UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Lucas May Petry

MULTIPLE-ASPECT TRAJECTORY SIMILARITY MEASURING

Florianópolis

2017

Lucas May Petry

MULTIPLE-ASPECT TRAJECTORY SIMILARITY MEASURING

Trabalho de Conclusão de Curso submetido ao Departamento de Informática e Estatística para a obtenção do Grau de Bacharel em Ciência da Computação. Orientadora: Prof. Dr[a]. Vania Bogorny

Florianópolis

2017

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Petry, Lucas May
Multiple-Aspect Similarity Measuring / Lucas May Petry ; orientadora, Vania Bogorny, 2017.
65 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Ciências da Computação, Florianópolis, 2017.

Inclui referências.

1. Ciências da Computação. 2. Multiple-Aspect Trajectory. 3. Similarity Measure. I. Bogorny, Vania. II. Universidade Federal de Santa Catarina. Graduação em Ciências da Computação. III. Título.

Lucas May Petry

MULTIPLE-ASPECT TRAJECTORY SIMILARITY MEASURING

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para a obtenção do Título de "Bacharel em Ciência da Computação" e aprovado em sua forma final pela banca examinadora.

Florianópolis, 4 de dezembro de 2017.

Prof. Dr. Eng. Rafael Luiz Cancian
Coordenador do Curso

Banca Examinadora:

Prof. Dr[a]. Vania Bogorny
Orientadora

Prof. Dr. Ronaldo dos Santos Mello

MSc. André Salvaro Furtado

AGRADECIMENTOS

Em primeiro lugar, eu gostaria de agradecer aos meus pais por todo o suporte fornecido durante toda a minha graduação e por tornar possível a realização deste trabalho.

Quero levar meus agradecimentos também à professora Vania, por ter me acolhido, por ter acreditado em mim para a realização deste trabalho e pela excelente orientação durante toda a pesquisa e desenvolvimento.

Por último, mas não menos importante, quero agradecer aos meus colegas e amigos que sempre estiveram ao meu lado, e a todos os professores pelo conhecimento repassado e determinante para o meu sucesso acadêmico.

ABSTRACT

The wide use of mobile devices such as GPS, as well as the popularization of social media, has led to the generation of large amounts of movement data, called trajectories of moving objects. Trajectory data analysis and mining has become very important because of the variety of information that may be extracted/inferred from these data, such as the daily habits or the profile of individuals. Because of the complexity of the data, they must be analyzed not only from the spatial and temporal characteristics, but any other semantics that may be related to the data. Behind the large amount of information available about movement, trajectories may be analyzed from multiple points of view, that we call multiple aspect trajectories. Similarity measures are widely employed for trajectory data analysis and have a large impact on the analysis outcomes. Most existing works for trajectory similarity are limited to the space and time dimensions of trajectories, and only a few analyze some semantic characteristics of trajectories. Works such as LCSS, EDR and MD-DTW are very rigid and limited to the order of the trajectory points, and two trajectories are considered similar if they match on all dimensions. On the other hand, works such as MSM are too flexible, considering two trajectories as similar if they match in any dimension. In this work, we define the concept of *multiple-aspect trajectory*, proposing the use of several attributes regarding different aspects related to movement. We propose MUITAS, a novel similarity measure for multiple-aspect trajectory similarity analysis, which overcomes the described limitations of previous works. MUITAS is evaluated on a toy example and over a real dataset of user check-ins on a social network containing different aspects related to movement. The results show that MUITAS is more accurate than existing similarity measures for analyzing multiple-aspect trajectories, in addition to allowing the analysis of trajectories in ways not explored before.

Keywords: trajectory similarity, multiple-aspect trajectory.

RESUMO

O amplo uso de dispositivos móveis como GPS e a popularização das redes sociais têm gerado um grande volume de dados de movimento, também chamados de trajetórias de objetos móveis. A análise e a mineração de dados de trajetórias se tornou muito importante pela variedade de informações que podem ser extraídas/inferidas desses dados, como hábitos diários ou perfis de comportamento dos indivíduos. Devido a riqueza desses dados e sua complexidade, eles precisam ser analisados com relação a características espaciais, temporais e semânticas. Além disso, devido a grande variedade de informações disponíveis sobre trajetórias, faz-se necessária a análise de trajetórias sobre múltiplos pontos de vista, denominados neste trabalho como *trajetórias multiaspecto*. Medidas de similaridade são amplamente empregadas na análise de trajetórias e têm grande influência nos resultados obtidos na análise. Grande parte das medidas de similaridade existentes analisam apenas os atributos espaciais e temporais de trajetórias, e poucos utilizam alguns atributos semânticos na análise. Medidas de similaridade como LCSS, EDR e MD-DTW dependem fortemente da ordem dos pontos da trajetória e na similaridade de todos os atributos dos pontos. Por outro lado, trabalhos como MSM analisam a similaridade considerando quaisquer semelhanças entre pontos e ignorando a frequência em que um comportamento ocorre na trajetória. Neste trabalho, é introduzido o conceito de *trajetórias multiaspecto*, propondo o uso de vários atributos relacionados a diferentes aspectos do movimento na análise de similaridade. Ainda, este trabalho propõe uma nova medida denominada MUITAS, uma medida de similaridade para trajetórias multiaspecto que supera as limitações de trabalhos existentes. MUITAS é avaliada por meio de um exemplo e em um conjunto de dados de *check-ins* de usuários em uma rede social, contendo diferentes aspectos relacionados ao movimento. Os resultados mostram que MUITAS é mais robusta na análise de trajetórias multiaspecto em relação a trabalhos existentes, além de permitir a análise de trajetórias por meios não explorados anteriormente.

Palavras-chave: similaridade de trajetórias, trajetória multiaspecto.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CONTENTS

1 INTRODUCTION

Our daily movement is guided by goals and influenced by the environment where we move. For example, one may go to a coffee shop to work, while another person goes there to get coffee; people may go out to eat early in the evening by car because their neighborhood is too dangerous. Being able to understand and to capture the underlying factors behind movement data can help, for instance, recommendation systems to more accurately target their audience based on similar user profiles.

With the popularization of GPS (Global Positioning System) technology and social media, huge amounts of geotagged data are generated and collected about people lives. GPS devices, for instance, make it possible to know the spatial location of a person at a certain time. The collection of movement points of an object, i.e., the locations and *timestamps* collected by a GPS device, constitute the object trajectory. Any moving object, such as a person, an animal, a car, a tornado, can generate a trajectory.

The emergence of trajectories and means to collect them allowed their study and analysis for the past decade. Due to the explosion of social media and geolocation services, it became possible to enrich trajectories with semantic information, such as interesting places visited by the trajectory (e.g., park, restaurant, hotel). More recently, Bogorny et al. (2014) proposed a model for trajectory representation comprising different aspects related to movement, such as means of transportation, weather conditions and the trajectory goal, composing what we call *multiple-aspect trajectory*. Therefore, the existence of similarity measures for analyzing multiple-aspect trajectories has become essential for performing data mining tasks, such as extracting movement patterns, associating trajectories with certain profiles regarding how they move, what they do or where they go, etc.

Trajectory similarity measuring has been deeply investigated in the last few years (some examples are (BERNDT; CLIFFORD, 1994), (BOLLOBáS et al., 1997), (VLACHOS; KOLLIOS; GUNOPULOS, 2002), (CHEN; ÖZSU; ORIA, 2005), (HOLT; REINDERS; HENDRIKS, 2007), (LIU; SCHNEIDER, 2012) and (FURTADO et al., 2015)) and it is still a challenge when dealing with several aspects behind movement data. Besides the space and time information that is intrinsic to trajectory data, large amounts of information from social media, sensors and weblogs can be used for trajectory analysis. The semantic enrichment of trajectories (ALVARES

et al., 2007; KRUEGER; THOM; ERTL, 2014) with context and social information leads to *multiple-aspect trajectories*. Examples are trajectories enriched with visited places, weather and traffic conditions during the movement, means of transportation, among others. All aforementioned information provide a large amount of relevant features regarding trajectory similarity, but such data have not been explored so far.

The analysis of trajectories with respect to their different aspects, i.e., distinct representations of the same trajectory, has become extremely important to better understand movement and to discover more interesting patterns. Multiple-aspect trajectory analysis can be widely applied from a simple GPS trajectory to the life trajectory of a person (NOËL et al., 2015), and it is one of the main challenges in current trajectory research (FERRERO; ALVARES; BOGORNY, 2016). Moreover, existing approaches for trajectory analysis and mining, in general, have limitations regarding semantic information of trajectories and they do not analyze these multiple aspects of trajectories in conjunction.

To the best of our knowledge, existing approaches have addressed the similarity of trajectories by comparing their points in two ways: (i) considering only points that match in all dimensions (e.g., space, time) (VLACHOS; KOLLIOS; GUNOPULOS, 2002; CHEN; ÖZSU; ORIA, 2005; HOLT; REINDERS; HENDRIKS, 2007) or; (ii) that match in at least one dimension (FURTADO et al., 2015). Besides that, no work in the literature has analyzed the similarity of trajectories from their multiple aspects.

Figure 1 presents different representations of the same trajectories P, Q and R, in order to illustrate the problem addressed in this work. Figure 1 (a) shows three raw trajectories, the simplest portrayal of trajectories. Figure 1 (b) displays the same trajectories represented by their stops, i.e., the places visited by the users for a minimal amount of time. Figure 1 (c) encompasses the previous representations and other aspects related to movement data, such as means of transportation, weather conditions and social media interaction, i.e., *multiple-aspect trajectories*.

From the point of view of raw trajectories or stops, it is clear that trajectories P and Q are more similar. In Figure 1 (a) P and Q are closer in space and in Figure 1 (b) they visit the same category of places (Home, Restaurant and Mall), thereby being more similar. When analyzing their multiple aspects (see Figure 1 (c)), however, it might not be trivial to infer their similarity. Regarding weather conditions, for example, P and R are more similar. When looking at means of transportation, Q and R are more similar. If we analyze the social media interactions, we are able to derive information not available in

raw movement data. For example, the user of trajectory P is less happy than the user of trajectory R, according to their postings on Twitter and on Facebook. We can also infer from the Facebook post of R that the user is eating, and similarly is the user of trajectory Q that checked in at the restaurant on Foursquare.

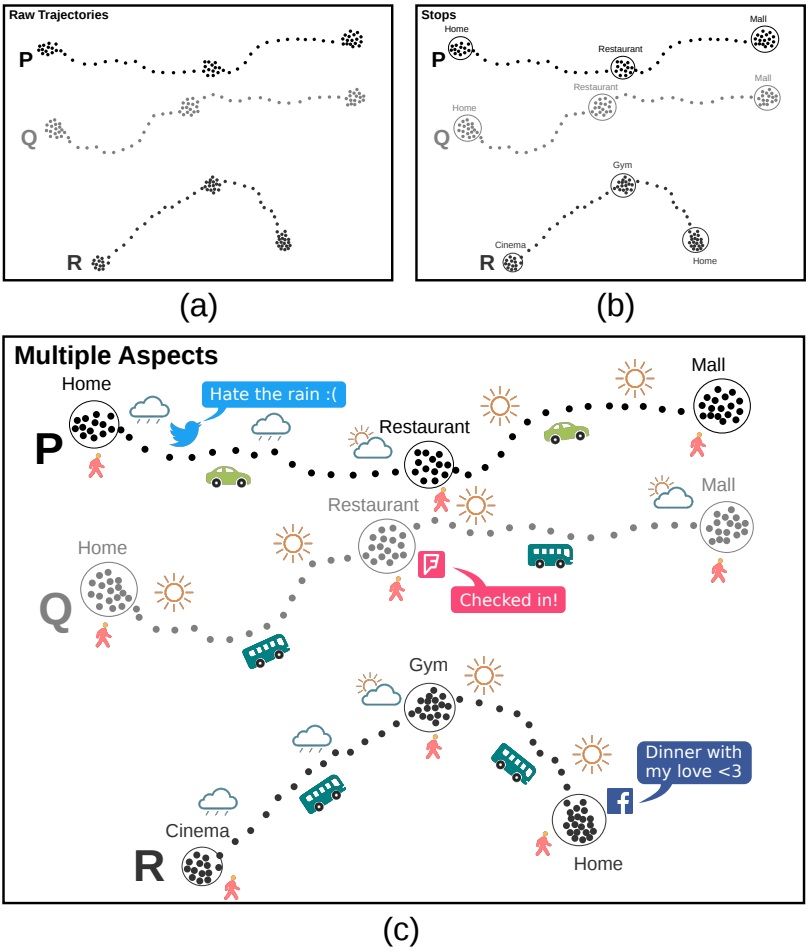

Figure 1 – Different trajectory representations.

The problem becomes even more complex when there is the need to combine different information, so a deeper analysis can be performed.

For instance, which users visit the same places when it is raining? Do the means of transportation change if the weather conditions change? Do people only go to highly rated places, regardless if they are cheap or expensive? From the users trajectories, which ones are of tourists and which ones are of residents? The multiple-aspect representation of trajectories opens a whole new world of possibilities never explored before. The main challenge here is how to group, store and deal with these different aspects all together, so that it becomes possible to analyze trajectories considering all the information available. In the following section we detail the problem addressed in this work.

1.1 PROBLEM STATEMENT

Suppose we have a dataset of trajectories of user check-ins on Foursquare. Every user check-in is connected to a venue on Foursquare, which is the place visited by the user. There is a lot of information and statistics about venues on Foursquare, such as their rating by users, how expensive they are, whether or not they take credit cards, and many others.

There are many factors behind user choices of places to visit, and these factors may be related to each other. For instance, let us say a user only goes to well-rated restaurants. Because of the high violence index in the area where the user lives, he only carries some cash and his credit card whenever he goes out. Therefore, for him to eat at an expensive restaurant they must take credit cards. Otherwise, he would not be able to afford it due to the small amount of cash he has. Notice that, in this example, there are a few *dependency relations* on the characteristics of the places he visits. A restaurant *must be cheap* or, *if it is expensive* then it *must take credit cards*. Additionally, the restaurant *must be well-rated*. When we aim to discover users whose trajectories have the described behaviour, the similarity measure employed must be able to capture these relationships, so that only the trajectories we are interested on, i.e., the ones that satisfy these conditions, are retrieved.

Existing similarity measures, however, are still limited to a few trajectory features and do not capture all of these dependency relationships. Measures like LCSS (VLACHOS; KOLLIOS; GUNOPULOS, 2002), EDR (CHEN; ÖZSU; ORIA, 2005) and MD-DTW (HOLT; REINDERS; HENDRIKS, 2007), for instance, only compare trajectories regarding their space and time dimensions. Additionally, they do not take into consideration partial similarity among dimensions, i.e., points of two trajecto-

ries only match if they are similar in both space and time, so becoming very rigid measures. As for our problem, if we extended these measures to work with semantic information, they would be able to capture the relationship between the rating, the price tier and the credit cards option of places. However, because they are rigid measures, they would not be able to retrieve trajectories that semantically match if they do not match in space and time.

More recently, Furtado et al. (2015) proposed a new similarity measure considering the partial matching for points of two trajectories and supporting all three dimensions of space, time, and semantics. The proposed measure, MSM, is very flexible, but such flexibility may lead to an inaccurate similarity analysis in some cases. For the problem described, MSM would not necessarily retrieve trajectories that follow the described relationships. If a point of a trajectory was similar to another point in everything, except for the price tier of the place, MSM would still result in a high similarity score, which is not the desired behaviour. Moreover, MSM always accounts the best match for every point of two given trajectories, regardless if their points were already considered in a previous match. For this reason, it may assign a high similarity score even if two trajectories are only similar for a small portion of their length. In other words, suppose a user A, whose trajectory has one check-in at a restaurant and 9 check-ins at the gym, and another user B who checked in at the same restaurant 9 times, but only once at the gym. The similarity computed by MSM, considering time, space and semantics, could be close to 100%[1]. In sum, MSM is not effective when analyzing repetitive behaviour, leading to inaccurate results.

In multiple-aspect trajectory similarity some characteristics or attributes[2] might be related and should be analyzed together, while others may be optional. In our example, when a place is expensive its price tier must be considered together with the credit cards option. Due to the fact that for some cases EDR, LCSS and MD-DTW may be too rigid and MSM too flexible, a new similarity measure that combines such characteristics becomes necessary.

[1]Given that semantic and spatial dimensions are given more importance than the time dimension.

[2]We use the terms *dimension* and *attribute* interchangeably.

1.2 OBJECTIVE

The general objective of this work is to propose a new similarity measure that allows the combination of different attributes of different aspects for analyzing *multiple-aspect trajectories*.

In order to overcome the limitations of previous works, the specific objectives to be achieved include:

(i) The design of a new flexible similarity measure for *multiple-aspect trajectories*;

(ii) The definition of *dependency relations* between attributes of points, so that the matching of points is more flexible or stricter, depending on the nature of the problem;

(iii) The computation of trajectory similarity observing the single matching of their points;

(iv) The use of different aspects of trajectories in the similarity analysis.

1.3 OUTLINE

The rest of this document is organized as follows. Chapter 2 describes basic concepts relevant to our work and summarizes related works and their limitations. In Chapter 3 we introduce the proposed similarity measure, its properties and results on a running example based on the problem statement defined in this chapter, and present an evaluation and comparison of our measure with existing ones, validating the accuracy and improvements made by our approach. Finally, in Chapter 4 we conclude by describing advantages and limitations of this work, in addition to potential future work.

2 BASIC CONCEPTS AND RELATED WORK

In this chapter we introduce the basic concepts that are fundamental to our work. We begin by defining distance and similarity, followed by definitions of raw, semantic and multiple-aspect trajectories. Subsequently, we describe related work in the literature.

2.1 BASIC CONCEPTS

2.1.1 Distance and Similarity

Distance and similarity are concepts largely used in database queries and data mining techniques, such as top-k queries, nearest neighbor classification methods, clustering analysis, among others. Similarity and distance quantify how similar or distant two objects are, respectively. In fact, they can be obtained directly from the objects - for example, "subjects in a food tasting experiment may be asked to state similarities between flavors of ice-cream", or indirectly measured from the objects characteristics. Once there is a formal definition of one of the measures, the other one can be easily computed (HAND; MANNILA; SMYTH, 2001). For instance, suppose we have objects A, B and C located at points $(1, 3)$, $(4, 7)$ and $(1, 5)$, respectively. Let us compute the distances among them using the *Euclidean distance*. The distances are presented below, where Equation 2.1 shows the distance between A and B; Equation 2.2 between A and C; and Equation 2.3 between B and C.

$$d(A, B) = \sqrt{(1-4)^2 + (3-7)^2} = 5 \tag{2.1}$$

$$d(A, C) = \sqrt{(1-1)^2 + (3-5)^2} = 2 \tag{2.2}$$

$$d(B, C) = \sqrt{(4-1)^2 + (7-5)^2} \approx 3.6 \tag{2.3}$$

We can use the *Euclidean distance* to see how similar the objects are. Thus, A is more similar to C than to B, because the distance between A and B is 5, but only 2 for A and C. The greater the distance of two objects is, the less similar they are.

2.1.2 Raw, Semantic and Multiple-Aspect Trajectory

A *raw trajectory* is essentially a sequence of points composed by a pair of spatial coordinates and a *timestamp*. Alvares et al. (2007) define a raw trajectory as "a *list of space-time points* $\langle (x_0, y_0, t_0), (x_1, y_1, t_1), \ldots, (x_N, y_N, t_N) \rangle$, where $x_i, y_i, t_i \in \mathbf{R}$ for $i = 0, \ldots, N$ and $t_0 < t_1 < \cdots < t_N$." Raw trajectories can be directly collected from GPS devices and may be generated by any moving object, such as a person, an animal, a car, a tornado, among others. Figure 1 (a) illustrates three raw trajectories P, Q and R.

By enriching raw trajectories with semantic information, we then create *semantic trajectories* (SPACCAPIETRA et al., 2008). One approach for semantic enrichment of trajectories is introduced by Alvares et al. (2007), whose work introduces an algorithm to extract stops and moves of trajectories. Stops are collections of sample points (x, y, t) that are very close in space and time, representing interesting spatial locations called *Points of Interest* (POIs). Every stop has a start and an end time, a spatial location and a minimal duration. Moves, on the other hand, are composed by the sample points between stops (or at the beginning or at the end of a trajectory). Figure 1 (b) depicts the stops of three trajectories, labeled with categories of POIs representing the semantics of the trajectories.

The use of semantics in the context of trajectories, including aspects such as weather conditions, social media interactions, means of transportation, etc, defines what we call *multiple-aspect trajectory*. There is no formal definition for multiple-aspect trajectory in the literature. Bogorny et al. (2014), for instance, introduced a model for trajectory representation comprising several aspects related to movement, but they are limited to means of transportation, environment conditions, places visited by the object, and events related to the trajectory goal. Figure 1 (c) illustrates the same three trajectories previously presented, including aspects of weather, means of transportation, the visited POIs and interactions with social media. Ferrero, Alvares and Bogorny (2016) later emphasized the need for similarity analysis of trajectories regarding their multiple representations. In this work, we introduce a formal definition for *multiple-aspect trajectory*, which is described in Definition 1.

Definition 1. A multiple-aspect trajectory is a sequence of points $\mathcal{T} = \langle p_1, p_2, \ldots, p_n \rangle$, with $p_i = \{A_1, A_2, \ldots, A_j\}$ being the i-th point of the trajectory composed of j aspects, where $A_i = \{a_1, a_2, \ldots, a_l\}$ is

an aspect composed by l characterizing attributes, and these attributes may refer to space, time or any other attributes, and aspects referring to space and time are mandatory.

A point of a multiple-aspect trajectory can be, for instance, a sample point of a raw trajectory with the attributes space and time, as shown in Figure 1 (a); the stops of a trajectory (space, time and the POI category) as presented in Figure 1 (b); or a complex element with any other aspects. As commented before, an aspect may be the means of transportation, weather conditions, the POI and its characteristics, the person's humor or feelings, among others. The attributes of the aspect means of transportation could be the name (bus, car, foot) and the speed; for weather conditions there could be the temperature and humidity; and the POI could have its category, its rating in a social media network, the price tier, among others.

As an example, one could define aspects A_1 as space, A_2 as time, A_3 as POI information and A_4 as weather conditions, as follows:

$$A_1 \begin{cases} x \\ y \end{cases} \qquad A_2 \{t \qquad A_3 \begin{cases} POI\ category \\ rating \\ price\ tier \end{cases} \qquad A_4 \begin{cases} temperature \\ humidity \\ weather\ condition \end{cases}$$

A trajectory T_1 could then be instantiated, with each point holding information about every aspect.

$$\begin{aligned} T_1 = \langle\ & \{(2,4),(1),(Restaurant, Good, Cheap),(31.2, 0.70, Sunny)\}, \\ & \{(5,9),(2),(Market, Medium, Cheap),(33.2, 0.75, Cloudy)\}, \\ & \{(22,2),(3),(Hotel, Good, Expensive),(25.5, 0.80, Rainy)\}, \\ & \{(27,3),(4),(Nightclub, Good, Expensive),(22.0, 0.79, Rainy)\}\ \rangle \end{aligned}$$

2.2 RELATED WORK

The similarity of sequences and time series was the primary problem discussed in the literature, long before first works started analyzing actual trajectories. Agrawal, Faloutsos and Swami (1993) discussed the

problem of similarity among sequenced data and proposed an indexing technique to efficiently process similarity queries in sequence databases. Faloutsos, Ranganathan and Manolopoulos (1994) extended their work, so they were able to search for similar sequences that not necessarily fully match.

Despite of the pioneering of these works, a well-known algorithm for similarity measurement between time series was designed by Berndt and Clifford (1994), called *Dynamic Time Warping (DTW)*. The DTW algorithm aligns two sequences in order to minimize the distance between their elements. A matrix with the distances between elements of both series is created, which is then used to find the contiguous path with the minimum total distance between the series. Given DTW's limitation to uni-dimensional data, Holt, Reinders and Hendriks (2007) extended DTW to create *Multidimensional Dynamic Time Warping (MD-DTW)*. MD-DTW normalizes the distance of elements in all dimensions and then builds the distance matrix, whose elements are the sum of the distances in all dimensions for every two elements in the sequences. DTW and MD-DTW tend to be sensitive to noise because all elements of the sequences being compared are taken into consideration. If at least one of the elements of a sequence is far away from all the ones in the other sequence, the whole similarity may be affected.

The *Longest Common Subsequence (LCSS)*[1] was introduced as a robust similarity measure for trajectories (VLACHOS; KOLLIOS; GUNOPULOS, 2002). It is based on the longest common subsequence concept, in which two sequences are considered to be similar if they have similar behavior for a large part of their length. Differently than DTW and MD-DTW, LCSS reduces the impact of noisy data by defining distance and matching thresholds. Two elements match and are assigned a similarity value of 1 if their distance lies below the matching threshold; otherwise, they do not match and have a similarity of 0. Although it works well with noise, LCSS has some disadvantages. First, two elements match only if they are close in all dimensions. Additionally, LCSS ignores possible gaps in sequences, which, for certain problems, would mean giving the same similarity value for different pairs of trajectories.

Chen, Özsu and Oria (2005) measured the similarity of trajectories similarly to LCSS. *Edit Distance on Real sequence (EDR)*, as the measure was named, is based on Edit Distance (ED), widely used for

[1] Even though LCSS was first designed for time series by Bollobás et al. (1997), we only consider the most recent approach proposed by Vlachos, Kollios and Gunopulos (2002) for trajectory data, since it is more robust than the first one.

measuring similarity between strings. The underlying idea in EDR is that, being A and B two trajectories, $EDR(A, B)$ is given by the minimum number of insert, delete and replace operations needed to change A into B. Clearly, a matching threshold must be defined. Like LCSS, EDR assigns 1 when two elements are similar and 0 otherwise. Besides reducing the effects of noise and dealing with local time shifting, EDR overcomes a major drawback present in LCSS. It assigns penalties according to the length of the gaps between two matched sub-sequences, which results in more accurate results than those reached by LCSS. However, a match occurs only if all dimensions match for any two elements.

Another work, by Liu and Schneider (2012), computes the similarity of trajectories by considering both geographic and semantic features. They segment trajectories into sub-trajectories based on changes in direction and in the speed of the movement. The authors then compute the geographic distance between two sub-trajectories, using the distance between their centroids, the difference of their length, and the cosine similarity of their directions. Afterwards, a semantic ratio is calculated based on the LCSS measure, to compare the sequence of visited places. Finally, the semantic ratio is combined with the geographic similarity and a constant value to form the total distance between trajectories. One issue with this approach is that it ignores the time dimension.

More recently, Furtado et al. (2015) presented a new similarity measure that overcame most limitations of previous works. Essentially, given two trajectories A and B, for every point of A, the *Multidimensional Similarity Measure (MSM)* looks for the best match in sequence B. Subsequently, the weighed scores of the matches are added to compose the parity of A with B. Since the parity is not symmetric, $MSM(A, B)$ is computed by the "average" of $parity(A, B)$ and $parity(B, A)$. Rather than considering pairs of elements only if they match in all dimensions, MSM treats all dimensions separately and assigns partial similarity according to the number of dimensions in which the elements match. MSM also allows one to define different weights for every dimension, given that a dimension might be of more or less importance for different problems. The high flexibility of MSM, however, may not be appropriate to certain problems. For instance, two semantic trajectories A and B that visit the same places in a different order are identified to be very similar. In a traffic management application, for example, objects moving in the opposite direction should be very dissimilar. In addition, MSM disregards any dependency relations

that might exist between dimensions and that could be important in multiple-aspect analysis. For instance, if we want to analyze the means of transportation under certain weather conditions, the attributes of means of transportation and weather conditions are related and should be analyzed together. Lastly, MSM always accounts the best match for every element of two given trajectories, regardless if elements were already considered in a previous match. For that reason, it may assign a high similarity score even if two trajectories are only similar for a small portion of their length.

Another measure, proposed by Furtado (FURTADO et al., 2017), is a new distance and similarity measure for raw trajectories called Uncertain Movement Similarity (UMS), which is more robust than previous works regarding different sampling rates and the heterogeneity of trajectory data. Although UMS is more robust than works such as DTW, LCSS and EDR, it is also limited to the spatial dimension, presenting the same limitations of previous works regarding semantic information.

Sharif and Alesheikh (2017) define a context model for trajectories, subdivided in four specific contexts called *motivation*, *movement*, *modality* and *milieu*. Subsequently, they propose a new similarity measure based on DTW (BERNDT; CLIFFORD, 1994), which essentially computes DTW on each context and adds it all together according to specific weights defined by the user. Even though weights are defined for each context, if two trajectories are very distant in one of the contexts, the whole similarity may be affected. Moreover, this approach disregards any relationships that may exist between different contexts.

Ferrero, Alvares and Bogorny (2016) focus on the need for analyzing trajectories from different points of view, which they called aspects. According to their work, only a few trajectory data analysis and mining methods in the literature have considered several dimensions of trajectories, and none of them have taken into account distinct representations of a single trajectory for similarity analysis. Multiple-aspect similarity evaluation would allow us to consider different paradigms when measuring trajectory similarity, such as spatial location, semantics of the visited places, weather conditions, means of transportation used throughout the trajectory, and others.

So far, to the best of our knowledge, there is no work in the literature on multiple-aspect trajectory similarity. Indeed, previously mentioned works address trajectory similarity, regarding trajectory dimensions, either in a too restrictive or too flexible manner. For instance, while DTW, MD-DTW, LCSS and EDR consider that elements should match in all dimensions together, MSM considers pairs of elements that

Table 1 – Features and limitations of existing similarity measures.

Feature	**MD-DTW**	**LCSS**	**EDR**	**MSM**	**Proposed Measure**
Noise (outliers)		✓	✓	✓	✓
Gaps in trajectories			✓	✓	✓
Order	✓	✓	✓	✓[2]	✓
Semantic trajectories	✓[3]	✓[3]	✓[3]	✓	✓
Dimensional weighing				✓	✓
Unordered events[4]				✓	✓
Restrictive match of elements[5]	✓	✓	✓		✓
Partial match of elements[6]				✓	✓
Multiple-aspect trajectories					✓
Attributes dependency relations[7]					✓

match in a single dimension.

Table 1 compares characteristics of the main discussed approaches and our work, such as robustness to noise, ability to handle various dimensions or aspects, account of partial similarity, among others. As shown in Table 1, the similarity measure being proposed has the challenge to group together the main characteristics of other works, and support multiple-aspect trajectories.

[2]Order may be considered partially.
[3]The measure may be extended for semantic information.
[4]The measure accounts similarity for different sequences of the same events.
[5]The measure enforces the full match of dimensions between pairs of elements.
[6]Ability to account partial similarity between elements.
[7]Ability to define relations of dependence between attributes, mixing partial and restrictive matching.

3 THE PROPOSED SIMILARITY MEASURE

In this chapter we propose a new similarity measure for *multiple-aspect trajectories*. As stated before, existing approaches address trajectory similarity by comparing their points either by considering only points that match in all the dimensions, or points that match for any dimension independently. Therefore, in the next section we introduce a flexible multiple-aspect similarity measure that allows the definition of *dependency relations* between attributes of aspects. The measure is flexible to the point that it can behave similarly to MSM (FURTADO et al., 2015); or be stricter and behave like LCSS (VLACHOS; KOLLIOS; GUNOPULOS, 2002), EDR (CHEN; ÖZSU; ORIA, 2005) and MD-DTW (HOLT; REINDERS; HENDRIKS, 2007), if necessary.

3.1 MUITAS: MULTIPLE-ASPECT TRAJECTORY SIMILARITY MEASURE

In this section we introduce the basic definitions that are essential to our work, followed by the proposed similarity measure, MUITAS. Afterwards, the similarity problem is described as the General Assignment Problem (GAP) (KUHN, 1955) and an algorithm to compute the similarity is introduced.

In (FURTADO et al., 2015), trajectory similarity measuring was performed with *each dimension* corresponding to a *single attribute* of a trajectory, and the dimensions were analyzed independently. This means, for example, that two trajectories P and Q that visit the same places at different times still have 50% of similarity. When dealing with multiple-aspect trajectories, however, some dimensions should not be treated separately and must be aggregated as a complex feature. For instance, to compare if two objects travel by the same *means of transportation* under the same *weather conditions*, these multiple dimensions must be considered together, because they are related (or dependent) to each other.

In order to measure the similarity between two multiple-aspect trajectories it is necessary to quantify the distance between points. For each point we must quantify the distance for each attribute, since different attributes may belong to different aspects, and so having distinct natures they require different distance functions. We define the attribute distance measuring as follows.

Definition 2. *Attribute matching.* Let P and Q be two multiple-aspect trajectories $P = \langle p_1, p_2, \ldots, p_m \rangle$ and $Q = \langle q_1, q_2, \ldots, q_n \rangle$, and let $A = \{a_1, a_2, \ldots, a_l\}$ be an aspect with l attributes. Subsequently, for any two elements $p \in P$ and $q \in Q$, the distance between p and q on attribute a_i is given by the function $dist_{a_i} : P \times Q \to \mathbb{Q}$. Two elements $p \in P$ and $q \in Q$ will *match* on attribute a_i if $dist_{a_i}(p, q) \leq \delta_i$, where δ_i is a distance threshold for attribute a_i.

For our problem definition in Chapter 1, there were attributes like *rating*, *price tier*, and the *credit cards* option of payment at places. For the rating attribute ranging from 4 to 10, one could define a distance function as the one in Equation 3.1, and a threshold $\delta_{Rating} = 1$. The dot in $p.rating$ means we are accessing the attribute *rating* of point p.

$$dist_{Rating}(p, q) = |p.rating - q.rating| \tag{3.1}$$

Thus, two check-ins on venues rated 7.8 and 8.4 would match on attribute *rating*, because $|7.8 - 8.4| = 0.6$, which is less than 1 (δ_{Rating}). *Credit cards*, on the other hand, is a binary attribute with values Yes and No. A possible distance function is shown in Equation 3.2.

$$dist_{CreditCards}(p, q) = \begin{cases} 0, & \text{if } p.credit\_cards = q.credit\_cards \\ \infty, & \text{otherwise} \end{cases} \tag{3.2}$$

In multiple-aspect trajectory similarity measuring we tackle a very important issue that refers to some sets of attributes that must be considered together in the matching process. In other words, there might be a dependency between a subset of attributes such that they should all match and must be considered as a whole, while others should be considered independently. We saw in the research problem in Chapter 1 that the user goes to restaurants that are *well-rated and cheap*, *or* the ones *well-rated*, *expensive and* that accept payments with *credit cards*. The rating, the price tier and the credit cards option are related to each other, but the space and time attributes are not related to anything. To address the relationships of attributes, we introduce the concept of feature, which will be the unit of analysis for measuring similarity.

Definition 3. *Feature.* A feature $f = \{a_1, a_2, \ldots, a_z\}$ is a nonempty set of attributes that describe a unit of analysis of a multiple-aspect trajectory. Let $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ be the set of features through which trajectories are analyzed. For each feature $f_i \in \mathcal{F}$ we define a

corresponding weight w_i, so that $\sum_{i=1}^{|\mathcal{F}|} w_i = 1$.

The weight w_i of a feature f_i represents the importance of that feature for computing the similarity between trajectories for a specific application. For our research problem, we should define a feature $f_1 = \{Rating, PriceTier, CreditCards\}$ in order to express the attributes relationships discussed before. Other features $f_2 = \{Space\}$ and $f_3 = \{Time\}$ could be defined, so the similarity of check-ins would be a bit lower if they occurred in different locations or different times (according to distance functions $dist_{Space}$ and $dist_{Time}$, and thresholds δ_{Space} and δ_{Time}) and also because space and time are independent attributes. To avoid misunderstanding and conflict of concepts, we hereafter refer to *attribute* as an atomic view of a point, and to *feature* as the unit of analysis of a trajectory.

As the important features for similarity analysis are application dependent, we give the formal definition of *Application* in Definition 4. We define an application according to the features, the distance functions and distance thresholds used in the analysis.

Definition 4. *Application.* An application $\mathbb{A}$ is defined by a tuple $\mathbb{A} = (\mathcal{F}, \mathcal{D}, \Delta)$, where $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$ is a nonempty set of features, $\mathcal{D} = \{dist_{a_1}, dist_{a_2}, \ldots, dist_{a_l}\}$ is a nonempty set of distance functions and $\Delta = \{\delta_1, \delta_2, \ldots, \delta_l\}$ is a nonempty set of distance thresholds.

An application essentially defines the context of the problem, i.e., how trajectories will be analyzed. Different applications may imply different features, distance functions and/or different thresholds. We must now define how to measure the similarity between points of trajectories, and the trajectories themselves, under a defined application $\mathbb{A}$.

Definition 5. *Score.* Given two points $p \in P$ and $q \in Q$ and an application $\mathbb{A} = (\mathcal{F}, \mathcal{D}, \Delta)$, the matching score between p and q is given by the function $score : P \times Q \to [0, 1]$, defined as follows

$$score(p, q) = \sum_{k=1}^{|\mathcal{F}|} (match_{f_k}(p, q) * w_k),$$

$$\text{where } match_{f_k}(p, q) = \begin{cases} 1, & \text{if } \forall a_l \in f_k,\ dist_{a_l}(p, q) \leq \delta_l \\ 0, & \text{otherwise} \end{cases}$$

At this point, we have the basic definitions necessary to, indeed, propose the similarity measure. The work of Furtado et al. (2015) (MSM) defines a parity function which is the basis of the similarity measure. The parity function adds the scores of the best matches of the points of one trajectory with points of another trajectory, regardless if a point is matched more than once. As stated in Chapter 1, if a user A checked in once at a restaurant and 9 times at the gym, and another user B checked in at the same restaurant 9 times, but only once at the gym, the similarity computed by MSM could be close to 100%. Thus, differently from MSM, we define a *map* function, which is an important concept behind our measure and fundamental for overcoming this limitation of MSM.

Definition 6. *Map function.* Given two multiple-aspect trajectories P and Q such that $|P| \leq |Q|$, a map function is an injection $map : P \rightarrow Q$ that maps all the elements of P to distinct elements in Q.

The goal is to map all the points in the shorter trajectory to distinct points in the possibly longer trajectory. Figure 2 (a) shows a mapping example for two trajectories of the same length, and Figure 2 (b) portrays the mapping of two trajectories of different lengths. If P and Q had the same length, *map* would be a bijection, because all the points of both sequences would be in the mapping. Otherwise, some points would be left out (as displayed in Figure 2 (b)).

It is important to emphasize that the shorter trajectory (if the trajectories are not of the same length) is kept in the domain of function *map* so that *map* may be an injection (only for definition matters) and, therefore, it is guaranteed that there is no pair of elements p_1 and p_2 in P, such that $map(p_1) = map(p_2)$. Similarly, one could define *map* as a surjection with the shorter trajectory in the co-domain.

Finally, Definition 7 details the similarity of two multiple-aspect trajectories.

Definition 7. *MUITAS Similarity.* Let $\mathcal{M}$ be the set of all *map* functions for two given multiple-aspect trajectories P and Q, with $|P| \leq |Q|$. The similarity of P and Q is defined as

$$MUITAS(P,Q) = max\left(\left\{\frac{2 \cdot \sum_{p \in P} score(p, m(p))}{|P| + |Q|} : m \in \mathcal{M}\right\}\right)$$

Definition 7 states that the similarity of two trajectories is given

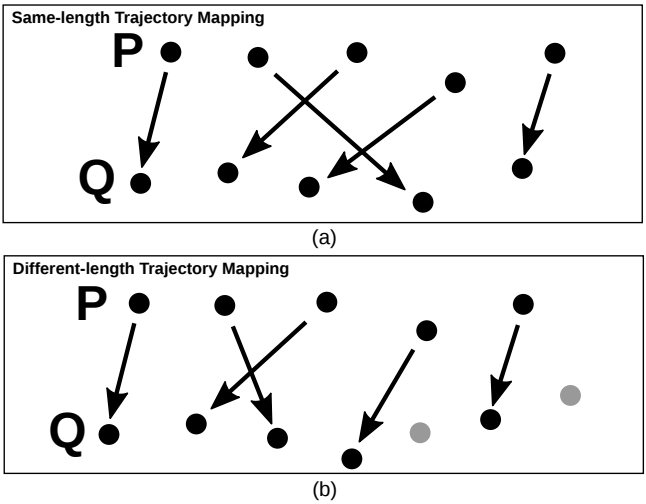

Figure 2 – Trajectory mapping example.

by the mapping of scores between points with the highest global similarity value. In the example of users A and B from the problem statement in Chapter 1, while MSM could assign the users a similarity score close to 100%, MUITAS would result in at most 20% of similarity.

Given two trajectories of length n, there are $n!$ possible *map* functions for these trajectories. Computing the scores of the trajectories for all of these $n!$ functions is not computationally feasible. Therefore, we address the problem through an analogous problem known in the literature, which is the Generic Assignment Problem (GAP) (KUHN, 1955).

3.1.1 The Similarity Problem as The General Assignment Problem (GAP)

Kuhn (1955) introduces the General Assignment Problem (GAP) as the problem of assigning n tasks to n men, such that the performance of the men in their assigned tasks is the maximum. Also, every men must be assigned a job and no job can be assigned to two different men. Subsequently, a dual problem is defined where there is a cost for every men to perform each of the n tasks, and the goal is to assign the tasks to the men in order to minimize the total cost of performing the jobs.

The number of possible assignments is $n!$ and, therefore, it is not viable to analyze every possible assignment. In his work, Kuhn describes a method for solving the problem in polynomial time.

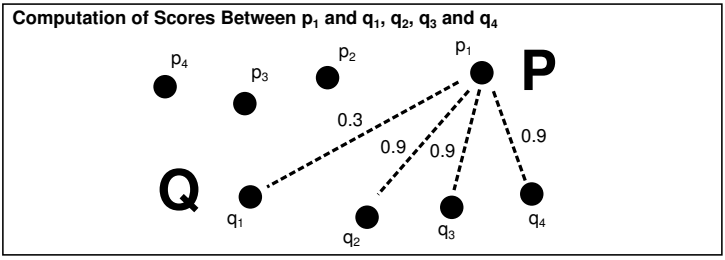

Figure 3 – Computation of scores of between pairs of points.

Now, suppose we have two trajectories P and Q containing 4 points each, as shown in Figure 3. After computing the scores between all pairs of points from distinct trajectories, we can build a matrix S of similarities, which is a $|P| \times |Q|$ matrix, with $s_{ij} = score(p_i, q_j)$, where s_{ij} is the element on row i and column j of S, p_i is the i-th point of P and q_j is the j-th point of Q. Figure 3 illustrates the computation of the scores between point p_1 of P and all points of Q, which is then stored in S, presented as follows.

$$S = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{cccc} q_1 & q_2 & q_3 & q_4 \end{array} \\ \begin{bmatrix} 0.3 & 0.9 & 0.9 & \mathbf{0.9} \\ 0.9 & \mathbf{0.4} & 0.2 & 0.1 \\ 0.9 & 0.1 & \mathbf{0.3} & 0.2 \\ \mathbf{0.9} & 0.1 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

Similarly to the GAP presented by Kuhn (1955), given the matrix of similarities, we intend to "assign" points of P to points of Q, such that the sum of the total similarity between points is the maximum possible and no point from both trajectories is assigned twice. We can easily verify that the scores highlighted in S are the mapping of points with the highest global similarity score. By that we conclude that the set of scores $\{score(p_1, q_4),\ score(p_2, q_2),\ score(p_3, q_3),\ score(p_4, q_1)\}$ are the mapping with the greater similarity score for P and Q, resulting in a similarity of 0.625 (Equation 3.3).

$$MUITAS(P,Q) = \frac{2 \cdot (0.9 + 0.4 + 0.3 + 0.9)}{4 + 4} = 0.625 \tag{3.3}$$

MSM (FURTADO et al., 2015), on the other hand, would assign P and Q a similarity of 0.9 (Equation 3.4), because it makes use of a parity function that finds the best match for every point in one trajectory, regardless if a point was matched more than once.

$$MSM(P,Q) = \frac{4 \cdot 0.9 + 4 \cdot 0.9}{4 + 4} = 0.9 \tag{3.4}$$

For this specific problem, notice that point p_1, unlike other points in P, is very similar to q_2, q_3 and q_4 (their similarity is 0.9). Similarly, there is a high similarity between q_1 and p_2, p_3 and p_4. Hence, MSM matches points q_2, q_3 and q_4 with p_1, and points p_2, p_3 and p_4 with q_1, assigning an inaccurate similarity score between P and Q.

Lastly, it is important to highlight that even though the two trajectories in our example have the same length, it is possible to extend the problem for trajectories of different lengths (or non-square matrices). The problem is addressed by Bourgeois and Lassalle (1971). The next section presents and describes an algorithm for computing the similarity of two trajectories. We also describe the experimental evaluation in the following sections, showing that the behavior of MSM may be an issue when dealing with semantic information of trajectories for certain types of problems.

3.1.2 An Algorithm for Computing the Similarity

Algorithm 1 shows how to compute the similarity score of two trajectories. It takes as input the two multiple-aspect trajectories and outputs the similarity degree. The first step is to compute the similarity matrix (lines 6-12). The function $score()$ in line 10 is the score function of Definition 5. For two trajectories with n points each, the similarity matrix S is computed in $O(n^2)$ time.

Subsequently, the indexes of the similarities that are the solution to the problem, i.e., the mapping of points with the highest possible total similarity, are calculated by the function $computeMaximumIndexes()$ (line 12). This function is essentially the execution of an algorithm that solves the GAP. For the similarity matrix S presented in the previous section, for example, indexes would be the set $\{(1,4),(2,2),(3,3),(4,1)\}$.

Algorithm 1 MUITAS

```
1  Input: The multiple-aspect trajectories T1 and T2
2  Output: the similarity score
3  Begin:
4      Let S be a matrix of similarities |T1|x|T2|;
5
6      For i from 0 to |T1| - 1 do:
7          For j from 0 to |T2| - 1 do:
8              Let Pi be the i-th point of T1;
9              Let Pj be the j-th point of T2;
10             Let S[i][j] = score(Pi, Pj);
11         End For
12     End For
13
14     Let indexes = computeMaximumIndexes(S);
15     Let total = 0;
16
17     For each (i, j) in indexes do:
18         Let total = total + S[i][j];
19     End For
20
21     Return 2*total / (|T1| + |T2|);
22 End
```

Algorithms for solving the GAP are proposed in the works of Kuhn (1955), Munkres (1957) and Bourgeois and Lassalle (1971). The best algorithms are executed in $O(n^3)$ time.

Finally, the total similarity can be computed by adding the similarity values indicated by the vector of indexes (lines 17-21). For two trajectories with n points each, the total similarity is computed in $O(n)$ time. Therefore, the computation of our similarity measure takes $O(n^3)$ time.

3.2 RUNNING EXAMPLE

In this section we present a running example for which existing measures give undesired results. We consider a generic application that consists of five user check-ins on Foursquare. Figure 4 presents the five trajectories that will be analyzed. The check-ins are presented along with the POI name, how expensive the POI is (\$), its rating ($\star$), whether or not the place takes credit cards (CC) and the time that the user checked in (HH). The POI type is implied from its name.

We are interested on Anna's trajectory, because she has a similar

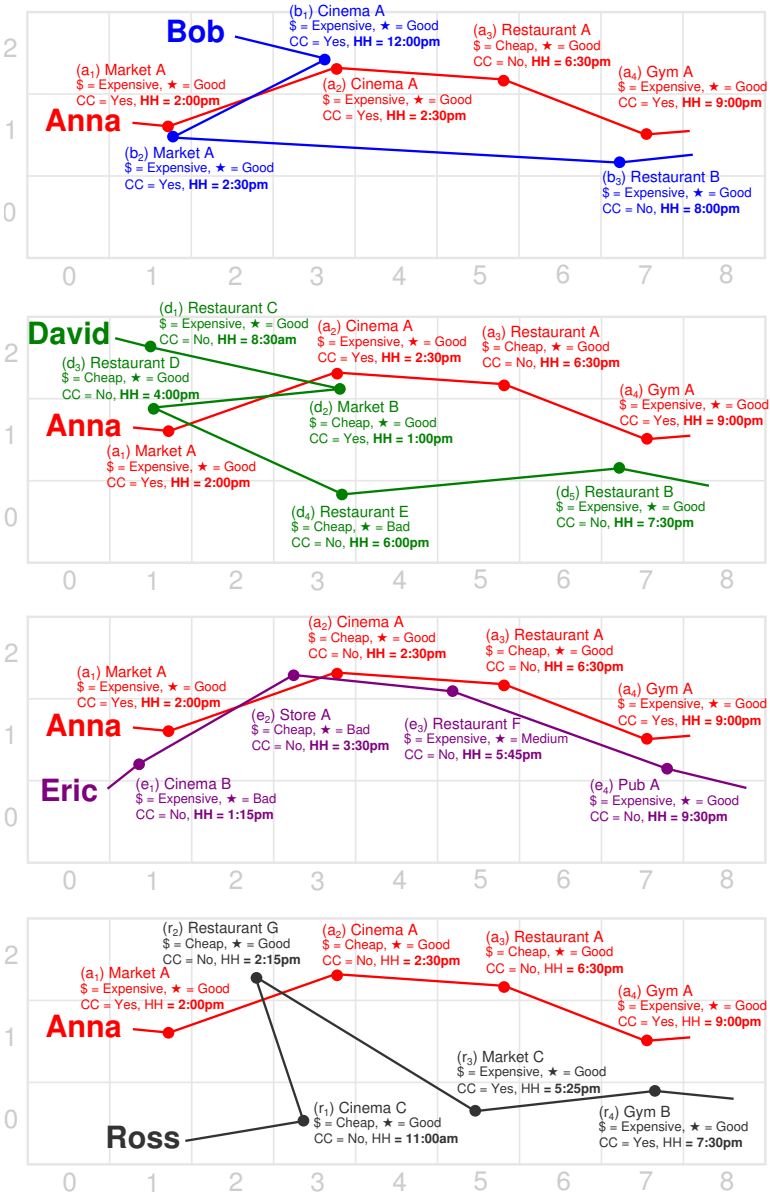

Figure 4 – Users check-ins on Foursquare.

behaviour to the one described in the problem statement in Chapter 1. She only goes to places that are *well-rated*, and they must be *cheap* or, *if they are expensive*, then they must *accept credit cards*. We start by computing the similarity of Anna and Bob, step by step, in order to illustrate the computation of the similarity measure. Afterwards, we evaluate the similarity of all trajectories against Anna's trajectory and compare with existing works.

3.2.1 MUITAS Step by Step

Let us compute the similarity of trajectories of Anna and Bob in Figure 4. The first step is to define the distance functions and distance thresholds of the attributes. Table 2 shows the functions and thresholds that will be used in our running example. For the sake of simplicity, we defined the *price tier* as a binary attribute, and most of the distances are based on the equality of the attributes. For the space, we considered the *manhattan distance* according to the cell line (y) and column (x) numbers, and two points are similar if their distance is zero, i.e., they are in the same cell.

Table 2 – Distance functions and thresholds for the attributes.

Attribute	**Distance Function**	δ
POI Type	$dist(p,q) = \begin{cases} 0, & \text{if } p.poi\_type = q.poi\_type \\ \infty, & \text{otherwise} \end{cases}$	0
Price Tier	$dist(p,q) = \begin{cases} 0, & \text{if } p.price\_tier = q.price\_tier \\ \infty, & \text{otherwise} \end{cases}$	0
Rating	$dist(p,q) = \begin{cases} 0, & \text{if } p.rating = q.rating \\ \infty, & \text{otherwise} \end{cases}$	0
Credit Cards	$dist(p,q) = \begin{cases} 0, & \text{if } p.credit\_cards = q.credit\_cards \\ \infty, & \text{otherwise} \end{cases}$	0
Space	$dist(p,q) = \lvert p.x - q.x \rvert + \lvert p.y - q.y \rvert$	0
Time	$dist(p,q) = interval(p.time, q.time)$[1]	3600

Subsequently, we must define the features that will be used for the similarity analysis. In order to ensure that the objects similar to Anna only visit well-rated places (because Anna only goes to well-rated places), we must have the *rating* attribute in every feature where the semantics is relevant, otherwise a matching of other attributes but not in the *rating* could result in a high similarity. Additionally, the places

[1] Function $interval(t_1, t_2)$ returns the number of seconds between time t_1 and t_2.

must take credit cards if they are expensive. Therefore, the corresponding attributes *price tier* and *credit cards* need to be analyzed together. Here the *rating* should also be analyzed with the *price tier* and the *credit cards* attribute, so there will be no match of check-ins if the *rating* is not good for whichever value *price tier* and *credit cards* have.

Table 3 presents the defined features and the corresponding weights. We also define features for analyzing the *space*, *time* and *POI type* attributes. They are kept separate from any other features because they do not hold any relationship with other attributes. However, we would still like to account some similarity if there is a match in *space*, in *time* or for the *POI type*, since that may indicate a similar behaviour. For that reason, we set lower weights for features f_2, f_3 and f_4, than for feature f_1. Given that feature f_1 represents the relationship between attributes that we are most interested, we give it a weight of 0.7.

Table 3 – Attributes and weights of features.

Feature	**Attributes**	w
f_1	*Rating*, *Price Tier*, *Credit Cards*	0.7
f_2	*POI type*	0.1
f_3	*Space*	0.1
f_4	*Time*	0.1

The application $\mathbb{A}$ can now be defined, according to definition 4, as the tuple $(\mathcal{F}, \mathcal{D}, \Delta)$, where $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$, $\mathcal{D} = \{dist_{POIType}, dist_{PriceTier}, dist_{Rating}, dist_{CreditCards}, dist_{Space}, dist_{Time}\}$ and $\Delta = \{\delta_{POIType}, \delta_{PriceTier}, \delta_{Rating}, \delta_{CreditCards}, \delta_{Space}, \delta_{Time}\}$. Once set the application $\mathbb{A}$, we are ready to compute the scores between every pair of points of Anna and Bob's trajectories.

Let us start by computing the score between points a_1 and b_1 (see Figure 4). The matches on each feature are presented as follows:

- $match_{f_1}(a_1, b_1) = 1$, because the price and the rating of the POI in a_1 and b_1 are the same, and they both take credit cards;

- $match_{f_2}(a_1, b_1) = 0$, because the POI of a_1 is a market, while the one of b_1 is a cinema;

- $match_{f_3}(a_1, b_1) = 0$, because $dist(a_1, b_1) = |1 - 2| + |1 - 3| = 3$, which is greater than 0 (δ_{Space});

- $match_{f_4}(a_1, b_1) = 0$, because $interval$(2:00pm,12:00pm) = 7200, which is greater than 3600 (δ_{Time}).

Lastly, Equation 3.5 shows the score between a_1 and b_1.

$$\begin{aligned} score(a_1, b_1) &= match_{f_1}(a_1, b_1) * 0.7 + match_{f_2}(a_1, b_1) * 0.1 \\ &\quad + match_{f_3}(a_1, b_1) * 0.1 + match_{f_4}(a_1, b_1) * 0.1 \quad (3.5) \\ score(a_1, b_1) &= 0.7 \end{aligned}$$

We can compute the score between points a_1 and b_2 in a similar manner. The matches on each feature are given as follows.

- $match_{f_1}(a_1, b_2) = 1$, because the POI in a_1 is the same as the one in b_2;
- $match_{f_2}(a_1, b_2) = 1$, because the POIs in a_1 and in b_2 are the same, which is a market;
- $match_{f_3}(a_1, b_2) = 1$, because $dist(a_1, b_2) = |1 - 1| + |1 - 1| = 0$ and $0 \leq \delta_{Space}$;
- $match_{f_4}(a_1, b_2) = 1$, because $interval$(2:00pm,2:30pm) = 1800, which is less than 3600 (δ_{Time}).

Equation 3.6 shows the score between a_1 and b_2.

$$\begin{aligned} score(a_1, b_2) &= match_{f_1}(a_1, b_2) * 0.7 + match_{f_2}(a_1, b_2) * 0.1 \\ &\quad + match_{f_3}(a_1, b_2) * 0.1 + match_{f_4}(a_1, b_2) * 0.1 \quad (3.6) \\ score(a_1, b_2) &= 1 \end{aligned}$$

After computing the scores between every pair of points of the two trajectories, we can build a matrix S of similarity scores. The full matrix S for the similarity analysis of Anna and Bob is presented as follows.

$$S = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \left[\begin{array}{ccc} 0.7 & 1.0 & 0.0 \\ 0.9 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.1 \\ 0.7 & 0.7 & 0.2 \end{array}\right] \end{array}$$

With the matrix of similarities, we solve the maximization problem using one of the approaches aforementioned (KUHN, 1955; MUNKRES, 1957; BOURGEOIS; LASSALLE, 1971). The scores selected by the algorithm will be $score(a_1, b_2)$, $score(a_2, b_1)$ and $score(a_4, b_3)$. Figure 5

shows the attributes considered in the analysis of the similarity of Anna and Bob, and any other attributes not considered are faded. The resulting similarity score of Anna and Bob is

$$MUITAS(Anna, Bob) = \frac{2 \cdot (1.0 + 0.9 + 0.2)}{4 + 3} = 0.6$$

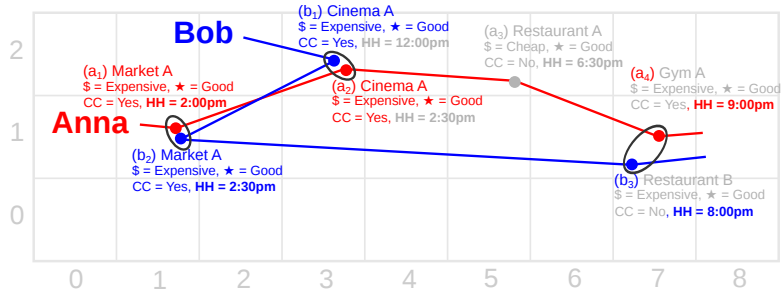

Figure 5 – Mapping of Anna and Bob check-ins.

If we take a look on the trajectories of Anna and Bob, the result represents exactly what we expected. Bob visited only one place that was not the behaviour we were looking for (check-in b_3 was at a place that is expensive and does not accept credit cards). However, he went to two places where Anna also went to, which explains the 0.6 score. In the next section we evaluate the similarity results for existing works and the proposed measure for the same application on the same set of trajectories shown in Figure 4.

3.2.2 Evaluation with the Running Example

Now we evaluate the accuracy of the proposed similarity measure on the running example. Our objective is to find out which of the four users in our toy example is the most similar to Anna, according to application $\mathbb{A}$; and to compare the effectiveness of our work with existing similarity measures.

Besides MUITAS, we also implemented LCSS (VLACHOS; KOLLIOS; GUNOPULOS, 2002), EDR (CHEN; ÖZSU; ORIA, 2005), MSM (FURTADO et al., 2015) and a modified version of EDR, which we call EDRM. EDRM was implemented based on the work of (FURTADO et al., 2015), which runs EDR on each attribute and composes the total similarity

according to the attribute weights, analogous to MSM and our work.

For the first experiment, we set up the same weights for all attributes when running EDRM and MSM, i.e., each of the six attributes (POI type, price tier, rating, credit cards, space and time) has a weight of $1/6 \approx 0.17$. MUITAS, as stated, was set up according to application $\mathbb{A}$ previously defined, with weights as defined in Table 3. Table 4 portrays the results of the first experiment run.

As it is noticeable, EDR resulted in a zero score in all comparisons. EDR is strictly dependent on the order of the trajectory points and will only score similarity if all the attributes of two points match, thereby outputting the results shown in Table 4 for our running example. Similarly, LCSS assigned zero scores for all comparisons, except for Anna and Bob, in which it was able to capture some similarity. LCSS has the same limitations mentioned of EDR, but it also ignores the existence of gaps in trajectories. Thus, it assigned a 0.33 similarity value for Anna and Bob, because they both visit *Market A* around the same time. Due to the limitations and poor performance of LCSS and EDR, we focus our analysis on EDRM, MSM and MUITAS.

Table 4 – Similarity results of the first experiment.

Similarity	**LCSS**	**EDR**	**EDRM**	**MSM**	**MUITAS**
Anna x Bob	0.33	0.00	**0.58**	0.79	**0.60**
Anna x David	0.00	0.00	0.50	0.61	**0.24**
Anna x Eric	0.00	0.00	0.54	0.54	**0.23**
Anna x Ross	0.00	0.00	0.38	**0.63**	**0.63**

Anna and Bob visited two places in common with some time discrepancies. Bob, however, checked in at a restaurant that does not represent the behaviour pattern we are looking for. Also, Anna checked in at four places while he visited only three. For these reasons, a score of 0.6 is appropriate. MSM, on the other hand, computed a higher score for Anna and Bob, because it disregards the relationships between attributes. Besides the two check-ins in common, MSM also assigned check-in b_3 of Bob to check-ins a_3 and a_4 of Anna with a high score. EDRM resulted in a score close to the one computed by MUITAS, because, unlike MSM, EDRM considers the order of the points, so its computed score is basically the score computed by MSM penalized by the wrong order of points. Even though the score of 0.58 is appropriate, the reasons behind it are not the same reasons that MUITAS is based on. We shall see that in further comparisons.

It is clear from Figure 4 that Ross is the closest user to Anna regarding the semantic aspects. He behaves just like Anna throughout his entire trajectory: he visits places that are expensive and they accept credit cards; and all the places are well-rated. Note that the similarity score computed by our work for Anna and Ross is 0.63, while for Anna and Bob it is 0.6. The small difference comes from a trade-off between space and time matching, and the matching of the behaviour pattern. The score of Anna and Ross was not higher because their check-ins did not match in space and time. MSM resulted in the same similarity score for Anna and Ross. However, Ross would be disregarded in MSM approach, because Bob has a significantly higher similarity score with Anna. EDRM performed poorly in this case due to the fact that the most similar check-ins of Anna and Ross are in a different order.

The trajectory of Eric is visibly the one most similar to Anna in space and time. They checked in at places spatially close and around the same time. Eric's behaviour, however, is not like Anna's. He went to places that are not well-rated and expensive ones that do not take credit cards. As a result, the similarity score between Anna and Eric computed by MUITAS is 0.23, because they are similar only in space, time and a few POI types. MSM and EDRM, however, assigned them a high score (0.54). Again, MSM and EDRM are not able to capture the dependencies between attributes, so they considered any attributes that were similar to build the total similarity score.

David, like Eric, went to places with a bad rating and expensive ones that do not accept payments with credit cards. In addition, the location of the POIs and the time of the check-ins differ for most parts in comparison to Anna. Analogously, the similarity score of our work for Anna and David is 0.24. MSM, however, gave them a 0.61 similarity value. If we take a closer look at the check-ins of David and Anna, we can find the root of the problem. The best match for all the four restaurants visited by David, for MSM, is *Restaurant A* visited by Anna. Therefore, because it considered the same check-in several times and due to the nonobservance of relationships between attributes in the analysis, the resulting score is overestimated. The analysis of EDRM does not consider the points more than once, but it still computed a relatively high score for Anna and David, because of the dependencies between attributes that it cannot analyze.

We performed a second experiment aiming to get better results with EDRM and MSM. The weights of the attributes rating, price tier and credit cards were increased to 0.25 each, because they are the most important ones. The attributes POI type, space and time

were decreased to $1/12 \approx 0.083$ each, because they are less important. Table 5 summarizes the new results of EDRM and MSM. MUITAS results were repeated for matters of comparison.

Table 5 – Similarity results of the second experiment.

Similarity	EDRM	MSM	MUITAS
Anna x Bob	**0.67**	0.83	**0.60**
Anna x David	0.55	0.68	**0.24**
Anna x Eric	0.44	0.53	**0.23**
Anna x Ross	0.44	0.81	**0.63**

As Table 5 shows, the change of weights caused an increase in all the scores for both EDRM and MSM, except for Anna and Eric. Both measures still cannot precisely compute and capture the relationships in attributes that are essential to our problem.

MUITAS performs a significantly better and more accurate similarity analysis than existing measures. EDRM and MSM have difficulty in capturing the attributes relationships, explaining the discrepant similarity scores in all comparisons. LCSS and EDR performed poorly, because they were not designed to deal with semantic information and multiple-aspect trajectories. In summary, the experiments demonstrate the improvements by our approach especially developed for analyzing the similarity of multiple-aspect trajectories.

3.3 EVALUATION ON A REAL DATASET

We evaluated the similarity measure over a real dataset. We use a dataset of Foursquare check-ins at the city of New York between April 2012 and February 2013 (YANG et al., 2015). The dataset contains users check-ins and their corresponding venue IDs on Foursquare. We then collected venue information, such as rating, price tier, parking information, tips, etc, using the Foursquare API[2], in August 2017. Subsequently, historical weather data were collected via the Weather Wunderground API[3] and added to the check-in data. The data were stored in a PostgreSQL database and Figure 6 shows the relational model of the database. The proposed measure was implemented in Python and the experiments were conducted on a PC running Linux (Ubuntu

[2]https://developer.foursquare.com/
[3]https://www.wunderground.com/weather/api/

16.04) equipped with an Intel Core i7-3630QM CPU 2.4GHz × 8 and 6GB RAM.

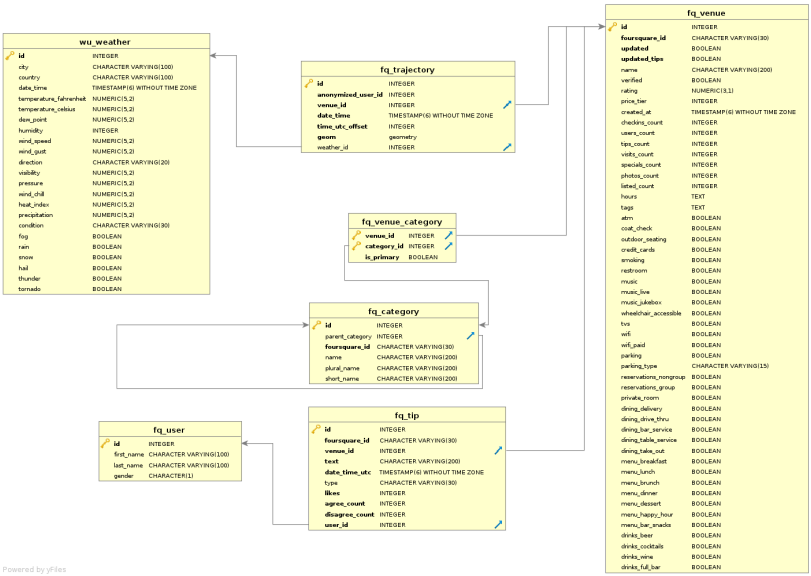

Figure 6 – Relational modeling of the Foursquare check-ins dataset.

3.3.1 Preliminary Experiment

We performed a simple preliminary experiment with only semantic data in order to facilitate the result interpretation. The rating and the price tier of the places users checked in at were the two attributes used. The price tier of places ranges from 1 to 4, with 4 being the most expensive. The rating, on the other hand, ranges from 4 to 10. We collected statistics about the places the users visited and then built a set of trajectories with 10 check-ins each, for the sake of simplicity. Afterwards, we retrieved a ranking of users, according to the number of places visited with a price tier of 1, and a rating of 8 or more, simultaneously. We then labeled the first trajectory in our ranking as the reference trajectory, i.e., we are interested on finding users with a similar behaviour as the reference user. Table 6 shows the trajectory of user 165, the reference user.

The reference user only went to places rated as 8.4 or more, and

Table 6 – Trajectory of the reference user (user ID 165).

#	Date	Time	POI Type	Rating	Price Tier
1	2012-07-03	05:21:57 PM	Cafe	8.4	1
2	2012-07-11	07:42:48 PM	Restaurant	9.2	3
3	2012-07-12	01:57:00 PM	Cafe	8.4	1
4	2012-07-18	11:22:37 AM	Cafe	9.0	1
5	2012-08-09	01:40:36 PM	Cafe	8.4	1
6	2012-08-13	12:40:37 PM	Cafe	9.0	1
7	2012-11-26	08:42:55 AM	Cafe	9.0	1
8	2012-12-10	08:43:27 AM	Cafe	9.0	1
9	2012-12-19	08:44:41 AM	Cafe	9.0	1
10	2013-02-03	11:42:05 AM	Cafe	9.0	1

all of them are cheap (price tier of 1), except for the restaurant the user visited, which has a price tier of 3. We labeled the ten trajectories right below user 165 in our ranking as relevant, because they were the ones that checked in at the highest number of cheap and well-rated places. In the same manner, ten trajectories after the first eleven in our ranking were labeled as irrelevant. We now want to compute the similarity between the trajectories labeled as relevant and the ones labeled as irrelevant against the reference trajectory, using MUITAS and existing works. We claim that a good measure would be able to assign high similarity scores between the reference trajectory and the relevant ones, and low scores to the ones that are not relevant.

Table 7 displays the distance functions and distance thresholds used for each attribute. Table 8 shows the features and the weights applied in our experiment. We have a feature of rating and price tier, because we want users that go to cheap and well-rated places, just like user 165. We set an equal weight for each of the two attributes rating and price tier, when running MSM.

Table 7 – Distance functions and thresholds for the attributes.

Attribute	Distance Function	δ
Price Tier	$dist(p,q) = \lvert p.price\_tier - q.price\_tier \rvert$	1
Rating	$dist(p,q) = \lvert p.rating - q.rating \rvert$	1

Table 9 presents the results for LCSS, EDR, MSM and MUITAS over the small set of trajectories. The column "Rlv." is the relevance

of that comparison, according to what we had previously labeled. The IDs of the trajectories being compared are displayed in the first column. The results are ordered in descending order of the value computed by MUITAS.

Table 8 – Attributes and weights of features.

Feature	**Attributes**	w
f_1	*Rating*, *Price Tier*	1.0

Table 9 – Similarity results of the experiment on the Foursquare dataset.

Similarity	**Rlv.**[4]	**LCSS**	**EDR**	**MSM**	**MUITAS**
165 x 886	R	1.00	1.00	1.00	1.00
165 x 707	R	0.90	0.90	1.00	1.00
165 x 980	R	0.90	0.90	1.00	1.00
165 x 1021	R	0.90	0.90	1.00	1.00
165 x 151	R	0.90	0.80	1.00	1.00
165 x 264	R	0.90	0.80	0.98	0.90
165 x 818	R	0.90	0.80	0.98	0.90
165 x 294	R	0.90	0.80	0.95	0.90
165 x 315	R	0.90	0.80	0.95	0.90
165 x 953	R	0.70	0.60	1.00	0.80
165 x 54	I	0.80	0.80	0.95	0.80
165 x 66	I	0.70	0.70	0.93	0.70
165 x 4	I	0.70	0.60	0.93	0.70
165 x 10	I	0.60	0.60	0.93	0.70
165 x 51	I	0.50	0.50	0.90	0.60
165 x 55	**I**	**0.30**	**0.20**	**0.95**	**0.30**
165 x 23	I	0.30	0.20	0.83	0.30
165 x 21	I	0.30	0.20	0.83	0.30
165 x 17	I	0.20	0.20	0.80	0.20
165 x 65	I	0.00	0.00	0.50	0.00

LCSS, EDR and MUITAS had a similar behaviour, being able to correctly classify most of the trajectories. The comparisons with trajectories of IDs 953 and 54 were confused by LCSS and EDR, i.e.,

[4]Relevance of the similarity. R means relevant and I stands for irrelevant.

despite the fact that trajectory 953 is relevant and 54 is not, 165 and 54 were computed by LCSS and EDR as more similar than 165 and 953. MUITAS, however, computed the same score of 0.80 for both trajectories.

MSM was the only measure that was not able to accurately classify most of the irrelevant trajectories. For instance, if we take a look at the similarity of trajectory 165 and 55, highlighted in Table 9, we notice that MSM resulted in a 0.95 score, while all the other measures computed a 0.20 or 0.30 score. We further investigate the similarity of users 165 and 55. Table 10 shows the trajectory of user 55.

Most of the places visited by user 55 have a rating above or equal to 8. However, most of the places are also expensive (price tier of 3 and 4). The reasons why MSM resulted in such a high similarity score for users 165 and 55 is that (i) it matched 6 of the 7 expensive places that user 55 checked in at with the only expensive restaurant visited by user 165; and (ii) it is not able to consider the dependency between the rating and the price tier of the places.

Table 10 – Trajectory of the irrelevant user 55.

#	Date	Time	POI Type	Rating	Price Tier
1	2012-05-21	06:47:21 PM	Restaurant	8.6	3
2	2012-07-04	03:22:10 PM	Restaurant	8.2	3
3	2012-07-08	01:12:03 AM	Restaurant	8.4	4
4	2012-07-08	09:13:04 PM	Restaurant	9.3	3
5	2012-07-09	10:56:13 AM	Cafe	9.4	2
6	2012-10-11	06:15:03 PM	Restaurant	7.0	1
7	2012-10-12	10:22:52 PM	Nightclub	8.0	3
8	2012-12-06	08:44:14 PM	Restaurant	9.4	3
9	2012-12-11	01:59:31 PM	Restaurant	8.8	3
10	2012-12-22	08:13:06 AM	Cafe	7.4	1

The results of this experiment once again showed that MSM cannot handle strong relationships between attributes. In addition, the fact that it may match the same points of a trajectory multiple times may lead to inaccurate results.

3.3.2 Clustering Analysis

Clustering analysis is one of several applications for similarity measures in data mining. A clustering algorithm allows us to group user trajectories according to similar profiles, so a recommendation system, for instance, more accurately targets its audience.

We run a second experiment on the same dataset of Foursquare check-ins, using a random subset of 150 user trajectories of length ranging from 10 to 356 check-ins. Our objective was to group users according to their habits, i.e., the types of places they visit and spend money on, how much money they spend, and when during the day they visit these different types of places. We run a clustering algorithm for MSM and MUITAS, in order to further analyze the results achieved by each measure. EDR and LCSS were not analyzed because they perform poorly when more semantic data is considered, as shown in the evaluation of the running example. The attributes considered in the analysis were *Space*, *Time*, the *POI Type*, its *Rating* and *Price tier*. Table 11 describes the format of the attributes and Table 12 presents the distance functions and distance thresholds employed for each attribute.

Table 11 – Attributes format.

Attribute	**Format**
Space	Latitude and longitude
Time	Hours, minutes and seconds, from 00:00:00AM to 11:59:59PM
POI Type	String, from a set of 10 root categories[5]
Rating	Decimal number, from 4.0 to 10.0
Price Tier	Integer number, from 1 to 4

For MUITAS we define features as presented in Table 13, in order to capture the behaviour previously described - to group users according to the types of places they visit and spend money, how much money they spend, and when during the day they visit these different types of places. The *POI Type* is considered together with each of the attributes *Rating*, *Price Tier* and *Time*, because we only would like

[5]https://developer.foursquare.com/docs/resources/categories

[6]Function $haversine(s_1, s_2)$ returns the *haversine* distance between s_1 and s_2 in meters.

[7]Function $interval(t_1, t_2)$ returns the number of seconds between time t_1 and t_2.

Table 12 – Distance functions and thresholds for the attributes.

Attribute	**Distance Function**	δ
Space	$dist(p,q) = haversine(p.space, q.space)$[6]	1000
Time	$dist(p,q) = interval(p.time, q.time)$[7]	5400
POI Type	$dist(p,q) = \begin{cases} 0, & \text{if } p.poi\_type = q.poi\_type \\ \infty, & \text{otherwise} \end{cases}$	0
Rating	$dist(p,q) = \lvert p.rating - q.rating \rvert$	1
Price Tier	$dist(p,q) = \lvert p.price\_tier - q.price\_tier \rvert$	1

to consider check-ins of the same type when analyzing these attributes, as they describe habits of the user. The space is kept alone because two check-ins close in space will probably share commonalities, regardless of the type of place they are. Certain areas within a city, for example, have their own characteristics regarding public transportation, cleanliness, safety and general quality of life.

We define an application $\mathbb{A}$, with the features shown in Table 13 and the distance functions and thresholds in Table 12. For MSM the five attributes are considered independently, with the same weight of 0.2. For MUITAS all features were computed with an equal weight of 0.25.

Table 13 – Attributes and weights of features.

Feature	**Attributes**	w
f_1	*POI Type*, *Rating*	0.25
f_2	*POI Type*, *Price Tier*	0.25
f_3	*POI Type*, *Time*	0.25
f_4	*Space*	0.25

We computed a matrix of similarity scores between every pair of trajectories for each similarity measure, and subsequently applied hierarchical clustering on the matrix of similarities using the agglomerative strategy. The linkage criteria employed was the *complete-linkage clustering*. Complete-linkage clustering is more appropriate for our analysis than other methods, because it allows us to achieve more consistent clusters, i.e., the maximum distance between a pair of elements in a cluster is the minimum possible, as opposed to criteria such as single-linkage and average-linkage clustering.

Figures 7 and Figure 8 show the dendrograms of the computed clusters. The horizontal axis shows the trajectory identifiers and the

vertical axis represents the distance between trajectories. The vertical lines of the dendrogram tree represent clusters, while the horizontal lines of the tree represent merged clusters. For purposes of analysis and visualization, only the last 8 levels of the dendrogram tree are shown. The cut-off point is at 0.4, i.e., only clusters with a maximum intracluster distance of 0.4 were accepted (minimum intracluster similarity of 0.6).

Through a first analysis of the dendrograms, a few things come to our attention. MSM resulted in 6 dense clusters, while 32 small clusters were formed with MUITAS. Besides that, when using MSM the maximum distance computed between two trajectories was close to 0.6 (last merge in the dendrogram tree), while for MUITAS it was close to 1. These statistics are a result of the extremely high flexibility of MSM, which, as we stated earlier, may lead to inaccurate or overestimated results. MSM considers any similarity between two points, regardless of any relationships between attributes such as the ones we are interested on, and such behaviour is reflected on the high computed scores.

Let us take a closer look on the trajectories of users 48 and 250, which are both in the yellow cluster computed by MSM (see Figure 8). The similarity computed by MSM for users 48 and 250 was 0.78, while MUITAS assigned them a similarity of 0.31. Table 14 summarizes the trajectories of users 48 and 250. Note that the trajectory of user 48 is almost *four times longer* than user 250 trajectory. Both users only visited food and nightlife places, such as restaurants and pubs. The majority of check-ins of user 48 are at nightlife places, while user 250 visited more food places than nightlife ones. These numbers tell us that, even though user 48 and user 250 visit the same types of places, they do not visit these places with the same frequency. Such aspect is not captured by MSM, but it is considered by MUITAS when computing similarity.

Table 14 – Summary of users 48 and 250 trajectories.

	User 48	**%**	**User 250**	**%**
Food Check-ins	20	31	10	59
Nightlife Check-ins	45	69	7	41
Total of Check-ins	**65**	**100**	**17**	**100**

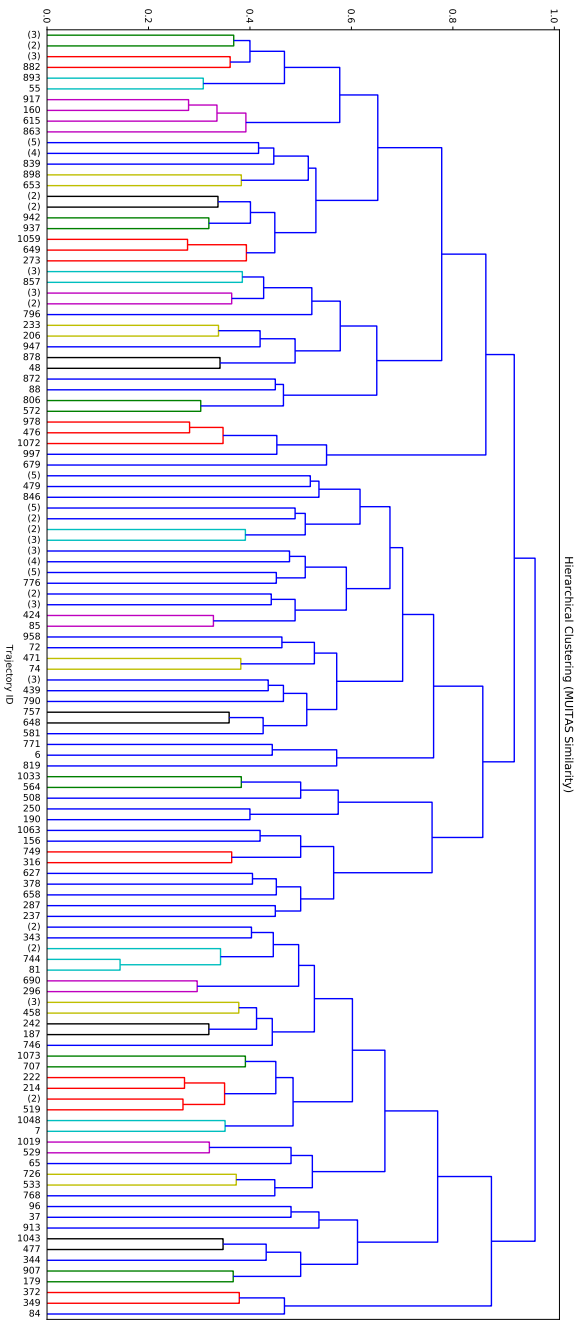

Figure 7 – Hierarchical clustering with MUITAS similarity.

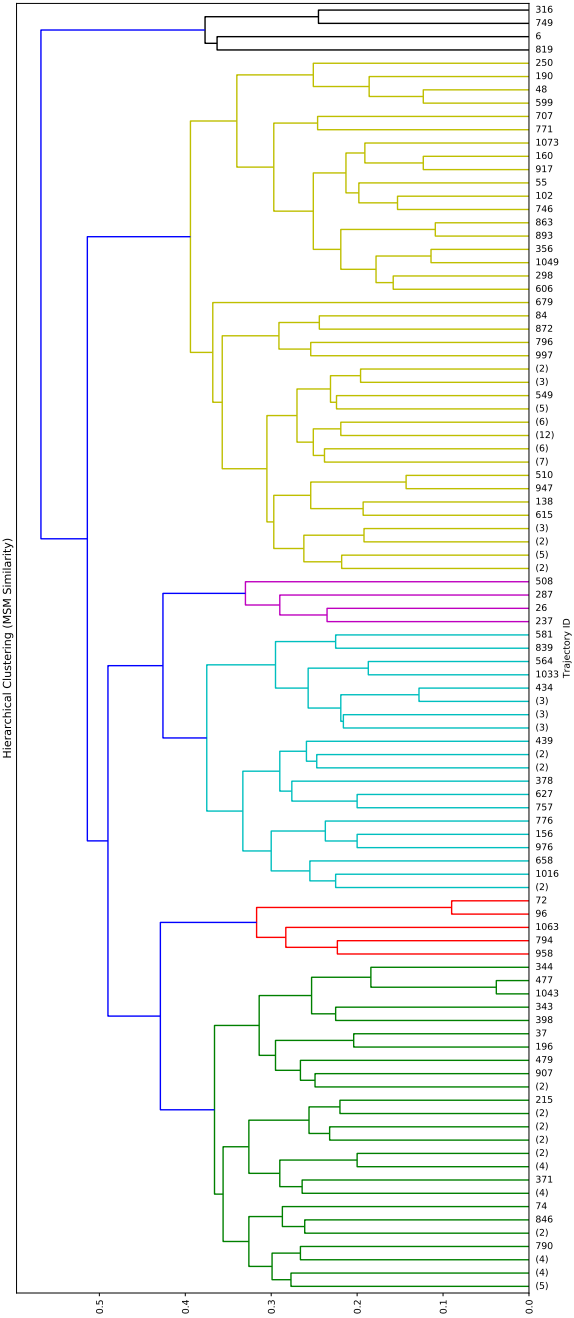

Figure 8 – Hierarchical clustering with MSM similarity.

Figure 9 shows the locations of the places where users 48 and 250 checked in. It is clear that there is not match for most of the check-ins. User 250 mostly visited places in the Brooklyn area (bottom of the map in Figure 9), while user 48 stayed for most parts in Manhattan and Queens (center and top areas in Figure 9).

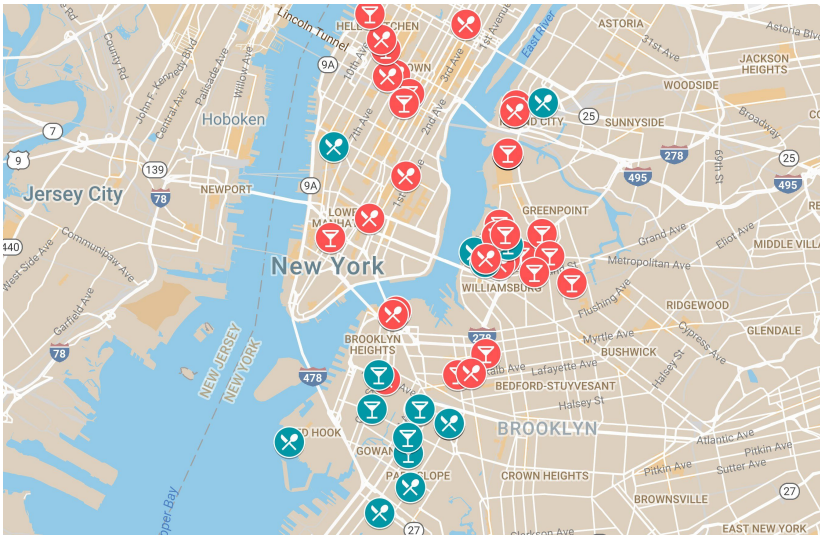

Figure 9 – Check-ins of user 48 (red) and 250 (blue).

Figure 10 shows the distributions of check-ins for the *Price Tier* and *Rating* attributes for users 48 and 250. For each attribute, the data is presented according to the *POI Type* (food or nightlife). The price tier of nightlife places are similar between both users. For food places, however, they are not as similar, but the differences are within the threshold defined for *Price Tier*. The *Rating*, on the other hand, differs more for both users. While user 250 goes, for the most part, to places rated approximately as 8.25 or more, user 48 goes to several places with a lower rating than 8.25, i.e., there should not be a match for many of those places. Additionally, the fact that user 48 goes to nightlife spots more often than user 250 should be significant for their similarity.

In conclusion, the similarity computed by MSM for user 48 and 250 is overestimated, because (i) their trajectories differ by almost four times in length; (ii) user 48 visits way more nightlife places than food ones, while user 250 visits more food places than nightlife ones; (iii)

most of their check-ins do not match for attribute *Space*; and (iv) there are some discrepancies for the *Price Tier* and *Rating* attributes for the same *POI Types*.

Even though MSM groups some of the users also clustered together by MUITAS, MSM includes many other trajectories that are not relevant or similar to each other in the clusters, such as users 48 and 250 previously analyzed. The results presented once more demonstrated that MUITAS is more adequate for analyzing multiple-aspect trajectories than existing similarity measures. Due to time constraints, we were not able to further develop experiments on this dataset.

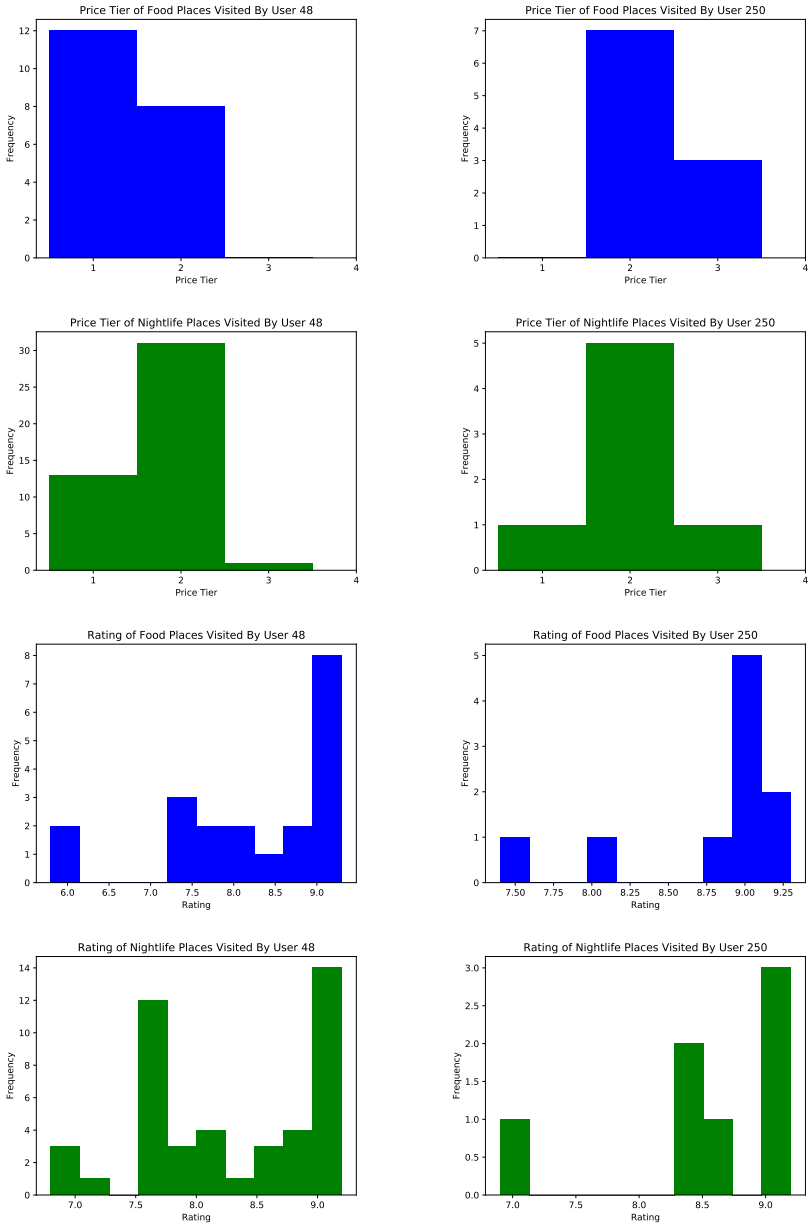

Figure 10 – *Price Tier* and *Rating* of places visited by users 48 (left) and 250 (right).

4 CONCLUSION

The analysis of multiple-aspect trajectories has become very attractive and necessary because of the high volume of geotagged information available nowadays. In this work, we proposed a formal definition of multiple-aspect trajectories, as well as a new similarity measure for computing the similarity of these trajectories. Our similarity measure overcame limitations of previous works, such as the ability to define dependency relationships between attributes and the observance of repetitive behaviour when considering the points of two trajectories in the similarity analysis.

In order to evaluate the relevance and effectiveness of our work, we performed and presented an experimental evaluation based on a toy example, using multiple aspects that are present on real world data. We then compared previous works with MUITAS, the proposed measure, demonstrating the improvements made by our work. We also described a simple experiment on a real dataset of Foursquare check-ins, but due to time constraints, we were not able to further develop the experiment.

Even though we focused on multiple-aspect trajectories, the proposed similarity measure can be applied to any sequenced data including any number of attributes. Potential future work include, but is not limited to:

- Analyzing the similarity of multiple-aspect trajectories regarding their global attributes (e.g., distance traveled, average speed);
- Analyzing the similarity of heterogeneous points of trajectories, i.e., points that have different aspects and different attributes throughout the trajectory;
- Proposing a clustering technique based on the similarity measure for multiple-aspect trajectories;
- Algorithmically finding the best relationships between attributes, in order to automatically set up the features of an application.

REFERENCES

AGRAWAL, R.; FALOUTSOS, C.; SWAMI, A. N. Efficient similarity search in sequence databases. In: *FODO '93 Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*. London, UK: Springer-Verlag, 1993. p. 69–84.

ALVARES, L. O. et al. A model for enriching trajectories with semantic geographical information. In: ACM. *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*. [S.l.], 2007. p. 22.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: *AAAIWS'94 Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. Seattle, WA: AAAI Press, 1994. p. 359–370.

BOGORNY, V. et al. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, Wiley Online Library, v. 18, n. 1, p. 66–88, 2014.

BOLLOBáS, B. et al. Time-series similarity problems and well-separated geometric sets. In: *SCG '97 Proceedings of the thirteenth annual symposium on Computational geometry*. New York, NY: ACM, 1997. p. 454–456.

BOURGEOIS, F.; LASSALLE, J.-C. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Commun. ACM*, ACM, New York, NY, USA, v. 14, n. 12, p. 802–804, dez. 1971. ISSN 0001-0782.

CHEN, L.; ÖZSU, M. T.; ORIA, V. Robust and fast similarity search for moving object trajectories. In: *SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. New York, NY: ACM, 2005. p. 491–502.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. Fast subsequence matching in time-series databases. In: *SIGMOD '94 Proceedings of the 1994 ACM SIGMOD international conference on Management of data*. New York, NY: ACM, 1994. p. 419–429.

FERRERO, C. A.; ALVARES, L. O.; BOGORNY, V. Multiple aspect trajectory data analysis: Research challenges and opportunities. In:

Proceedings XVII GEOINFO. Campos do Jordão, SP: [s.n.], 2016. p. 56–67.

FURTADO, A. S. et al. Unveiling movement uncertainty for robust trajectory similarity analysis. *International Journal of Geographical Information Science*, Taylor & Francis, p. 1–29, 2017.

FURTADO, A. S. et al. Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, v. 20, p. 280–298, 2015.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. [S.l.]: MIT press, 2001.

HOLT, G. A. ten; REINDERS, M. J. T.; HENDRIKS, E. A. Multi-dimensional dynamic time warping for gesture recognition. In: *Thirteenth annual conference of the Advanced School for Computing and Imaging*. Heijen: [s.n.], 2007.

KRUEGER, R.; THOM, D.; ERTL, T. Semantic enrichment of movement behavior with foursquare–a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics*, v. 21, p. 903 – 915, 2014.

KUHN, H. W. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, Wiley Subscription Services, Inc., A Wiley Company, v. 2, n. 1-2, p. 83–97, 1955. ISSN 1931-9193.

LIU, H.; SCHNEIDER, M. Similarity measurement of moving object trajectories. In: *IWGS '12 Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming*. New York, NY: ACM, 2012. p. 19–22.

MUNKRES, J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, Society for Industrial and Applied Mathematics, v. 5, n. 1, p. 32–38, 1957. ISSN 03684245.

NOËL, D. et al. Modeling semantic trajectories including multiple viewpoints and explanatory factors: application to life trajectories. In: ACM. *Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*. [S.l.], 2015. p. 107–113.

SHARIF, M.; ALESHEIKH, A. A. Context-awareness in similarity measures and pattern discoveries of trajectories: a context-based

dynamic time warping method. *GIScience & Remote Sensing*, v. 54, n. 3, p. 426–452, 2017.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. *Data & knowledge engineering*, Elsevier, v. 65, n. 1, p. 126–146, 2008.

VLACHOS, M.; KOLLIOS, G.; GUNOPULOS, D. Discovering similar multidimensional trajectories. In: *ICDE '02 Proceedings of the 18th International Conference on Data Engineering*. Washington, DC: IEEE, 2002. p. 673–684.

YANG, D. et al. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, IEEE, v. 45, n. 1, p. 129–142, 2015. ISSN 2168-2216.