimensions: POI, event, and eventually architectural style for the POI.

The dataset of touristic activities is taken from DATAtourisme[4], a national standardized tourism business ontology used for describing touristic entities.

DATAtourisme is a large ontology mainly organized in two major parts, *POI* and *Event*. Thereby, a touristic instance (i.e. an activity) is described in terms of a POI and an Event, for example "a Sound and light show at a Renaissance castle".

In practice, we split DATAtourisme in smaller focused ontologies, each one serving as a knowledge graph for the three semantic dimensions of activities.

The first one is extracted from the PlaceOfInterest node (POI) and all its sub-classes, the second one regroups several nodes related to events. Finally, we take advantage that some POI have architectural details and create a third ontology constructed around the architectural styles of the buildings. Figure 1 provides a summarized representation of the extracted POI ontology, including the nodes relevant to our experiments. The complete datasets can be consulted in our GitHub[5].

[4]https://framagit.org/datatourisme/ontology/
[5]https://github.com/Anonymous-codeLab/SAC2021

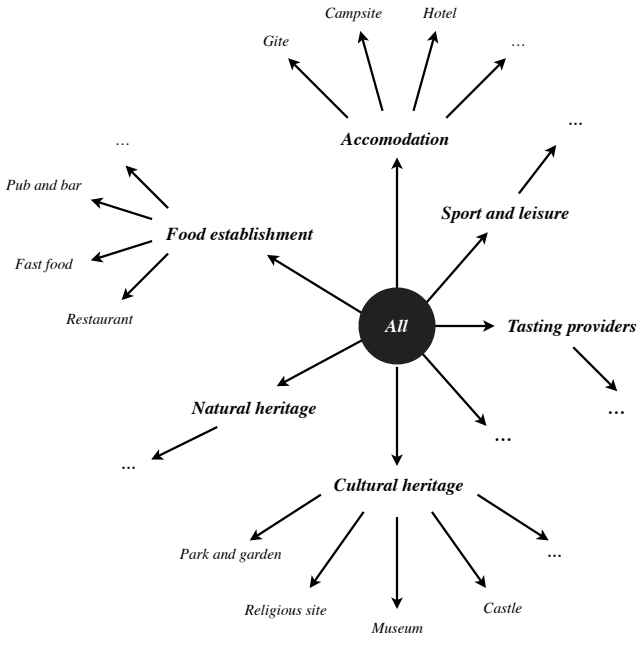


Figure 1: POI ontology extracted from DATAtourisme

Category	Concept	Number of instances
Accomodation	Hotel	50
	Gite	50
	Camping	50
Food establishment	Fast Food	1
	Bar	29
	Restaurant	50
Nature Heritage	Nature	16
Cultural Heritage	Park and garden	24
	Religious site	29
	Museum	46
	Castle	47
Sport and leisure	Sport	50
Other (...)	...	50
Tasting Providers	Tasting	50

Table 1: Number of used instances in POI ontology

id	State
1	*Start*
2	Morning activity
3	Lunch Time
4	Afternoon activity
5	Accomodation
6	Night activity
7	*End*

(a)

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & p_1 & p_2 & 0 & p_3 & 0 & 0 \\ 0 & p_4 & p_5 & p_6 & p_7 & 0 & 0 \\ 0 & 0 & 0 & p_8 & p_9 & 0 & p_{10} \\ 0 & 0 & 0 & p_{11} & p_{12} & p_{13} & 0 \\ 0 & 0 & 0 & 0 & 0 & p_{14} & p_{15} \\ 0 & 0 & 0 & 0 & 0 & p_{16} & p_{17} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

(b)

Figure 2: Description of the random Markov walker for touristic sequence generation. (a) Table of states (b) Stochastic matrix

4.2 Data Generation and touristic profiles

As the SmartLoire project is currently being deployed, we do not have any real trace of tourists' activities, apart from prototypical profiles described by tourism agents. Thus, we instead choose to use artificial sequences, realistically derived from real activities, to validate our methodology and analyse its behavior and the sensibility of its parameters in a controlled environment.

Thereby, for generating artificial touristic sequences, we used real activities, that are instances in the DATAtourisme ontology. To this end, we selected all identified instances in a area of 50km around the city of Amboise in the Loire Valley. We obtained around 2500 different instances. Among them, we selected the top 50 most described instances for the most used concepts of the POI ontology illustrated in Figure 1. Table 1 lists the number of instances selected for each concept of the POI ontology, totalizing **542** instances used for data generation.

The artificial sequences were generated using a random walker marching on an absorbing Markov chain which gives a credible structure to the daily activities of our virtual tourists. The stochastic process is defined over the states described Figure 2.

To understand the generation process, let O be the set of concepts in the POI ontology, and consider a set of predefined prototypical profiles Ψ. Each probability p_k in the matrix is setting up according to a profile $\psi \in \Psi$. Moreover, each state s contains a set of concepts $O_s \subset O$ such that, for each $c \in O_s$, the probability to select an instance σ of c depends of the given profile ψ. Subsequently, σ is chosen in c with a uniform probability $\frac{1}{|c|}$. Thus, for a given state s and profile ψ, the probability that the instance $\sigma \in c$ is selected is:

$$p(\sigma) = \frac{1}{|c|} \times p(c|\psi) \tag{7}$$

For the experiments in section 4.3, we select ***5 touristic profiles*** partly inspired by ones identified by [1] where each one of them has a natural inclination for doing some activities:

- The *Night Owl*: Has a strong tendency to sleep in the morning, visits pub and practises night activities.
- The *Cultural Interest*: Visits mainly museums, castles and other remarkable buildings. Always do a morning and an afternoon activity.
- The *Camping enthusiast*: Accommodates only in camp sites and like visiting natural areas.
- The *Young couple*: Scattered and varied activities. Accommodates mostly in gites.
- The *Fine Dinning Amateur*: Sleeps only in hotels, has lunch in restaurants and has some predispositions for tasting activities.

Note that the state 7 is absorbing. Thus, this process generates broad finite patterns of sequences. In each step of the random walk, an instance is chosen according to Equation 7. The concatenation of such instances produces an artificial sequence representing a one day touristic trip. The details about the setting up of the probabilities for each profile is given on our code lab[6].

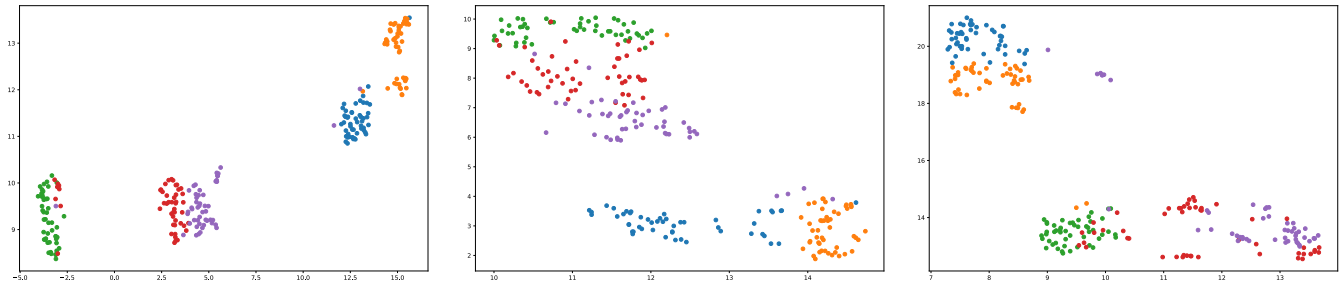
4.3 Clustering process

Considering the main use cases for a multi symbol semantic distance, recommendation of items or full item sequences and analysis of existing data sets to extract actionable insights. We can safely assume that the main feature required of such distance in a practical application such as the Smart Loire project is its ability to discriminate between trajectories emanating from groups exhibiting significant behavioral differences (i.e., classes or clusters).

4.3.1 Choice of clustering algorithms. As motivated in Section 2, we aim to test several clustering algorithms and empirically find the most adapted to sequences of multi-dimensional semantic elements.

Concretely, we test a Hierarchical clustering algorithm (Hierarchical for short, the one originally paired with CED [31]), DBSCAN (the Agglomerative clustering algorithm used in [34] and combined with Levenshtein [25]), K-Medoids, and a Spectral clustering algorithm (Spectral for short). Remember that K-Means (the algorithm used in [21, 45]) cannot be used in non-Euclidean spaces. We also explore the fairly recent UMAP dimensionality reduction technique, projecting our dataset into a 2D Euclidean space. Then, we can test K-Means, Spectral and DBSCAN on the projected dataset. Note that on the 2D euclidean space produced as a projection by UMAP K-Means replaces Hierarchical using Ward criterion and K-Medoids as they all minimize intra-cluster variance but K-Means has a low time complexity.

[6] https://github.com/Anonymous-codeLab/SAC2021



(a) (b) (c)

Figure 3: UMAP 2D Projection of the 250 sequences using (a) CED, (b) Lev. + Ontology and (c) Levenstein as similarity measure. Colors: blue *Camping enthusiast*, orange *Night Owl*, green *Fine Dinning Amateur*, red *Cultural Interest*, purple *Young Couple*

4.3.2 Setting of algorithm parameters. In this paper we use the experimental setup proposed by Moreau et al. [31], namely

- α coefficient is set to 0 to give maximum weight to context
- The contextual vector is encoded by the Gaussian kernel: $f_k(i) = \exp\left(-\frac{1}{2}\left(\frac{i-k}{\beta}\right)^2\right)$ where β is a coefficient which controls the flatness of the curve around the activity at position k. The bigger is β, the more context around index k is take into account. For our experiment, we take $\beta = \frac{m}{2}$ where m is the median size of sequences. Thus, $v_i(e) = f_k(i)$.

Concerning UMAP settings, we use the *umap-learn* python library version 0.4.3 using default parameters if not specified, *min_dist* is set to 0.01, *n_neigbohrs* to 25 and the pseudo random number generator seed is 12. All experiments can be reproduced by running the python notebook[7] in Google Colab or a Jupyter environment.

4.3.3 Protocol. To validate our approach, we use the clustering algorithms to cluster the generated touristic sequences. The main goal of our experiments is to assess to what extent the clustering algorithms, paired with our similarity measure, are able to regroup sequences corresponding to a same profile. Remark that an advantage of using generated data is that the labels of the clusters are known, conforming a ground truth.

We propose to evaluate our approach using the performance of the clustering algorithm as quality metric. This allows to use metrics such as the Adjusted Rand Index (ARI) [38]. Unlike internal quality metrics (e.g. Silhouette), this measure lets us get a better picture of which approach should be the most performing.

Furthermore, in order to assess our similarity measure, we also evaluate clustering algorithms paired with two additional similarity measures: (i) *Levenstein* distance and (ii) an extension of Levenstein using Halkidi similarity as the cost for operations. We refer to it as *Lev. + Ontology*.

4.4 Results

Table 2 shows our experiment results. The best ARI score, 0.833, is obtained by UMAP combined with K-Means or Spectral, paired with CED. Without UMAP projection, Spectral, paired with CED, outperforms other combinations with an ARI of 0.649. Notably,

[7] https://colab.research.google.com/drive/1ndEudixznlYKjtSyaOuAl6ghU08W-YVF?usp=sharing

	Levenstein	Lev. + Ontology	CED
DBSCAN	0.128	0.203	0.409
Hierarchical	0.315	0.171	0.483
K-Medoids	0.300	0.170	0.550
Spectral	0.510	0.590	**0.649**
UMAP + DBSCAN	0.549	0.636	0.733
UMAP + K-Means	0.659	0.673	**0.833**
UMAP + Spectral	0.665	0.680	**0.833**

Table 2: Adjusted Rand Index for different similarity measuresand clustering methods

CED outperforms the other measures for all reported clustering algorithms. We remark that when paired with CED, the second worst result is achieved by Hierarchical.

Interestingly, working on a raw similarity matrix (without UMAP projection), Spectral outperforms the other methods, for the three similarity measures. The obtained ARI scores are sensibly higher.

A more surprising result is that UMAP boosts performance above any algorithm and distance in the original space even using the naive Levenstein distances. This highlights the great potential of the UMAP technique for semantic sequence clustering. Furthermore the ability to project such complex object in a simple 2D representation opens the possibility of using it as visualisation tool for experts exploring the data.

To highlight and compare the behavior of the three metrics when coupled with UMAP we plot, in Figure 3, a 2D representation of the touristic sequences using the same UMAP parameters. Each color represents a touristic profile. For the CED projection (a) we notice that profiles are quite well separated, the only class being difficult to separate from others is *Cultural Interest* with several points mixed with *Fine dinning* and being relatively close to *Young Couple*. In plot (b), corresponding to Lev. + Ontology, we immediately see less densely packed points. We think Lev. + Ontology distance is not appropriate for sequences with missing values. Again we notice that *Cultural interest* points are the hardest to separate from other classes. Finally, the basic Levenstein distance (c) exhibits the most mixing of points belonging to different classes not only *Young couple* and *Fine dinning* but also *Night Owl* and *Camping enthusiast* are difficult to separate.

5 CONCLUSION AND FUTURE WORK

In this paper we proposed a new approach for clustering sequences of multi-dimensional semantic elements. This type of complex sequences is commonly used in many domains for purpose of recommendation or behavior extraction.

The introduced approach comprises a similarity measure, based on business ontologies, and especially designed to compare multi-dimensional semantic element. We combined the Wu-Palmer similarity with Halkidi similarity, $average\_if$ aggregation function and the CED measure for sequences. Nonetheless, the approach is modular and can be used with other techniques listed in the related work. Finally, we tested several clustering algorithms in order to identify the best technique to deal with our proposal. We found that the UMAP dimensional reduction technique combined with a Spectral clustering outperforms others methods. To our knowledge, this is the first use of this technique for semantic sequences.

Experiments have been applied on a touristic domain using the DATAtourisme ontology and real instances. A random Markov model is proposed in order to generate artificial touristic trips according to prototypical defined profiles. The results are very promising and show an ARI score of 0.83.

As future work, we plan to apply our approach to real sequences of multi-dimensional semantic elements. Also, we envisage to expand the methodology to incorporate the time dimension, for example integrating the duration of activities.

Acknowledgements

This work is supported by Smart Loire project which has received funding from the Centre-Val de Loire region.

REFERENCES

[1] Tripbarometer 2016 – traveler trends & motivations global findings. Technical report, Tripadvisor., 2016.
[2] A. Abbott and A. Tsay. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000.
[3] L. Alvares, V. Bogorny, B. Kuijpers, J. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. *Proc. of the 15th annual ACM international symposium on Advances GIS*, (22):1–8, 2007.
[4] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.
[5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.
[6] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1):66–88, 2014.
[7] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International Conference on World Wide Web*, pages 757–766, New York, NY, USA, 2007.
[8] M. Boulakbech, N. Messai, Y. Sam, T. Devogele, and L. Etienne. Smartloire: A web mashup based tool for personalized touristic plans construction. In *WETICE*, pages 259–260, 2016.
[9] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection for discrete sequences: A survey. *KDE*, 24(5):823–839, 2010.
[10] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3):370–383, 2007.
[11] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The youtube video recommendation system. In *REXIS*, pages 293–296, 2010.
[12] M. M. Deza and E. Deza. *Encyclopedia of distances*. Springer, 2016.
[13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
[14] R. Fileto, C. May, C. Renso, N. Pelekis, D. Klein, and Y. Theodoridis. The baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98:104–122, 2015.
[15] A. S. Furtado, D. Kopanaki, L. O. Alvares, and V. Bogorny. Multidimensional similarity measuring for semantic trajectories. *Trans. in GIS*, 20(2):280–298, 2016.
[16] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou. Mobile recommender systems in tourism. *Journal of network and computer applications*, 39:319–333, 2014.
[17] M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
[18] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis. Thesus: Organizing web document collections based on link semantics. *The VLDB Journal*, 12:320–332, 2003.
[19] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343, 1975.
[20] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A.-C. Ngonga Ngomo. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920, 2017.
[21] S. Jiang, J. Ferreira, and M. C. González. Clustering daily patterns of human activities in the city. *DMKD*, 25(3):478–510, 2012.
[22] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. 2009.
[23] C. Leacock and M. Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet*, 49(2):265–283, 1998.
[24] A. L. Lehmann, L. O. Alvares, and V. Bogorny. SMSM: a similarity measure for trajectory stops and moves. *International Journal of Geographical Information Science*, 33(9):1847–1872, 2019.
[25] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
[26] Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882, 2003.
[27] D. Lin et al. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, volume 98, pages 296–304, 1998.
[28] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.
[29] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 281–297, 1967.
[30] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
[31] C. Moreau, T. Devogele, V. Peralta, and E. Laurent. Contextual edit distance for semantic trajectories. In *Proc. of the 35th Annual ACM SAC*, pages 635–637, 2020.
[32] C. Moreau, V. Peralta, P. Marcel, A. Chanson, and T. Devogele. Learning analysis patterns using a contextual edit distance. In *DOLAP*, pages 46–55, 2020.
[33] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in NIPS*, pages 849–856, 2002.
[34] L. Pappalardo and F. Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.
[35] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):1–8, 2015.
[36] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. G. Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4), 2013.
[37] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.
[38] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
[39] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.
[40] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.
[41] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
[42] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.
[43] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh. Indexing multidimensional time-series. *The VLDB Journal*, 15(1):1–20, 2006.
[44] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.
[45] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.
[46] Z. Wu and M. Palmer. In *ACL*, pages 133–138, USA.
[47] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proc. of EDBT*, pages 259–270, 2011.
[48] G. Zhu and C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Trans. on Knowledge and Data Engineering*, 29(1):72–85, 2016.