ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 10(18), DOI: 10.17485/ijst/2017/v10i18/103400, May 2017* ISSN (Online) : 0974-5645

Towards a Semantic Trajectory Similarity Measuring

Francisco Javier Moreno Arboleda[1*], Santiago Román Fernández[1] and Vania Bogorny[2]

[1]Facultad de Minas, Universidad Nacional de Colombia, Medellín, Colombia; fjmoreno@unal.edu.co, sromanf@unal.edu.co
[2]Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis, Brasil; vania.bogorny@ufsc.br

Abstract

Objectives: To propose a new similarity function to determine trajectory similarity considering semantic aspects. **Methods/Analysis:** We propose different methods to calculate the similarity according to visited sites or activities performed: the first one considers only the sites included in the trajectories and the second considers the activities performed by the trajectories in the sites. A third method is proposed to find the similarity bitten trajectories based on both sites and activities. **Findings:** The similarity measure presented in this work allows us to make comparisons and user analysis according to trajectory data generated by users, which represents their routines, likes and preferences. This could be a key element for recommender systems, clustering or social networks. **Novelty/Improvements:** Our methods consider semantic aspects for finding the similarity of trajectories, considering visited sites and activities performed in these sites.

Keywords: Moving Objects, Semantic Trajectories, Similarity Measures, Trajectory Similarity

1. Introduction

Due to the price reduction and the increase use of GPS technologies and social media in daily life, large amounts of trajectory data are now available as spatio-temporal databases. Trajectory data are collected as raw trajectories, represented as a sequence of space-time points (x, y, t) that correspond to the position (x, y) of an object in a space at instant t.

The discovery of similar movement behavior from trajectory data is interesting for several domains, such as trajectory clustering and nearest neighbor queries. During the last few years, several approaches have been proposed to measure the similarity of raw trajectories. Among the main approaches it will-known DTW (Dynamic Time Warping)[1,2], developed for time series, LCSS (Longest Common Subsequence)[3], and EDR (Edit Distance on Real Sequences)[4].

More recently, an enormous effort is being made to add more data to raw trajectories, i.e., transforming a raw trajectory into a semantic trajectory[5,6]. A semantic trajectory has more data associated than a raw trajectory. In addition to space and time, a semantic trajectory has data, such as the name and the type of the visited sites by a moving object, and the activities performed at each site [7]. Several definitions can be found in the literature for a semantic trajectory, such as[5] and [7], but for the sake of simplicity, it considers a semantic trajectory as a sequence of visited sites called stops, as originally introduced in[8]. Figure 1 shows an example of two semantic trajectories, considering both, the type of the visited site and the activity performed there. Trajectory A visits Hotel X, Bank K, and University Y, while trajectory B visits Bed-and-breakfast Z, School U, Bank K, University Y, and Restaurant W. Trajectory A visits a hotel, while trajectory B visits a bed-and-breakfast, which are different sites but with the same semantic type, i.e., *accommodation*. Notice that trajectory A visits a hotel for working, while trajectory B visits a bed-and-breakfast as a client, so both visited sites refer to accommodation but with different activities performed by the moving objects. Both trajectories also visit *educational* sites, a university and a school, but with the same activity, teaching. Considering that both trajectories visit similar types of sites, but may perform different activities there, the question that arises is: how similar are trajectories A and B from a semantic point of

* *Author for correspondence*

view? How similar are both trajectories considering the visited sites? Considering the activities? Considering both sites and activities?

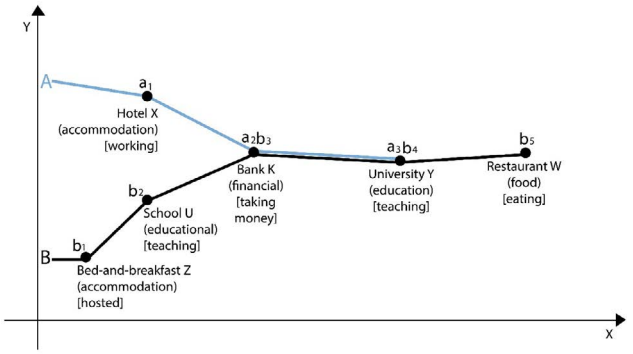


Figure 1. Example of two semantic trajectories A and B.

To the best of our knowledge, there is no approach in the literature that focuses on the similarity of trajectories considering both sites and activities. An approach, proposed[9] splits a semantic trajectory into *sub-trajectories* and computes the semantic similarity of two trajectories based on the longest common subsequence of visited sites. In their approach only a full match is considered, i.e., 1 if there is a match on the name of the site and 0 otherwise (the activities are not considered in their work).

In[10] proposed a semantic similarity measure that considers the semantics of the stops, the sequence of the visited sites (*stops*), the travel time between the sites, and the frequency that a site is visited. Two trajectories are considered similar if they visit the same sequence of sites, several times, and with similar travel time. Notice that their approach is different from existing similarity measures because it considers the frequency of the visited sites, what is more related to trajectory patterns.

More recently, in[11] it is proposed the MSM (*Multidimensional Similarity Measure*), which measures the similarity of semantic trajectories in several dimensions, including a semantic one. In their approach the similarity of each dimension is given by a different distance function, and the specific function to measure the similarity of each dimension is not the focus of that work. For instance, the spatial distance is measured by the Euclidean distance, while the semantic similarity is given by the full match on the name of the site (1 if there is a match on the name of the site, and 0 otherwise).

Issues such as classification and clustering of trajectories are of special interest due to the social or collective information they can generate. Different clustering techniques have been proposed in order to discover similar trajectories. For instance, in[12] it is proposed a method for grouping trajectories based on their shape: two trajectories are considered similar if they have *sub-trajectories* in common (with the same shape).

In[13] proposed a progressive refinement clustering algorithm, where different clustering strategies are defined to discover similar trajectories according to proximity in time and space. The algorithm creates a Boolean matrix where the columns are the stops and the rows are the trajectories, and uses the dynamic time warping distance [14] to measure the similarity between the trajectories according to their chronological sequence of stops.

In[15] proposed a method to calculate the similarity between users, considering their location and the sites they visit. It relies on a category hierarchical graph, where each site visited by a user is associated with a node of the graph (called location node).

In[16], raw trajectories become semantic trajectories through *stay cells*. A *stay cell* represents a geographic region where the user made a stop (exceeding a time threshold). Subsequently, it assigns semantic terms (such as school, park, bank, etc.) to these cells and defines a measure of semantic similarity between trajectories called Maximal Semantic Trajectory Pattern Similarity (MSTP-Similarity) based on the stay cells of each trajectory.

In[17] proposed a method to calculate the similarity between users based on their data location history. Through a framework called HGSM (hierarchical-graph-based similarity measurement) and a hierarchical grouping of the sites it is possible to explore the visited sites by users in different layers of similarity, where the finest layer contains the users with higher similarity.

In[18] defined a similarity measure between two trajectories based only on spatio-temporal features. In[19] it is defined the dissimilarity between two trajectories based on the Euclidean distance and their timestamps. Similarly, in[12] it is considered sub-trajectories to establish the dissimilarity.

In[20] proposed an algorithm that determines when a trajectory is similar to a sub-trajectory of another trajectory. It relies on the Euclidean time-Uniform distance function[21], a variant of the Euclidean distance that considers the time in which the events occur.

In[22] defined two measures of similarity between two trajectories, one based on space, and the other based on time; which can be combined to obtain an overall measure of similarity; thus, the user can obtain the similarity

between two trajectories by these three criteria.

In this paper, we propose a new similarity function for semantic trajectories, which supports both the semantics of the visited sites by the trajectories and the activities performed at each site, what to the best of our knowledge has not been addressed before. This new function may be incorporated in previous works, such as MSM, to compute the similarity of the semantic dimension. While previous works do only consider the full match on the semantic dimension, it proposes a taxonomy of sites and activities to consider *partial matching* of sites and activities performed at a site.

2. Trajectories Similarity

In[11] define a multidimensional similarity measure that considers the distance between two sets of elements in different dimensions, based on a score between two elements *a* and *b* as defined in Equation (1). In the particular case of trajectories, these elements are *episodes*[11].

$$score(a,b) = \sum_{k=1}^{|D|} (match_k(a,b) * w_k) \quad (1)$$

where,*D* is a set of dimensions, $w_k$is a weight assigned to each dimension, and $match_k(a, b)$ is given by Equation (2).

$$match_k(a,b) = \begin{cases} 1, & if\ dist_k(a,b) \leq maxDist_k \\ 0, & otherwise \end{cases} \quad (2)$$

where,$maxDist_k$ is a distance threshold for dimension *k*. In this way, different dimensions can be considered such as time, spatial, and semantic, for calculating the similarity between trajectories as shown in Figure 2. The focus of Furtado's work is to define a *multidimensional similarity measure, but not the similarity function for each dimension*; as a way of example, they define a very simple similarity measure for the semantic dimension, which is given by Equation (3).

$$dist_k(a,b) = \begin{cases} 0, & if\ a.type = b.type \\ 1, & otherwise \end{cases} \quad (3)$$

That is, the similarity between two episodes *a* and *b* is 0 (full match) if the episodes have the same type of visited site, 1 otherwise. It proposes a new similarity measure for the semantic dimension based on the sites visited and the activities performed there, defining a similarity measure for the semantic dimension between two trajectories, i.e., $dist_k$ for *k* = Semantic; as highlighted in Figure 2.

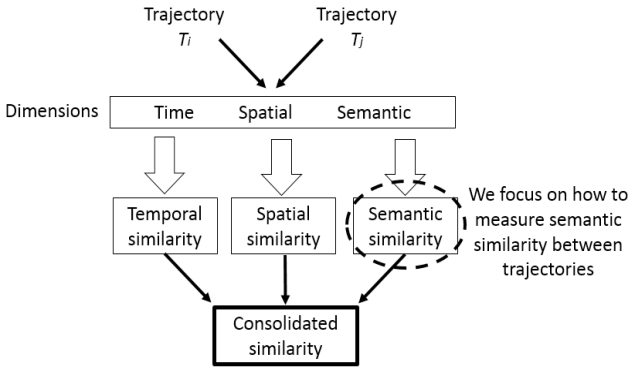


Figure 2. Focus of our proposal.

This section introduces the new concepts and the proposed similarity measure for semantic trajectories. Similarly, to[15], it considers a Category (concept) Tree for the Classification of the sites (CTCS), where a site is a Point Of Interest (POI) for the application. For simplicity, each site is associated with a single category (its main category) corresponding to a leaf node of the tree. The CTCS is a set of nodes having a parent-child relationship and satisfies that: the CTCS has a special node r called "Site" (root), which does not have parent node; and each node *ns*∈ CTCS, such that *ns* ≠ *r*, has a single parent node *p*∈ CTCS, *p* ≠ *ns*. Figure 3, it shows an example of a CTCS. The relationship between the CTCS nodes is hierarchical, where a child node represents a more specialized category than the category represented by its parent node.

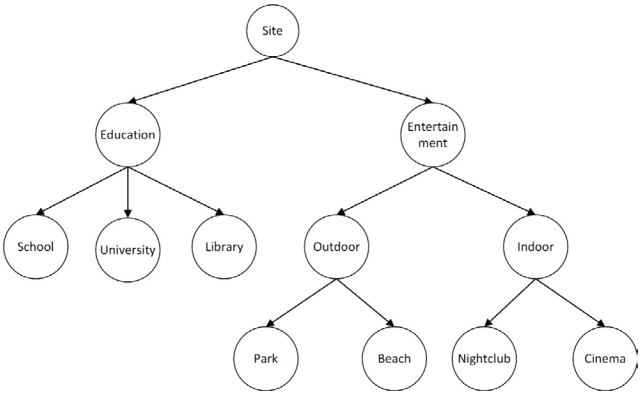


Figure 3. CTCS example.

Similarly, it considers a Category Tree for the Classification of the Activities (CTCA). Note that some combinations of sites and activities might not make sense, e.g., studying in a nightclub. Valid combinations could be

specified and controlled by the analysts. Figure 4, it shows an example of a CTCA (adapted from [23]). Likewise, the CTCS, the analyst can define the CTCA as required by its application.

Note that an activity may be associated with more than one parent node (e.g., the activity "Dancing" could also be associated with the node "Motor"); for simplicity it considers only one parent for each activity. It plans as future work to extend our proposal for supporting sites/activities with several parents.

In the following it introduces the main definitions of semantic trajectories and activities, using as examples the hierarchies shown in Figures 3 and 4.

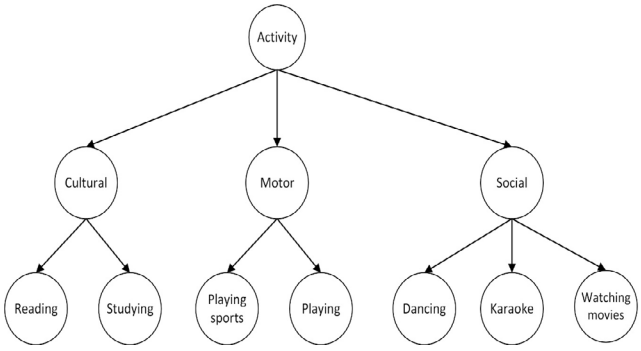


Figure 4. CTCA example.

Let *S* be a set of *m* sites $S = \{s_1, s_2, \ldots, s_m\}$, where $s_i = (s_{id}, s_{name}, s_{cat})$, where,*s_id* is the site identifier, *s_name* its name, and *s_cat* represents the CTCS category (leaf node) which is associated with the site. Thus, one site is (*directly*) associated with one leaf node of the CTCS and (*indirectly*) with all the ancestor nodes of that leaf node in the CTCS.

Example. Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ be the set of sites, where,s_1 = (1, Cinema Central, Cinema), s_2 = (2, Bocagrande, Beach), s_3 = (3, University of Cartagena, University), s_4 = (4, El Rosario, Beach), s_5 = (5, Golden Disco, Nightclub), s_6 = (6, University of Bolívar, University), and s_7 = (7, Botanical Garden, Park).

Similarly, it defines a set of *p* activities $A = \{a, a_2, \ldots, a_n\}$, where, $a_i = (a_{id}, a_{name}, a_{cat})$, where,*a_id* is the activity identifier, *a_name* its name, and *a_cat* represents the CTCA category (leaf node) which is associated with the activity.

Example. Let $A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$be the set of activities, where,a_1 = (1, Studying math, studying), a_2 = (2, Bicycling, Playing sports), a_3 = (3, Reading science fiction, Reading), a_4 = (4, Dancing electronic, Dancing), a_5 = (5, Studying Spanish, Studying), a_6 = (6, Swimming, Playing sports), a_7 = (7, Singing rock, Karaoke), and a_8 = (8, Watching adventure movies, watching movies).

On the other hand. a trajectory *T* is a set of *n* episodes $T = \{e_1, e_2, \ldots, e_n\}$, where $e_i = (s_i, a_i, t_i)$ where, $s_i \in S$ represents the site where the episode occurred, $a_i \in A$ represents the activity performed at site s_i, and $t_i = (t_{ini}, t_{fin})$ represents the start time (t_{ini}) and the end time (t_{fin}) of the episode, $t_{ini} < t_{fin}$.

Example. Consider the trajectory $T_1 = \{e_1, e_2, e_3, e_4\}$, where, $e_1 = (s_6, a_5, t_1), e_2 = (s_4, a_6, t_2), e_3 = (s_1, a_8, t_3)$, $and\ e_4 = (s_7, a_3, t_4)$. Table 1 details the episodes of T_1.

To calculate the similarity between trajectories, it extend the proposal of Zhao, Han[24]. They propose a formula to determine whether two trajectories are spatial similarity complete based on the set of POI of each trajectory and a threshold θ.

Let POI_{ns,T_i} be the set of all sites (either directly

Table 1. Events of the trajectory T_1

e_i	s_i			a_i			t_i	
	s_{id}	s_{name}	s_{cat}	a_{id}	a_{name}	a_{cat}	t_{ini}	t_{fin}
							February 18th, 2016	
e_1	6	University of Bolívar	University	5	Studying Spanish	Studying	8am	12m
e_2	4	El Rosario	Beach	6	Swimming	Playing sports	3pm	4pm
e_3	1	Cinema Central	Cinema	8	Watching adventure movies	Watching movies	4pm	5:30pm
e_4	7	Botanical Garden	Park	3	Reading science fiction	Reading	8pm	9pm

or indirectly) associated with a node *ns*∈ CTCS included in the episodes of trajectory T_i. The similarity between two trajectories T_i and T_j with regard to *ns*, C_{ns,T_i,T_j}, is calculated by Equation (4).

$$C_{ns,T_i,T_j} = \frac{|POI_{ns,T_i} \cap POI_{ns,T_j}|}{|POI_{ns,T_i} \cup POI_{ns,T_j}|} \quad (4)$$

That is, C_{ns,T_i,T_j} is the relationship between the total number of sites *common* to the two trajectories associated with the node ns and the total number of sites of the two trajectories associated with that node. $C_{ns,T_i,T_j} = Undef$ (Undefined) if $POI_{ns,T_i} \cup POI_{ns,T_j} = \varnothing$, i.e., when none of the two trajectories have sites associated with the node *ns*. Note that, Equation (4) is based on the Jaccard index, whose range is between the interval [0, 1] and its value is 1 when both sets are empty; in our proposal when this situation arises it assigns an *Undef* value.

Note that in our proposal if the same site is included in several episodes of a trajectory, this similarity measure considers it only once. Another aspect to keep in mind is the following. Suppose a trajectory T_3 that has a single episode which includes site s_3 = (3, University of Cartagena, University) and a trajectory T_4 that has a single episode which includes site s_6 = (6, University of Bolívar, University), i.e., both trajectories included a university in their respective episodes, but since the universities are different then $C_{University,T_3,T_4} = 0$. Note that these two trajectories included in their episodes two different sites which are associated with the same node (University). It refers to these sites as *non-matching sites*. This situation may deserve a similarity greater than zero. Thus, to incorporate these sites in our measure of similarity, it proposes a parameter called *non-matching sites weight nmsw*∈ [0, 1]. This parameter acts as a *weight* by which the user sets the degree of contribution of the non-matching sites for the similarity. Thus, the formula for the similarity is modified according to Equation (5) (the *nnms* parameter, called *number of non-matching sites* is explained).

$$C_{ns,T_i,T_j,nmsw} = \frac{|POI_{ns,T_i} \cap POI_{ns,T_j}| + nmsw*nnms}{|POI_{ns,T_i} \cup POI_{ns,T_j}| - nmsw*nnms} \quad (5)$$

Similarly to Equation (4), Equation (5) has a range in the interval [0, 1] and is *Undef* when $POI_{ns,T_i} \cup POI_{ns,T_j} = \varnothing$. Note that when *nmsw* = 0, then Equation (5) is equal to Equation (4). For instance, considering again trajectories T_3 and T_4 it obtained $C_{University,T_3,T_{4,0}} = 0$. Furthermore, when *nmsw* = 1 (and *nnms* = 1 as is explained below), $C_{University,T_3,T_{4,1}} = 1$, i.e., it is considered that trajectories T_3 and T_4 are 100% similar with regard to University node because although they visited different sites (s_3 and s_6), both belong to the same site category (University).

Now, it explains the *nnms* parameter. Consider trajectories T_5 and T_6. T_5 includes in its episodes the following sites associated with University node: $POI_{University,T_5} = \{s_{10}, s_{11}, s_{12}, s_{13}\}$, where,$s_{10}$ = (10, University A, University), s_{11} = (11, University B, University), s_{12} = (12, University C, University), and s_{13} = (13, University D, University). T_6 includes in its episodes the following sites also associated with University node: $POI_{University,T_6} = \{s_{10}, s_{14}, s_{15}\}$, where,$s_{14}$= (14, University E, University) and s_{15}= (15, University F, University).

Thus, trajectories T_5 and T_6 have a common site (site s_{10}, University A), i.e., $|(POI_{ns,T_5} \cap POI_{ns,T_6})| = 1$. On the other hand, T_5 has in its episodes three different universities in comparison with T_6, while T_6 has in its episodes two different universities in comparison with T_5. So although T_5 visited more different universities in comparison with T_6, it can conclude that each of these trajectories has in its episodes at least two universities (other than the one they have in common, University A). This is the value of *nnms*. Formally, *nnms* is calculated according to Equation (6).

$$nnms = \text{Min}(\left|POI_{ns,T_i} - POI_{ns,T_j}\right|, \left|POI_{ns,T_j} - POI_{ns,T_i}\right|) \quad (6)$$

Consider again trajectories *T5* and *T6*, *nmsw* = 1 and *ns* = University. Note that if the term *nmsw*nnms* is not considered in the denominator of Equation (5), the similarity would be given by $C_{ns,T_5,T_{6,1}} = \frac{1+1*2}{6} = \frac{3}{6} = 0.5$, since it is considered that the two universities visited by T_6 (s_{14}, s_{15}) are "equal" to the two universities visited by T_5 (two out of s_{11}, s_{12}, s_{13}), i.e., that the trajectories have in common two more universities (aside from s_{10}, so the numerator is 3), then the total of "different" universities between the two trajectories should be four and not six just as the intersection of the visited sites increases from one to three. Therefore, the term *nmsw*nnms* is subtracted in the denominator and

$C_{ns,T_5,T_{6,1}} = \frac{1+1*2}{6-1*2} = \frac{3}{4} = 0.75$, (*in practical terms it means that* T_5 *and* T_6 *have three out of four universities in common*). In addition, $C_{ns,T_5,T_{6,0}} = \frac{1}{6} = 0.167$ (i.e., T_5 and T_6 strictly only have one out of six universities in common) and $C_{ns,T_5,T_6,0.5} = \frac{1+0.5*2}{5} = \frac{2}{5} = 0.6$ (with *nmsw* = 0.5 it means in practical terms that T_5 and T_6 have two out of five universities in common).

Initially, it proposes two methods for calculating the similarity between two trajectories considering *only* the sites included in the trajectories episodes, i.e., based on CTCS. Subsequently, itconsiders the activities performed by the trajectories in the sites to establish their similarity.

2.1 Method 1

Consider two trajectories *Ti* and *Tj*. In this method, it computes the similarity of each node *ns*∈ CTCS by Equation (5), i.e., $SIM_{ns} = C_{ns,T_i,T_j,nmsc}$. In this way, the user can analyze the trajectories similarity with regard to each CTCS node. For instance, if *ns* is the root of CTCS, then $C_{ns,T_i,T_j,nmsc}$ indicates the similarity of the trajectories from a general point of view (node "Site"). The user can then analyze the similarity from a more specific point of view as he descends through the levels of the CTCS (a "drill-down").

Note that in this method, to calculate the similarity of a non-leaf node, it is not required to calculate the similarity of its child nodes (confront with method 2).

Example: Consider trajectory $T_2 = \{e_1, e_2, e_3, e_4, e_5\}$, where, $e_1 = (s_3, a_1, t_1), e_2 = (s_2, a_4, t_2), e_3 = (s_4, a_2, t_3), e_4 = (s_1, a_8, t_3), and\ e_5 = (s_5, a_7, t_3)$. Table 2 details the episodes of T_2.

With *nmsw* = 0.5 and considering trajectories T_1 and T_2, the CTCS with the similarity of each node is shown in Figure 5. For example, the calculation of $SIM_{University}$ (a leaf node) is obtained in this way: the trajectories do not have common sites with regard to this node, i.e., $|(POI_{ns,T_1} \cap POI_{ns,T_2})| = 0$, where,*ns* = University. Furthermore, each trajectory included in its episodes a university, i.e., *nnms* = 1; therefore, $SIM_{University} = \frac{(0 + 0.5 * 1)}{(2 - 0.5 * 1)} = 0.33$.

For calculating $SIM_{Entertainment}$(a non-leaf node) its leaf nodes are considered (Park, Beach, Nightclub, and Cinema). The trajectories have two sites in common (s_1 and s_4), *nnms* = 1, and $|POI_{ns,T_1} \cup POI_{ns,T_2}| = 5$, where, *ns* = Entertainment. Hence, $\frac{(2 + 0.5 * 1)}{(5 - 0.5 * 1)} = 0.55$.

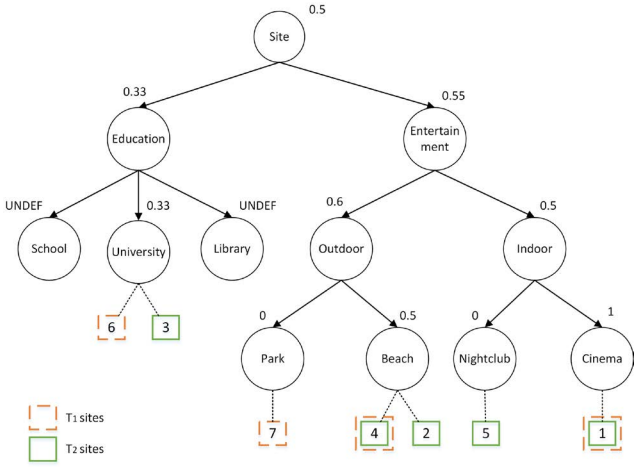


Figure 5. CTCS with similarity values for T_1 and T_2 using method 1.

2.2 Method 2

In this method, each node *ns*∈ CTCS will have a similarity SIM_{ns}, where, $SIM_{ns} = C_{ns,T_i,T_j,nmsw}$ for a *leaf node*, i.e., Equation (5). Unlike method 1, for *a non-leaf node* SIM_{ns} is calculated by Equation (7).

Table 2. Events of the trajectory T_2

e_i	s_i			a_i			t_i	
	s_{id}	s_{name}	s_{cat}	a_{id}	a_{name}	a_{cat}	t_{ini}	t_{fin}
							February 18th, 2016	
e_1	3	University of Cartagena	University	1	Studying math	Studying	7am	10am
e_2	2	Bocagrande	Beach	4	Dancing electronic music	Dancing	11am	1pm
e_3	4	El Rosario	Beach	2	Bicycling	Playing sports	2pm	3pm
e_4	1	Cinema Central	Cinema	8	Watching adventure movies	Watching movies	9pm	11pm
e_5	5	Golden Disco	Night club	7	Singing rock	Karaoke	10:30pm	11:30pm

$$SIM_{ns} = \sum(\text{weight}_{nl} * SIM_{nl}), \forall nl \in H \wedge SIM_{nl} \neq Undef \quad (7)$$

Where, H is the set of the child nodes of node *ns* and weight_{nl}, named *weight of node nl*, is the weight assigned by the analyst to node *nl*, i.e., the analyst can specify the weight with which each child node nl contributes to the similarity of its parent node *ns*. For example, a user might consider for node "Outdoor" that the beaches should "weight" (contribute) more in the similarity than the parks. To do this, he could specify that weight_{Beach} = 0.8 and weight_{Park}= 0.2. Note that the sum of the weights of the children of a node must be equal to 1, i.e., $\sum weight_{nl} = 1, \forall nl \in H$.

Example: Consider trajectories T_1 and T_2. The CTCS with the similarity of each node is shown in Figure 6. The same weight was considered for the child nodes of a node. For instance, for *nmsw* = 0, SIM_{Beach} (a leaf node) is obtained in this way: since both trajectories included in their episodes the site s_1 and T_2 also included site s_2, then $SIM_{Beach} = C_{Beach, T_1, T_{2,0}} = (1 + 0 * 0) / 2 = 0.5$.

To calculate $SIM_{Outdoor}$ (a non-leaf node), it considers the similarity of leaf nodes Park (SIM_{Park} = 0) and Beach (SIM_{Beach} = 0.5); by applying Equation (7) with $weight_{Beach}$ = $weight_{Park}$= 0.5 it obtained: (0.5 * 0 + 0.5 * 0.5) = 0.25. To calculate $SIM_{Education}$ (another non-leaf node) it considers only the similarity of University node, inasmuch as the similarity of nodes School and Library is *Undef*.

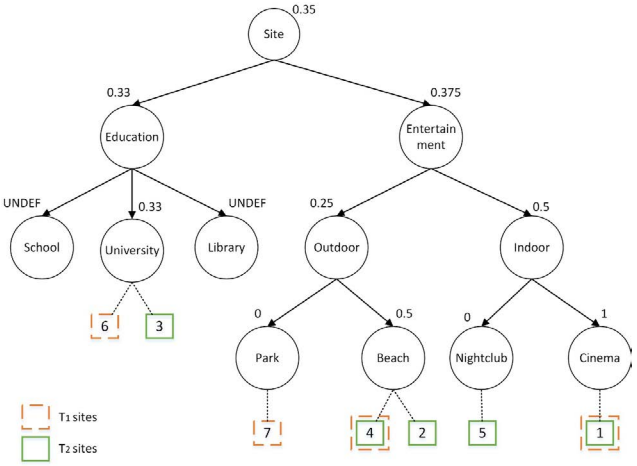
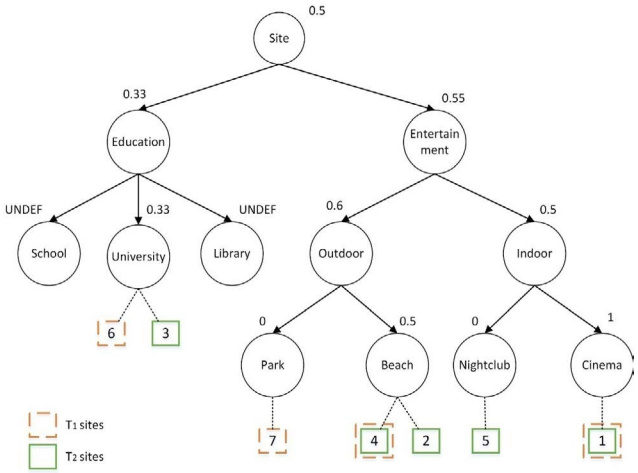


Figure 6. CTCS with similarity values for T_1 and T_2 using method 2.

2.3 Differences and Interpretation of the Two Methods

Figures 5 and 6 show that the similarity of two trajectories with regard to a *non-leaf node* may differ depending on the method to be applied (in both methods the similarity with regard to the leaf nodes is equal). For example, the similarity of trajectories T_1 and T_2 with regard to the root node (Site) is 0.5 with method 1 and 0.35 with method 2. This difference occurs due to the weights assigned to the child nodes and to the number of sites of the trajectories associated with the leaf nodes. For instance, if one considers the same weight *w* for the child nodes of a node *ns*, the difference of similarity obtained with the two methods with regard to *ns* becomes larger as the set of sites of a trajectory *Ti* associated with the leaf nodes descendants of ns becomes larger with regard to the corresponding set of sites of a trajectory *Tj*. This is because method 1 considers for each node (whether it is a leaf or not) all the sites associated with it (directly or indirectly), whereas in method 2 after calculating the similarity for each leaf node, the similarity of *ns* is calculated considering only the similarity of its children and the weights w assigned to these.

Example: Consider the CTCS of Figure 7 and two trajectories T_7 and T_8. Consider Nightclub node (a leaf node), *nnms* = 0, and suppose that $|(POI_{ns,T_7} \cap POI_{ns,T_8})| = 0$ and $|(POI_{ns,T_7} \cup POI_{ns,T_8})| = 1$, where, *ns* = Nightclub. Furthermore, consider Cinema node (a leaf node), *nnms* = 0, and suppose that $|(POI_{ns,T_7} \cap POI_{ns,T_8})| = 7$ and $|(POI_{ns,T_7} \cup POI_{ns,T_8})| = 10$, where, *ns* = Cinema. Considering *weight* = 0.5, with method 1 it obtained $SIM_{Nightclub}$ = 0, SIM_{Cinema} = 0.7, and SIM_{Indoor} = 0.6363. With method 2 it obtained $SIM_{Nightclub}$ = 0, SIM_{Cinema} = 0.7, and SIM_{Indoor} = 0.35. Note that the difference of the similarities with regard to Indoor node is 0.28.



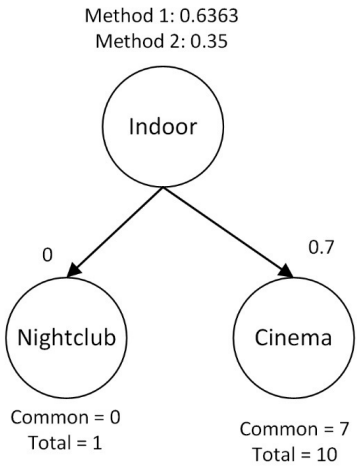


Figure 7. Different similarity values obtained with methods 1 and 2.

Let U_1 and U_2 be two users with trajectories *Ti* and *Tj*, respectively. Taking as example node *ns* = "Entertainment", and assuming that U_1 visited more outdoor entertainment sites and U_2 visited more indoor entertainment sites, it will be determined when it is appropriate to apply method 1 or 2. To do this, consider the question: is it important to consider the type of entertainment experienced by a user or only whether a user has been entertained (i.e., regardless of the type of entertainment)? If in the application domain is important to differentiate the type of entertainment experienced by the users, i.e., that the similarity measure is affected because each user visited most sites in different categories, it is appropriate to use method 2 since this considers all the subcategories (and even it is possible to give weights to the different types of entertainment); if it only want to obtain a similarity measure regardless of the type of entertainment is appropriate to use method 1.

Note that in method 1 all sites remain "with the same level of importance", e.g., in Figure 7 a nightclub is as important as a cinema since the specific type of site is not of interest; whereas in method 2, a nightclub becomes relevant (it weights more in the calculation of the similarity). Consequently, it decreases the similarity value with regard to method 1.

2.4 Similarity Algorithms of Two Trajectories

Next, it proposes two algorithms to find the similarity between two trajectories corresponding to the methods explained.

Listing 1. Algorithm SimMethod1 for method 1

SimMethod1(T1, T2, nmsw, G, ns)
Input: T1, T2: Trajectories, nmsw, G: CTCS, ns: Node ∈ G
Output: Node ns with its similarity
BEGIN
1. ST = G.subTree(ns); //Extract the subtree with ns as root
2. L = leafNodes(ST); //Extract the set of leaf nodes of ST
3. S1 = {}; //Set of sites of T1 related to nodes of interest for calculating similarity
4. S2 = {}; //Set of sites of T2 related to nodes of interest for calculating similarity
5. **FOREACH** nsAux∈ L
6. Add to S1 the sites of T1 related to nsAux node
7. Add to S2 the sites of T2 related to nsAux node
8. **END FOR**
9. **IF** |S1| = 0 **AND** |S2| = 0 **THEN**
10. ns.sim = Undef;
11. **ELSE**
12. nnms = MIN(|S1 - S2|, |S2 - S1|);
13. ns.sim = (|S1 ∩ S2| + nmsw * nnms) / (|S1 ∪ S2| - nmsw * nnms);
14. **END IF**
15. **END** SimMethod1

Listing 2. Algorithm SimMethod2 for method 2

SimMethod2(T1, T2, nmsw, G, ns)
Input: T1, T2: Trajectories, nmsw, G: CTCS, ns: Node ∈ G
Output: Node ns with its similarity
BEGIN
1. **IF** ns.isLeaf**THEN**
2. SimMethod1(T1, T2, nmsw, G, ns);
3. **ELSE**
4. H = G.children(ns); //Extract the set of children of node ns
5. sum = 0;
6. **FOREACH** nsAux∈ H
7. Sim2(T1, T2, nmsw, G, nsAux);
8. **IF** nsAux.sim ≠ Undef**THEN**
9. sum += nsAux.weight * nsAux.sim;
10. **END IF**
11. **END FOR**
12. ns.sim = sum;

13. **END IF**
14. **END** SimMethod2

Note that algorithm SimMethod2 calculates the similarity to all descendants of the node of interest *ns*, which allows access to the similarity value of any of these nodes without calculating it again. It also allows us finding the similarity of each node of CTCS when invoked with their root node.

It can also use algorithms 1 and 2 to find the similarity between two trajectories with regard to the activities performed by using a CTCA instead of a CTCS. For instance, if SimMethod1 is invoked with the CTCA of Figure 4, i.e., *SimMethod1 (T_1, T_2, 0.5, CTCA, Activity)*, the results are shown in Figure 8.

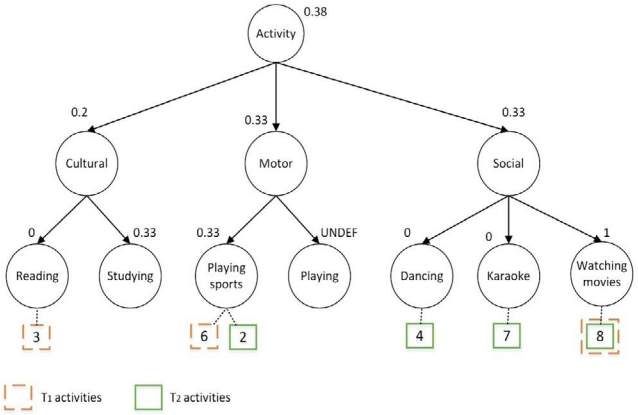


Figure 8. CTCA with similarity values for T_1 and T_2 using method 1.

2.5 Combined Similarity: Sites and Activities

So far the similarity has been calculated based on the visited sites or in the activities performed at these sites, but both criteria have not been considered simultaneously. The following is a method proposed for this.

Let *ns*∈ CTCS be the node of interest and T'_1 be a subset of episodes corresponding to T_1 episodes whose sites are associated with *ns* or with a descendant of *ns*. The similarity with regard to the activities is obtained by applying method 1 or 2 sending *ns* = Activity as parameter. It is then obtained a CTCA with a value for each one of its nodes, which represents the similarity between two trajectories based on the activities performed at site *ns*; the value of the Activity node represents the similarity with regard to all the activities performed by the users on site *ns*.

Note that for each node *ns*∈ CTCS, a CTCA is generated with similarity values for each CTCA node, which indicates the similarity of each activity performed at site *ns*. If *ns* is the root of CTCS, then the CTCA generated represents the similarity of all activities performed regardless of the site where they were performed.

Example: Figure 9, it shows the CTCA with the similarity values when applying the method 1 for node *ns* = Entertainment and *nmsw* = 0.5. To calculate the similarity in the Cultural node, it only use a_3 since a_1 and a_5 were not performed at Entertainment sites; therefore, $SIM_{Cultural}$= 0. It is concluded that trajectories T_1 and T_2 are similar in 0.4 with regard to Entertainment sites, and 0.25 with regard to the activities performed at such type of sites.

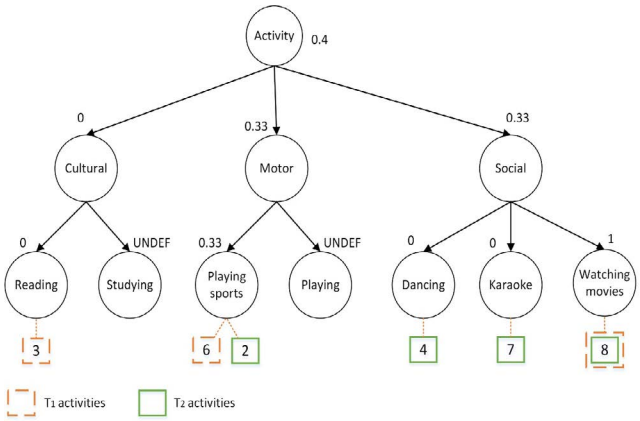


Figure 9. CTCA with similarity values for node *ns* = Entertainment for T_1 and T_2 using method 1.

2.6 Algorithms for Combined Similarity

The algorithm for extracting *T'*, a subset of *T* episodes associated with a node *ns*, is presented in Listing 3. The algorithm to calculate the general similarity of the two trajectories (including sites and activities) is shown in Listing 4.

Listing 3. Algorithm for extracting *T'*
Extract(T, G, ns)
Input: T: Trajectory, G: CTCS, ns: Node ∈ G
Output: T': Trajectory with episodes related to ns
BEGIN
1. T' = {};
2. H = G.allDescendants(ns); //Extract the set of descendants of node ns
3. **FOREACH** e ∈ T
4. **IF** e.s.s_cat∈ H **THEN**
5. T'.add(e);

6. **END IF**
7. **END FOR**
8. **END** Extract

Listing 4. Algorithm for calculate the similarity including sites and activities

Sim(T1, T2, GS, nmswS, GA, nmswA, ns)
Input: T1, T2: Trajectories, GS: CTCS, nmswS: nmsw for sites, GA: CTCA, nmswA: nmsw for activities, ns: Node $\in$ GS
Output: Node ns with the sites similarity, CTCA with the activities similarity
BEGIN
1. SimMethod1(T1, T2, nmswS, GS, ns); // Alternatively SimMethod2 can be invoked
2. T1' = Extract(T1, GS, ns); //Extract the episodes of T1 related to ns node
3. T2' = Extract(T2, GS, ns); //Extract the episodes of T2 related to ns node
4. SimMethod1(T1', T2', nmswA, GA, GA.root); // Alternatively SimMethod2 can be invoked.
5. **END** Sim

Note that when method 1 is invoked with CTCA only the similarity value of the root node is calculated. If it wants to calculate the value of the other CTCA nodes it is necessary to make more calls to *SimMethod1*. This is not necessary if the similarity is calculated with method 2, due to its recursive nature.

3. Results and Discussion

In this section, the results of our experiments are presented and we compare them with Zhao's proposal[24]. For the experiments, it used a database that keeps track of Foursquare users in NYC between October 24th, 2011 and February 20th, 2012. The sites were classified according to the labels indicated by the users and the CTCS shown in Figure 10. Similarly, the activities were classified according to the comments left by the users when they made a check-in; the CTCA shown in Figure11 was used. Since not all check-in records had comments, it was necessary to assume some activities according to the most likely activity the user did on the site. For the analysis, it chose the 51 pairs of trajectories that had more sites in common.

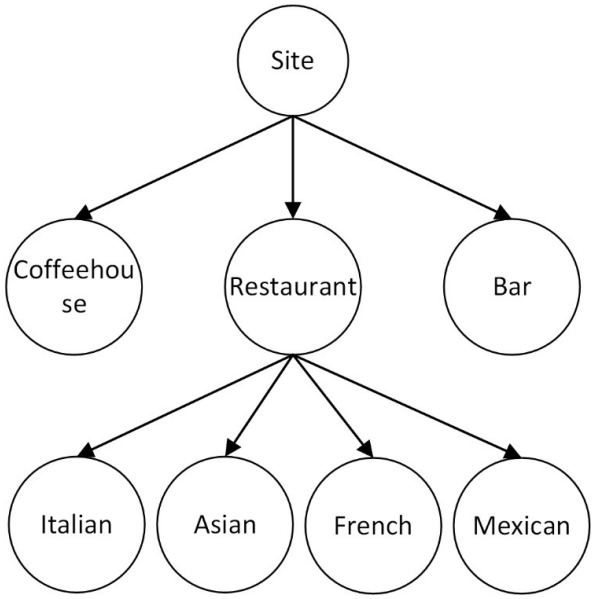


Figure 10. CTCS for the experiments.

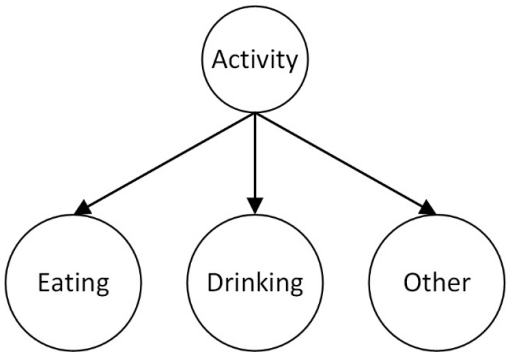


Figure 11. CTCA for the experiments.

Initially, methods 1 and 2 were applied to find the similarity based on the sites and on the activities separately, with *nmsw* = 0.5, and the same weight for each node of the same level in method 2. Tables 3 and 4,it shows the results obtained by each method.

Table 3. Results of methods 1 and 2 with regard to sites

Node	Similarity method 1			Similarity method 2		
	Average	Max	Min	Average	Max	Min
Site	0.35	0.87	0.07	0.29	0.65	0.09
Bar	0.31	0.92	0	0.31	0.92	0
Coffeehouse	0.28	0.85	0	0.28	0.85	0
Restaurant	0.34	0.92	0.05	0.28	0.54	0.06
Italian	0.27	0.73	0	0.27	0.73	0
Asian	0.32	0.75	0	0.32	0.75	0
French	0.22	0.8	0	0.22	0.8	0
Mexican	0.3	1	0	0.3	1	0

Table 4. Results of methods 1 and 2 with regard to activities

Node	Similarity method 1			Similarity method 2		
	Average	Max	Min	Average	Max	Min
Activity	0.41	0.91	0.07	0.41	0.83	0.07
Eating	0.39	0.9	0.07	0.39	0.9	0.07
Drinking	0.42	0.98	0.09	0.42	0.98	0.09
Other	0.42	1	0	0.42	1	0

Figure 12 shows the results obtained by the pairs of trajectories that obtained the highest similarity values on the nodes Site and Activity. Note that even though the results in the two methods were different, the pair that obtained the highest similarity in both cases was the same.

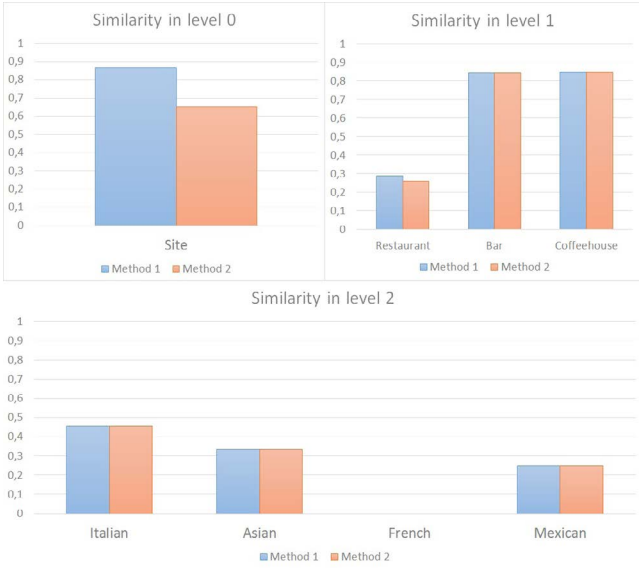


Figure 12. Results for the pair of trajectories that obtained the highest similarity with regard to sites.

Note that due to the nature of the methods, in the leaf nodes the same similarity value will be obtained regardless of the method used. Subsequently, the similarity based on sites and activities was applied. The similarity for each node belonging to CTCS was calculated and also the similarity obtained in the Activity node is shown. Tables 5 and 6 show the average similarity for each site node and the average similarity of users that performed activities on such site.

Next, it used the same 51 pairs of trajectories for comparison with Zhao's proposal, and because their proposal *does not consider activities*, it applied methods 1 and 2 considering *only* sites. Experiments for different values of *nmsw* were conducted and it observed how the similarity measure changed.

Table 5. Combined similarity results using method 1

Node	Combined similarity with method 1					
	Site			Activity Node		
	Average	Max	Min	Average	Max	Min
Site	0.35	0.87	0.07	0.41	0.91	0.07
Bar	0.31	0.92	0	0.19	0.5	0
Coffeehouse	0.28	0.85	0	0.22	0.79	0
Restaurant	0.34	0.92	0.05	0.3	0.65	0.03
Italian	0.27	0.73	0	0.15	0.36	0
Asian	0.32	0.75	0	0.22	0.47	0
French	0.22	0.8	0	0.13	0.5	0
Mexican	0.3	1	0	0.18	0.45	0

Table 6. Combined similarity results using method 2

Node	Combined similarity with method 2					
	Site			Activity Node		
	Average	Max	Min	Average	Max	Min
Site	0.29	0.65	0.09	0.41	0.83	0.07
Bar	0.31	0.92	0	0.13	0.43	0
Coffeehouse	0.28	0.85	0	0.2	0.64	0
Restaurant	0.28	0.54	0.06	0.26	0.58	0.02
Italian	0.27	0.73	0	0.11	0.44	0
Asian	0.32	0.75	0	0.16	0.37	0
French	0.22	0.8	0	0.09	0.63	0
Mexican	0.3	1	0	0.12	0.53	0

Figure 13 shows the similarity measure obtained in each of the three methods when *nmsw* = 0, it can see that both method 1 and Zhao's proposal have equal values in all cases since when *nmsw* = 0, the similarity equations of both are equal. Method 2 presents different values because it uses the ACCS (its hierarchical structure) and the weights assigned to each node for determining the similarity.

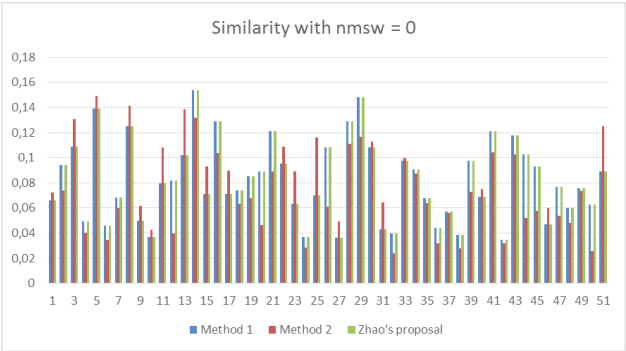


Figure 13. Comparison results when *nmsw* = 0.

Figure 14 shows the similarity measure obtained when *nmsw* = 0.5 and Figure 15 when *nmsw* = 1. Here, the similarity obtained by methods 1 and 2 is greater than that obtained by Zhao's proposal since these values positively affect the similarity when considering sites that are of similar categories, and as expected when *nmsw* = 1, the value of the similarity was higher with regard to Zhao's proposal. In our methods the similarity is greater when the trajectories are more similar, because of the relationship between sites (through the categories), which is right according to expectations. It is also noted that in most cases, the value of the similarity of method 1 is inferior than that of method 2, this is determined by the structure and weights of ACCS of method 2 which can increase or decrease the similarity of the root node (Site node) whereas in method 1 all sites have the same importance for calculating the similarity.

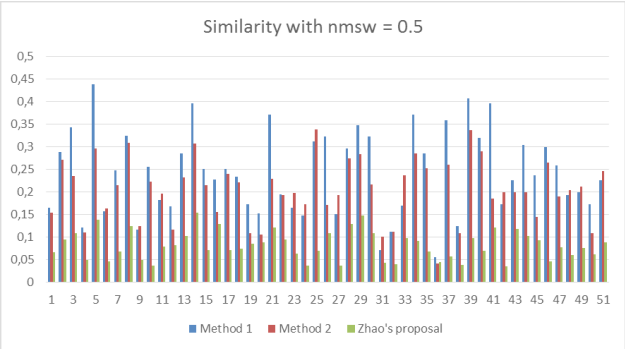


Figure 14. Comparison results when *nmsw* = 0.5.

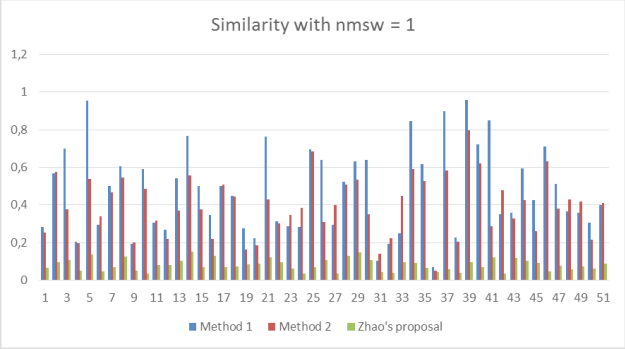


Figure 15. Comparison results when *nmsw* = 1.

4. Conclusion

We proposed a novel approach to measure the semantic similarity among trajectories of moving objects. To the best of our knowledge, our proposal is the first that considers the visited sites and the activities performed at each site. Our approach includes two methods for computing similarity and is flexible because it allows the analysts to define their own category trees for the classification of the sites and the activities. In addition, in method 2 it is possible to assign weights to the nodes of the trees in order to establish their importance when computing the similarity.

As future works,the order (*sequence*) of the visited sites, the frequency, and the duration of the visits for computing the trajectories similarity will be considered. The order is important for trajectories that visit the same type of sites but not in the same order. For instance, consider three users U_1, U_2, and U_3, where,U_1 and U_2 swim in the morning, study in the afternoon, and go shopping at night. U_3 goes shopping in the morning, swims in the afternoon, and studies at night. Although these three users perform the same activities because of their order, trajectories of users U_1 and U_2 may be considered more similar. The frequency is interesting for similarity analysis when objects visit similar sites and with similar frequency, what has not been considered so far. The duration of the visits will be interesting to discover similar trajectories that visit the same type of sites but with similar visiting duration

5. References

1. Kruskal JB. An overview of sequence comparison: time warps, string edits, and macromolecules. Society for Industrial and Applied Mathematics SIAM Review. 1983 Apr; 25(2):201-37. Crossref
2. Keogh E, Ratanamahatana AC. Exact indexing of dynamic time warping. Knowledge and Information Systems. 2004 Mar; 7(3):358–86. Crossref
3. Vlachos M, Kollios G, Gunopulos D. Discovering similar multidimensional trajectories. Proceedings 18th International Conference on Data Engineering, USA; 2002. p. 673–84. Crossref
4. Chen L, Özsu MT, Oria V. Robust and fast similarity search for moving object trajectories. Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Maryland; 2005. p. 491–502. Crossref
5. Alvares LO, Bogorny V, Kuijpers B, de Macedo JAF, Moelans B, Vaisman A. A Model for enriching trajectories with semantic geographical information. Proceedings of the 15th annual ACM International Symposium on Advances in Geographic Information Systems, Washington; 2007. p. 22:1–8. Crossref
6. Parent C, Spaccapietra S, Renso C, Andrienko G, Andrienko N, Bogorny V. Semantic trajectories modeling and analysis. ACM ComputSurv. 2013 Aug; 45(4):1–32. Crossref

7. Bogorny V, Renso C, de Aquino AR, de Lucca Siqueira F, Alvares LO. CONSTAnT – A conceptual data model for semantic trajectories of moving objects. Transactions in GIS. 2014 Feb; 18(1):66–88. Crossref
8. Spaccapietra S, Parent C, Damiani ML, de Macedo JA, Porto F, Vangenot C. A conceptual view on trajectories. Data & Knowledge Engineering. 2008 Apr; 65(1):126–46. Crossref
9. Liu H, Schneider M. Similarity measurement of moving object trajectories. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming, California; 2012. p. 19–22. Crossref
10. Xiao X, Zheng Y, Luo Q, Xie X. Inferring social ties between users with human location history. Journal of Ambient Intelligence and Humanized Computing. 2012 Dec; 5(1):3–19. Crossref
11. Furtado AS, Kopanaki D, Alvares LO, Bogorny V. Multidimensional similarity measuring for semantic trajectories. Transactions in GIS. 2015 Jul; 20(2):280–98. Crossref
12. Lee J-G, Han J, Whang K-Y. Trajectory clustering: A partition-and-group framework. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, China; 2007. p. 593–604. Crossref
13. Zhao X. Progressive refinement for clustering spatio-temporal semantic trajectories. Proceedings of 2011 International Conference on Computer Science and Network Technology, China; 2011. p. 2695–9. Crossref
14. Berndt D, Clifford J. Using dynamic time warping to find patterns in time series. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining; 1994. p. 359–70.
15. Lee M-J, Chung C-W. A user similarity calculation based on the location for social network services. Proceedings of the 16th International Conference on Database Systems for Advanced Applications - Volume Part I, China; 2011. p. 38–52. Crossref
16. Ying JJ-C, Lu EH-C, Lee W-C, Itng T-C, Tseng VS. Mining user similarity from semantic trajectories. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, California; 2010. p. 19–26. Crossref
17. Li Q, Zheng Y, Xie X, Chen Y, Liu W, Ma W-Y. Mining user similarity based on location history. Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, California; 2008. p. 1–10. Crossref
18. Yanagisawa Y, Akahani J-i, Satoh T. Shape-based similarity query for trajectory of mobile objects. Proceedings of the 4th International Conference on Mobile Data Management, Australia; 2003. p. 63–77. Crossref
19. Kreveld MV, Luo J. Trajectory similarity of moving objects. Young Researchers Forum Proceedings of the 5th Geographic Information Days, Germany; 2007. p. 229–32.
20. Trajcevski G, Ding H, Scheuermann P, Tamassia R, Vaccaro D. Dynamics-aware similarity of moving objects trajectories. Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, Washington; 2007. p. 11:1–8. Crossref
21. Cao H, Wolfson O. Trajcevski G. Spatio-temporal data reduction with deterministic error bounds. The VLDB Journal. 2006 Sep; 15(3):211–28. Crossref
22. Tiakas E, Papadopoulos AN, Nanopoulos A, Manolopoulos Y, Stojanovic D, Djordjevic-Kajan S. Searching for similar trajectories in spatial networks. Journal of Systems and Software. 2009 May; 82(5):772–88. Crossref
23. Clasificación actividades recreativas. [Internet]. [cited 2015 Aug 20]. Available from: http://cmapspublic3.ihmc.us/rid=1074811614949_1572273732_232077/Clasificacion.%20Actividades%20recreativas.cmap.
24. Hwang JR, Kang HY, Li KJ. Spatio-temporal similarity analysis between trajectories on road networks. Perspectives in Conceptual Modeling. Lecture Notes in Computer Science. Berlin, Heidelberg; 2005. p. 280–9. Crossref

Appendix

Proof:

$$C_{ns,T_i,T_j,nmsw} = \frac{|POI_{ns,T_i} \cap POI_{ns,T_j}| + nmsw*nnms}{|POI_{ns,T_i} \cup POI_{ns,T_j}| - nmsw*nnms} \in [0,1]$$

Let $n = |POI_{ns,T_i}|$ and $m = |POI_{ns,T_j}|$

- If $POI_{ns,T_i} \cap POI_{ns,T_j} = \emptyset$ then $C_{ns,T_i,T_j,nmsw} = \frac{nmsw * nnms}{n + m - nmsw * nnms}$ and $nnms$ can be n or m depending on which set (POI_{ns,T_i} or POI_{ns,T_j}) is bigger.

  Suppose that $|POI_{ns,T_j}| > |POI_{ns,T_i}|$ then $nnms = n$ and $C_{ns,T_i,T_j,nmsw} = \frac{nmsw * n}{n(1 - nmsw) + m}$, $0 \leq nmsw \leq 1$.

  If $nmsw = 0$ then $C_{ns,T_i,T_j,nmsw} = 0$ and if $nmsw = 1$ then $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$ because $n < m$.

  Similar when $POI_{ns,T_i} \cap POI_{ns,T_j} = POI_{ns,T_j}$

- If $POI_{ns,T_i} \cap POI_{ns,T_j} = POI_{ns,T_i}$ then $nnms = 0$ because $POI_{ns,T_i} - POI_{ns,T_j} = \emptyset$ $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$ because $n < m$.

  Similar when $POI_{ns,T_i} \cap POI_{ns,T_j} = POI_{ns,T_j}$

- If $POI_{ns,T_i} \cap POI_{ns,T_j} = X$ where X is a subset of POI_{ns,T_i} and POI_{ns,T_j} and $x = |X|$

  When $nmsw = 0$, $C_{ns,T_i,T_j,nmsw}$ is the Jaccard index definition, which $\in [0,1]$.

  When $nmsw = 1$, $C_{ns,T_i,T_j,nmsw} = \frac{x + nnms}{n + m - x - nnms}$ and $nnms = n - x$ or $nnms = m - x$ depending on which set (POI_{ns,T_i} or POI_{ns,T_j}) is bigger.

  Suppose that $|POI_{ns,T_j}| > |POI_{ns,T_i}|$ then $nnms = n - x$ and $C_{ns,T_i,T_j,nmsw} = \frac{n}{m} < 1$ because $n < m$.

  Similar when $|POI_{ns,T_j}| < |POI_{ns,T_i}|$.