UNIVERSIDADE FEDERAL DE SANTA CATARINA

BIBLIOTECA UNIVERSITÁRIA

André Luiz Lehmann

SMSM: A SIMILARITY MEASURE FOR TRAJECTORY STOPS AND MOVES

Florianópolis

2019

André Luiz Lehmann

SMSM: A SIMILARITY MEASURE FOR TRAJECTORY STOPS AND MOVES

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Ciência da Computação.

Orientadora: Prof$^{\underline{a}}$. Vania Bogorny, Dra.

Florianópolis

2019

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

André Luiz Lehmann
SMSM: a Similarity Measure for Trajectory Stops and Moves / André Luiz Lehmann; Orientadora, Profª. Vania Bogorny, Dra.,2019.
89 p.

Dissertação - Universidade Federal de Santa Catarina, Departamento de Informática e Estatística, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação.

Inclui referências

1. Trajectory similarity measures. 2. Semantic trajectory similarity. 3. Similarity measure framework I. Profª. Vania Bogorny, Dra. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação III. SMSM: a Similarity Measure for Trajectory Stops and Moves

André Luiz Lehmann

SMSM: A SIMILARITY MEASURE FOR TRAJECTORY STOPS AND MOVES

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre em Ciência da Computação, e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática e Estatística, Centro Tecnológico da Universidade Federal de Santa Catarina.

Florianópolis, 22 de Maio de 2019.

Prof. José Luís Almada Güntzel, Dr.
Coordenador do Programa de Pós-Graduação em
Ciência da Computação

Vania Bogorny

Prof^a. Vania Bogorny, Dra.
Orientadora
Universidade Federal de Santa Catarina – UFSC

Banca Examinadora:

P/

Prof^a. Ticiana Linhares Coelho da Silva, Dra.
Universidade Federal do Ceará - Campus Quixadá –
UFC (videoconferência)

JOSÉ LUÍS ALMADA GÜNTZEL
Coordenador do PPGCC/UFSC
Portaria nº 2092/2017/GR, de 27/10/2017

Prof. Renato Fileto, Dr.
Universidade Federal de Santa Catarina — UFSC

Prof. Ronaldo dos Santos Mello, Dr.
Universidade Federal de Santa Catarina — UFSC

This work is dedicated to adult children who,
When small, dreamed of becoming scientists.

Acknowledgements

I would like to thank my wife for all love, patience and motivation. Carolina, I love you. I would also like to thank my advisor, Profa. Dra. Vania Bogorny for all patience and dedication for encouraging me to learn both the technical details and the academic world. To Prof. Dr. Luis Otavio Alvares for all great conversations and discussions. Also, I must thank my lab colleagues, Lucas May Petry and Camila Silva who somehow helped me in the development of this work. Special thanks also to André Salvaro Furtado, who provided the source code of several similarity measures, and for Carlos Andres Ferrero, an always present colleague on the lab, helping me and teaching me about the fundamentals of data mining. This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

"The true sign of intelligence is not knowledge but imagination."

Albert Einstein

Resumo

Medidas de similaridade são a base para a maioria dos métodos de mineração de dados e extração de conhecimento. Na área de trajetórias de objetos móveis, por muitos anos a pesquisa em similaridade de trajetórias focou nas trajetórias brutas, considerando somente a informação de espaço e tempo. Com o enriquecimento das trajetórias com informações semânticas, como o nome e a categoria dos locais visitados, meio de transporte utilizado durante o movimento, o nome das ruas percorridas, etc, emergiu a necessidade por medidas de similaridade que suportem espaço, tempo e semântica. Apesar de algumas medidas de similaridade para trajetórias lidarem com todas estas dimensões, elas consideram somente os locais onde o objeto móvel faz paradas, denominados *stops*, ignorando o movimento que ocorre entre as paradas, denominado *move*. Acredita-se que, para algumas aplicações, o movimento entre os *stops* é tão importante quanto o *stop* em si, e ele deve ser levado em consideração na análise da similaridade, como em sistemas de transporte público, turismo, planejamento urbano, entre outros. Nesta dissertação é proposta a medida Similarity Measure for trajectory Stops and Moves (SMSM), um nova medida de similaridade para trajetórias semânticas que considera tanto os *stops* quanto os *moves*. O SMSM é avaliado em três conjuntos de dados: (i) um conjunto de dados de trajetórias sintéticas criadas com o gerador de trajetórias semânticas Hermoupolis, (ii) um conjunto de trajetórias reais de táxis do projeto CRAWDAD, e (iii) o conjunto de dados de trajetórias reais chamado Geolife, com trajetórias de pessoas na cidade de Pequim. Os resultados mostram que o SMSM supera as medidas de similaridade do estado da arte desenvolvidas tanto para trajetórias brutas quanto semânticas.

Palavras-chaves: Medidas de similaridade para trajetórias. Simila-

ridade de trajetórias semânticas. Framework de medição de similaridade.

Resumo Expandido

INTRODUÇÃO

A análise de similaridade é um tema importante na área de mineração de dados nas mais diversas áreas de aplicação, e não é diferente na área de análise de dados de trajetórias de objetos móveis. Na área de trajetórias, o cálculo da similaridade é importante para responder perguntas como "Dado um conjunto M de trajetórias, quais são as mais parecidas com uma trajetória s?" ou "Quais são os pares de trajetórias mais semelhantes em um conjunto M de trajetórias?", ou ainda, dado um conjunto de trajetórias quais são os diferentes grupos de trajetórias que possuiem maior semelhança entre si? Para tanto, é importante a criação de medidas de similaridade para trajetórias. Por muitos anos a pesquisa em similaridade de trajetórias focou nas trajetórias brutas, que são sequências de pontos com informações de localização e tempo. Estas medidas de similaridade somente consideravam a informação espaço-temporal, limitando a comparação das trajetórias às suas características geo-espaciais. Com o advento das redes sociais e o enriquecimento das trajetórias com informações semânticas, como o nome e a categoria dos locais visitados, meio de transporte utilizado no deslocamento, o nome das ruas percorridas, etc, emergiu a necessidade por medidas de similaridade que suportem as *trajetórias semânticas*, onde cada ponto possui informações de espaço, tempo e semântica. Apesar de algumas medidas de similaridade para trajetórias lidarem com este novo tipo de trajetória, elas consideram somente os locais visitados pelo objeto móvel, denominados *stops*, ignorando aquilo que ocorre entre os locais, denominado *move*. Acredita-se que para algumas aplicações, o movimento entre os *stops* é tão importante quanto o *stop* em si, e ele deve ser levado em consideração na análise de similaridade.

Por exemplo em sistemas de gerenciamento de tráfego, sistemas de transporte público, planejamento urbano, entre outros. Nesta dissertação é proposta a medida SMSM (Similarity Measure for trajectory Stops and Moves), uma nova medida de similaridade para trajetórias semânticas que considera tanto os *stops* quanto os *moves* nas trajetórias semânticas. O SMSM é avaliado em três conjuntos de dados: (i) um conjunto de dados de trajetórias sintéticas criadas com o gerador de trajetórias semânticas Hermoupolis, (ii) um conjunto de trajetórias reais de táxis do projeto CRAWDAD, e (iii) o conjunto de dados de trajetórias reais chamado Geolife, com trajetórias de pessoas na cidade de Pequim. Os resultados mostram que o SMSM supera as medidas de similaridade do estado da arte desenvolvidas tanto para trajetórias brutas quanto semânticas.

OBJETIVOS

O objetivo geral deste trabalho é a proposição de uma nova medida de similaridade para trajetórias semânticas. Mais especificamente esta dissertação visa propor uma nova medida de similaridade para trajetórias semânticas que trate tanto os *stops* quanto os *moves* das trajetórias. A nova medida provê suporte a múltiplas dimensões como espaço, tempo, semântica e quaisquer outras dimensões adicionais, atribuindo diferentes pesos e permitindo o uso de diferentes funções de distância para cada dimensão, além de considerar parcialmente a ordem dos pontos na trajetória semântica.

METODOLOGIA

Inicialmente foi realizada uma revisão da literatura em tópicos relacionados à similaridade de trajetórias semânticas, através de ferramentas de pesquisa como Google Scholar e em periódicos e conferências de alto impacto (como TKDE, IJGIS, TGIS, VLDB, DKE, ACM-SIGSpatial, entre outros). Através da análise e implementação das medidas existentes foram identificadas algumas das suas limitações e com isto foi possível propor uma nova medida de similaridade que seja mais robusta para trajetórias semânticas e que acima de tudo seja capaz de tratar todas as partes das trajetórias e suas dimensões.

A medida proposta denominada SMSM permite definir graus de importância às diferentes partes da trajetória como: (i) o grau de importância entre os *stops* e os *moves*, e (ii) o grau de importância de cada atributo que compõe os *stops* e os *moves* e também os limiares (*thresholds*) utilizados em cada atributo para definir se houve casamento (*matching*) ou não entre os pontos.
A medida foi avaliada em conjuntos de trajetórias reais, já utilizadas na literatura como (PIORKOWSKI; SARAFIJANOVIC-DJUKIC; GROSSGLAUSER, 2009) e também (ZHENG et al., 2009), assim como um conjunto de dados sintéticos, gerados com a ferramenta Hermoupolis (PELEKIS et al., 2013). Inicialmente as bases de dados de trajetórias brutas foram enriquecidas com informações sobre os *stops* e os *moves* que ocorreram durante cada trajetória. Com as trajetórias semanticamente enriquecidas, foi possível avaliar e comparar a medida proposta. Para isto foi utilizada a abordagem de precisão em diferentes níveis de cobertura (BAEZA-YATES; RIBEIRO-NETO, 2011) em tarefas de recuperação da informação. Também foram avaliados o impacto dos parâmetros de grau de importância e limiares para a medida proposta e o tempo de execução da tarefa de recuperação de informação.

RESULTADOS E DISCUSSÃO

Os resultados obtidos evidenciam que a medida SMSM mostrou-se a mais robusta para avaliar a similaridade de trajetórias semânticas onde tanto as informações sobre os *stops* quanto os *moves* são relevantes. A medida também mostrou-se flexível para suportar múltiplas dimensões de dados tanto nos *stops* quanto nos *moves* e flexível também ao permitir a definição de diferentes limiares (*thresholds*) para cada dimensão e graus de importância.

CONSIDERAÇÕES FINAIS

A principal contribuição desta dissertação é uma medida de similaridade para trajetórias semânticas que considera tanto os *stops* quantos os *moves*, suportando dimensões espaciais, temporais e semânticas, permitindo o uso de diferentes funções de distância para

cada dimensão. A medida de similaridade é flexível o suficiente para considerar parcialmente a ordem dos *stops*, e suporta diferentes pesos para os *stops*, os *moves* e as dimensões constituintes de cada elemento, possibilitando atribuir maior ou menor importância para cada elemento. A medida proposta nesta dissertação foi publicada no periódico International Journal of Geographical Information Science.

Palavras-chaves: Medidas de similaridade para trajetórias. Similaridade de trajetórias semânticas. Framework de medição de similaridade.

Abstract

For many years trajectory similarity research has focused on raw trajectories, considering only space and time information. With the trajectory semantic enrichment, using information as the name and type of the visited places, the transportation mean, the name of the streets, etc, emerged the need for similarity measures that support space, time, and semantics. Although some trajectory similarity measures deal with all these dimensions, they consider only the places where the moving object stays for a certain time, called *stop*, ignoring the movement between stops. We claim that, for some applications, as traffic management systems, urban planning, public transportation, etc, the movement between stops is as important as the stops, and it must be considered in the similarity analysis. In this thesis we propose the similarity measure called Similarity Measure for trajectory Stops and Moves(SMSM), a novel similarity measure for semantic trajectories that considers both stops and moves. We evaluate SMSM with three trajectory datasets: (i) a synthetic trajectory dataset generated with the Hermoupolis semantic trajectory generator, (ii) a real trajectory dataset of taxis from the CRAWDAD project, and (iii) the Geolife trajectory dataset, with raw trajectories of persons around Beijing. The results show that SMSM overcomes state-of-the-art measures developed for both raw and semantic trajectories.

Keywords: Trajectory similarity measures. Semantic trajectory similarity. Similarity measure framework

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Trajectory similarity measuring has received significant attention in the last few years, and several measures have been proposed to deal either with raw trajectories or semantic trajectories. A *raw trajectory* is generally represented as a sequence of points $T =< p_1, p_2, ..., p_n >$, with $p_i = (x_i, y_i, t_i)$ where x, y is the position of the object in space at time instant t. Figure 1.1 presents an example of a raw trajectory T, where the first point of the trajectory is located at the coordinates $(2, 3)$ at time instant 1.

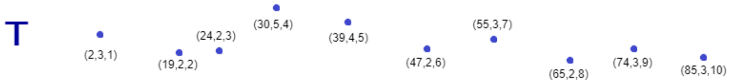

Figure 1.1 – Example of a raw trajectory T

Examples of similarity measures for raw trajectories are LCSS (Longest Common SubSequence) (VLACHOS; KOLLIOS; GUNOPULOS, 2002), EDR (Edit Distance on Real sequences) (CHEN; ÖZSU;

ORIA, 2005), NWED (Normalized Weighted Edit Distance) (DODGE; LAUBE; WEIBEL, 2012) and UMS (Uncertain Movement Similarity) (FURTADO; ALVARES, et al., 2018). LCSS and EDR consider the sequence, but they force a match in all dimensions, not allowing partial similarity between trajectory points. UMS is a parameter free method that considers only the spatial dimension, and it was developed to deal with data of varied or low sampling rate.

Existing works for raw trajectory similarity are limited to the spatio-temporal properties of raw trajectories, basically considering trajectories as data with only space information or space and time.

Similarity measures are the basis of several data processing and analysis techniques, such as information retrieval, location prediction, nearest neighbour queries, outlier detection, clustering, etc. A clustering algorithm, for instance, uses a similarity measure for grouping objects with similar trajectories. Outlier detection methods use similarity to find groups of trajectories with normal behavior, and the objects that are dissimilar to the majority, are the outliers. To detect specific trajectory patterns such as flocks (LAUBE; KREVELD; IMFELD, 2005), for instance, a raw trajectory similarity measure could be applied to find a minimal number of objects moving together in space and time.

In 2008 emerged the concept of semantic trajectories, introduced by (SPACCAPIETRA et al., 2008), where trajectories are represented as *sequences of stops and moves*. *Stops* are the most important parts of trajectories, representing the places that an object has visited for a minimal amount of time, and the *moves* are the trajectory points between stops. In several works, stops are called points of interest (POIs), episodes, or stay points. Semantic trajectories are more complex than raw trajectories, because they have at least three dimensions: space, time, and semantics. We consider in this thesis that semantics is any type of information associated to mobility data other than spatial location and time.

The enrichment of trajectories with semantic information is a well studied topic in the literature, and a number of methods have been developed for this purpose. Some of these methods are summarized in (BOGORNY; BRAZ, 2012). The first work for detecting stops was (ALVARES et al., 2007), followed by (PALMA et al., 2008), and (ROCHA et al., 2010). In (FILETO et al., 2013) an architecture is proposed to enrich trajectories with linked open data. Applications as DayTag (RINZIVILLO et al., 2013) can detect stops and moves and the user can annotate the semantics to his/her stops and moves. Examples of semantic information related to the stops can be, for instance, the name of the stop (e.g. Ibis Hotel) and the category (e.g. Hotel, Museum, Restaurant), while the semantic information related to the move could be, for instance, the name of the streets followed by the moving object, and the category of the transportation mode. An example of semantic trajectory is shown in Figure 1.2, which has three *stops* (Hotel, Museum, and Restaurant) and two *moves* (points between stops).

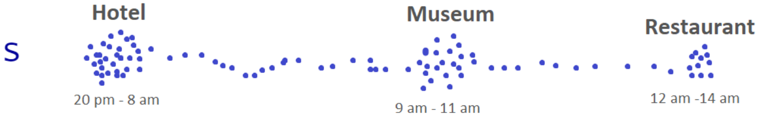

Figure 1.2 – Example of a semantic trajectory S

With the explosion of social media data, internet channels, and the facility to enrich trajectories with more context information as linked open data (FILETO et al., 2013), it is possible to represent the movement in a more meaningful way. From social media data, for instance, a stop at a hotel can be enriched with the information of the number of stars, the price average, evaluation rate, facilities, parking, wifi, etc. In this thesis we assume that semantic trajectories are represented as sequences of stops and moves, as originally defined in (SPACCAPIETRA et al., 2008), and the way how these trajectories are generated or enriched is out of the scope of this thesis.

Similarity measures that consider both stops and moves can be important in a vast number of applications such as public transportation systems, traffic management, fraud detection, tourism, urban planning, car sharing, etc. For instance, a similarity measure that considers both stops and moves can be used to cluster the trajectories of buses of the same line considering the traveled distance between two consecutive stops. The trajectories not included in a cluster are outliers, and may characterize a deviation from the scheduled itinerary. For the same application, the similarity measure could be used to check if the buses of the same line follow the sequence of stops and roads of the pre-defined itinerary. Another application is for car sharing, where the similarity of stops and moves can be used to find groups of people that follow the same sequence of stops and moves at similar times.

Only a few similarity measures were proposed for semantic trajectories, as (KANG; KIM; LI, 2009), (LIU; SCHNEIDER, 2012), (YING et al., 2010), and (FURTADO; KOPANAKI, et al., 2016). The main problem of these measures is that they do not address all three dimensions (space, time, and semantics), as the works of (KANG; KIM; LI, 2009) and (LIU; SCHNEIDER, 2012); or they exclusively address the stops, systematically ignoring all information about the moves, as the works of (YING et al., 2010) and (FURTADO; KOPANAKI, et al., 2016). To the best of our knowledge, none of the existing similarity measures for semantic trajectories have considered both stops and moves. The measure MSM (Multidimensional Similarity Measure) (FURTADO; KOPANAKI, et al., 2016), for instance, considers only the stops, and they are treated as elements that are independent from each other, without considering the order/sequence as they appear in the trajectories. As MSM ignores the moves between stops, it can only be used to answer questions like: *how similar are two trajectories P and Q considering their stops?*

To better understand the need for considering both stops and moves in trajectory similarity analysis, let us consider the example in Figure 1.3, for a tourism application, which shows three trajectories of tourists visiting Paris. These tourists visited four places, in this order: Arc de Triomphe (first stop - S1), Place de la Concorde (second stop - S2), the Louvre Museum (third stop - S3), and the Notre Dame Cathedral (the last stop - S4). The three tourists visited the same places at the same order, but the tourists of trajectories $T2$ (green trajectory) and $T3$ (red trajectory) moved on foot, following the shortest path, while the tourist of trajectory $T1$ used a city tour hop on and hop off bus to appreciate the view. The question we want to answer in this thesis is *how similar are trajectories $T1$, $T2$ and $T3$ considering both stops and moves?* From the figure it is clear that trajectories $T2$ and $T3$ are more similar, because they used almost the same paths between stops and moved on foot, while trajectory $T1$ has a spatially different move, performed with a different transportation mode. Now suppose that a tourism manager wants to recommend a trip to a new tourist arriving in Paris, and this new tourist wants to visit the same places visited by the tourists in the figure, but he wants to move on foot and follow the path used by the majority of the tourists. For this case, we need to retrieve trajectories $T2$ and $T3$. Another example is the evaluation of the flow of tourists moving on foot between these four stops in order to eventually propose a new and direct hop on hop off tourist bus line.

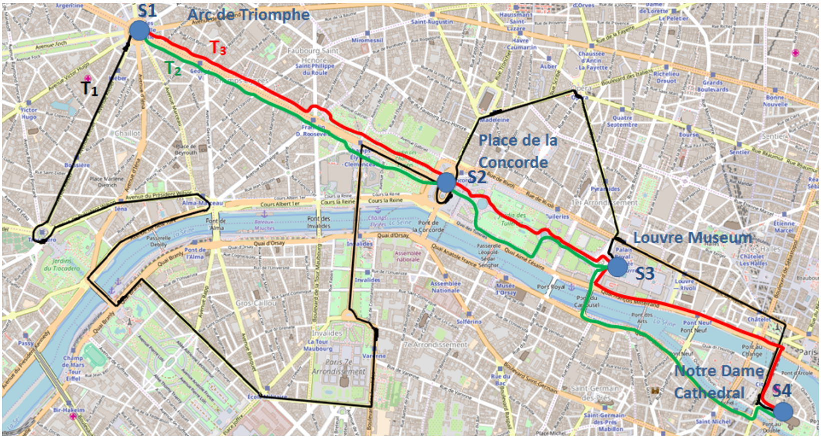

Figure 1.3 – Tourist trajectories in Paris with four stops

In taxi fraud detection, for instance, a similarity measure will help to answer questions like: given two regions of interest, which is the standard path followed by the majority of the taxis and which are the outliers? A real example of an outlier taxi trajectory is shown in Figure 1.4, in the San Francisco dataset of the

CRAWDAD project (PIORKOWSKI; SARAFIJANOVIC-DJUKIC; GROSSGLAUSER, 2009). In this example, given the stops Airport and Westfield San Francisco Centre (WSFC), a similarity measure must consider both stops and moves to find the black trajectories as the most similar movements between the Airport and WSFC, and the trajectory with purple dots as the most dissimilar trajectory, which made a completely different and longer trip.

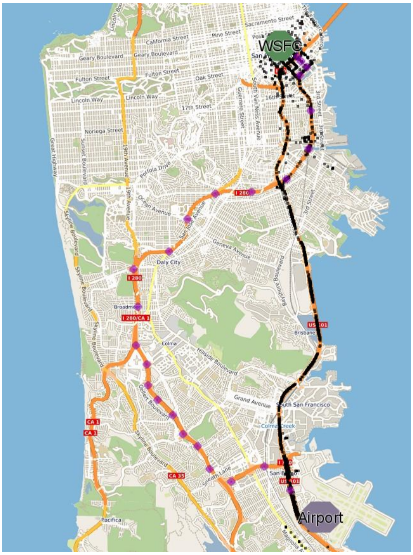

Figure 1.4 – An outlier trajectory (purple dots) going from Airport to downtown of San Francisco

In all previous examples, MSM cannot distinguish the trajectories, because it ignores the moves, and gives a similarity degree of 100% for the trajectories in both scenarios. Given the need of spatio-temporal similarity measures that consider both stops and moves, in this thesis we propose a new semantic trajectory similarity measure that extends MSM, proposed in (FURTADO; KOPANAKI, et al., 2016), to support both stops and moves. Our approach considers the partial sequence of the stops, what is not supported by MSM, allows different semantics for the moves, and uses weights to provide importance degrees for stops, moves, and their attributes.

In summary, we make the following contributions, as published in (LEHMANN; ALVARES; BOGORNY, 2019): (i) we propose a new similarity measure for multidimensional sequences treating elements with heterogeneous dimensions, which is the case of stops and moves; (ii) the semantic similarity measure considers both *stops* and *moves*, as well as their space, time, and semantic dimensions, allowing the use of different distance functions for each dimension, making the measure robust for several applications; (iii) the measure is flexible enough to partially consider the order between stops and to support different weights for stops, moves, and dimensions, allowing to give more or less importance to different trajectory parts; (iv) we evaluate the proposed measure with experiments over synthetic and real data, comparing our proposal to a large number of measures developed either for raw or semantic trajectories.

1.1 OBJECTIVE

The general objective of this thesis is the proposal of a novel semantic trajectory similarity measure that takes into account both *stops* and *moves*. The measure supports multiple dimensions, such as space, time, and semantics, and allows distinct distance functions for the dimension comparison.

1.2 METHODOLOGY

The methodology adopted in this thesis has 9 main steps:

Step 1: Perform a review of the literature in related subjects such as trajectory similarity and semantic trajectory similarity using Google Scholar.

Step 2: Study and implement related semantic trajectory similarity measures to find and understand limitations in related similarity measuring approaches.

Step 3: Define a new semantic trajectory similarity measure to overcome limitations of the state of the art.

Step 4: Select, organize, and pre-process several datasets with real trajectory data for the experimental evaluation.

Step 5: Define a set of experiments and establish a ground truth dataset for evaluating and comparing the proposed measure and state of the art.

Step 6: Study and define a set of measures to be used for evaluating similarity.

Step 7: Comparison of the results obtained in all datasets with the most related approaches in the literature.

Step 8: Write an article describing the proposed semantic similarity measure.

Step 9: Write the thesis describing the problem, the state-of-the-art, and the contribution for the problem solution with the advances over the state-of-the-art.

1.3 SCOPE AND OUTLINE

This thesis is limited to the proposal of a novel similarity measure for semantic trajectories, evaluation and comparison of this new similarity measure with the most related approaches in the literature.

The rest of this thesis is organized as follows: *Chapter* 2 presents the basic concepts and the related works for this thesis. *Chapter* 3 presents the proposed similarity measure with a running example. *Chapter* 4 presents experiments over real and synthetic trajectory data, and a discussion about the choice of a measure in face of application problems, and *Chapter* 5 concludes the thesis, presents its limitations, and points out future steps of the present research.

Chapter 2

Basic concepts and Related works

In this chapter we present the basic concepts for this thesis in Section 2.1. In Section 2.2 we present a review on trajectory similarity measures, where the section 2.2.1 presents measures for raw trajectory similarity and the section 2.2.2 presents measures for semantic trajectory similarity.

2.1 BASIC CONCEPTS

In this section we present basic concepts related to trajectories in Section 2.1.1 and basic concepts about similarity measures and some evaluation techniques in Section 2.1.2.

2.1.1 Trajectories

There are two important concepts that need to be explained and defined: raw trajectory and semantic trajectory. A raw trajectory is a discrete representation of the movement of an object that can be defined as a time-ordered finite sequence of space-time points, as formalized in Definition 1.

Definition 1 (Raw Trajectory)**.** A raw trajectory is a time-ordered sequence of points in the form $\mathrm{T} = < p_1, ..., p_n >$ where point $\mathrm{p_k} \in \mathrm{T}$ is a tuple $\mathrm{p_k} = (x, y, t)$, where x, y represent the spatial location of the moving object at a time instant t.

Figure 1.1 illustrates an example of a raw trajectory. The spatial coordinates are annotated next to the trajectory points and the time instants can be seen as the index associated to each point. For instance, the first point of the trajectory in the figure is located at the coordinates $(2, 3)$ at time instant 1.

In 2007, Alvares (ALVARES et al., 2007) and Spaccapietra (SPACCAPIETRA et al., 2008) proposed a new representation for trajectories, called semantic trajectory. A semantic trajectory is a time-ordered sequence of *stops* and *moves*, where the *stops* are the most relevant parts of the trajectory. In this work we formally define semantic trajectory considering its sequence of stops and moves, which is an enriched extension of the definition presented in (SPACCAPIETRA et al., 2008):

Definition 2 (Semantic Trajectory)**.** A semantic trajectory $S = \langle s_1, m_1, s_2, m_2, s_3, m_3, ..., s_k, m_k, s_{k+1} \rangle$ is a time ordered sequence of stops and moves, where each stop s_i has a set of attributes $\{d_{s1}, d_{s2}, ..., d_{sq}\}$ (including *space* and *time* as mandatory dimensions) characterizing it according to q-dimensions, and each move m_j has a set of attributes $\{d_{m1}, d_{m2}, ..., d_{mr}\}$ characterizing it according to r-dimensions.

Figure 2.1 shows a semantic trajectory S representing the movement of a professor. In this example the semantic trajectory is enriched with the name of the place where the *stop* occurred, the category of the place, its spatial coordinates, and the time interval that the *stop* happened. The *move* is enriched with the name of the street where the object moves, the traveled distance, and the average speed during the *move*.

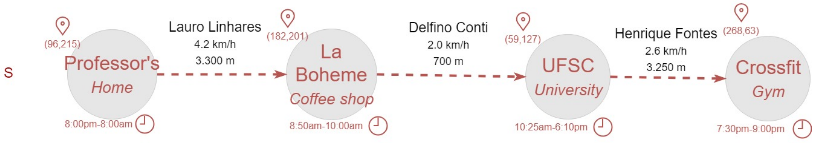

Figure 2.1 – A semantic trajectory S representing the movement of a professor

2.1.2 Similarity measures and evaluation techniques

To compare two trajectories we use a similarity measure. In this thesis we use the intuitive concept of similarity stated in (LIN et al., 1998), where two objects A and B are more similar as the commonality between each other increases, and they are less similar as their differences increase. We formalize the similarity measure concept according to Definition 3 introduced by (LIN et al., 1998):

Definition 3 (Similarity Measure)**.** A similarity measure on two objects A and B is a function $sim : A \times B \to [0, 1]$, such that the objects are more similar when the score returned by $sim(A, B)$ increases.

To evaluate how well a measure computes the similarity of two trajectories we use information retrieval evaluation techniques. In this thesis we use the Precision-Recall approach, computing the Mean Average Precision (MAP) and the Area Under the Curve (AUC) values, as stated in (BAEZA-YATES; RIBEIRO-NETO, 2011). In the Precision-Recall approach, the measure is evaluated as how precise is the information retrieval in each recall level. In this sense, the recall is the fraction of the relevant trajectories that are successfully retrieved. In the context of this thesis, a relevant trajectory is a trajectory of the same class of the one that is being evaluated. So, for each relevant trajectory retrieved, the precision increases. The process is repeated for the entire ground truth dataset. The Mean Average Precision (MAP) value of a Precision-Recall measure is the average precision in the recall of all relevant trajectories. The Area Under the Curve (AUC) value is calculated by constructing the Precision-Recall curve and calculating the area under this curve.

2.2 RELATED TRAJECTORY SIMILARITY MEASURES

Similarity measures have been proposed for several data processing and analysis techniques, such as outlier detection, top-K similarity queries, clustering, and others. In the context of trajectories, several similarity measures were proposed for both raw trajectories and semantic trajectories. In this section, we present a literature review on similarity measures of raw trajectories in Section 2.2.1 and Section 2.2.2 presents the similarity measures for semantic trajectories.

2.2.1 Related works on raw trajectory similarity measures

As presented in Section 2.1, a raw trajectory is a time-ordered sequence of points containing a spatial coordinate and a timestamp. For this reason, existing measures developed for generic time-ordered sequences or time-series can be applied to raw trajectories, even though they were not originally developed for this. At the beginning of this section, we present a distance measure proposed for time-series called Dynamic Time-Warping (DTW)(BERNDT; CLIFFORD, 1994) that was adapted to work with raw trajectory data in the work of (HOLT; REINDERS; HENDRIKS, 2007), creating the Multidimensional DTW (MD-DTW). Then we present similarity measures developed for raw trajectories which were adapted from more general similarity measures such as *Discrete Fréchet Distance* (EITER; MANNILA, 1994), *w-constrained discrete Fréchet Distance* (wDF) (DING; TRAJCEVSKI; SCHEUERMANN, 2008), *Longest Common Subsequence* (LCSS) (VLACHOS; KOLLIOS; GUNOPULOS, 2002), *Edit Distance on Real sequence* (EDR) (CHEN; ÖZSU; ORIA, 2005) and after, the similarity measure proposed exclusively for raw trajectories called *Uncertain Movement Similarity* (UMS) (FURTADO; ALVARES, et al., 2018).

Throughout this section, we use a set of symbols to denote hypothetical trajectories. Table 2.1 summarizes the symbols used in this section.

An early proposed distance measure is the *Dynamic Time Warping* (DTW) (BERNDT; CLIFFORD, 1994), developed for time-

Symbol	Meaning
P, Q and R	Trajectories
m and n	Number of points of trajectories P and Q, respectively
d_i	ith-dimension of data in a point
$window$	Size of the window
k	Number of *moves* in a semantic trajectory
ϵ	Distance threshold between two points matching
x, y	Spatial coordinates
$dist()$	Distance function

Table 2.1 – Symbol meanings

series. DTW is used to find the best match between the points of two time-series independent of their sizes. It creates a matrix with all possible pairs of points of the time-series with the pairwise distances as the entries. The distance between two trajectories is given by the sum of the entries of the minimum contiguous path in the matrix, where the minimum contiguous path is the best alignment between the sequences. Because DTW sums the distances between all points, it is sensitive to noise. For example, when a trajectory P has a point that is significantly distant from all points of the trajectory Q, even if all the other points of P and Q are close, their distance will be dominated by the distant point. A recursive formalization of DTW is presented in Equation 2.1.

$$DTW(P, Q) = \begin{cases} 0 & \text{if } m = n = 0 \\ \infty & \text{if } m = 0 \text{ or } n = 0 \\ dist(p_1, q_1) + min(& otherwise \\ DTW(< p_2 ... p_m >, < q_2 ... q_n >), & \\ DTW(< p_2 ... p_m >, Q), & \\ DTW(P, < q_2 ... q_n >)) & \end{cases} \tag{2.1}$$

The *Multidimensional DTW* (MD-DTW) (HOLT; REINDERS; HENDRIKS, 2007) extends DTW for dealing with sequences whose points have more than one dimension. MD-DTW normalizes the distance in the different dimensions and then creates a matrix with entries as the sum of the distances in all dimensions. Finally, it runs DTW over the matrix and finds the minimum contiguous path, that is, the path in the matrix connecting all points of both trajectories with minimum distance. Figure 2.2 illustrates the computation of MD-DTW between trajectories P and Q. Its distance is calculated as the sum of the minimum contiguous path between points of P and Q, i.e. the sum of all dashed lines.

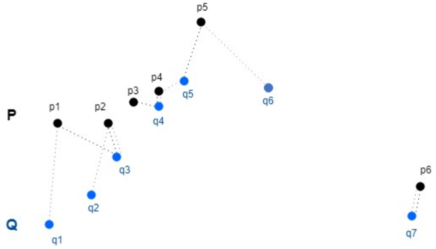

Figure 2.2 – MD-DTW score is the sum of distances of the minimum contiguous path between P and Q trajectories (dashed lines)

In the work of (SHOKOOHI-YEKTA et al., 2017) is proposed an *adaptive* DTW (DTWa) to multidimensional data. This adaptive approach is based on how the DTW computes the distance between two multidimensional sequences: (i) if the distance in each dimension is computed independently and summed at end; or (ii) if the distance between each multidimensional point is computed taking into account all dimensions together. The *adaptive* term comes from the decision of which approach is more reliable, by using a training dataset of multidimensional sequences and performing an evaluation.

Discrete Fréchet Distance was proposed in (EITER; MANNILA, 1994) as an adaption of the classical Fréchet Distance (FRÉCHET, 1906) to work with trajectories. This distance is also called the *coupling distance*, where the distance of two trajectories is the maximum distance of all aligned trajectory points on both trajectories. In this sense, an aligned trajectory point is a pair of points, where each point of one trajectory is *coupled* with one and only one point of the other trajectory, taking into account the order of the points in each trajectory. Due to this characteristic, the *Discrete Fréchet Distance* demands that both compared trajectories have the same number of points, what is a problem for real data.

Ding in (DING; TRAJCEVSKI; SCHEUERMANN, 2008) proposes *w-constrained discrete Fréchet Distance* (wDF), which extends the Discrete Fréchet distance (EITER; MANNILA, 1994) by adding a temporal window, in order to consider only the pairs of points that are within a given *window* time window. As DTW, wDF calculates the distance between the trajectory points by a continuous distance function (e.g. Euclidean distance), making it sensitive to noise. Indeed, this measure makes the assumption that the two trajectories have the same number of points, making point interpolation when necessary. This is a strict assumption and not good for real trajectories, that normally have very different sizes. The wDF distance is given by the minimum distance of all possible time windows over two trajectories, where the distance of each window is the maximum distance between all pairs of points of P and Q inside the window, as shown in Equation 2.2.

$$\begin{aligned} wDF(P,Q) &= min(\forall_{i,j=0} max(dist(P_i, Q_j))) \\ &\Rightarrow i \leq j + window \wedge j \leq P_m - window \end{aligned} \tag{2.2}$$

Figure 2.3 shows trajectories P and Q and a *window* time-window. The wDF distance between the trajectories is computed as the lower distance found among all *window*-constrained time-windows. As the time-window shifts over trajectories, the maximal distance between their points is computed using the Euclidean distance.

In 2002, Vlachos (VLACHOS; KOLLIOS; GUNOPULOS, 2002) proposed the *Longest Common Subsequence* (LCSS) for raw trajectory similarity measuring, considering the spatial distance between two points. In LCSS, given a point p of a trajectory P and a point q of a trajectory Q, they *match* if the distance between them is less or equal to a given *threshold* ϵ, as can be seen in Equation 2.3.

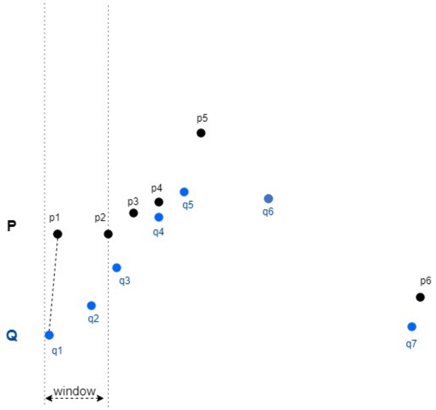

Figure 2.3 – wDF score is the minimal distance of all maximal distances between two points within the given time window *window*

LCSS reduces the effect of noisy data by quantifying the similarity between two points to binary values: 1 if the points match, and 0 otherwise. The longer the common subsequence of point matches between two trajectories, the more similar they are. A recursive formalization of LCSS is presented in Equation 2.4, which gives the total similarity of two trajectories P and Q.

$$match(p,q) = \begin{cases} true & dist(p_x, q_x) \leq \epsilon \\ & \text{and } dist(p_y, q_y) \leq \epsilon \\ false & otherwise \end{cases} \tag{2.3}$$

$$LCSS(P,Q) = \begin{cases} 0 & \text{if } m = n = 0 \\ 1 + LCSS(< p_2...p_m >, < q_2...q_n >) & \text{if } match(p_1, q_1) \\ max(LCSS(< p_2...p_m >, Q), & otherwise \\ LCSS(P, < q_2...q_n >)) & \end{cases} \tag{2.4}$$

A drawback of LCSS is its subsequence specificity, causing a inability to take into account gaps of any size in the trajectory. A gap is a subsequence of points in a trajectory P that is not close to any subsequence of points in a trajectory Q. Since the LCSS computation only takes into account the common/close points on both trajectories, this gap subsequence will not impact the computed similarity score. Figure 2.4 shows three trajectories P, Q, and R, with 3, 4, and 5 points, respectively. As can be seen, the first point of trajectory P matches with the first point of the trajectories Q and R, since the distance between the points is less than the threshold ϵ. The total LCSS similarity of P and Q is $LCSS(P, Q) = 1$, while the similarity of P and R is also $LCSS(P, R) = 1$, even though two points of R do not match any points of P.

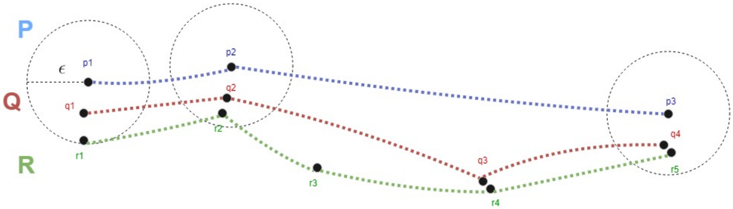

Figure 2.4 – Trajectories P, Q and R have 3 points matching, while trajectories Q and R have 4 points matching.

The LCSS similarity score is given by the size of the longest common subsequence ($LCSS(P, Q)$) over the size of the shortest trajectory, i.e., $\dfrac{LCSS(P,Q)}{min(m,n)}$. Figure 2.5 shows the matching of points of trajectories P and Q considering a threshold $\epsilon = 15$. The LCSS similarity score of P and Q is the number of points that match (solid black points) normalized by the size of the shortest trajectory, i.e. $\dfrac{4}{6} \approx 0.67$.

Chen in (CHEN; ÖZSU; ORIA, 2005) proposes the Edit Distance on Real sequence (EDR), another similarity measure for raw trajectories. EDR calculates the distance between two trajectories by computing the edit distance between their points. The edit distance between two trajectories is given by summing the distance between their points quantified as 1 if their points do not match, and 0 when they match (Equation 2.5). Using this approach, EDR solves

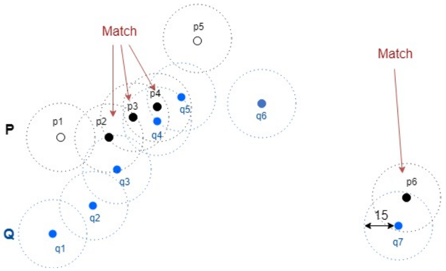

Figure 2.5 – LCSS similarity score is the number of matched points normalized by the size of shortest trajectory.

the problem of the gaps in LCSS, by taking into account points that do not match. However, to enforce a match between two points EDR requires that their distance is below a given threshold in all dimensions. A recursive formalization of EDR is presented in Equation 2.6.

$$match(p,q) = \begin{cases} 0 & dist(p,q) \leq \epsilon \\ 1 & otherwise \end{cases} \tag{2.5}$$

$$EDR(P,Q) = \begin{cases} 0 & \text{if } m = 0 \\ 0 & \text{if } n = 0 \\ min(EDR(< p_2 ... p_m >, < q_2 ... q_n >)+ & otherwise \\ match(p_1, q_1), EDR(< p_2 ... p_m >, Q) + 1, & \\ EDR(P, < q_2 ... q_n >) + 1) & \end{cases} \tag{2.6}$$

The EDR similarity score is given by the inverse of the number of non-matched points over the size of the longest trajectory, i.e., $1 - \dfrac{EDR(P,Q)}{max(m,n)}$. In the example of Figure 2.6, trajectories P and Q match in 4 of their points when using a threshold $\epsilon = 15$. As an edit distance, EDR takes into account how many changes in one of the trajectories are necessary to transform one trajectory in the other. In this case, the trajectories have four common/close points. To make the trajectories look similar, the trajectory P needs three changes in its points. For instance, 1) adding a new point similar to $q1$; 2) changing the point $p1$ to be close to the point $q2$; and 3) moving the point $p5$ closer to the point $q6$. The EDR similarity score of P and Q is the inverse of the total of non-matched points over the size of the longest trajectory, i.e. $1 - \frac{3}{7} \approx 0.57$. This similarity score shows that EDR is robust to compare trajectories of different sizes, by giving distinct similarity scores for trajectories of different sizes, solving the drawback of LCSS. Moreover, EDR maintains the robustness to noise of LCSS by using a threshold value in all dimensions.

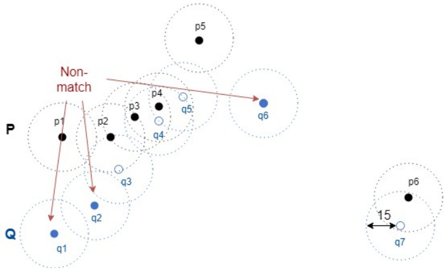

Figure 2.6 – EDR distance score is the number of non-matched points normalized by the size of largest trajectory, subtracted by 1.

Very recently in 2018, Furtado proposed the *Uncertain Movement Similarity* (UMS) in (FURTADO; ALVARES, et al., 2018). UMS is a parameter-free similarity measure designed exclusively for raw trajectories, using only the spatial dimension. The main contribution of UMS is the elimination of parameters for similarity measuring, by defining a dynamic spatial threshold that is computed automatically according to the distance between the pairs of points of a trajectory. As a consequence, it solves the problem of irregular distribution of trajectory points. UMS represents trajectories as a sequence of movement ellipses, covering the space between two sampled trajectory points. By using a dynamic ellipse size, UMS avoids the definition of a radius of fixed size around each point, which is a problem for real applications where the sampling rate is normally irregular as the object changes its movement speed.

Figure 2.7 shows the trajectories P and Q represented as two elliptical trajectories according to UMS. UMS computes the similarity score taking into account three premises: i) *alikeness*: the number of P ellipses that have some intersection with Q ellipses plus the number of Q ellipses that have some intersection with the P ellipses; ii) *shareness*: the space covered by ellipses of trajectories P and Q have a big shared area; and iii) *continuity*: the ellipses order represents moving objects traveling continually in the same direction. The limitation in this method lies in its inability to handle trajectories with higher sampling rate, because the higher the sampling rate of the points is, the smaller will be the generated ellipses, making the shared area between trajectories shorter.

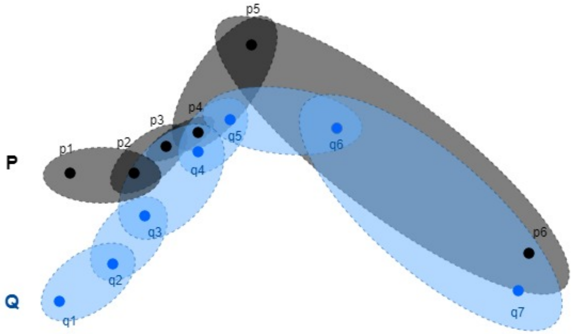

Figure 2.7 – UMS similarity score is given by: i) the shape *alikeness* of ellipses; ii) the *shared* area of ellipses; and iii) the *continuity* of points inside ellipses

2.2.2 Related works on semantic trajectory similarity measures

With the definition of semantic trajectory, the creation of new semantic-aware similarity measures became necessary. These new measures may analyze, besides the semantic information, any other information about the trajectory, as for instance the temporal duration of *stops* and *moves*, the moves spatial points, the average speed of the *moves* and so on. In the following we describe existing semantic trajectory similarity measures, as well as their limitations and applications.

An early similarity measure considering semantic trajectories is Common Visit Time Interval (CVTI) proposed in (KANG; KIM; LI, 2009). It defines a measure for integrating the semantics and the temporal dimensions of the stops. It finds the Longest Common Subsequence of two semantic trajectories in which all semantic aspects are the same and there is a time intersection between the stops. CVTI gives as similarity score the proportion of time that two trajectories share in the same *stops*. As CVTI is strongly based on LCSS, it presenting the same drawback of LCSS: the inability to penalize gaps of any size in the trajectory. Although CVTI uses different data dimensions, the measure is not extensible for other data dimensions associated with *stops* and *moves*, since it handles exclusively the semantic and the time dimensions of *stops*.

In (YING et al., 2010) the measure *Maximal Semantic Trajectory Pattern Similarity* (MSTP) was proposed. It identifies the Longest Common Sequence (LCS) between two semantic trajectories, which are sequences of labels describing the types of places such as $< School, Park, Cinema >$. MSTP uses only the semantic dimension of the trajectories, not being extensible for multiple dimensions, as time and space. MSTP differentiates from LCSS because it computes a ratio between each trajectory and their common patterns, i.e. the sequence of places visited by both trajectories. The average ratio is used to compute the similarity score, avoiding the drawback of LCSS that does not differentiate matching gaps of different sizes. The main limitation of the method lies in the exclusively semantic focus, being not extensible to multiple dimensions.

The work of (LIU; SCHNEIDER, 2012) proposed a semantic similarity measure that combines two distances: geographic and semantic. The geographic distance considers three aspects: (i) the distance between the centroids of the trajectories; (ii) the difference in the length of the trajectories; and (iii) the cosine similarity of the directions of subtrajectories. The measure uses the speed variation between the *stops* to split trajectories into subtrajectories. The semantic distance is based on LCSS to find the longest common subsequence of *stops* that were visited by the individual. Limitations of this approach include: i) sensibility to noise in the geographic distance; ii) the time distance is not considered; and iii) the prevalence of the geographic distance, i.e. two trajectories are similar only if they are similar in space.

The *Maximal Travel Match* (MTM)(XIAO et al., 2010) analyzes the trajectory similarity in the semantic dimension constrained by time. In order to do that, MTM takes into account the semantics of the visited places (e.g., restaurant, university etc.), the sequence of these places, the traveled time between places, and the frequency that a place was visited. Two trajectories are more similar if they visited places of the same type, in the same order, with similar travel times, according to a time threshold. Limitations of this approach include: i) two semantic trajectories are similar only if they visit the places in the same order; ii) the space dimension is not considered; and iii) MTM measures the similarity considering the whole dataset in order to obtain the frequency of the visited places, what makes the result dependent of the other trajectories in the dataset.

Furtado in (FURTADO; KOPANAKI, et al., 2016) proposed the MSM (Multidimensional Similarity Measure). This measure is the first in the literature working with multiple dimensions of stops, including space, time, and semantics. MSM was designed to handle multidimensional sequences, in which each dimension is independent and each dimension may have its own distance function. MSM uses a threshold value for defining if two elements match in a dimension or not, as can be seen in Equation 2.7.

$$match_d(a,b) = \begin{cases} 1 & dist_d(a,b) \leq maxDist_d \\ 0 & otherwise \end{cases} \tag{2.7}$$

Equation 2.8 presents how MSM computes the similarity score between all possible pairs of *stops* of two trajectories. For each pair of *stops*, MSM sums the matching value for all D dimensions and multiplies it by a pre-defined importance weight w_d for each dimension. With this, MSM supports to assign more or less relevance for each dimension, based on the application needs.

$$score(a,b) = \sum_{d=1}^{D} match_d(a,b) * w_d \tag{2.8}$$

To compute the similarity score between two trajectories, initially MSM calculates the *parity* between them, as shown in Equation 2.9. The parity score is given by summing the highest similarity scores of all stops a of the trajectory A when compared with all stops b of the trajectory B.

$$parity(A,B) = \sum_{a \in A} \mathbf{max}\{score(a,b) : b \in B\} \tag{2.9}$$

As the parity value is the number of commonalities between two trajectories A and B, MSM computes the final similarity score between them as the average of their parity values by the number of *stops* in both trajectories, as presented in Equation 2.10.

$$MSM(A,B) = \begin{cases} 0 & if|A| = 0 \vee |B| = 0 \\ \frac{parity(A,B)+parity(B,A)}{|A|+|B|} & otherwise \end{cases} \tag{2.10}$$

With this approach, MSM can take into account distinct dimensions, such as space, time, and semantics, to be scored in a single similarity value and it supports to define individual importance weights for each dimension. Some limitations of this approach include: i) the elements homogeneity allows MSM to handle *stops* only, since *stops* and *moves* have distinct attributes, and ii) the order of the elements is not taken into account during the similarity calculation.

Figure 2.8 shows the comparison of two semantic trajectories P and Q. In this figure, MSM scores the similarity in a pair-wise fashion, comparing all *stops* of trajectory P with all *stops* of trajectory Q. Its compares each dimension of the stops (space, time, and semantics), using a specific distance function for each dimension. After all stop-to-stop comparisons, MSM computes the similarity score as the sum of the best matching score of each *stop* of P and Q, divided by the sum of the trajectories length. In this example, MSM highly scores the similarity of the two trajectories, since both trajectories basically visit the same places, both spatially and semantically, and stay on the *stops* at approximately the same time. Notice that the sequence of the visited *stops* is very different between the two trajectories, but MSM does not take this into account, and what distinguishes these trajectories is the sequence of the *stops* and the *moves*.

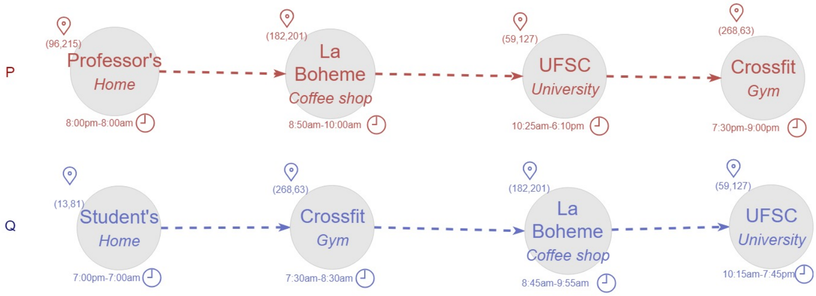

Figure 2.8 – MSM similarity measure computes the similarity score of P and Q using multiple dimensions with partial matching.

Cai in (CAI; K. LEE; I. LEE, 2016) proposed a measure that combines the strategies of LCSS (VLACHOS; KOLLIOS; GUNOPULOS, 2002) and MSM(FURTADO; KOPANAKI, et al., 2016) for semantic trajectories. It finds the longest common subsequence between two semantic trajectories. The difference to LCSS is that it does not require the matching in all dimensions, and it does separate the dimensions in two types: required or optional. While all required dimensions should be similar for two points to match, the optional ones are used only to increase the score. It uses weights for the optional dimensions, so if two points match in one required dimension, but do not match in other dimensions, their similarity is greater than when compared with other points that match only in the optional dimensions.

Table 2.2 summarizes the main characteristics of most related measures in comparison to the measure proposed in this thesis. We group the measures in two distinct categories: raw or semantic trajectory similarity.

In the first half of Table 2.2 are the similarity measures for raw trajectories. These similarity measures take into account only the raw data about trajectories, i.e., the space and the time dimensions. Not all similarity measures for raw trajectories handle the space and time dimensions. For instance, the wDF (DING; TRAJCEVSKI; SCHEUERMANN, 2008), DWTa (SHOKOOHI-YEKTA et al., 2017), and UMS (FURTADO; ALVARES, et al., 2018) only take into account the space. On the other hand, LCSS (VLACHOS; KOLLIOS; GUNOPULOS, 2002) and EDR (CHEN; ÖZSU; ORIA, 2005) were developed to consider both space and time.

In the last half of the Table 2.2 are the similarity measures for semantic trajectories. All similarity measures for semantic trajectories only consider the stops, and not all of them consider all three stops dimensions (space, time, and semantics). The sequence of the stops is only taken into consideration in CVTI (KANG; KIM; LI, 2009) , MSTP (YING et al., 2010), and the work of Liu and Schneider (LIU; SCHNEIDER, 2012). Only MSM (FURTADO; KOPANAKI, et al., 2016) and the work of Cai(CAI; K. LEE; I. LEE, 2016) make the definition of weights when measuring trajectory similarities, because they consider all three dimensions of stops. The work of this thesis considers all these features when measuring the similarity of two trajectories: i) the stops and the moves are taken into account, using all data dimensions, as space, time, and semantics; ii) there is a weight definition for each element (stop and move) and for each data dimension; and iii) the sequence of the elements is partially considered when comparing two trajectories.

Related Works	Raw trajectories	Semantic trajectories						Weights	Sequence
		Stops			Move				
		Space	Time	Semantics	Space	Time	Semantics		
LCSS (Vlachos, 2002)	✓								✓
EDR (Chen, 2005)	✓								✓
wDF (Ding, 2008)	✓								✓
DTWa (Shokoohi-Yekta, 2017)	✓								✓
UMS (Furtado, 2018)	✓								✓
CVTI (Kang, 2009)			✓	✓					✓
MSTP (Ying, 2010)		✓	✓	✓					✓
Liu, Schneider (2012)		✓		✓					✓
MSM (Furtado, 2016)		✓	✓	✓				✓	
Cai (2016)		✓	✓	✓				✓	
SMSM (our)		✓	✓	✓	✓	✓	✓	✓	✓

Table 2.2 – Comparative table

Chapter 3

Proposed measure

Stops and moves by definition are different and heterogeneous trajectory elements. A stop may have a spatial position, a start and end time, a category, and a set of attributes related to the category. For example, a stop at a hotel may have the attributes spatial location of the hotel, the start and end time the moving object stayed at the hotel, the number of stars, rate, price of the hotel, etc. A move starts and ends in a stop and may be characterized by different attributes as the average speed, traveled distance, sequence of followed streets, duration, the sequence of raw points, etc. These attributes are defined according to the needs of the application. From these examples we notice that stops and moves are characterized by different attributes, and they must be treated as different trajectory elements.

In order to deal with these heterogeneous elements (stops and moves), we introduce the concept of *movement element*. A movement element is a new representation that is not treated by other measures, mainly MSM, which supports only stops. Indeed, MSM does not consider the order of trajectory elements, while in our approach we want to preserve the partial sequence of both stops and moves in a movement element. With this approach, we say that two trajectories P and Q are more similar the more similar their movement elements are, i.e., the more similar the attributes of the stops (e.g. space, time, and semantics) and the moves are.

Definition 4 (Movement element)**.** A movement element e = *(stopS, move, stopE)* is a tuple formed by a start stop *stopS*, the move between *stopS* and *stopE*, and the end stop *stopE*, where *stopS* and *stopE* are two consecutive stops.

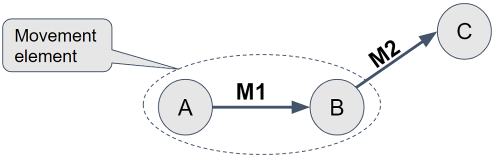

Figure 3.1 – A movement element from stop A to stop B passing over the move $M1$

Figure 3.1 exemplifies how the movement elements are built: the first movement element is formed by the sequence: stop A, move $M1$, and stop B. The second movement element is formed by the sequence stop B, move $M2$, and stop C.

Hereafter we will consider a semantic trajectory as a sequence of *movement elements*, as follows: $P = \langle e_1 = (s_1, m_1, s_2), e_2 = (s_2, m_2, s_3), ..., e_n = (s_n, m_n, s_{n+1})\rangle$.

Notice that we define a movement element as a trajectory part, or subtrajectory, and this structure will be used for the proposed similarity measure, where one trajectory will be compared with another one based on their movement elements. In a movement element we preserve the sequence of two stops and the move that connects the stops.

We analyze the similarity of a movement element $a \in A$ with another movement element $b \in B$, where A and B are semantic trajectories, in two parts: their stops and their moves. The basis for measuring the similarity of these two parts is the *match* function, given in Equation 3.1. The function returns 1 if the distance between an attribute i (also called dimension) of two movement elements is less than a given threshold *maxDist* for the dimension i, and zero otherwise. This function is used for measuring the distance of all dimensions of both the stops and the moves. For analyzing the spatial distance between two stops, for instance, if considering $maxDist = 100$, two stops match when their spatial distance is less or equal to 100.

$$match_i(a,b) = \begin{cases} 1 & dist_i(a,b) \leq maxDist_i \\ 0 & otherwise \end{cases} \tag{3.1}$$

To compute a total score for two movement elements a and b, we define the function *score(a,b)* in Equation 3.2, where w_{stop} and w_{move} are the weights of the stops and the moves, respectively, and their sum should be one. The importance of either stops or moves can vary from one application to another, so we can use the weights to give the respective importance.

$$score(a,b) = scoreStop(a,b) * w_{stop} + scoreMove(a,b) * w_{move} \tag{3.2}$$

In our measure we consider a score for the stops (scoreStop) and a score for the move (scoreMove). The functions *scoreStop(a,b)* and *scoreMove(a,b)* are defined in Equations 3.3 and 3.4, respectively. In both equations, r and q are the number of dimensions (attributes) of stops and moves, respectively. The score of the stops, computed according to Equation 3.3, is given by the weighted sum of all dimension matches of the start and end stops of two movement elements a and b. Some examples of computing *scoreStop()* and *scoreMove()* are presented in the next section. We observe in Equations 3.3 and 3.4 that, as MSM, we also give a weight for the dimensions. For instance, the spatial dimension or the semantic dimension of a stop could have a higher weight than the time.

$$\begin{aligned} scoreStop(a,b) = \sum_{i=1}^{r} (&match_i(a_{stopS}, b_{stopS}) + \\ &match_i(a_{stopE}, b_{stopE})) \div 2 * w_i \end{aligned} \tag{3.3}$$

$$scoreMove(a,b) = \begin{cases} \sum_{i=1}^{q} match_i(a_{move}, b_{move}) * w_i & if matchStops(a,b) \\ 0 & otherwise \end{cases} \tag{3.4}$$

Note in Equation 3.4 that the *scoreMove* depends on the function *matchStops(a, b)*. The intuition is that the moves of two trajectories should be compared only if their starting positions (starting stops) are spatially close and the ending positions (ending stops) are close as well. The function *matchStops(a,b)* is true when the spatial distance between a_{stopS} and b_{stopS} as well as between a_{stopE} and b_{stopE} is less than or equal to $maxDist$.

Let us consider the example of Figure 3.2 with two trajectories P and Q. The movement elements of trajectory P are $P_{e1} =< A, M1, B >$ and $P_{e2} =< B, M2, C >$. The trajectory Q has the movement elements $Q_{e1} =< A, M1, B >$ and $Q_{e2} =< B, M3, D >$. Considering P_{e1} and Q_{e1}, the function $scoreMove(P_{M1}, Q_{M1})$ will only be executed if the function $matchStops(P_{e1}, Q_{e1})$ is true, i.e., if the spatial distance between the stops P_A and Q_A and between the stops P_B and Q_B are both less than $maxDist_{space}$. Here, as both start stops and end stops are close in space, the function $matchStops(P_{e1}, Q_{e1})$ returns true, leading the function $scoreMove(P_{M1}, Q_{M1})$ to be executed, by computing the similarity of both moves.

For the movement elements $P_{e2} =< B, M2, C >$ and $Q_{e2} = < B, M3, D >$ shown in Figure 3.2, the start stops B and B match in space, but the end stops C and D do not have a spatial match (suppose their spatial distance is higher than $maxDist_{space}$). In this comparison the value of $scoreMove(M2, M3)$ is zero because the moves will not be compared.

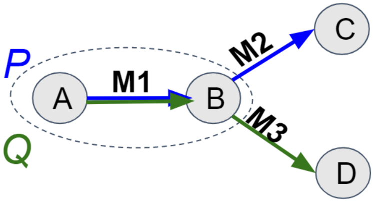

Figure 3.2 – Movement elements: (i) stop A to stop B through the move $M1$, (ii) stop B to stop C passing through the move $M2$, and (iii) stop B to stop D by the move $M3$

The function $scoreMove()$ guarantees the order of the stops inside the movement elements, what partially includes the order of stops in the similarity analysis. Suppose the example in Figure 3.2 is a real scenario, where A, B, C and D represent places as Spain, France, Germany, and Italy respectively. We believe that the movement elements going from Spain (A) to France (B) must have their move analyzed because they visit the same sequence of places, while the trajectory with the movement element that goes from France (B) to Germany (C) does not share the same destination of the trajectory with the movement element that goes from France (B) to Italy (D), so the moves of both movement elements are not compared, and the function $scoreMove(M2, M3)$ has value zero.

Having defined the score for stops and moves for comparing movement elements, Equation 3.5 defines the parity of two semantic trajectories P and Q. The parity of P with Q is the sum of the highest score of all the elements $p \in P$ when compared with all the elements of Q.

$$parity(P, Q) = \sum_{p \in P} \mathbf{max}\{score(p, q) : q \in Q\} \tag{3.5}$$

Finally, we can define the global similarity of two trajectories P and Q with $SMSM$. Equation 3.6 defines the stops and moves similarity measure $SMSM(P, Q)$ by the average parity of P with Q and of Q with P. The average parity is given by the sum of both parities over the sum of the number of elements in P ($|P|$) and the number of elements in Q ($|Q|$).

$$SMSM(P,Q) = \begin{cases} 0 & if |P| = 0 \lor |Q| = 0 \\ \frac{parity(P,Q)+parity(Q,P)}{|P|+|Q|} & otherwise \end{cases} \tag{3.6}$$

SMSM holds the same properties as MSM(FURTADO; KOPANAKI, et al., 2016), namely: non-negativity (Lemma 1); relaxed identity of indiscernibles (Lemma 2); and symmetry (Lemma (3).

Lemma 1. *Non-negativity* Given any two semantic trajectories A and B, then in all cases $SMSM(A,B) \geq 0$.

Proof. Direct from Equations 3.5 and 3.6 □

Lemma 2. *Relaxed identity of indiscernibles* Given two semantic trajectories A and B, and matching thresholds $maxDist_k^{stop}$ and $maxDist_j^{move}$, then $SMSM(A,B) = 1$ if and only if $A = B$ or
$(\forall a \in A \exists b \in B \mid (dist_k(a_{stop\_start}, b_{stop\_start}) \leq maxDist_k) \land$
$(dist_j(a_{move}, b_{move}) \leq maxDist_j) \land (dist_k(a_{stop\_end}, b_{stop\_end})$
$\leq maxDist_k)) \land (\forall b \in B \exists a \in A \mid (dist_k(a_{stop\_start}, b_{stop\_start})$
$\leq maxDist_k) \land (dist_j(a_{move}, b_{move}) \leq maxDist_j) \land$
$(dist_k(a_{stop\_end}, b_{stop\_end}) \leq maxDist_k))$

Proof. Equation 3.2 denotes that when an element a is within maximum distance thresholds for each dimension of the start stop, the end stop, and the move of another element b, *score*$(a,b) = (1 \times w_{stop} +$
$1 \times w_{move}) = 1$. Considering that the sum of the weights for the stop dimensions is 1, i.e. $\sum_{i=1}^{r} w_i = 1$, the sum of the weights for the move dimensions is 1, i.e., $\sum_{i=1}^{q} w_i = 1$, and the sum of the weights of stop and move is 1, i.e., $(w_{stop} + w_{move}) = 1$, we would have that for each element $a \in A$, at least one element $b \in B$ would be within the threshold in all dimensions of both stops and move

(and vice-versa) hence $score(a, b) = 1$. Therefore, *parity* (A, B) and *parity*(B, A) computed by Equation 3.5 will receive the maximum value (the number of movement elements of A and B, respectively). Finally, if $A = B$ then we have that $\frac{|A| + |B|}{|A| + |B|} = 1$. On the other hand, if at least one element of $a \in A$ does not match with at least one element $b \in B$ in all dimensions (or vice-versa), then its best matching score will be less than one. Hence, since $SMSM(A, B)$ is the average parity of A with B and of B with A (Equation 3.6) we have that in any other case $SMSM(A, B) < 1$.

□

Lemma 3. *Symmetry* Given any two semantic trajectories A and B, then in all cases $SMSM(A, B) = SMSM(B, A)$.

Proof. Direct from Equation 3.6 □

The complexity of SMSM is defined by the function *parity* (P, Q) that executes the function *score*() $m \cdot n$ times, where m and n are the length of trajectories P and Q. This is the same complexity of other similarity methods as MSM, LCSS and EDR, when the function *match*() inside *scoreMove*() is constant in time, that is the case when *match*() compares the name of the streets of the moves, the transportation means of the moves, the duration of the moves, etc. The processing time of SMSM is higher when the raw trajectories of the moves are considered for *match*(). When this is the case then the complexity of *match*() can become $O(r \cdot s)$, where r and s are the number of points of the two moves being compared. On the other hand, SMSM is faster than methods applied to raw trajectories.

3.1 EVALUATION OVER A RUNNING EXAMPLE

In this section we present a running example, comparing SMSM and MSM, since SMSM is an extension of MSM. Let us consider the two trajectories shown in Figure 3.3. Trajectory Q represents the daily routine of a professor, that starts his day at the gym in the morning, while trajectory P is the daily routine of a student, that starts his day at a coffee shop. Both trajectories visit the same places, sharing some streets, but in totally different order. The trajectories are annotated with the stop category, start and end time of the stop, an hypothetical geographic position (x, y) of the stop and the main street followed during the moves. So considering the notation *stop name ((x, y), [start timestamp - end timestamp])*, the student has the following movement behavior: stays at *Home* (*(96,215)*, [*8pm-8am*]), then he goes via Edu Vieira street to have breakfast at the *Coffee shop* (*(182,201)*, [*8:50am-10am*]), and from there goes via Delfino Conti street to the *University* (*(59,127)*, [*10:25am-6:10pm*]), finishing the day moving via Henrique Fontes street to the *Gym* (*(268,63)*, [*7:30pm-9pm*]). The professor (trajectory Q) goes from *Home* (*(13,81)*, [*7pm-7am*]) via Beira-mar avenue for jogging at the *Gym* (*(268,63)*, [*7:30am-8:30am*]). After he goes via Edu Vieira street to the *Coffee shop* (*(182,201)*, [*8:45am-9:55am*]), and via Delfino Conti reaches the *University* (*(59,127)*, [*10:15am-7:45pm*]) to teach his classes until the end of the day. We have two trajectories P and Q with their stops and moves annotated with the category of the place, the spatial position of the visited place, the time of the visit, and the name of the street to represent the move.

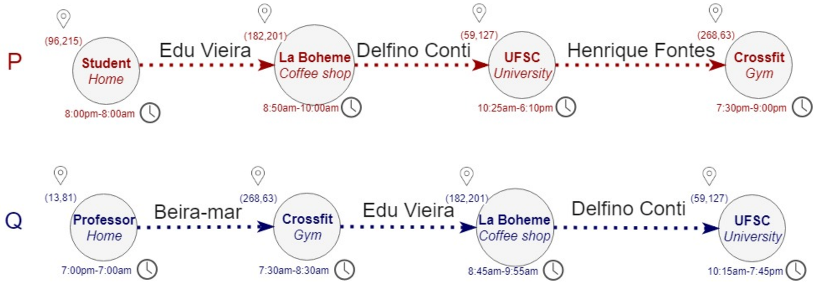

Figure 3.3 – Semantic Trajectories P (Student) and Q (Professor) with stops and moves

In order to calculate the SMSM similarity value, we first need to construct all movement elements for each trajectory. Table 3.1 lists these elements, where each element contains the start stop, the name of the street followed during the move, and the next stop.

Student (P)	Professor (Q)
<Home, Edu Vieira, Coffee shop>	<Home, Beira-mar, Gym>
<Coffee shop, Delfino Conti, University>	<Gym, Edu Vieira, Coffee shop>
<University, Henrique Fontes, Gym>	<Coffee shop, Delfino Conti, University>

Table 3.1 – Movement elements

To measure the distance between two movement elements, in this example we use the following distance functions for each stop dimension:

- Space: the Euclidean distance between the centroids of the stops. In this running example, the centroid is the central point of the stop;
- Time: let $[t1, t2]$ be the time interval of a stop. The time distance of two stops is given by:

$$dist_t(a,b) = 1 - \frac{duration([a.t1, a.t2] \cap [b.t1, b.t2])}{duration([min(a.t1, b.t1), max(a.t2, b.t2)])} \tag{3.7}$$

  We use this formula in order to have a proportion of the time intersection and not an absolute value;
- Semantics: the distance is equal to 0 in case of exact match and equal to 1 otherwise.

For the sake of simplicity, for the move, in this example we consider only the semantic information, i.e., the name of the followed street, where the distance is equal to 0 in case of exact match of street name and equal to 1 otherwise. We consider that stops and moves have the same weight and also the dimensions space, time and semantics of the stops.

In this running example we use as thresholds $maxDist_{space}$ = 100 and $maxDist_{time}$ = 0.5, i.e. two stops are said as matched in time when both share half of their period in that stop. With the distance functions and threshold values defined and elements constructed, we use Equation 3.2 to measure the similarity values between all element dimensions, computing first the match in both start and end stops and, if the stops match, we compute the match for the move.

To better understand how to measure the movement element similarity let us consider the following two movement elements:

- $element_P =$ $< Home_{[8pm-8am]}, EduVieira,\ Coffee\ shop_{[8{:}50am-10am]} >$
- $element_Q =$ $< Home_{[7pm-7am]}, Beira-mar, Gym_{[7{:}30am-8{:}30am]} >$

First, we apply the function *match*() (Equation 3.1) for the stops. In this case, the start stops have some degree of similarity: their semantics is the same and the time distance of $Home_P$ and $Home_Q$ is ≈ 0.15, lower than our defined threshold of 0.5. However, the spatial distance is $dist_{eucl}(Home_P, Home_Q) \approx 158$, higher than the defined threshold (100), so not matching in space, only in time and semantics, leading to a similarity score of 2/3 between start stops $Home_{[8pm-8am]}$ and $Home_{[7pm-7am]}$. The end stops (Gym and Coffee Shop) are dissimilar in space (with a distance of ≈ 163), in time (no overlap), and in semantics, then the similarity score between both end stops is 0.0.

Equation 3.3 computes the stops similarity as the average similarity of start stops and end stops as: *scoreStops*(*element*$_P$, *element*$_Q$) $= (2/3 + 0)/2 \approx 0.33$. As the function *matchStops*() is false in this example since $dist_{eucl}(Coffee\ shop_P, Gym_Q) > 100$, when applying Equation 3.4, the function *scoreMove*() $= 0$. Then, with Equation 3.2 we compute the movement element similarity as the sum of stops similarity weighted by w_{stops} and the move similarity weighted by w_{move}. Keeping this running example simple, we chose the same weights for the stops and the moves, i.e., 0.5 for w_{stops} and 0.5 for w_{move}. In this case, *score*(*element*$_P$, *element*$_Q$) $= (0.33 * 0.50) + (0.00 * 0.50) \approx 0.17$. Table 3.2 summarizes SMSM similarity scores between all movement elements.

After computing the similarity scores of both trajectories,

Q \ P	<**Home**, Edu V., **Coffee**> [8pm-8am] [8:50am-10am] (96,215) (182,201)	<**Coffee**, Delfino C., **University**> [8:50am-10am] [10:25am-6:10pm] (182,201) (59,127)	<**University**, Henrique F., **Gym**> [10:25am-6:10pm] [7:30pm-9pm] (59,127) (268,63)
<**Home**, Beira-mar, **Gym**> [7pm-7am] [7:30am-8:30am] (13,81) (268,63)	0.17	0	0.25
<**Gym**, Edu V., **Coffee**> [7:30am-8:30am] [8:45am-9:55pm] (268,63) (182,201)	0.25	0	0
<**Coffee**, Delfino C., **University**> [8:45am-9:55am] [10:15am-7:45pm] (182,201) (59,127)	0.08	1	0

Table 3.2 – Similarity scores for SMSM

with Equation 3.5 we compute the parity of trajectories, summing the highest scores of all movement elements of one trajectory when compared with all elements of the other trajectory. The parity calculus of $parity(P, Q) = (0.25+1.00+0.25) = 1.50$ and $parity(Q, P) = (0.25 + 0.25 + 1.00) = 1.50$. The final SMSM score is given by Equation 3.6 with $(parity(P, Q) + parity(Q, P))/(|P| + |Q|) = (1.50 + 1.50)/(3 + 3) = 0.50$, indicating that the trajectories have some degree of similarity, since the two trajectories have several common stops at similar time, move across the same streets, but the most important is that the order of the stops is different. Notice from Table 3.2 that movement elements where either the start stops or the end stops do not match, still have a degree of similarity, which is the case of the movement elements $< Home, EduVieira, Coffeeshop >$ and $< Home, Beira - mar, Gym >$, because they have a partial stop matching, i.e., their starting stops.

To compare SMSM with MSM, which is the closest work to our approach, we used for MSM the same thresholds for the stops and the same weights for space, time and semantics. MSM will measure the similarity between all stops using the same dimensions: space, time, and semantics. Let us consider the two stops at *Home*. Both stops have the same semantics and their time overlap is ≈ 0.15, lower than the defined threshold of 0.5. As the spatial distance between both (≈ 158) is higher than the defined threshold (100), in this dimension they do not match. The similarity score between both *Home* stops is the average of matched dimensions, leading to a similarity score of 2/3, the same as SMSM. The MSM similarity scores between all stops of trajectories P and Q are shown in Table 3.3.

Q P	Home [7:00pm-7:00am] (13,81)	Gym [7:30am-8:30am] (268,63)	Coffee shop [8:45am-9:55pm] (182,201)	University [10:15am-7:45pm] (59,127)
Home [8:00pm-8:00am] (96,215)	2/3	0	1/3	1/3
Coffee shop [8:50am-10:00pm] (182,201)	0	0	1	0
University [10:25am-6:10pm] (59,127)	1/3	0	0	1
Gym [7:30pm-9:00pm] (268,63)	0	2/3	0	0

Table 3.3 – Similarity scores for MSM

MSM calculates the parity between both trajectories by summing the highest scores of all stops of one trajectory compared with all stops of the other trajectory. The similarity value of MSM is given by $(\mathit{parity}(P,Q) + \mathit{parity}(Q,P))/(|P| + |Q|) = (3.33 + 3.33)/(4+4) \approx 0.83$, indicating that the two trajectories have a high similarity degree, what is not the case of the trajectories in the example. The high similarity given by MSM is due to the fact that the order of the stops is not important and the moves are not considered.

As we claimed initially, in some applications the movement sequence can be very important. In this example, SMSM evidences that, beside a strong similarity in the spatial dimension and stop semantics, the sequence of stops (i.e person routine) and the moves is very dissimilar. As SMSM compares the move between two consecutive stops using all data dimensions, or any information defined by the user, it is suitable for use with any kind of semantic trajectory, where the moves have multiple data dimensions to be analyzed, such as the spatial GPS points, the name of the followed streets, the traveled distance or travel time, etc. In the following section we compare our measure with other state-of-the-art approaches, considering both real and synthetic trajectories.

Chapter 4

Experimental Evaluation

To evaluate the proposed measure we performed three different experiments. The two first experiments use real and well-known trajectory datasets: the taxi trajectories in San Francisco from the CRAWDAD project (PIORKOWSKI; SARAFIJANOVIC-DJUKIC; GROSSGLAUSER, 2009) and the Geolife dataset (ZHENG et al., 2009). The third experiment uses a synthetic trajectory dataset, created using the Hermoupolis (PELEKIS et al., 2013) trajectory generator, which generates semantic trajectories with *stops* and *moves*. In the Geolife and taxi datasets we evaluate the similarity of both stops and moves, where the *moves* similarity is evaluated considering its raw points, while in the synthetic dataset we consider several types of semantic information associated to the *moves*.

We also evaluated how SMSM is impacted by changing its parameters and compare the running time of SMSM and the other similarity measures, for both raw and semantic trajectories.

We evaluate the precision of SMSM by the retrieval-based approach (*precision and recall*), computing the Area Under the Curve (AUC) and Mean Average Precision (MAP). To calculate the precision and recall, the trajectories are segregated into T_{class} by their classes and were used as the ground truth trajectories. For each ground truth trajectory, the $|T_{class}|$ most similar trajectories should also belong to T_{class}. For each one, a similarity search over the dataset is performed, ranking the trajectories until all T_{class} trajectories are found. Ideally, a similarity measure should return all trajectories in the ground truth between 1 to $|T_{class}|$ positions. The results of precision at each recall level are the average obtained for all T_{class} trajectories at that recall level. We compared SMSM with the following state-of-the-art similarity measures: MSM, LCSS, EDR, MSTP, CVTI, DTWa, wDF, and UMS.

Section 4.1 describes the experiment with the taxi dataset, Section 4.2 details the experiments with the Geolife dataset, Section 4.3 details the experiments with the synthetic dataset, Section 4.4 details the experiments changing the SMSM parameters, Section 4.5 evaluates the running time of the similarity measures, and Section 4.6 presents a discussion about the choice of a measure in face of application problems.

4.1 EXPERIMENT WITH THE TAXI DATASET

The epfl/mobility dataset contains taxi trips in San Francisco collected between May and June 2008, with an average sampling rate of about one point per minute. Each trajectory has several days of duration, what is not useful to determine similar movements around the town. For that reason, we split each taxi trajectory into short trajectories: (i) splitting when the occupation status of the taxi changed (taken or free); and (ii) splitting when a 5 minutes gap between two consecutive points was found, i.e. if a taxi does not send a GPS signal for 5 minutes we assume that the car was not moving around the city.

4.1.1 Ground truth generation

In this experiment we defined six distinct regions in San Francisco with high density of trajectories moving between these regions, which are shown in Figure 4.1(left). The regions are the

Park, the **Fisherman's Wharf**, the **Pier**, the Westfield San Francisco Center (**WSFC**), the **Intersection** between highways 280 and 101, and San Francisco **Airport**. In total, there are 6940 trajectories moving between these regions.

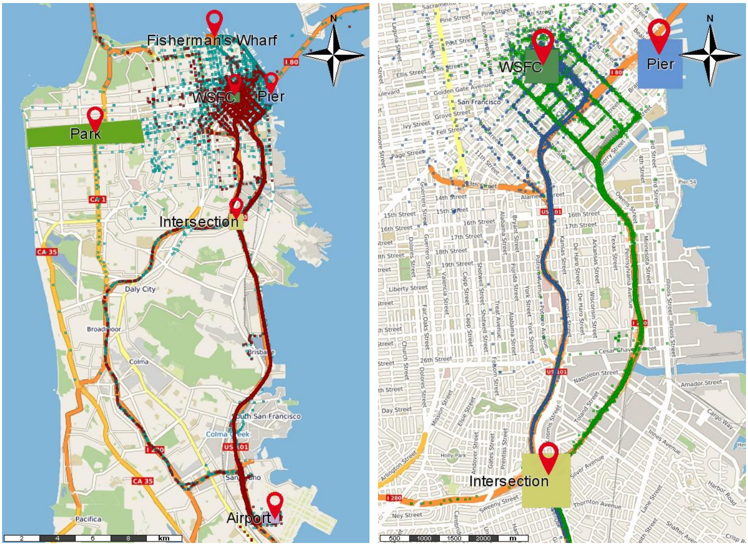

Figure 4.1 – (left) All trajectories moving between the six regions, where the red points are the ground truth trajectories and the light blue points are the remaining trajectories. (right) Ground truth trajectories, where dark blue points are the trajectories moving on highway 101 and green points are trajectories using highway 280

As ground truth we consider the subset of trajectories moving between Airport, Intersection, and WSFC. These trajectories were selected because they have different moves. Figure 4.1 (right) shows a zoom over the trajectories moving between **Airport** and **WSFC** using the **Intersection** of highways 101 (blue) and 280 (green), where we can visualize that the trajectories have different *moves*. We consider as the ground truth the four distinct paths followed by the trajectories that move between these regions, which are shown in Table 4.1. This ground truth definition is very spatial-based, with the classes defined as the spatial location of the stops and the moves.

Direction	Highway	Trajectories	Class
Airport to WSFC	101	145	A1
WSFC to Airport	101	1242	A2
Airport to WSFC	280	531	B1
WSFC to Airport	280	704	B2

Table 4.1 – Ground truth trajectories

4.1.2 Results for the taxi trajectories

In this experiment we considered the following dimensions for stops and moves: as spatial dimension of the stops we considered the centroid of the stop; as temporal dimension we used both start and end time of the stop; and as semantic information we used the name of the region (WSFC, Pier, Fisherman's Wharf, Park, Intersection, and Airport). For the moves, we analyze the real movement, and use as spatial dimension the moves raw points.

For measuring the stop similarity we use: (i) the Euclidean distance for space; (ii) distance 0 in case of exact match for semantics and 1 otherwise; and (iii) for the time dimension, where $[t1, t2]$ is the time interval of a stop, the time distance of two stops is given by Equation 3.7. For the moves, we consider the raw trajectory points and use the UMS measure for the move spatial similarity because it is the most appropriate for low sampled trajectories, which is the case for this dataset.

In this experiment we consider the same weights for each dimension and for stops and moves, so 0.5 for the stops and 0.5 for the moves, and 0.33 for the dimensions space, time, and semantics. Later, in the parameter analysis section, we show how the results change as we vary the weights of the stops and moves, as they are the central contribution of this thesis.

As several measures were not developed for semantic trajectories, for a more fair comparison we apply existing measures over semantic trajectories and over raw trajectories. For doing so we split the experiment in two parts: 1) a *precision and recall* evaluation using only semantic trajectories; and 2) a *precision and recall* evaluation using the raw trajectories.

Table 4.2 summarizes the dimensions used in each measure.

To general multidimensional similarity measures as MSM, we provide as input all dimensions of each stop, namely: 1) spatial information; 2) time interval; and 3) semantic information. We extend LCSS and EDR to support multiple dimensions, as in (FURTADO; KOPANAKI, et al., 2016): given two multidimensional trajectories, two points match when all dimensions match, where each dimension has a distinct distance threshold. With those adaptations, both LCSS and EDR are used to measure similarity using the dimensions of space, time and semantics for stops. For CVTI, we provide as input the time interval of the stops and the stop names. For MSTP, we provide the stop names only.

	Semantic trajectories				Raw trajectories
	Stop			Move	
	Space	Time	Semantics	Trajectory points	Space
SMSM	✓	✓	✓	✓	
MSM	✓	✓	✓		
MSTP			✓		
CVTI		✓	✓		
LCSS	✓	✓	✓		✓
EDR	✓	✓	✓		✓
DTWa					✓
UMS					✓
wDF					✓

Table 4.2 – Dimensions used for each measure

Table 4.3 shows the thresholds used for each measure. To define threshold values for the stops we experimented a range of values on each dimension as follows: for space (distance between stop centroids) we varied the distance from 100 meters to 500 meters in a 100 meters range; and for the time distance we tested a proportion of intersection from 0% to 100% varying in ranges of 10 %. For the move threshold we varied the UMS similarity for two moves from 0 to 1 in a 0.1 unit step.

Table 4.4 shows the comparison of SMSM with approaches developed either for raw or semantic trajectories. For semantic trajectories, SMSM (MAP=0.84) outperformed the other measures in 50% or more. This occurs because state-of-the-art measures do not take into account the move between two stops or consider only stops. We may notice that the second best measure for semantic trajectories is MSM, but it reaches only 39% of precision. This shows that

MSM is not robust when considering both stop and move similarity, because MSM cannot deal with moves and does not distinguish the order of the stops, i.e., the direction of the trajectories. To compare SMSM with similarity measures for raw trajectories, we use the SMSM MAP result (0.84) for semantic trajectories. SMSM was 27 % better than UMS (MAP=0.62) and DTWa (MAP=0.61), that were the most accurate measures apart from SMSM. Both UMS and DTWa perform worse than SMSM because, on the contrary to MSM, they consider only raw trajectories, and cannot deal with stops and their semantics.

	Semantic trajectories			Raw trajectories
	Space (meters)	Time proportion	Move	Space (meters)
SMSM	100	0.3	0.8	-
MSM	100	0.1	-	-
LCSS	100	0.1	-	100
EDR	100	0.1	-	100

Table 4.3 – Thresholds used for each measure

	Semantic		Raw	
	MAP	AUC	MAP	AUC
SMSM	**0.84**	**0.87**	**0.84**	**0.87**
UMS	-	-	0.62	0.66
DTWa	-	-	0.61	0.65
wDF	-	-	0.47	0.51
MSM	0.39	0.42	-	-
EDR	0.26	0.30	0.36	0.40
LCSS	0.29	0.33	0.34	0.38
MSTP	0.30	0.33	-	-
CVTI	0.28	0.32	-	-

Table 4.4 – MAP and AUC evaluation for the taxi dataset

Figure 4.2 (left) and Figure 4.2 (right) summarize the results of *precision and recall* of all similarity measures. On the left, SMSM was better to recover trajectories of the same class than the other methods in almost all recall levels, being around 60% more precise than state-of-the-art measures. On the right, we may notice that the measures developed for raw trajectories performed well because the trajectories of the different classes are partially discriminated by the moves, characterized by the trajectory raw points. These measures do not perform better than SMSM because they ignore the semantic dimensions and the stops.

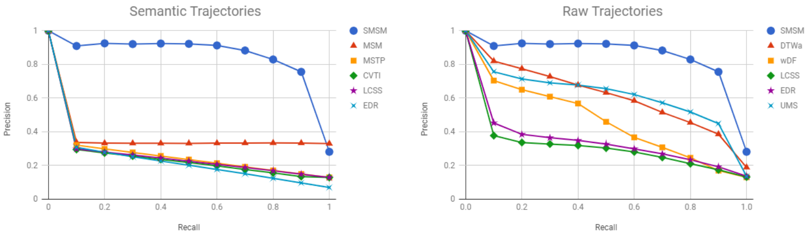

Figure 4.2 – Precision and recall results for semantic and raw trajectories

4.2 EXPERIMENT WITH THE GEOLIFE DATASET

The Geolife is a well-known trajectory dataset, created by Microsoft Research Asia (ZHENG et al., 2009) containing trajectories of 182 users, moving around Beijing, collected between April 2007 and August 2012. As a preprocessing step, we split trajectories when a 5 minutes gap between two consecutive points was found, since the trajectories of this dataset are highly sampled (lower than 2s).

4.2.1 Ground Truth Definition

As in the previous experiment, the Geolife dataset has no ground truth for evaluating similarity measures, so we had to generate a ground truth. We had to find stops where the objects make different moves between the stops in order to distinguish the trajectories. To build the ground truth, we chose an area in Beijing, where pedestrians move between the University Dormitories and Microsoft Research Office. We considered five places as stops (Microsoft, Starbucks, Market, Park and Dormitory), that are shown in Figure 4.3 (left). We selected 1976 trajectories that pass over two or more of these places. Among the 1976 trajectories we defined as ground truth the 337 trajectories that go from Microsoft to Dormitory (and vice versa) passing by Park and Market or Starbucks. We considered 5 distinct paths connecting the stops, labeled as A, B, C, D, and E, as shown in Figure 4.3 (right).

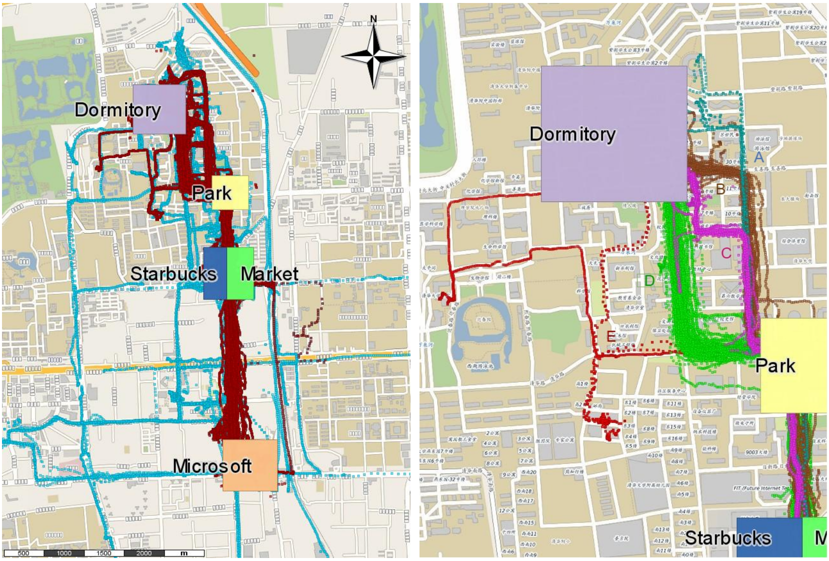

Figure 4.3 – (left) All trajectories moving between the five regions, where the red points are the ground truth trajectories and the light blue points are the remaining trajectories. (right) zoom over the ground truth trajectories moving between Park and Dormitory to observe their moves

In Table 4.5 we define as ground truth 8 distinct classes of movement based on the sequence of stops and the followed path: Microsoft to Dormitory via Market and Park by path A with 5 trajectories named as class A, Microsoft to Dormitory via Market and Park by path B with 40 trajectories named as class B, Dormitory to Microsoft via Park and Starbucks by path C with 11 trajectories named as class C, Dormitory to Microsoft via Park and Starbucks by path D with 115 trajectories named as class D1, Dormitory to Microsoft via Park and Market by path D with 7 trajectories named as class D2, Microsoft to Dormitory via Market and Park by path D with 149 trajectories named as class D3, Microsoft to Dormitory via Starbucks and Park by path D with 6 trajectories named as class D4 and Microsoft to Dormitory via Market and Park by path E with 4 trajectories named as class E.

It is worth mentioning that in this experiment, similar to the previous one, the moves are characterized by the raw trajectory points and we use these points to compare stops and moves similarity.

Direction	Path	Trajectories	Class
Microsoft to Dormitory via Market and Park	A	5	A
Microsoft to Dormitory via Market and Park	B	40	B
Dormitory to Microsoft via Park and Starbucks	C	11	C
Dormitory to Microsoft via Park and Starbucks	D	115	D1
Dormitory to Microsoft via Park and Market	D	7	D2
Microsoft to Dormitory via Market and Park	D	149	D3
Microsoft to Dormitory via Starbucks and Park	D	6	D4
Microsoft to Dormitory via Market and Park	E	4	E

Table 4.5 – Classes representing distinct paths of the ground truth

4.2.2 Results with the Geolife dataset

Following a similar methodology used for the previous experiment, we calculate the precision and recall for all classes in the ground truth (8), comparing the SMSM results to the other measures. The dimensions used for stops are: a) space; and b) the region name (Dormitory, Park, Starbucks, Market and Microsoft). For the moves we used the raw points of the move. The time dimension was not taken into account because in this experiment there are classes with few trajectories and most of them do not match in time.

For measuring the stop similarity we use: (i) the Euclidean distance for space; and (ii) 1 and 0 for the semantics in case of exact match or no match, respectively. For the moves, we consider the raw trajectory points. In this experiment we use DTW distance for analyzing the spatial similarity of the moves because the trajectory points are highly sampled and trajectory points are very near in space. For this dataset UMS is not the best measure for distinguishing the moves, since it was developed for irregular sampling. When the points are highly sampled, UMS tends to build small ellipses, so giving low similarity degree for very similar/close trajectories. For the weights we give the same importance for stops and moves, so we use 0.5 for stops and 0.5 for moves and for the dimensions, we use 0.5 for space and 0.5 for semantics.

Table 4.6 presents the thresholds used for each measure. We defined the thresholds by running each experiment over a range of possible threshold values and the best results for each method are reported. For raw trajectories, we evaluated as threshold the values 2, 4, 6, 8 and 10 meters because this dataset is highly sampled and is of pedestrian trajectories. The threshold for the move dimension was defined as follows: two moves are said to match if the DTW distance between them is less than the sum of the Euclidean distance of the moves. We used this approach for the move comparison because the DTW distance of two point sequences is not in the closed range between 0 and 1 as UMS, but an unbounded value, since DTW uses the Euclidean distance function to compare the points.

	Semantic trajectories	Raw trajectories
	Space (meters)	Space (meters)
SMSM	100	-
MSM	100	-
LCSS	100	8
EDR	100	8

Table 4.6 – Thresholds used for each measure

Table 4.7 shows the experimental results. SMSM (MAP=0.94) outperforms all measures for semantic trajectories, being significantly better than MSM (MAP=0.66), which ignores the moves and the sequence of stops, so it is not able to distinguish trajectories that move in the opposite direction. On the other hand, EDR (MAP=0.72) performs better than MSM because it considers the sequence, and the order of the stops distinguishes the classes. The measures for raw trajectories perform very well because of the low number of stops in this dataset and because the raw trajectories are similar in terms of space, and the classes were build based on the moves spatial similarity, so benefiting DTWa (MAP=0.92), LCSS (MAP=0.81), and EDR (MAP=0.81).

Figure 4.4 shows the precision and recall results. Figure 4.4 (left) shows that SMSM was better to recover semantic trajectories of the same class in all recall levels, while Figure 4.4 (right) shows that all measures developed for raw trajectories, except wDF, performed well, but the closest results to SMSM were achieved with DTWa.

	Semantic trajectories		Raw trajectories	
	MAP	AUC	MAP	AUC
SMSM	**0.94**	**0.95**	**0.94**	**0.95**
DTWa	-	-	0.92	0.93
EDR	0.72	0.73	0.81	0.83
LCSS	0.27	0.30	0.81	0.83
UMS	-	-	0.70	0.73
MSM	0.66	0.68	-	-
wDF	-	-	0.54	0.58
CVTI	0.30	0.34	-	-
MSTP	0.28	0.32	-	-

Table 4.7 – MAP and AUC evaluation with the Geolife dataset

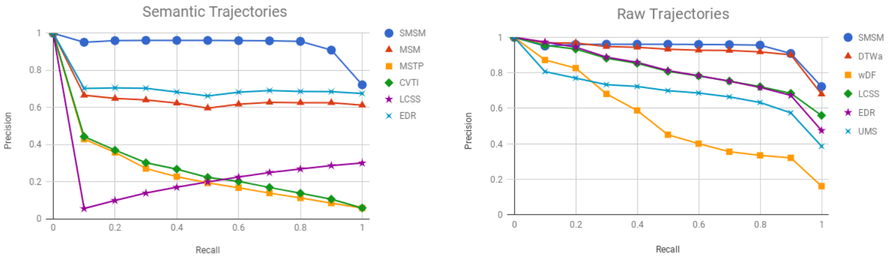

Figure 4.4 – precision and recall results to semantic (left) and raw (right) trajectories

4.3 EVALUATION WITH A SYNTHETIC DATASET

The objective of this experiment with synthetic data is to evaluate trajectory similarity considering trajectories with different number of stops and the semantic dimensions of the moves, instead of the raw points of the moves that were evaluated in the previous experiments.

We generate the trajectories with the Hermoupolis trajectory generator (PELEKIS et al., 2013), that allows creating trajectories based on pre-defined profiles. It generates semantic trajectories where both *stops* and *moves* are semantically enriched with annotations defined by the user. Therefore, it is possible to enrich the *moves* between the *stops* with semantic information, such as the transportation mean, the goal of the *move*, the name of the streets, etc. Hermoupolis has several parameters to simulate real trajectories, such as the definition of the average time of the moving object at each *stop*, the standard deviation of the time of each *stop*, the speed of the moves, sampling rate, and so on. We generated 440 trajectories with several stops and moves, using as semantics of the *stops* the POI category and the activity performed at the *stop*. For the moves we generated the raw points with the following attributes:(i) the transportation mode; (ii) the goal of the *move*; (iii) the traveled distance; (iv) the average speed; and (v) the duration of the move.

There are two main differences of this experiment w.r.t. the previous ones: (i) the trajectories of the ground truth have a different number of stops; and (ii) the moves have several and heterogeneous dimensions.

4.3.1 Ground Truth Definition

In this experiment we defined two classes, summarized in Table 4.8: (i) students with a job (80 trajectories); and (ii) students without a job (60 trajectories). What distinguishes the classes in this experiment is not the followed paths, but mainly the time and the duration of the stops, as well as the transportation mode and the goal of the move. The trajectories of class (Student Workder) start at $Home$, stay some time at the $Mall$ working, after they go to the $University$ to study and then they go back $Home$. The primary transportation mean is the public bus, but some trajectories move on foot or by car. The trajectories of class (Only Student) also start at $Home$, but instead of going to the $Mall$ to work, they go to the $Mall$ either for shopping or for lunch. After this, the trajectories go to $University$ and end at $Home$. Besides the 140 trajectories of the ground truth, we generated 300 other trajectories in the same area, with distinct behaviors, to make the retrieval task more challenging.

Class	Trajectories	number of stops
Student worker	80	4 and 5
Only student	60	between 4 and 7

Table 4.8 – Ground truth trajectories

Figure 4.5 (left) shows the generated trajectories, where the points in red are the trajectories of the ground truth, while the light blue points represent the remaining trajectories of the dataset. Some stops are Home, University, Supermarket, Restaurant, Bar, Mall, etc. Figure 4.5 (right) shows the trajectories of the ground truth.

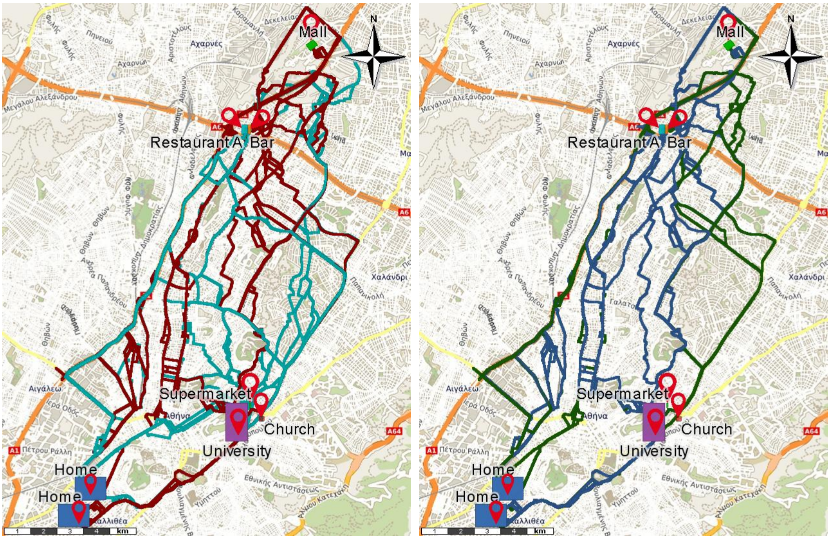

Figure 4.5 – (left) Hermoupolis generated trajectories in red (ground truth) and light blue (remaining); and (right) Ground truth trajectories in green (student worker) and blue (only student).

4.3.2 Results with the Synthetic Dataset

Following the methodology used in the previous experiments, we calculate the precision and recall for the two classes in the ground truth, comparing the results of SMSM with the other measures. The dimensions used for the *stops* are: a) the spatial centroid of the *stop*; b) the duration of the *stop*; and c) the *stop* name (Home, Mall, University, etc). For the *moves* we used several dimensions, including the transportation mode, the goal of the move, the traveled distance, average speed, and duration.

For measuring the *stops* similarity, we use: (i) the Euclidean distance for space; (ii) the proportion of the intersection between the time intervals of two *stops*, as given by Equation 3.7; and (iii) 1 for the semantics in case of exact match and 0 for no match. For the *moves*, the transportation mode and the goal have similarity as 1 in the case of exact match and 0 otherwise. The weights for the *stops* and the *moves* are the same (0.5), and 0.33 for of the dimensions space, time, and semantics. The thresholds used for each measure are shown in Table 4.9. As in the previous experiments, we define the thresholds for each dimension after having tested a range of values in each dimension (space and time), and the best results are reported for each measure.

	Semantic trajectories		Raw trajectories
	Space (meters)	Time (intersection proportion)	Space (meters)
SMSM	300	0.1	-
MSM	300	0.2	-
LCSS	100	0.2	8
EDR	100	0.2	8

Table 4.9 – Spatial and temporal thresholds used for each measure

Table 4.10 presents the experimental results, in which we tested SMSM with several attributes over the moves, and the best results where achieved with the dimensions transportation mode and goal. Comparing with measures for semantic trajectories, SMSM (MAP=0.95) was around 14% more precise than LCSS (MAP=0.82), EDR (MAP=0.82) and MSM (MAP=0.81). The other measures for semantic trajectories had worse results, with MAP scores of 0.40 and 0.33 for MSTP and CVTI, respectively. As can be noticed in the results, SMSM did not perform well for the dimensions duration, average speed, and traveled distance, because these dimensions are extracted from the moves raw points. As this experiment is characterized by the semantic dimensions of the moves, we can also notice that the measures for raw trajectories (EDR (MAP=0.51), DTWa (MAP=0.45), LCSS (MAP=0.44), UMS (MAP=0.29)) that performed well in the previous datasets where the moves raw points were considered, achieved at maximum 51% of precision in the current experiment.

Figure 4.6 (left) shows the precision and recall curves for all similarity measures developed for semantic trajectories. SMSM, using the transportation mode as the semantic dimension of the moves, is more accurate at each level, but EDR, LCSS, and MSM had good results despite considering only the stops. All the remaining mea-

	Semantic		Raw	
	MAP	AUC	MAP	AUC
SMSM (Transportation Mode)	**0.95**	**0.95**	-	-
SMSM (Goal)	**0.93**	**0.94**	-	-
SMSM (Duration)	0.73	0.75	-	-
SMSM (Average Speed)	0.72	0.75	-	-
SMSM (Traveled Distance)	0.72	0.74	-	-
EDR	0.82	0.85	0.51	0.54
LCSS	0.82	0.85	0.44	0.47
MSM	0.81	0.83	-	-
DTWa	-	-	0.45	0.48
MSTP	0.40	0.44	-	-
wDF	-	-	0.35	0.39
CVTI	0.33	0.37	-	-
UMS	-	-	0.29	0.33

Table 4.10 – MAP and AUC evaluation for the Hermoupolis dataset

sures present worse results. Figure 4.6 (right) shows the precision and recall curves for similarity measures developed for raw trajectories. In this dataset, all similarity measures developed for raw trajectories had poor results, because each class has several paths for each move, but what distinguishes the moves are not the raw points, but the transportation mode.

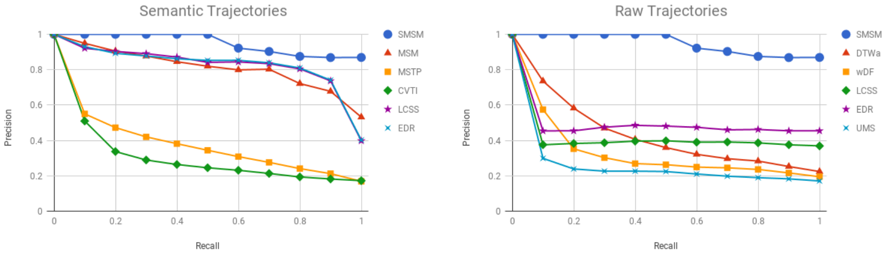

Figure 4.6 – (left) precision and recall results for semantic trajectories and (right) precision and recall results for raw trajectories

4.4 PARAMETER ANALYSIS

SMSM has two important groups of parameters: i) the *weights* to define the importance of stops/moves and dimensions; and ii) the spatial *threshold* to define if two stops match in order to analyze the moves. With the weights framework, SMSM is flexible to give more or less importance to the stops, moves, and dimensions.

To show the impact of the weight parameters on the similarity analysis, we evaluate the *weight* of the *stops* and the *moves* in all experimental datasets. Table 4.11 shows the results for the weight varying from 0 to 1, where 0 means that the moves will be ignored and 1 means that just the moves will be considered. As can be seen, if the influence of the *moves* is ignored, i.e., weight = 0, and all importance is given to the stops, SMSM reaches a MAP score of only 0.49 in the Taxi dataset, 0.72 in the Geolife dataset, and 0.75 in the synthetic dataset, indicating a confusion between the similarity of the trajectories that are also discriminated by the moves. The best average result is achieved when the weight of the *moves* is set as 0.5, i.e, half of the weight goes for the stops and the other half to the moves. This is because the move information, such as the raw points for the Taxi and Geolife dataset or the transportation mean for the Hermoupolis, play a decisive role in the information retrieval task.

Stop weight	Move weight	Taxi (MAP)	Geolife (MAP)	Hermoupolis (MAP)
0.00	1.00	0.66	0.93	0.67
0.25	0.75	0.66	0.94	0.96
0.50	**0.50**	**0.70**	**0.94**	**0.96**
0.75	0.25	0.71	0.94	0.92
1.00	0.00	0.49	0.72	0.75

Table 4.11 – Impact of the weights over the moves in trajectory similarity

We also evaluate the behavior of SMSM as the spatial distance threshold of the stops varies, from very low values (50 meters) up to very high values (2, 000 meters). As SMSM uses the spatial match between the start *stops* and between the end *stops* of two movement elements, the definition of this threshold has a great impact in the similarity score of the trajectories because the number of stop matches will increase. Table 4.12 shows the impact of the spatial threshold in the MAP results for the three experiments, i.e., Geolife, Taxi, and Hermoupolis datasets. As can be observed, there is not much variation in the results for the Geolife and Taxi datasets, where the stops are spatially distant from each other. For the synthetic dataset, when the distance threshold increases to 2, 000 meters, the precision decreases significantly, from 0.96 to 0.69, because the stops of different classes will have a spatial match and so the moves will be analyzed and many will match, so confusing the classes.

Threshold (meters)	Geolife (MAP)	Taxi (MAP)	Hermoupolis (MAP)
50	0.94	0.70	0.95
100	0.94	0.70	0.95
300	**0.94**	**0.70**	**0.96**
500	0.93	0.70	0.96
1000	0.93	0.70	0.93
2000	0.92	0.68	0.69

Table 4.12 – The MAP score for the spatial threshold variation from 50 up to 2,000 meters

4.5 RUNNING TIME EVALUATION

The computation time of a similarity analysis is affected directly by two points: (i) the similarity measure employed; and (ii) the number of points of the trajectories. In a similarity analysis task, we compute a similarity matrix between all trajectories of a dataset.

Table 4.13 presents the average running times of 5 separate runs for each similarity measure for semantic trajectories in seconds. The running times of SMSM were higher than all other measures. Comparing SMSM with the measures in the literature, the running time of SMSM in the Taxi dataset (354 seconds) was about 4 times higher than the CVTI (71 seconds) running time. In the Geolife dataset, SMSM (135 seconds) was approximately 13 times higher than MSM (7.8 seconds). In the Hermoupolis dataset, the SMSM running time (4.5 seconds) is about 9 times higher than the running time of CVTI (0.5 seconds). These differences rely on the *move* analysis performed by SMSM. In the Taxi and Geolife datasets, the *move* is evaluated through the raw points of the movement, and in this kind of analysis the quantity of GPS points affects proportionally the running time. The Taxi dataset has $132,680$ GPS points in $6,940$ trajectories, with $11,936$ *moves*. The similarity analysis task computes the similarity between all trajectories and, consequently, all *moves*, leading to perform $142,468,096$ *move* spatial points comparisons. The Geolife dataset has $806,688$ GPS points in $1,976$ trajectories, with $6,494$ *moves*, leading to $42,172,036$ *move* spatial points comparisons.

The focus of SMSM is its accuracy in measuring the similarity between semantic trajectories, and not on the running time. However, there are some techniques that can be used to speed up the running time, as the method proposed by Furtado in (FURTADO; PILLA; BOGORNY, 2018).

On the Hermoupolis dataset where we used the semantic information for the *move* comparison (transportation mode), the SMSM running time was lower. The comparison of the semantic dimension is less computationally intensive, reducing the total running time of the similarity analysis task.

	Taxi	Geolife	Hermoupolis
CVTI	**71**	11	**0.5**
EDR	77	14	1.2
LCSS	79	13	1.0
MSM	191	**7.8**	2.8
MSTP	100	13	1.0
SMSM	354	135	4.5

Table 4.13 – The average running times (in seconds) of the similarity analysis for the similarity measures for semantic trajectories

Table 4.14 shows the average running times for the similarity measures for raw trajectories. SMSM performed much faster than other measures on the Geolife and Hermoupolis datasets. That is directly related to the trajectory number of points, since SMSM compares semantic trajectories, with less points (as the *stops* and the *moves*), and similarity measures for raw trajectories compare each spatial point of each trajectory with the spatial points of all other trajectories. For instance, in the Taxi dataset the average number of points of the raw trajectories ≈ 19.12. When enriched with *stops* and *moves*, each trajectory has ≈ 2.72 *stops* and each move about 8.43 points. The difference in the number of points (raw points or stops) is about 1 order of magnitude. In the Geolife dataset that difference is bigger, since the average number of points of the raw trajectories is around 408, while the average number of stops in the semantic trajectories ≈ 4.29, leading to the difference being about 2 order of magnitude. The Hermoupolis dataset has the biggest difference on average number of points of the two kinds of trajectories: the raw trajectories have on average $\approx 3,232.84$ points, while the semantic trajectories have ≈ 5.23 *stops* and the *moves* raw points are not analyzed, only their semantics.

	Taxi	Geolife	Hermoupolis
DTWa	789	26,501	82,542
EDR	268	5,886	19,867
LCSS	255	5,919	20,337
UMS	729	5,219	17,086
wDF	**240**	1,302	3,177
SMSM	354	**135**	**4.5**

Table 4.14 – The average running time (in seconds) of the similarity measures developed for raw trajectories and the running time of SMSM

4.6 DISCUSSION

Trajectory data can have several formats. Depending on the format and the application requirements, different trajectory analysis and mining methods will be needed, and so different similarity measures can be applied. For applications that use raw GPS data, as trajectories of taxis, buses, or cars, with the intend to detect, for instance, traffic conditions or traffic jams, the most appropriate measures are UMS and DTWa. UMS is robust for trajectories with different sampling rates or different distances between trajectory points (the case when a trajectory varies the speed in a city), because instead of using a radius around each trajectory point to find the similar trajectories in the spatial neighbourhood, it uses ellipses between every two trajectory points, and the size of the ellipses is dinamically defined based on the distance between two trajectory points. UMS is not the best measure in highly sampled trajectories, where the ellipses are very small. In this case, DTWa is a good choice.

For applications that use GPS trajectories annotated only with stops or where the moves are not important, or trajectories extracted from social media data, which are more sparse and that do not have moves, the best measure is MSM. MSM is useful in applications where one is interested in finding users that visit the same places, at similar times, but where the order of the visits is not important. In tourism applications where the analyst wants to find similar tourist trajectories to either predict or to recommend the next place to be visited, MSM is not appropriate because it ignores the order. When the sequence of the visited places is important, even when the details about the moves are not available, SMSM is more appropriate, because it considers the order of the stops.

For dealing with GPS trajectories enriched with both stops and moves, and the spatial, temporal or semantic characteristics of both stops and moves are important, SMSM is the most appropriate measure. In a tourism application, for instance, where the tourist has a time constraint to visit a city, a sequence of visits can be recommended based on the similarity analysis of other tourists that visited the same city. SMSM is also robust in applications that focus on the most similar paths or popular routes between stops.

It is important to emphasize that for applications where the spatial movement of the moves is important, i.e., the raw trajectory points, SMSM can use UMS for the move similarity when trajectories have low sampling rate, and DTW when trajectories have high sampling rate.

Chapter 5

Conclusion

In this work we proposed SMSM, a new similarity measure for semantic trajectories that supports both stops and moves. To the best of our knowledge, SMSM is the first semantic trajectory similarity measure that deals with both stops and moves and their space, time and semantic dimensions. Our similarity measure is robust to consider multiple dimensions of stops and moves, where a move, for instance, can be represented as raw points, the traveled distance, the major direction, the names of streets, the transportation mode, etc.

SMSM is framework that supports the definition of weights for stops, moves and dimensions, so the measure is flexible to give more or less importance for specific parts of trajectories. On the other hand, these weights may be difficult to estimate from the user point of view, but in case he has no knowledge about the domain, the best option is to define the same weight for all elements.

We performed experiments using real and synthetic data of distinct contexts, including car trajectories and pedestrian trajectories. By evaluating SMSM with an information retrieval approach, we show that SMSM was more accurate than other measures developed either for raw or semantic trajectories.

We also evaluate the impact of the matching thresholds and weights in SMSM similarity, as well as its running time with respect to other measures.

SMSM requires a full spatial match between the start and end stops of two movement elements to evaluate the move. In future works we will study an extension of SMSM to evaluate the move in cases where the final stops of two movement elements do not have a match. Another future work is the expansion of the movement element in order to look ahead, not only to the final stop of a movement element. In other words, the idea is to allow movement elements with noise/gaps. For instance, a trajectory J_1 that goes from A to B and a trajectory J_2 that goes from A to X to B. According to SMSM, both trajectories will have little match, because J_2 has a stop X between A and B. An extension of SMSM to allow a more flexible construction of the movement element could benefit the semantic trajectory similarity.

Bibliography

ALVARES et al. A Model for Enriching Trajectories with Semantic Geographical Information. In: *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*. Seattle, Washington: ACM, 2007. (GIS '07), 22:1–22:8. ISBN: 978-1-59593-914-2. DOI: 10.1145/1341012.1341041. Address: <http://doi.acm.org/10.1145/1341012.1341041>. Cited 2 times on pages 27, 34.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. *Modern Information Retrieval*. 2. ed. Harlow, England: Pearson Addison Wesley, 2011. ISBN: 978-0-321-41691-9. Cited 2 times on pages 15, 35.

BERNDT, Donald J; CLIFFORD, James. Using dynamic time warping to find patterns in time series. In: SEATTLE, WA. *KDD workshop*. [S.l.]: [s.n.], 1994. vol. 10, pp. 359–370. Cited 2 times on page 36.

BOGORNY, Vania; BRAZ, Fernando Jose. *Introdução a Trajetórias de Objetos Móveis: conceitos, armazenamento e análise de dados*. Joinville: Univille, 2012. Cited 1 time on page 27.

CAI, Guochen; LEE, Kyungmi; LEE, Ickjai. Discovering Common Semantic Trajectories from Geo-tagged Social Media. In: FUJITA, Hamido et al. *Trends in Applied Knowledge-Based Systems and Data Science*. Ed. by Hamido Fujita. Cham: Springer International Publishing, 2016. pp. 320–332. ISBN: 978-3-319-42007-3. Cited 2 times on pages 48, 49.

CHEN, Lei; ÖZSU, M. Tamer; ORIA, Vincent. Robust and Fast Similarity Search for Moving Object Trajectories. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. Baltimore, Maryland: ACM, 2005. (SIGMOD '05), pp. 491–502. ISBN: 1-59593-060-4. DOI: 10.1145/1066157.1066213. Address: <http://doi.acm.org/10.1145/1066157.1066213>. Cited 4 times on pages 25, 36, 41, 49.

DING, Hui; TRAJCEVSKI, Goce; SCHEUERMANN, Peter. Efficient Similarity Join of Large Sets of Moving Object Trajectories. In: *Proceedings of the 2008 15th International Symposium on Temporal Representation and Reasoning*. Washington, DC, USA: IEEE Computer Society, 2008. (TIME '08), pp. 79–87. ISBN: 978-0-7695-3181-6. DOI: 10.1109/TIME.2008.25. Address: <https://doi.org/10.1109/TIME.2008.25>. Cited 3 times on pages 36, 39, 49.

DODGE, Somayeh; LAUBE, Patrick; WEIBEL, Robert. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Geographical Information Science*, vol. 26, no. 9, pp. 1563–1588, 2012. DOI: 10.1080/13658816.2011.630003. eprint: https://doi.org/10.1080/13658816.2011.630003. Address: <https://doi.org/10.1080/13658816.2011.630003>. Cited 1 time on page 26.

EITER, Thomas; MANNILA, Heikki. *Computing discrete Fréchet distance*. [S.l.], 1994. Cited 3 times on pages 36, 38, 39.

FILETO, Renato et al. Baquara: A holistic ontological framework for movement analysis using linked data. In: SPRINGER. *International Conference on Conceptual Modeling*. Springer, Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2013. pp. 342–355. Cited 2 times on page 27.

FRÉCHET, M. Maurice. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, vol. 22, no. 1, pp. 1–72, Dec. 1906. ISSN: 0009-725X. DOI: 10.1007/BF03018603. Address: <https://doi.org/10.1007/BF03018603>. Cited 1 time on page 38.

FURTADO, Andre Salvaro; ALVARES, et al. Unveiling Movement Uncertainty for Robust Trajectory Similarity Analysis. *Int. J. Geogr. Inf. Sci.*, Taylor & Francis, Inc., Bristol, PA, USA, vol. 32, no. 1, pp. 140–168, Jan. 2018. ISSN: 1365-8816. DOI: 10.1080/13658816.2017.1372763. Address: <https://doi.org/10.1080/13658816.2017.1372763>. Cited 4 times on pages 26, 36, 43, 49.

FURTADO, Andre Salvaro; KOPANAKI, Despina, et al. Multidimensional Similarity Measuring for Semantic Trajectories. *Transactions in GIS*, vol. 20, no. 2, pp. 280–298, 2016. ISSN: 1467-9671.

DOI: 10.1111/tgis.12156. Address: <http://dx.doi.org/10.1111/tgis.12156>. Cited 9 times on pages 28, 30, 46, 48, 49, 56, 67.

FURTADO, Andre Salvaro; PILLA, Laercio Lima; BOGORNY, Vania. A branch and bound strategy for Fast Trajectory Similarity Measuring. *Data & Knowledge Engineering*, Elsevier, vol. 115, pp. 16–31, 2018. Cited 1 time on page 80.

HOLT, Gineke A ten; REINDERS, Marcel JT; HENDRIKS, EA. Multi-dimensional dynamic time warping for gesture recognition. In: *Thirteenth annual conference of the Advanced School for Computing and Imaging.* [S.l.]: [s.n.], 2007. vol. 300. Cited 2 times on pages 36, 37.

KANG, Hye-Young; KIM, Joon-Seok; LI, Ki-Joune. Similarity Measures for Trajectory of Moving Objects in Cellular Space. In: *Proceedings of the 2009 ACM Symposium on Applied Computing.* Honolulu, Hawaii: ACM, 2009. (SAC '09), pp. 1325–1330. ISBN: 978-1-60558-166-8. DOI: 10.1145/1529282.1529580. Address: <http://doi.acm.org/10.1145/1529282.1529580>. Cited 4 times on pages 28, 45, 49.

LAUBE, Patrick; KREVELD, Marc van; IMFELD, Stephan. Finding REMO — Detecting Relative Motion Patterns in Geospatial Lifelines. In: *Developments in Spatial Data Handling.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. pp. 201–215. ISBN: 978-3-540-26772-0. Cited 1 time on page 26.

LEHMANN, Andre; ALVARES; BOGORNY, Vania. SMSM: A Similarity measure for trajectory Stops and Moves. *Int. J. Geogr. Inf. Sci.*, Taylor & Francis, Inc., Bristol, PA, USA, 2019. Cited 1 time on page 31.

LIN, Dekang et al. An information-theoretic definition of similarity. In: CITESEER, 1998. *Icml.* [**sinelocosinenomine**], 1998. vol. 98, pp. 296–304. Cited 2 times on page 35.

LIU, Hechen; SCHNEIDER, Markus. Similarity Measurement of Moving Object Trajectories. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming.* Redondo Beach, California: ACM, 2012. (IWGS '12), pp. 19–22. ISBN: 978-1-4503-1695-8. DOI: 10.1145/2442968.2442971. Address: <http:

//doi.acm.org/10.1145/2442968.2442971>. Cited 4 times on pages 28, 45, 49.

PALMA, Andrey Tietbohl et al. A Clustering-based Approach for Discovering Interesting Places in Trajectories. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*. Fortaleza, Ceara, Brazil: ACM, 2008. (SAC '08), pp. 863–868. ISBN: 978-1-59593-753-7. DOI: 10.1145/1363686.1363886. Address: <http://doi.acm.org/10.1145/1363686.1363886>. Cited 1 time on page 27.

PELEKIS, Nikos et al. Hermoupolis: A Trajectory Generator for Simulating Generalized Mobility Patterns. In: BLOCKEEL, Hendrik et al. *Machine Learning and Knowledge Discovery in Databases*. Ed. by Hendrik Blockeel. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. pp. 659–662. ISBN: 978-3-642-40994-3. Cited 3 times on pages 15, 63, 73.

PIORKOWSKI, Michal; SARAFIJANOVIC-DJUKIC, Natasa; GROSSGLAUSER, Matthias. *CRAWDAD dataset epfl/mobility (v. 2009-02-24)*. [S.l.]: [s.n.], Feb. 2009. Downloaded from https://crawdad.org/epfl/mobility/20090224. DOI: 10.15783/C7J010. Cited 3 times on pages 15, 30, 63.

RINZIVILLO, Salvatore et al. Where Have You Been Today? Annotating Trajectories with DayTag. In: NASCIMENTO, Mario A. et al. *Advances in Spatial and Temporal Databases - 13th International Symposium, SSTD 2013, Munich, Germany, August 21-23, 2013. Proceedings*. Ed. by Mario A. Nascimento. [S.l.]: [s.n.], 2013. vol. 8098. (Lecture Notes in Computer Science), pp. 467–471. ISBN: 978-3-642-40234-0. DOI: 10.1007/978-3-642-40235-7_30. Address: <https://doi.org/10.1007/978-3-642-40235-7%5C_30>. Cited 1 time on page 27.

ROCHA, Jose Antonio M. R. et al. DB-SMoT: A direction-based spatio-temporal clustering method. In: *5th IEEE International Conference on Intelligent Systems, IS 2010, 7-9 July 2010, University of Westminster, London, UK*. [S.l.]: [s.n.], 2010. pp. 114–119. ISBN: 978-1-4244-5164-7. DOI: 10.1109/IS.2010.5548396. Address: <https://doi.org/10.1109/IS.2010.5548396>. Cited 1 time on page 27.

SHOKOOHI-YEKTA, Mohammad et al. Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 1–31, Jan. 2017. ISSN: 1573-756X. DOI: `10.1007/s10618-016-0455-0`. Address: <`https://doi.org/10.1007/s10618-016-0455-0`>. Cited 2 times on pages 38, 49.

SPACCAPIETRA, Stefano et al. A Conceptual View on Trajectories. *Data Knowl. Eng.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, vol. 65, no. 1, pp. 126–146, Apr. 2008. ISSN: 0169-023X. DOI: `10.1016/j.datak.2007.10.008`. Address: <`http://dx.doi.org/10.1016/j.datak.2007.10.008`>. Cited 4 times on pages 26, 27, 34.

VLACHOS, Michail; KOLLIOS, George; GUNOPULOS, Dimitrios. Discovering similar multidimensional trajectories. In: IEEE. *Data Engineering, 2002. Proceedings. 18th International Conference on.* [S.l.]: [s.n.], 2002. pp. 673–684. Cited 5 times on pages 25, 36, 39, 48, 49.

XIAO, Xiangye et al. Finding Similar Users Using Category-based Location History. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems.* San Jose, California: ACM, 2010. (GIS '10), pp. 442–445. ISBN: 978-1-4503-0428-3. DOI: `10.1145/1869790.1869857`. Address: <`http://doi.acm.org/10.1145/1869790.1869857`>. Cited 1 time on page 46.

YING, Josh Jia-Ching et al. Mining User Similarity from Semantic Trajectories. In: *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks.* San Jose, California: ACM, 2010. (LBSN '10), pp. 19–26. ISBN: 978-1-4503-0434-4. DOI: `10.1145/1867699.1867703`. Address: <`http://doi.acm.org/10.1145/1867699.1867703`>. Cited 4 times on pages 28, 45, 49.

ZHENG, Yu et al. Mining interesting locations and travel sequences from GPS trajectories. In: ACM. *Proceedings of the 18th international conference on World wide web.* [S.l.]: [s.n.], 2009. pp. 791–800. Cited 3 times on pages 15, 63, 69.