Pontifícia Universidade Católica
do Rio de Janeiro

Lívia Couto Ruback Rodrigues

Enriching and analyzing Semantic Trajectories with Linked Open Data

Tese de Doutorado

Thesis presented to the Programa de Pós-graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências - Informática.

Advisor: Prof. Marco Antonio Casanova

Co-Advisor: Dra. Chiara Renso

Rio de Janeiro
December 2017

Pontifícia Universidade Católica
do Rio de Janeiro

Lívia Couto Ruback Rodrigues

Enriching and analyzing Semantic Trajectories with Linked Open Data

Thesis presented to the Programa de Pós-graduação em Informática of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Ciências - Informática. Approved by the undersigned Examination Committee.

Prof. Marco Antonio Casanova
Advisor
Departamento de Informática – PUC-Rio

Dra. Chiara Renso
Co-advisor
ISTI/CNR – Pisa, Italy

Prof. Hélio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Prof. Bernardo Pereira Nunes
Departamento de Informática – PUC-Rio

Prof. Giseli Rabello Lopes
UFRJ

Prof. Luiz André Portes Paes Leme
UFF

Prof. Marcio da Silveira Carvalho
Vice Dean of Graduate Studies
Centro Técnico Científico da PUC-Rio

Rio de Janeiro, December 15th, 2017

Lívia Couto Ruback Rodrigues

Lívia Couto Ruback Rodrigues holds a master in computer science degree from Pontifical Catholic University of Rio de Janeiro (PUC-Rio), also a computer science degree from Federal University of Juiz de Fora (UFJF). Her main research topics areas include Semantic Web, Information Retrieval, Semantic Trajectories and Data Mining.

Bibliographic data

Rodrigues, Lívia Couto Ruback

Enriching and analyzing Semantic Trajectories with Linked Open Data / Lívia Couto Ruback Rodrigues; advisor: Marco Antonio Casanova. – 2017.

104 f.: il. ; 29,7 cm

1. Tese (Doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui bibliografia.

1. Informática – Teses. 2. Trajetórias Semânticas. 3. Dados Interligados. 4. Web Semântica. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CCD: 004

I dedicate this thesis to the light of my life:

My dear parents, my lovely sisters, and our super cute cats
Thor, Lica and Milka (in memory).

Acknowledgments

This thesis becomes a reality with the support of many people. I would like to extend my gratitude to all of them.

Foremost, I would like to express my special and deep gratitude to my advisor, Prof. Marco Antonio Casanova, as I said four years ago, in the acknowledgments of my master dissertation, the best advisor I could ever ask for. Thank you for your unwavering support, wisdom, sense of humor, inspiring ideas, and kindness. It was an honor to me to work with you these six years.

I would like to thank Chiara, also the best co-advisor I could ask, for her kindness, patience, always willing to help and encouraging me with her best research advices and personal support. I am deeply grateful for your support during my staying in Italy. I admire your strength and sensitive personality.

To my dear father Rubens for your love. This thesis is my gift for your 70' birthday! To my mom Izabel, for her infinite loving and caring. To my lovely sister Flávia for organizing all family gatherings to keep us together. To my little sister Bianca for your huge support even being six years younger. To Itamar for being the brother we did not have. To our cute cats for bringing happiness and joy to the house. I am a fortunate person to be a member of this family.

To all my friends and colleagues that somehow contributed for this work. To Sergio, for the wonderful moments we spent together and for helping me revealing the best of myself. To Higor, for your sincere friendship during almost twenty years. To my therapist Kelly, for her kindness on helping me in this crazy process of self-knowledge. To my dear friends Sharon, Thais, Ana, Mayra, Aline, Marília, Rachel, Maria Solara, Fernanda, Kris, Regis and Vinicius. I will keep all of you in my heart.

To PUC-Rio, CAPES, CNPq and Tecgraf for funding my research.

Abstract

Rodrigues, Lívia Couto Ruback; Casanova, Marco Antonio (Advisor). **Enriching and analyzing Semantic Trajectories with Linked Open Data**. Rio de Janeiro, 2017. 104p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The last years witnessed a growing number of devices that track moving objects: personal GPS equipped devices and GSM mobile phones, vehicles or other sensors from the Internet of Things but also the location data deriving from the Social Networks check-ins. These mobility data are represented as trajectories, recording the sequence of locations of the moving object. However, these sequences only represent the raw location data and they need to be semantically enriched to be meaningful in the analysis tasks and to support a deep understanding of the movement behavior. Another unprecedented global space that is also growing at a fast pace is the Web of Data, thanks to the emergence of the Linked Data initiative. These freely available semantic rich datasets provide a novel way to enhance trajectory data. This thesis presents a contribution to the many challenges that arise from this scenario. First, it investigates how trajectory data may benefit from the Linked Data Initiative by guiding the whole trajectory enrichment process with the use of external datasets. Then, it addresses the pivotal topic of the similarity computation between Linked Data entities with the final objective of computing the similarity between semantically enriched trajectories. The novelty of our approach is that the thesis considers the relevant entity features as a ranked list. Finally**,** the thesis targets the computation of the similarity between enriched trajectories by comparing the similarity of the Linked Data entities that represent the enriched trajectories.

Keywords

Semantic trajectories; Semantic similarity; Movement data; Linked Data; Semantic Web.

Resumo

Rodrigues, Lívia Couto Ruback; Casanova, Marco Antonio. **Enriquecendo e analisando Trajetórias Semânticas com Dados Abertos Interligados.** Rio de Janeiro, 2017. 104p. Tese de Doutorado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Os últimos anos testemunharam o uso crescente de dispositivos que rastreiam objetos móveis: equipamentos com GPS e telefones móveis, veículos ou outros sensores da Internet das Coisas, além de dados de localização de check-ins de redes sociais. Estes dados de mobilidade são representados como trajetórias, e armazenam a sequência de posições de um objeto móvel. Porém, estas sequências representam somente os dados de posição originais, que precisam ser semanticamente enriquecidos para permitir tarefas de análise e apoiar um entendimento profundo sobre o comportamento do movimento. Um outro espaço de dados global sem precedentes tem crescido rapidamente, a Web de Dados, graças à iniciativa de Dados Interligados. Estes dados semânticos ricos e livremente disponíveis fornecem uma nova maneira de enriquecer dados de trajetória. Esta tese apresenta contribuições para os desafios que surgem considerando este cenário. Em primeiro lugar, a tese investiga como dados de trajetória podem se beneficiar da iniciativa de dados interligados, guiando todo o processo de enriquecimento semântico utilizando fontes de dados externas. Em segundo lugar, aborda o tópico de computação de similaridade entre entidades representadas como dados interligados com o objetivo de computar a similaridade entre trajetórias semanticamente enriquecidas. A novidade da abordagem apresentada nesta tese consiste em considerar as características relevantes das entidades como listas ranqueadas. Por último, a tese aborda a computação da similaridade entre trajetórias enriquecidas comparando a similaridade entre todas as entidades representadas como dados interligados que representam as trajetórias enriquecidas.

Palavras-chave

Trajetórias semânticas; Similaridade semântica; Dados de movimento; Linked Data; Semantic Web.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LOD	Linked Open Data
RDF	Resource Description Framework
VoID	Vocabulary of Interlinked Datasets
LIMES	Link Discovery Framework for metric spaces
URI	Uniform Resource Identifier

1
Introduction

1.1.
Context and Challenges

The Ericsson Mobility report[1], that periodically forecast the growth of mobile technology, predicts that, for the next 6 years, more than 1 million new mobile broadband-subscribed will be added per day, which means an additional 2.6 billion subscribers by the end of 2022. Figure 1 shows the subscriptions of smartphones, from 2010 to 2017, and predictions up to 2023, around the world.

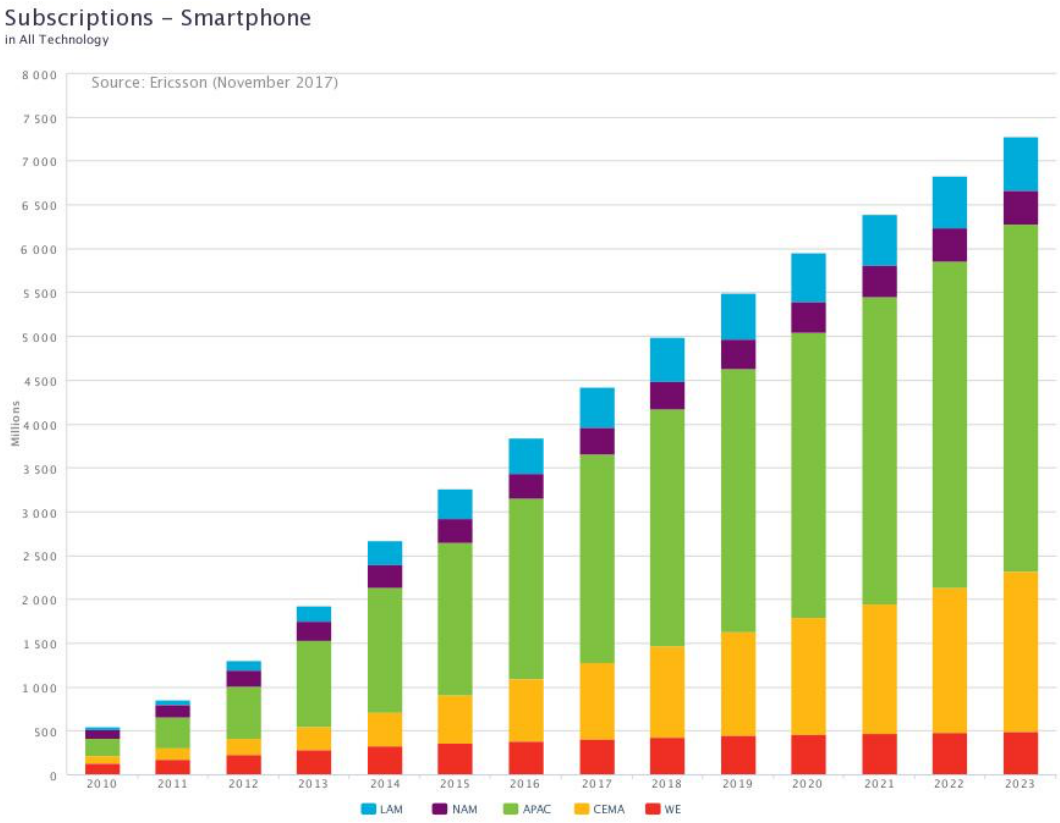

Figure 1: Data traffic per active smartphone (GB per month), from the Ericsson Mobility Report.

[1] https://www.ericsson.com/en/mobility-report

These location-enabled devices store spatio-temporal traces of their travel companions, in the form of GPS points, GMS cells, Wi-Fi or Bluetooth connections. In addition, social media activities, like check-ins and geo-tagged photos, record the position of moving users. In this light, tremendous efforts have been spent in the research literature on approaches that may take advantage of such data (YAN *et al.*, 2008; FILETO *et al.*, 2013; RENSO *et al.*, 2013). Analyzing large amount of human mobility data can provide interesting insights on the daily activities of people (RENSO *et al*, 2013) or the urban mobility patterns for example in the context of traffic management (LIU *el al.*, 2012).

Mobility data has a common representation in the form of movement trajectories: *the spatio-temporal evolution of a moving object*. Most of the literature focuses on the pure spatio-temporal aspects of these (*raw*) trajectories: *time* and *space* (PARENT *et al.*, 2013; RENSO *et al*, 2013). (Figure 2)

Figure 2: A spatio-temporal trajectory.

However, we have recently witnessed a growing research area where the representation of movement is by the so-called *semantic trajectories* (FILETO *et al.*, 2014; BOGORNY *et al.*, 2014; PARENT *et al.*, 2013). A semantic trajectory is a representation of a trajectory as a sequence of meaningful segments. A common case is where each segment is delimited by two stops (where the object has no movement) and a move (the actual change of position between the two stops). Labels can be associated to segments as annotations to represent their "semantics". Figure 3 shows an example of semantic trajectories with stop and moves and the associated semantic label.

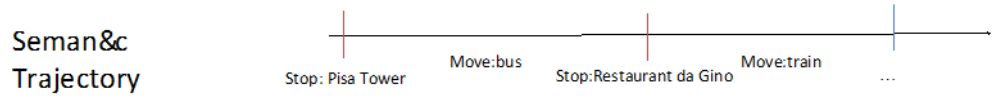

Figure 3: A semantic trajectory example.

The process of adding semantics to the trajectories is known as the *trajectory enrichment* process. Such enrichment may link, for instance, information about the points of interest visited by a tourist, the means of

transportation employed, or the goal of the movement pattern (BOGORNY *et al.*, 2014; PARENT *et al.*, 2013). However, the trajectory enrichment can be done - automatically or semi automatically - only when enough contextual and meaningful data is available to be properly linked to the movement data.

Nowadays, trajectories are enriched using only one dimension at a time (e.g. the stops, moves, the transportation means) and the enriching data is mainly a one-dimensional annotation or label. When we look at the semantic dimensions of the segments of each trajectory, we see that simple labels do not properly represent the richness and complexity of the data. Depending on the final application at hand, many different aspects can be properly exploited to enrich the spatio-temporal data. For example, as shown in Figure 4, the history of the leaning tower of Pisa taken from open source data (e.g. Wikipedia) can *enrich* the first stop in the trajectory. Similarly, other social media sources can add more semantic aspects, like the reviews and opinion of other users about the tower or the photos shared in social platforms like Flickr.

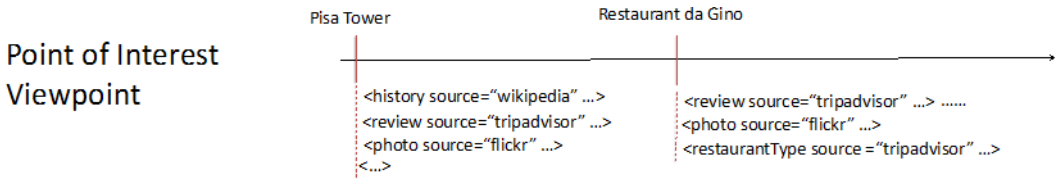

Figure 4: A viewpoint of the initial trajectory along the Point of Interest aspect.

Thanks to the emergence of the Linked Data initiative, an unprecedented global space is also growing fast: the Web of Data. The Linked Data principles have promoted in the last years the creation and publication of previously isolated datasets as interlinked and reusable data graphs. Figure 5 shows the current Linking Open Data cloud diagram, with 1,163 datasets from 9 different domains.

The availability of Linked datasets combined with the growing amount of mobility data opens up a variety of opportunities for exploiting Linked Data to enrich trajectories. The resulting enriched data will enable a wide range of improvements in several applications, from traffic management (LIU *el al.*, 2012) to animal behavior (BOGORNY *et al.*, 2014) to tourist recommendation (BRILHANTE *et al.*, 2003; QUERCIA *el al.*, 2014) just to name a few.

The challenge we want to cope with is to understand if and how the available Linked Data can be used to semantically enrich trajectories and which kinds of analysis this enrichment enables. In the next section, we present the hypothesis and the research questions analyzed in this thesis.

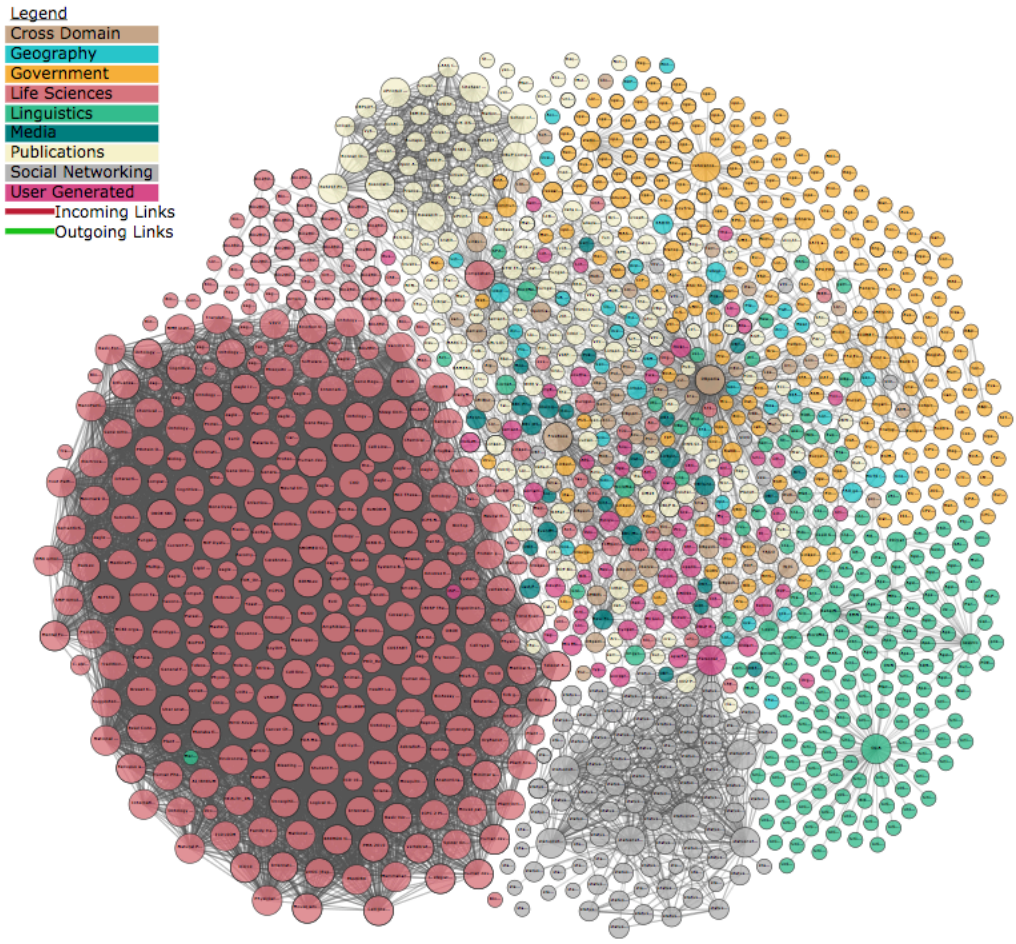

Figure 5: The Linking Open Data cloud diagram in August 2017[2].

1.2. Hypotheses and Research Questions

Considering the scenario described in the previous section, the main hypotheses that guided our research are the following:

1. The Linked Data Initiative can guide the whole trajectory enrichment process to generate semantic trajectories, from the representation of

[2] http://lod-cloud.net/

semantic trajectories using ontologies to the use of external datasets provided by the Linked Data Initiative;

2. The similarity between Linked Data entities, such as the entities that represent places like museums, is better captured if we extract and rank their relevant features, also obtained as Linked Data;

3. The computation of the similarity between trajectories can be improved, if we take advantage of the relevant categories of their Points-of-Interest (POIs), available as Linked Data.

The first hypothesis inspired us to bring up the following research question:

RQ1. *How can we use the Linked Data Principles and technologies to guide the semantic trajectory enrichment process?*

In this thesis, we investigate this problem by proposing a conceptual framework aiming at guiding the whole trajectory enrichment process to generate semantic trajectories. The framework takes advantage of the Linked Data principles in two aspects. First of all, representing trajectories according to the Linked Data principles offers a strategy to incorporate trajectories into this global data space in a way they can be easily shared and reused, which is the main motivation of the Linked Data initiative. As a second important aspect, the Web of Data might be used as the main source of contextual information to enrich movement data.

The output of the enrichment step is stored in the *semantic trajectory repository*, which enables a number of interesting analyses. For instance, one may need to find tourist routes that are somehow similar by comparing the points of interests (POIs) of their trajectories, with the assumption that trajectories with more POIs in common are more similar. Inspired by this context, we brought up the second research question:

RQ2. *How can we discoverer similar Linked Data entities by considering their relevant features?*

This research question led us to propose SELECTOR, a two-module framework that takes as input Linked Data entities, ranks the lists of entity

features according to their relevance for describing the entities, compares the ranked lists using rank correlation metrics, and outputs the entities similarity.

It is important to notice that the terms *semantic relatedness* and *similarity* are used interchangeably in the literature. However, we adopted the same perspective of (BUDANITSKY et al., 2006), that considers similarity as a special case of the broader concept semantic relatedness. They mention, as an example, that the terms "cars" and "gasoline" would seem to be more closely related than "cars" and "bikes", but the latter pair are certainly more similar, in the sense that both are transportation means.

This field has an enormous potential to find behavioral patterns and similarities, for instance, between travelers, which are especially useful for recommender systems. In this case, the characteristics (or features) of the museums or other POIs visited by a traveler can help understand his behavior. However, this approach may also be applied to different domains, such as the academic field, when comparing different conferences, and comparing different datasets available as Linked Data.

By comparing trajectory POIs we open up the opportunity to compare trajectories as a whole, but considering the features of their segments that represent POIs. Inspired by this scenario, we formulated the last research question:

RQ3. *How can we compare semantic trajectories by considering their semantic dimension, extracted from Linked Data?*

Our results from RQ2 gave us some insights regarding the computation of entity similarity, an important task when considering the analysis and understanding of a group of trajectories.

In this part of the thesis, we discuss different approaches to represent semantic trajectories: (i) the representation of the trajectories as set of POIs and (ii) the representation of trajectories as frequency vectors of their POIs categories. We then compared two different trajectories considering those two types of representations, considering that the trajectories to be compared may be geographically far from each other, for instance, in two different cities. Finally, we analyze a group of trajectories, considering their POI categories, and

comparing a group of trajectories with the places recommended by known trip guides.

1.3. Thesis Contribution

The contributions of this thesis are the following.

- We propose a conceptual framework for the semantic enrichment of trajectory data based on Linked Open Data. We clarify the whole enrichment process by introducing a running example in the tourism domain and instantiating the two steps of the framework, the *segmentation* step and the *enrichment* step. Comparing with other works dealing with the semantic enrichment process, our approach faces all steps involved in the trajectory enrichment process and analysis, taking advantage of the Linked Data principles, while some works focuses on the formalization task using ontologies or on the enrichment part, in isolation. The results were published in (RUBACK *et al*, 2016).
- We present a novel approach to compute Linked Data entity similarity, by proposing SELECTOR, a two-module framework that takes as input two Linked Data entities, ranks the lists of entity features according to their relevance for describing the entities, compares the ranked lists using rank correlation metrics, and outputs the similarity between the entities. Comparing with other approaches that compute similarity between Linked Data entities, our main contribution is the usage of the relevant features as ranked lists available as Linked Data in the comparison. The results were published in (RUBACK *et al*, 2017) and (RUBACK *et al*, 2017a).
- We propose a new approach to compare trajectories considering their segments that represent POIs. We analyze characteristics of: (i) a single trajectory; (ii) a pair of trajectories; and (iii) a group of trajectories. We use trajectories collected in Florence. Contrasting with other approaches that compare trajectories, our main contribution is the usage of the semantic dimension (taken from the Linked Open Data) in the comparison, while most of the literature focus on the spatial and temporal dimensions. The results will be submitted for publication.

1.4. Thesis Organization

The remainder of this thesis is structure as follows. In Chapter 2, we summarize the basic concepts and the work related to this thesis. More specifically, we first present related work on the field of semantic trajectories and semantic enrichment. Then, we present work related to the similarity between Linked Data entities. We then go through work on rank correlation metrics. Finally, we present work related to the computation of the similarity between a pair of trajectories. In Chapter 3, we introduce the first contribution of this thesis, a conceptual framework for the semantic enrichment of movement data based on Linked Open Data, and we describe a running example of the trajectory enrichment process of a tourist trajectory in Florence. In Chapter 4, we describe the second contribution of this thesis, SELECTOR, a two-module framework to compare Linked Data entities, and cover experiments on: comparing (i) trajectory stops – more specifically museums - available as Linked Data; (ii) comparing datasets from a Linked Data repository; and (iii) comparing computer science conferences. In Chapter 5 we discuss the representation of semantic trajectories, and introduce the third research contribution of this thesis describing approaches to compare semantic trajectories and to analyze a group of trajectories, all considering the trajectory POIs. Finally, in Chapter 6, we draw the conclusions and future directions of this thesis.

2
Basic Concepts and Related Work

This chapter summarizes work related to this thesis. Section 2.1 presents work related to the representation of trajectories, together with some basic foundations in the field. Section 2.2 discusses semantic trajectories enrichment and ontologies Section 2.3 covers works on the similarity of Linked Data Entities. Section 2.4 presents work on concepts concerning rank correlation metrics. Finally, Section 2.5 addresses trajectories similarity.

2.1.
Semantic Trajectories

In the last years, massive amounts of tracking data have been generated by GPS and other positioning devices, for the benefit of novel applications (PARENT *et al.*, 2013) that address human mobility (RENSO *et a.l*, 2013), traffic management (LIU *el al.*, 2012), and animal migration patterns (BOGORNY *et al.*, 2014), among others.

In this context, a *moving object O* represents the device (for instance, a GPS or a smartphone) that moves in geographical space over some period of time and records the temporal sequence of its spatio-temporal positions, known as the *raw data*. However, most of the applications do not keep exhaustive 24/7 records of movement; instead, they keep the segments of the movement track that are of interest for the application, called *trajectories* (PARENT *et al.*, 2013). Figure 6 shows two trajectories extracted from a movement track, both identified two positions of the movement track: *Begin* (the first position) and *End* (the last position) of the trajectory.

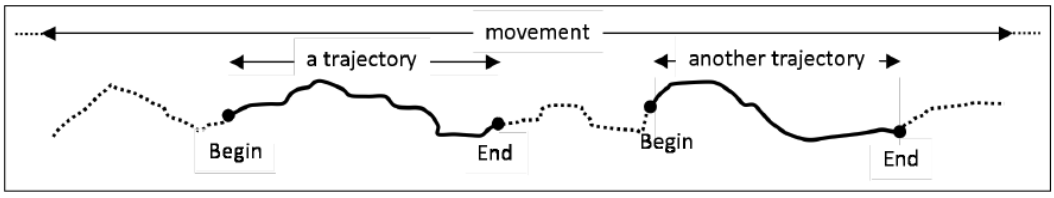

Figure 6: Trajectories extracted from a movement track (PARENT *et al.*, 2013).

A *raw trajectory* consists of the trajectory extracted from a raw movement track. The definition of raw trajectory we use in this thesis has been adapted from (RENSO *et al*, 2013).

Definition 1 (Raw Trajectory) Let O be a moving object. A *sample point* of O is a triple $(x_k,y_k,t_k) \in \Re^3$, where (x_k,y_k) are the *geographical coordinates* of the point and t_k is the *timestamp* of the point. A *raw trajectory* of O is a sequence $T_O = <p_1,...,p_n>$ of sample points of O, where n is the number of sample points of T_O. A sub-trajectory of $T_O = <p_1,...,p_n>$ is a subsequence of $<p_1,...,p_n>$.

Although the raw trajectory is built from the sample of points collected from tracking devices, the real movement is approximated using interpolation functions to fill the temporal gaps (the periods of time in which the movement of the object is missing) between two consecutive points (RENSO *et al*, 2013).

Raw trajectories can be useful for some applications that require only the movement track of the objects, but most applications require additional data from the application context (for instance, the city information, traffic conditions, weather data, among others). The process of adding knowledge to the raw trajectories from external repositories is known as *semantic enrichment* process (PARENT *et al.*, 2013).

A raw trajectory is *annotated* when it or any of its sub-trajectories is complemented with additional data, called an *annotation*. An example of annotation is the *transportation means* used by a person – e.g., bus, taxi, by foot – that may perhaps be inferred from the velocity and acceleration, and combined with external data about the transport networks (GUC *et al.*, 2008).

A *segmented trajectory* is a partition of the points of the trajectory into homogeneous segments, where a given property holds. For example, according to the stop-and-move model (SPACCAPIETRA *et al.*, 2008), a raw trajectory can be split into segments of two kinds: *stop* where the speed of the object is lower than a certain threshold and *move* where the speed is greater than such a threshold. Other segmentation criteria are sometimes used, such as the change of direction (ROCHA *et al.*, 2010).

A *semantic trajectory* is a segmented trajectory annotated with contextual information. The most common form of semantic trajectory is the *stop-and-move*s representation, adapted from (RENSO *et al*, 2013):

Definition 2 (Semantic Trajectory) A *semantic trajectory* for a segmented trajectory is a pair $\tau = <o,(<g_1,c_1>,...,<g_n,c_n>)>$ such that, for each $i=1,...,n$, $<g_i,c_i>$ is a pair indicating that segment g_i is enriched with *contextual information* c_i.

We leave open what exactly is the type of the contextual information: it is any kind of information that can be related to the object and its movement data and that can be taken from a contextual information repository in the style of (PARENT *et al.*, 2013). Some examples of contextual information are the places visited by the moving object, the events that the object took place and the meteorological conditions of the segment.

In this thesis, we extensively use the concepts raw trajectory and semantic trajectory, and the *stop-and-moves* segmentation criteria.

2.2. Semantic trajectory enrichment and ontologies

This section covers related work regarding the representation of semantic trajectories and the use of ontologies in this context, and is mainly related to the conceptual framework we propose in Section 3.

The formalization of the trajectory semantic enrichment process has been firstly outlined in (FILETO *et al.*, 2013), where Baquara has been proposed as a general all-inclusive ontology representing both the trajectory and the enriching concepts. This pioneering approach set the way to the use of ontologies to support the enrichment process with Linked Open Data. However, it has some limitations. For example, the proposed Baquara ontology, in trying to be general enough to cover a large range of applications, let to a "monolithic" approach, which is complex and difficult to personalize to different needs. Furthermore, the Linked Open Data sources that enrich the trajectory are predefined.

One of the first approaches that tried to conceptualize movement data as a trajectory ontology was proposed by (YAN *et al.*, 2008). The conceptual

framework was aimed at combining in a unique top-level ontology the different aspects of the movement embedded into three main ontologies representing application-related information, the trajectory as a sequence of stops and moves and the geography information. However, it offers a conceptual and top-level vision of trajectories without explicitly dealing with the problem of the enrichment process or adopting the Linked Open Data formalism.

Another approach trying to represent semantic trajectories based on ontologies was proposed by (ALVES *et al*., 2009) and was built on Ontology Engineering techniques to connect Generic Places Ontologies with POI instances. Different from our framework presented in Chapter 3, they focus on the enrichment of POIs using the proper Ontologies terms, while our approach faces all steps involved in the trajectory enrichment process and analysis.

A few years later, (RENSO *et al*., 2013) made a step towards employing the Athena ontology (structured into application ontology giving the application domain analysis concepts and core ontology representing the segmented trajectories) into a reasoning process based on OWL to support meaningful pattern interpretations of human behavior, combining inductive reasoning and deductive reasoning. However, their work still does not deal with the modeling of the enrichment step and it can be seen as complementary to our framework, presented in Chapter 3, since it focus on the analysis part, instead of the enrichment part.

The approach proposed by Yingjie et al. (HU *et al.*, 2003) introduced a geo-ontology design pattern for semantic trajectories that is very similar to our Segmented Trajectory Ontology. A formal encoding of the classes together with their properties is obtained by using OWL. Authors also define a number of interfaces to integrate related geographic information, domain knowledge and device data. We go a step beyond the approach proposed in Yingjie et al. (HU *et al.*, 2003) since we also face the issues of how to implement the enrichment step by using Linked Data Mashups.

2.3. Linked Data and entity similarity

The Linked Data principles (HEATH *et al.*, 2011; BIZER *et al.*, 2009) promote the creation and publication of previously isolated databases as interlinked, reusable data graphs using known Web standards. The principles recommend the use of ontologies and RDF (Resource Description Framework) to publish databases on the Web, thereby minimizing the problem of schema alignment, a difficult and error-prone task. Intuitively, following the Linked Data Principles facilitates the task of linking trajectories with external data sources. Due to the huge and heterogeneous amount of Linked Open Data available, it is critical to be able to properly select and integrate the relevant entities.

In Chapter 4, we focus on the problem of discovering similar entities on Linked Data by ranking and comparing their features, that depends on two basic definitions (*RUBACK et al.*, 2017):

Definition 3 (Ranked Features) The *ranked features* of a Linked Data entity e is a list $F = \{(f_1, s_1), \ldots, (f_n, s_n)\}$, where f_i is a feature of e and s_i is the score of relevance of f_i, for $j \in [1, n]$.

Intuitively, feature f_i is more relevant to describe e than feature f_{i+1} , for $1 \leq i \leq n$.

Definition 4 (Entities Similarity). Given two Linked Data entities e_i and e_j and their ranked features F_i, and F_j, the *similarity* between e_i and e_j is the distance between their respective lists F_i and F_j, according to a rank correlation metric m.

Note that entities e_i and e_j may have a different number of relevant features, i.e., their lists F_i and F_j may have different sizes. Furthermore, F_i and F_j may or may not have features in common.

Semantic Relatedness of Linked Data Entities

Several approaches focus on measuring the similarity between Linked Data entities. Such approaches can be further classified as social network theory approaches (ZHANG *et al.*, 2013), entity disambiguation (HULPUŞ *et al.*, 2010), ontology-based approaches (HAJMOOSAEI *et al.*, 2016; GRIESER *et al.*, 2011),

and Wikipedia structure-based approaches, among others. One can also combine these approaches to create other hybrid approaches to take advantage of several strategies. In the context of Linked Data, such associations are found in the relations between the Linked Data entities.

(PASSANT *et al.*, 2010) measured semantic distances on Linked Data in order to provide a new kind of self-explanatory recommendations, combining Linked Data and Artificial Intelligence principles. They considered only the links that can exist between Linked Data resources, using both direct and indirect links to provide Linked Data resource recommendations.

(LEME *et al.*, 2013) proposed a technique based on probabilistic classifiers to the dataset recommendation problem. They ranked the most relevant datasets to recommend based on the probability that links between datasets can be found.

In the last years, some measures have been proposed to semantically relate Linked Data entities. The two more relevant measures for this thesis are: *Wikipedia Link-based Measure* (WLM) and the *Semantic Connectivity Score* (SCS).

The *Wikipedia Link-based Measure* (WLM) is a measure proposed in 2003 by Milne and Witten (WITTEN *et al.*, 2008). Formally, the relatedness between two Linked Data entities (more specifically, Wikipedia articles) of interest, *a* and *b*, is defined as:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where *A* and *B* are the sets of all articles that link to *a* and *b*, respectively, and *W* is the entire Wikipedia.

The *Semantic Connectivity Score* (SCS), proposed in (NUNES *et al.*, 2013), between two entities, *a* and *b*, is based on Katz score (KATZ, L., 1953) and is defined as:

$$SCS(a,b) = \sum_{l=1}^{\tau} \beta^{l} \cdot |paths_{(a,b)}^{<l>}|$$

where $|paths_{(a,b)}^{<l>}|$ is the number of paths between entities *a* and *b* of length *l*, τ is the maximum path length considered, and β is a positive damping factor ranging from 0 to 1, responsible for exponentially penalizing longer paths.

Other similarity measures

Two widely used measures to compare entities are the Jaccard similarity measure (based on the Jaccard distance) and the Cosine similarity measure (based on the cosine distance).

In this thesis, we use all these measures to compare Linked Data entities, presented in Sections 4.2.1 and 4.3.

2.4. Rank Correlation metrics

This section presents related work on ranking Linked Data entities that exploit the semantic aspects of Linked Data datasets, and presents two rank correlation metrics, the *Average Overlap* (AO) and the *Rank-biased Overlap* (RBO).

(HOGAN *et al*., 2006) proposed ReConRank, a ranking method that adapts the well-known PageRank HITS algorithms to Semantic Web data.

(MIRIZZI et al., 2010) exploited semantic tagging on Linked Open Data to rank resources. Their methodology combined the graph-based nature of RDF, semantic relations in the graph, and search engine results to rank Linked Data resources.

(ROA-VALVERDE *et al.*, 2014) formalized the problem of ranking information in the Web of Data, unified the core concepts that characterize ranking algorithms, and compared different approaches to ranking Linked Data.

(MIRIZZI et al., 2010) is the most relevant work to our approach, proposed in Chapter 4. The main difference is that they did not exploit rank correlation metrics to rank resources; instead, they took advantage of results coming from search engines.

Correlating ranked lists is a common problem in several areas, such as graph analysis and information retrieval. Webber et al. (WEBBER *et al.*, 2010) categorize them according to two main characteristics: the *conjointness* (two conjoint lists consist of the same items) and the *weightedness* (a list is *weighted*

when the items have different relevancies and a list is *top-weighted* when the top of the list is more important than the tail).

For conjoint lists, some widely used rank correlation coefficients are Kendall's and Spearman's. For non-conjoint lists with items of different weights (ranks), there are some similarity measures that can be used, such as *Jaccard* (IOFFE, S., 2010), *Cosine Similarity*, and *Average Overlap* (FAGIN *et al.*, 2003; FAGIN *et al.*, 2004), and the *Rank-biased Overlap* (RBO). When dealing with non-conjoint lists, it is common to start from set intersection, considering the size of the intersection or the overlap between two rankings.

The *Average Overlap* (AP) is based on set intersection, but considers the overlap at increasing depths when comparing two rankings, and is defined as (WEBBER *et al*, 2010):

$$AO(S,T,k) = \frac{1}{k}\sum_{d=1}^{k} A_d$$

where S and T are two possibly infinite lists, k is the evaluation depth, and A_d is their agreement at depth d, defined as

$$A_d = \frac{1}{d}|S_d \cap T_d|$$

where S_d (or T_d) is the prefix of S (or T) up to depth d. For each d $\in$ {1,…,k}, it calculates the overlap at d, and then averages those overlaps to derive the similarity measure.

Table **1** shows a sample calculation of AO. Up to depth 2, the AO score is 0, since the common items in S and T first appear at depth 3. If we consider, for instance, depth 7, AO gives 0.312 as the similarity score between S and T. However, the higher AO score is at depth 6, 0.317.

Table 1: Average Overlap (AO) of two lists.

d	***S_d***	***T_d***	***A_d***	***AO(S,T,d)***
1	<a>	<z>	0.000	0.000
2	<ab>	<zc>	0.000	0.000
3	<abc>	<zca>	0.667	0.222

d	S_d	T_d	A_d	*AO(S,T,d)*
4	<abcd>	<zcav>	0.500	0.292
5	<abcde>	<zcavw>	0.400	0.313
6	<abcdef>	<zcavwx>	0.333	0.317
7	<abcdefg>	<zcavwxy>	0.286	0.312
k	<abcdefg…>	<zcavwxy…>	…	…

Webber et al. (WEBBER *et al*, 2010) proposed to extend this idea to incomplete ranks (i.e. they do not cover all elements in the domain) by adding a parameter that determines the importance of the weighting of the top ranks. They defined the *Rank-biased overlap* (RBO) as follows:

$$\mathrm{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d$$

The RBO measure handles non-conjoint lists and weights higher ranks more heavily than lower ranks (their top-k item is more relevant then the top-k+1, and so on). In addition, RBO has a parameter p, which ranges from 0 to 1, and determines the strength of the weighting of the top ranks, i.e., the smaller p, the more top-weighted the metric is. The choice of parameter k in this work is inspired by the experiments performed in [11], which consider k ranging from 0.9 to 0.998. If $p = 0$, only the top-ranked item is considered and the score is either 0 or 1 [11]. Rank-biased overlap ranges from 0 to 1, where 0 means disjoint, and 1 means identical.

Table 2 shows the same sample calculation as AO, but it includes the weight for each depth k. Up to depth 2, the RBO score is also 0, since the common items in S and T first appear at depth 3. If we consider, for instance, depth 5, RBO gives as similarity score 0.116, if $p = 0.98$, and 0.027, if $p = 0.9$.

Table 2: Rank-biased overlap (RBO) of two lists.

d	$S_{:d}$	$T_{:d}$	$A_{S,T,d}$	*AO (S, T, d)*	weight
1	<a>	<z>	0.000	0.000	p^0
2	<ab>	<zc>	0.000	0.000	p^1
3	<abc>	<zca>	0.667	0.222	p^2

d	$S_{:d}$	$T_{:d}$	$A_{S,T,d}$	*AO (S, T, d)*	weight
4	<abcd>	<zcav>	0.500	0.292	p^3
5	<abcde>	<zcavw>	0.400	0.313	p^4
6	<abcdef>	<zcavwx>	0.333	0.317	p^5
7	<abcdefg>	<zcavwxy>	0.286	0.312	p^6
k	<abcdefg…>	<zcavwxy…>	…	…	P^{k-1}

Although the ranked lists we use in the experiments are not incomplete, we use RBO as the main rank correlation metric for non-conjoint lists, since it allows imposing a stronger penalty on differences at the top of the ranking than on differences further down, the lists.

Normalized Discounted Cumulative Gain (NDCG)

Traditionally used in search engine results to evaluate ranks, NDCG (*Normalized Discounted Cumulative Gain*) (YILMAZ *et al.*, 2008) emphasis retrieving highly relevant documents.

Intuitively, the idea behind NDCG is that a recommender system returns some items and we would like to evaluate how good the list is. The items (or documents) have relevance scores (or *gains*) that are added up (*cumulative* gain). Since we prefer to find the most relevant items at the top of the list, before summing the gains, they are divided by a growing number (usually a logarithm of the item position) – that is, *discounting*. Since DCG are not directly comparable between users, we *normalize* it.

Let g_1, g_2, ..., g_Z be the gain values associated with the Z documents retrieved by a system in response to a query. Let DCG_I denote the DCG value for an ideal ranked list for the query. NDCG is defined as follows (YILMAZ *et al.*, 2008):

$$nDCG = \frac{DCG}{DCG_I}, \text{where } DCG = \sum_{i=1}^{Z} g_i / \lg(i+1)$$

NDCG returns a non-negative score ranging from 0 to 1 for the first *k* items on the list (NDCG@*k*), since a recommender system is mainly interested in a relatively few items to be considered as relevant and be shown to the users.

2.5. Trajectory similarity

In the literature, there are different similarity measures to compare trajectories, most of them dealing with the geo-spatial aspect of the trajectories. This section covers related work regarding the analysis and comparison of semantic trajectories, mainly related to the discussion presented in Chapter 5.

(VLACHOS *et al.*, 2002) formalized similarity functions based on the Longest Common Subsequence (LCSS) and compared them with Euclidean and Time Warping distance functions. They presented experimental studies to validate the accuracy and efficiency of their approach.

(SANKARARAMAN *et al.*, 2013) also presented new measures of trajectory similarity, capturing the advantages of both dynamic time warping (DTW) and sequence alignment-based approaches. The authors ran an extensive experimental study on three real datasets: (i) the Geolife project by Microsoft Research Asia consisting of 17,621 trajectories of 182 users in China; (ii) 145 trajectories of school buses in Athens, Greece; and (iii) 38 trajectories from road cycling exercises captured by a fitness GPS device.

(WANG *et al.*, 2013) also conducted a comparative experimental study on the effectiveness of widely used trajectory similarity measures based on a real taxi trajectories dataset: Euclidean Distance, Dynamic Time Warping (DTW), Edit Distance, Longest Common Subsequence (LCSS).

(BUCHIN *et al*, 2014) developed a simple and context-aware similarity measure for movement trajectories based on well-known similarity measures, such as Hausdorff and Fréchet and applied their approach to movement data of hurricanes and albatross.

(RANACHER *et. al*, 2014) reviews qualitative and quantitative methods to compare movement, considering trajectory attributes as speed, spatial path, temporal duration, among others.

(TOOHEY *et al., 2015*) introduces and compares, using real trajectory data, four of the most common measures of trajectory similarity: longest common subsequence (LCSS), *Fréchet* distance, dynamic time warping (DTW), and edit distance.

Recently, the *semantic* aspects of trajectories have also been taken into consideration to compare trajectories, which represents a step ahead in the field. (FURTADO *et al.*, 2016) proposed a novel similarity measure, called MSM, to compare multidimensional sequences, considering the three dimensions individually: time, space and semantic, including the possibility to set different weights for each dimension, according to the application domain. Although their work represented a step ahead by considering the semantic aspects of the trajectories, this part of their work is still preliminary, and they performed experiments by applying a set of transformations over a seed trajectory.

In the approach, we present in Chapter 5, we are mainly concerned with the semantic aspect of the trajectories, and we argue that our approach may be combined with the geo-spatial approaches previously mentioned.

3
A conceptual framework for the trajectory enrichment process

This chapter addresses the research question **RQ1**. *How can we use the Linked Data Principles and technologies to guide the semantic trajectory enrichment process?* We present a conceptual framework for the semantic enrichment of movement data, which benefits from the emerging Web of Data (or Linked Open Data) both as a unifying formalism and as the source of contextual data, which can be greatly useful for trajectories enrichment. Section 3.1 introduces a running example in the tourism domain. Section 3.2 gives an overview of the conceptual framework, which is structured in two main steps: *segmentation* and *enrichment*. Section 3.3 illustrates some interesting analysis on the semantic trajectory repository. Finally, Section 3.4 presents the conclusions.

3.1.
Introduction and running example

We clarify the whole enrichment process by introducing a running example from the tourism domain. We assume the use of position-enabled devices that track and collect the movement of tourists visiting Florence in Italy with the objective to offer personalized services. Let us consider the typical trajectory of a tourist that starts at her hotel, combines sightseeing and lunch during the day, before going to the train station to depart. Figure 7 shows the first part of the trajectory until the bridge “Ponte Vecchio”. The trajectory is represented as a sequence of samples, i.e., timestamped coordinates *(x,y,t)* as collected by the GPS device. The objective of the analysis is to have a better understanding of the tourists’ behavior, such as characterizing them based on the features of the attractions visited, transportation means used to identify the visiting profiles (e.g. the tourist spending profile). It is clear that, to reach this objective, we need to build a new kind of trajectory data by augmenting the pure location data with a large amount of contextual, semantically rich data.

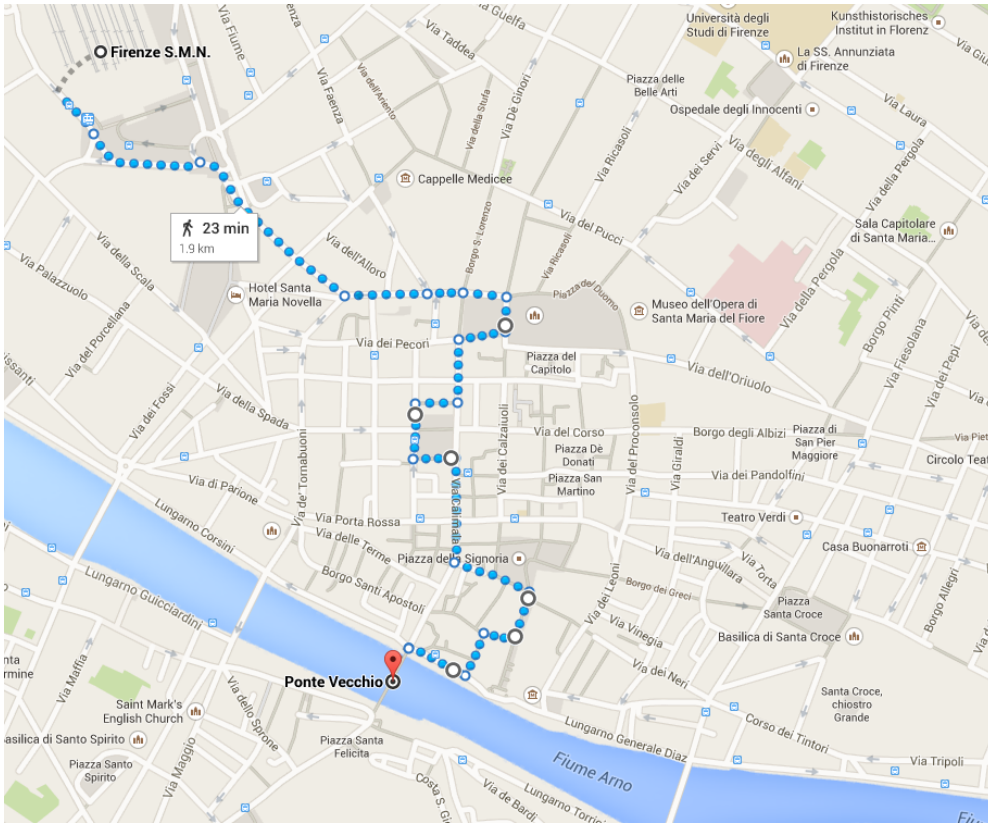

Figure 7: A raw trajectory of a tourist in Florence collected as GPS samples.

We describe in the next section the framework to build these semantically rich trajectory data by identifying the parts to be enriched and then exploiting Linked Open Data to give the actual meaning to the trajectory parts.

3.2. Overview of the conceptual framework

We first give a general overview of the proposed framework, that covers the trajectory segmentation and the Linked Open Data enrichment up to the construction of a semantic trajectory repository to be used for the analysis, as illustrated in Figure 8.

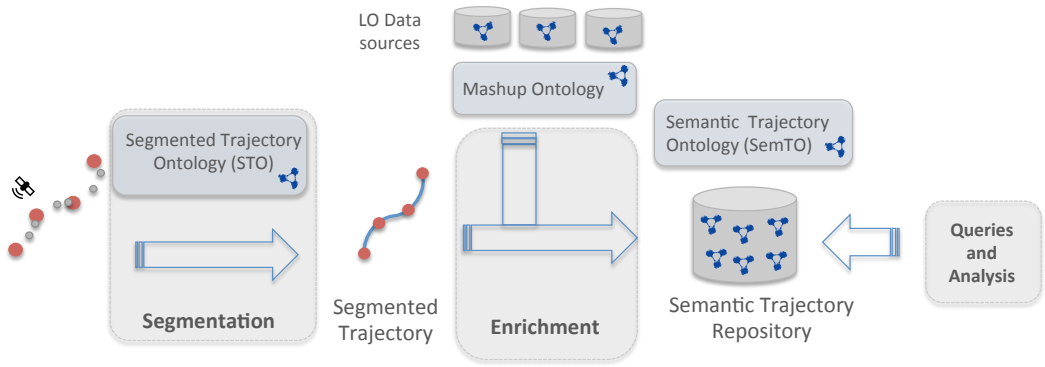

Figure 8: The trajectory enrichment process.

The semantic enrichment process takes as input a raw trajectory and a number of Linked Open Data sources and builds a semantic trajectory repository. This process is driven by the use of ontologies and it is structured into two main steps: segmentation and enrichment. The *segmentation step* partitions a raw trajectory into homogeneous segments, specifying the entities that will be enriched. This step is driven by a *Segmented Trajectory Ontology (STO)*, which identifies the different types of segmentation required by a specific application. The *enrichment step* matches the segments with the most appropriate semantic entities made available as a *linked data mashup view*, which provides cleaned and integrated data from selected linked data sources. In this step, the *mashup view ontology* specifies the concepts of the mashup view (i.e., the *conceptual model*), which is the common vocabulary for integrating data exported by the selected Linked Data sources.

The outcome of the enrichment step is the *semantic trajectory ontology (SemTO)* and an RDF repository containing semantically enriched trajectories. The semantic trajectory ontology contains all concepts and properties of the segmented trajectories and mashup ontologies, and also the definition of the semantic links between them. Research in semantic data mining (DOU *et al.*, 2015) attested that the domain knowledge formally encoded in an ontology is very helpful in all stages of the analysis process. In the proposed framework, each step is driven by the incremental use of ontologies that can be easily adapted to meet the application needs. We recall that the Linked Data principles provide the unifying formalism where the entities to be enriched (trajectories) and the

enriching contextual information (Linked Open data sources) are homogeneously modeled.

3.2.1. The Segmentation Step

The first step of our process is called *segmentation*: it takes as input a raw trajectory and some particular segmentation criterion and partitions the trajectory into segments, which are the entities to be enriched. This step is driven by the Segmented Trajectory Ontology (STO). This ontology aims at representing the trajectory entities featuring the parts to be enriched. There are many works in the literature (HU *et al.*, 2013; RENSO *et al.*, 2013; YAN *et al.*, 2008) proposing trajectory ontologies that can be easily adapted to the application needs by adding specializations of classes and properties.

Figure 9 shows the segmented trajectory ontology related to our running example. It has been inspired by (HU *et al.*, 2013) where the authors proposed a raw trajectory partitioned into a segmented trajectory and each segment class is identified by a begin and an end point.

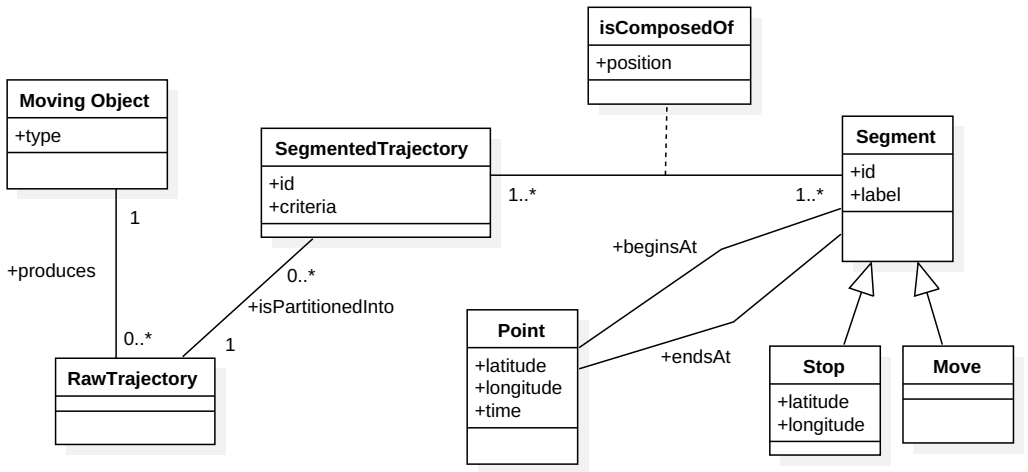

Figure 9: The Segmented Trajectory Ontology.

We notice that, here, the class *Segment* is specialized in *Stop* and *Move* to represent the two specific kinds of segments to be enriched. It is worth observing that a stop is characterized also by a spatial location, which could be the centroid of the segment. Naturally the specific ontology can be tailored to the application

needs and other specializations are possible, such as the transport mode segmentation (ZHENG *et al.*, 2013) or the activity segmentation (ZHENG *et al.*, 2013a). We remark how this incremental ontology construction is aimed at representing the entities to be enriched with the Linked Open Data.

Going back to our running example, we segmented the tourist raw trajectory following the "stop-and-move" model previously mentioned. The resulting segmentation is shown in Figure 10, where the tourist trajectory has been segmented into a "begin", seven "stops", eight "moves" and an "end".

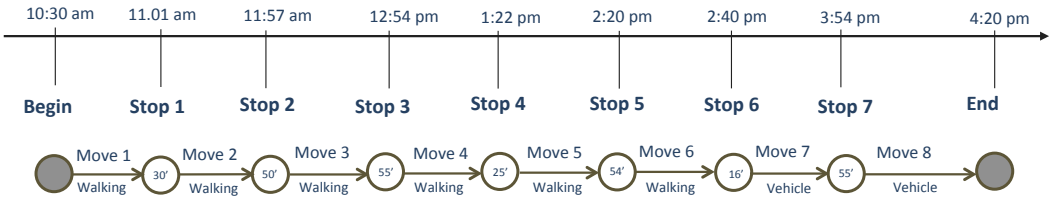

Figure 10: The running example segmented trajectory.

During a move segment, the tourist uses different transportation means: first she walks in the city center (from the begin to stop 6), then she uses a vehicle from Ponte Vecchio to Piazzale Michelangelo (e.g. a taxi) and another vehicle to go back to the railway station (e.g. a bus). The means of transportation (Walking or Vehicle) is the label associated with the move, whereas the arrival time at the stop (e.g. 10:30 am) and its duration (e.g. 30') are the labels related to the stop.

Being able to distinguish segments into subclasses, such as stops and moves, allows us to differentiate the enrichment of these two kinds of segments, as illustrated in the next section.

3.2.2. The Enrichment Step

The second step of the process shown in Figure 8 is called *enrichment* and it matches (or enriches) trajectory segments (e.g. stops or moves) to the most appropriate semantic entities made available by the external contextual information through the Linked Data Mashup (ENDRES-NIGGEMEYER, 2013). A linked data mashup tuned to the application domain is therefore fundamental since it simplifies the step of enriching trajectories, bringing all necessary data

tailored to the actual needs (YAN *et al.*, 2008). The enrichment step involves two main tasks:

1. Creation or selection (if it already exists) of the Linked Data Mashup (LDM), which integrates data from the relevant selected linked data sources.
2. Linking trajectory segments with Linked Data Mashup entities.

3.2.2.1. Creation or Selection of the Linked Data Mashup

The creation of the Linked Data Mashup is a complex and time-consuming task. We simply reuse a mashup built according to the previous mentioned views (CASANOVA *et al.*, 2014; VIDAL *et al.*, 2015).

Aimed at enriching both the stops and the moves, we present two mashup fragments in the tourism domain in Figure 11 and Figure 12. The entities available by the mashup add semantic to the places the tourists can visit during a trip, and the transportation means used to move between these places. The mashup aggregates entities from two different data sources:

1. the DBpedia dataset (AUER *et al.*, 2007), which constitutes a major part of the semantic data on the Web; and
2. the OpeNER Linked Dataset (OLD), part of the OpeNER project (BACCIU *et al.*, 2014), which consists of a repository for the tourism domain that covers the Tuscany region.

The mashup reuses terms from widely used vocabularies:

1. DBpedia vocabulary[3]
2. FOAF (Friend of a Friend)[4]
3. RDFs (RDF schema)[5] and
4. vCard vocabulary[6].

[3] http://wiki.dbpedia.org

[4] http://xmlns.com/foaf/spec/

[5] https://www.w3.org/TR/rdf-schema/

[6] https://www.w3.org/TR/vcard-rdf/

For the accommodations domain, the mashup reuses both HOntology (CHAVES *et al.*, 2012) and Accommodation Ontology[7].

The entities that represent museums, religious buildings, artists and art works come from DBpedia and are respectively represented by the classes `dbo:Museum`, `dbo:ReligiousBuilding` (both are also points of interest), `dbo:Artist` and `dbo:Artwork`. Some attributes of these entities are provided by the mashup itself, for example, the list of categories that a museum is related to (e.g.: Modern art museum and History museum). This part of the view can intuitively model, for instance, an art (category) museum that exhibits paintings (art work) of Botticelli (artist), belonging to the High Renaissance period (movement). Also a church (religious building) can be related to the gothic period (architecture style) and it can also have works of some artists.

[7] http://ontologies.sti-innsbruck.at/acco/ns.html

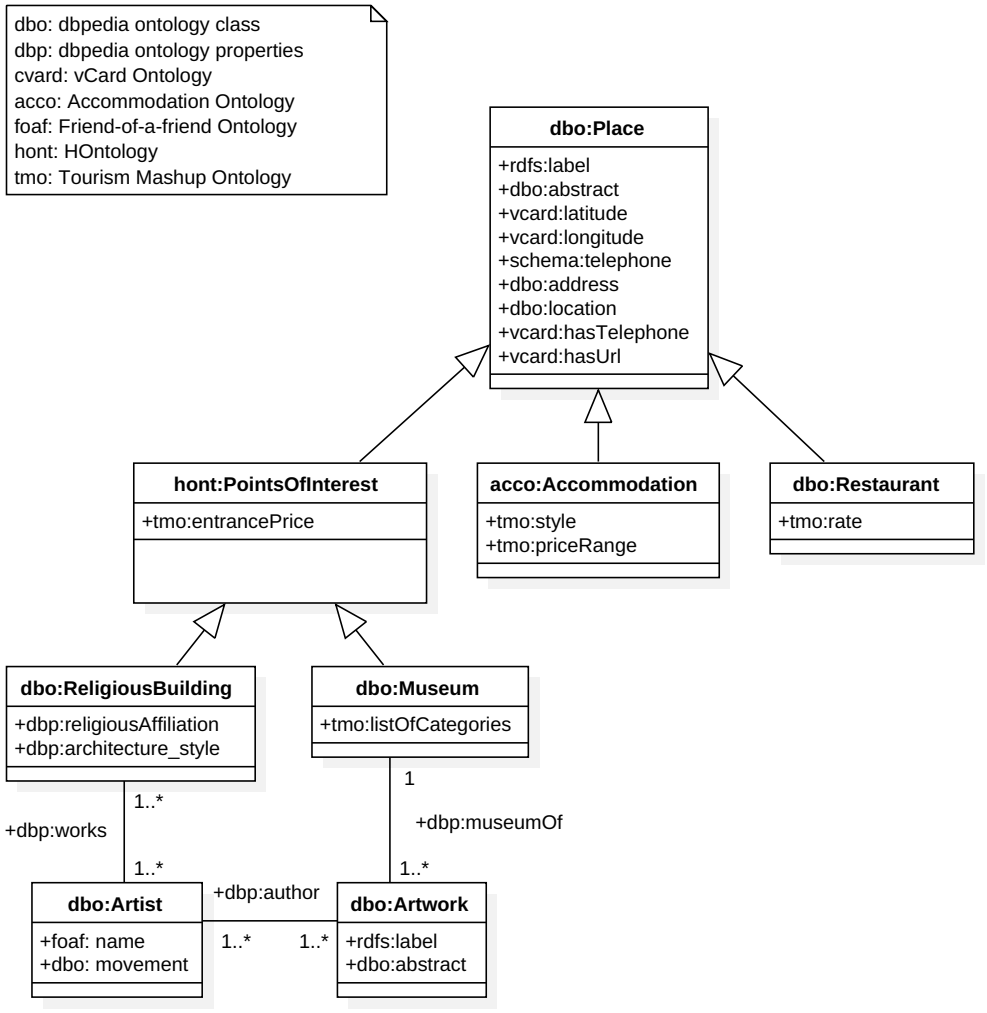

Figure 11: The tourism mashup view fragment representing places.

In turn, the entities that represent points of interests, accommodations and restaurants come from the OpeNER Linked Dataset (OLD) and are represented respectively by the classes `hont:PointsOfInterest`, `acco:Accommodation`, `dbo:Restaurant`.

The mashup view fragment, shown in Figure 12, contains the transportation features. The vocabulary extended by the mashup is the GTFS (General Transit Feed Specification)[8], a common format for public transportation schedules and associated geographic information, used by Google Maps, which also provides open transportation data.

[8] https://developers.google.com/transit/gtfs/reference

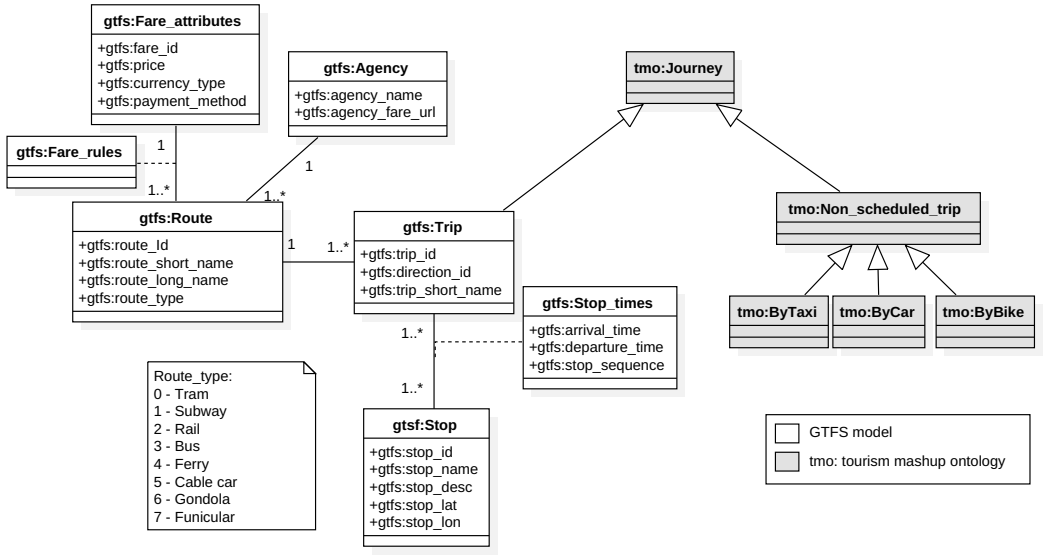

Figure 12: The Tourism Mashup view fragment about transportation.

The instances of `gtfs:Agency` represent the companies that provide the routes. The instances of `gtfs:Route` represent the entire journey made by an agency. The property `gtfs:route_type` holds the type of vehicle used on the route (tram, subway, rail, bus, ferry, cable car, or funicular). The class `gtfs:Trip` is a part of a route related to the direction (ex.: from the airport to the train station or vice-versa). Each trip is composed by `gtfs:Stop` instances, that have their respective latitude and longitude.

The white boxes are part of the GTFS data model and comprise the scheduled transport part of the mashup. Besides the scheduled transportation, travelers can also move by using other types of transportation. Classes `tmo:ByTaxi`, `tmo:ByCar` and `tmo:ByBike` respectively represent information about taxis companies, car and bicycle rentals/sharing, somehow available as Linked Open Data. In the case of bicycles, for example, one can use data available by a bike sharing system like Bicincittà, widely used in Italy. These two mashup views (Figure 11 and Figure 12) are the contextual sources used for the enrichment of the trajectory segments, presented in what follows.

3.2.2.2. Linking Trajectory Segments and Linked Data Mashups

The linking between trajectory segments and Linked Data mashups is specified as a view (VIDAL *et al.*, 2015) and it is actually performed as a matching between

entities to be enriched (segments expressed by the Segmented Trajectory Ontology) and the entities that provide the enrichment (which are instances of the classes of the Mashup Ontology). A typical example is the matching of stops with Points-Of-Interest (POIs), where the match predicate might be based on the distance between the stop and the POI (PELEKIS *et al.*, 2014; ZHENG *et al.*, 2010). In our example, we use the places mashup fragment to match the stops (Figure 11) and the transportation mashup fragment to enrich the moves (Figure 12).

The execution of this matching process produces one or more links between segments of a trajectory (stops and moves in our case) and mashup entities. In fact, some LOD tools help automate the matching process, such as Silk (VOLZ *et al.*, 2009) and Limes (NGOMO *et al.*, 2011).

We now discuss in detail the stop enrichment process for our example. First, as shown in Figure 13, we introduce the property `foaf:based_near`, which is part of FOAF and relates two "spatial things" being close to each other. This property is part of the Semantic Trajectory Ontology (SemTO) that reuses it from FOAF, following the Linked Data principles. Linking (i.e., enriching) instances of `sto:Stop` with instances of `dbo:Place` using this property we are stating that the stop is near a place.

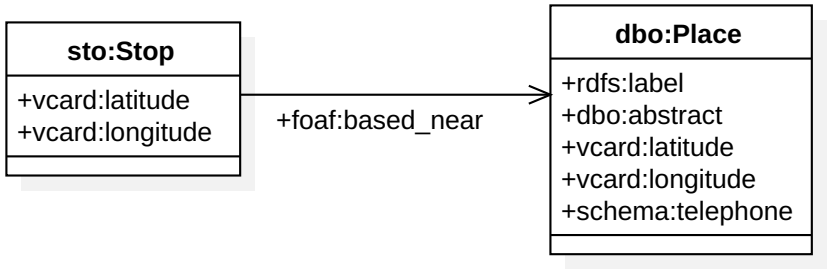

Figure 13: The Stop Enrichment.

The following triples (described in turtle notation) represent two instances for stop and place entities: (a) Stop 1 is an instance of Stop of the Segmented Trajectory Ontology (Figure 3); and (b) the church Basilica di Santa Maria Novella is a mashup instance of the `dbo:ReligiousBuilding` class (Figure 11).

```
1. sto-resource:stop1 rdf:type sto:Stop;
2.      vcard:lat "43.774836";
3.      vcard:long    "11.249375";
```

```
4.      foaf:based_near dbr:Basilica_of_Santa_Maria_Novella.
5.  dbr:Basilica_of_Santa_Maria_Novella rdf:type dbo:ReligiousBuilding;
6.      vcard:lat "43.774601";
7.      vcard:long    "11.249300";
8.      rdfs:label "Basilica of Santa Maria Novella";
9.      dbo:abstract "Santa Maria Novella is a church in Florence, Italy,
   situated just across from the main railway station…" .
```

Lines 1 to 4 describe the `sto-resource:stop1`, Stop 1. Likewise, lines 5 to 9 correspond to the church Basilica di Santa Maria Novella (other triples relating to the church were omitted). We note that `sto-resource` is the prefix of the Semantic Trajectory Repository resources (entities) and the `dbr` is the prefix of the `http://dbpedia.org/resource/namespace` for the DBPedia entities. Continuing our example, the result of the matching process will be a set of triples of the form (*s*, `foaf:based_near`, *o*), where *s* is the subject denoting a stop of the trajectory, `foaf:based_near` is the linking predicate introduced in Figure 13 and *o* is the object denoting a POI.

Similarly, we have the mapping of the other trajectory stops:

```
sto-resource:stop2 foaf:based_near dbr:Piazza_del_Duomo,_Florence
sto-resource:stop3 foaf:based_near mo-resource:OsteriaDellOlio
sto-resource:stop4 foaf:based_near dbr:Piazza_della_Repubblica,_Florence
sto-resource:stop5 foaf:based_near dbr:Palazzo_Vecchio
sto-resource:stop6 foaf:based_near dbr:Ponte_Vecchio
sto-resource:stop7 foaf:based_near dbr:Piazzale_Michelangelo
```

Besides the stops, we can also enrich the move segments with the transportation features taken from the mashup transportation fragment as illustrated in Figure 14.

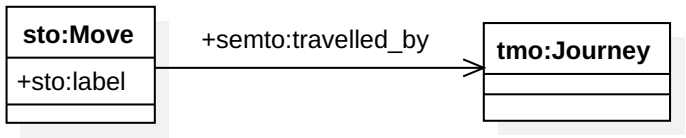

Figure 14: The Move Enrichment.

The property `semto:travelled_by` is part of the *semantic trajectory ontology (SemTO)* and links a move in the Segmented Trajectory Ontology to `tmo:Journey` (or to one of its subclasses `gtfs:Trip`, `tmo:ByCar`, `tmo:ByTaxi` or `tmo:ByBike`). The moving parts can be traveled by *scheduled journeys* (represented by the GTFS classes), which are trips made by tram, subway, rail,

bus, ferry, cable car or funicular, or *not scheduled journeys*, such as cars, taxis and bikes (see Figure 12).

The following triples represent the output of the move enrichment:

1. `Move8` is an instance of a move of the Segmented Trajectory Ontology (see Figure 9);
2. `trip13/1` is an instance of the `gtfs:Trip` class; and
3. `route13` is an instance of the `gtfs:Route` class.

```
1. sto-resource:move8 rdf:type sto:Move;
2.      sto:label "Vehicle";
3.      sto:travelled_by gtfs-resource:trip13/1.
4. gtfs-resource:trip13/1 rdf:type gtfs:Trip;
5.      gtfs:direction_id "1";
6.      gtfs:trip_short_name "ATAF Linea 13/Direzione Stazione FS SMN";
7.      gtfs:route gtfs-resource:route13.
8. gtfs-resource:route13 rdf:type gtfs:Route;
9.      gtfs:route_short_name "ATAF Linea 13";
10.     gtfs:route_type "3";
11.     gtfs:route_long_name "ATAF Linea 13 / Stazione Palazzo dei Con
   gressi - Cartoleria Il Gatto e La Volpe <=> Piazzale Michelangiolo".
```

The triples above add the missed semantics to `move8` (Lines 1-3), meaning that this segment of the trajectory was travelled with the line `ATAF Linea 13 / Direzione Stazione FS SMN` (lines 4-7) having as direction '1' (one way), that is part of the route with short name ATAF Linea 13 (lines 8-11) made by bus (`route_type 3`) (line 10).

Figure 15 illustrates the output of the stops and the move mapping phases for the running example, i.e., the trajectory of the tourist shown in Figure 15, having all its segments - stops and moves - enriched with the external data provided by the Linked Data mashups, as shown in the previous sections.

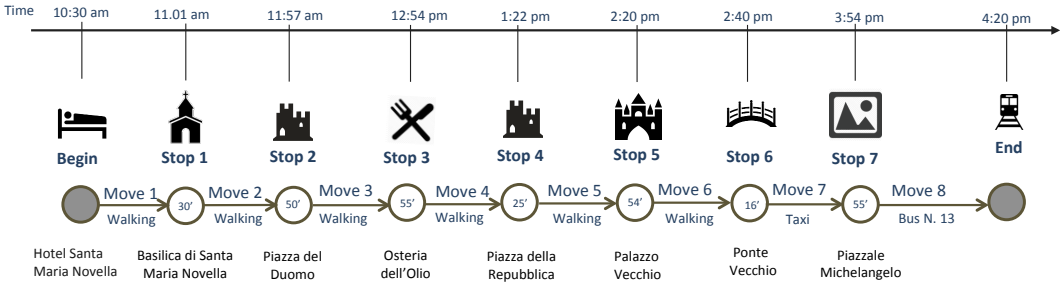

Figure 15: The running example trajectory enrichment.

3.3. Querying and analyzing Semantic Trajectories

The output of the enrichment step is stored in the semantic trajectory repository where movement data and the associated semantics are represented in a uniform formalism. We note that this repository is represented by the SEMantic Trajectory Ontology (SemTO), which is the union of the Segmented Trajectory Ontology (STO), the Mashup Ontology (MO) and the additional properties introduced during the enrichment step (e.g. `semto:based_near` and `semto:travelled_by` in our example).

The semantic trajectory repository enables a number of interesting analysis as we illustrate in the following SPARQL queries. In this section, we present explicative examples querying enriched trajectories involving.

1. A single trajectory (Q1);
2. A set of trajectories from the same traveler (Q2); and
3. All trajectories of the repository (Q3) and (Q4).

We remark that the following queries return entities made available by the Semantic Trajectory Repository, combining the movement data and the semantic data through the mashup view, as described in Section 3.2.2.2.

The first query characterizes the cultural tastes of the tourist during her trip through the city center by retrieving the artists and artworks information of the museums the traveler visited. We can notice the use of the matching property `foaf:based_near` and the use of property `tmo:listOfCategories` from the mashup view.

Q1) Which are the categories of the museums visited during the trip with id "tripFlorenceJune16" and the art movements related to them?

```
SELECT
       ?museumName ?museumCategory ?artmovement
WHERE{
       ?segmentedTrajectory
              stoid "tripFlorenceJune16";
              sto:isComposedOf ?stop.
       ?stop foaf:based_near ?museum.
       ?museum rdf:type dbo:Museum;
              rdfs:label ?museumName.
              tmo:listOfCategories ?museumCategories;
              dbp:museumOf ?artwork.
       ?artwork dbo:author ?artist.
       ?artist dbo:movement ?artmovement
}
```

The above query returns the categories of all the museums that were visited in trip "tripFlorenceJune16". For our running example, the query result is *Palazzo Vecchio* as museum, *Art museums and galleries* as category and *High Renaissance* as art movement. In fact, that museum contains the marble sculpture *The Genious Of Victory* (the artwork), by *Michelangelo* (the artist), that in turn is associated with *High Renaissance* (art movement).

Q2) Which was the average price spent in museums and churches by the traveler with id 257?

```
SELECT
       (AVG(?entrancePrice) AS ?avgEntrancePrice)
WHERE{
       ?movingObject sto:userId "257";
              sto:produces ?rawTrajectory.
       ?rawTrajectory sto:isTransformed ?segmentedTrajectory.
       ?segmentedTrajectory sto:isComposedOf ?stop.
       ?stop foaf:based_near ?attraction.
       ?attraction tmo:entrancePrice ?entrancePrice.
       {?attraction rdf:type dbo:Museum}
       UNION {
            ?attraction rdf:type dbo:ReligiousBuilding;
              dbp:architecture_style dbr:Church_(building)
       }.
}
```

This query adds up all money spent in museums and churches during all trips made by the traveler with id '257'. The UNION clause combines graph

patterns thus allowing that alternative possibilities match the same variable `?entrancePrice`, which refers to money spent in attractions that can be a `dbo:Museum` or a `dbo:ReligiousBuilding` of the type church.

The next query spans all the tourists' trajectories to retrieve those that used transportation from the local bus company called ATAF. Notice the use of the property `semto:travelled_by` (described in Section 3.2.2.2) of the semantic trajectory ontology that links the move to the actual transportation means used in that trajectory segment.

Q3) Which tourists used buses provided by the ATAF company?

```
SELECT
       DISTINCT ?travellerId
WHERE{
       ?movingObject sto:userId ?travellerId;
              sto:produces ?rawTrajectory.
       ?rawTrajectory sto:isTransformed ?segmentedTrajectory.
       ?segmentedTrajectory sto:isComposedOf ?move.
       ?move semto:travelled_by ?trip.
       ?trip rdf:type gtfs:Trip;
              gtfs:route ?route.
       ?route rdf:type gtfs:Route;
              gtfs:agency ?agency;
              gtfs:route_type "3".
       FILTER(?agency gtfs:agency_name "ATAF")
}
```

This query returns the identifiers of all tourists that traveled using buses of the ATAF Company, the local bus company of Florence.

Q4) Which travelers are interested in High Renaissance?

```
SELECT
       ?userId
WHERE{
       ?movingObject sto:userId ?userId;
              sto:produces ?rawTrajectory.
       ?rawTrajectory sto:isTransformed ?segmentedTrajectory.
       ?segmentedTrajectory sto:isComposedOf ?stop.
       ?stop foaf:based_near ?museum.
       ?museum rdf:type dbo:Museum;
              dbp:museumOf ?artwork.
```

```
        ?artwork rdf:type dbo:Artwork;
                dbp:author ?author.
        ?author rdf:type dbo:Artist;
                dbo:movement <http://dbpedia.org/resource/High_Renaissance>.
}
```

The above query filters out only travelers that may have an interest in the period of the Italian Renaissance art production called High Renaissance. Other examples of queries that can be easily answered by the semantic trajectory repository involve restaurant reviews and temporal filtering such as:

1. *How many travelers stayed in 5 stars hotels and ate at "very good" restaurants?*
2. *How many travelers went to catholic churches that contain Baroque artwork in Florence in the last 2 months?*

The SPARQL queries provide many answers to the requirements such as characterizing the visited venues or the cultural level of tourists. However, some more sophisticated questions can only be answered after analyzing data contained in the repository.

Going back to our running example, consider the discovery of the spending profile of the tourists. We may wish to distinguish between high spending and low spending tourist profiles. They define a visiting behavior based, for example, on the entry price of the visited venues, combined with the identification of prestigious accommodations and restaurants, properly joined with additional information such as the average life cost in the city.

The most natural way to perform these kinds of analyses is to exploit the structured format provided by the RDF model to perform inferences using OWL reasoning capabilities (POLLERES *et al.*, 2013). In this case we can combine the semantic trajectory ontology with a larger application ontology containing concepts related to the specific application representing the users' behaviors we want to infer from data in the style of (RENSO *et. al*, 2013).

Another challenging analysis is to identify groups of semantically similar trajectories. In the mobility field, similarity usually relies on the spatio-temporal characteristics of the raw trajectories, while here we can take advantage of the rich semantics coming from Linked Open Data that can be conveniently combined with the spatio-temporal component. This offers a new opportunity to define

innovative similarity measures, such as finding groups of tourists with similar cultural preferences, lifestyles and spending profiles, all useful information to design novel sophisticated recommendation systems.

3.4. Conclusions

We introduced a conceptual framework for the semantic enrichment of movement data based on Linked Open Data. Our proposal offers a flexible, reusable, application-oriented process based on ontologies that support the transformation of movement data into a Linked Open Data semantically enriched trajectory repository. We highlighted the different steps and how the availability of such repository improves the ability to formulate application analysis questions, thanks to the richness of the linked contextual data.

We discussed the process with the help of a running example in the tourism domain. It is important to stress that this process is meant to be general and support the semantic enrichment of several kinds of movement data in different domains. Thus, it includes not only GPS data but also social networks geo-located photos from Flickr, Twitter posts, FourSquare check-ins, and others.

Comparing with other works in the literature that deal with the semantic enrichment process, our approach faces all steps of the trajectory enrichment process but taking advantage of the Linked Data principles, while some of the works focus on some task in isolation, such as the formalization task, using ontologies for the enrichment part.

4
Computing the semantic similarity of Linked Data Entities

This chapter targets the research question **RQ2**. *How can we discoverer similar Linked Data entities by considering their relevant features?* We cope with this question by presenting a novel approach to estimate semantic entity similarity using entity features available as Linked Data. The key idea is to exploit ranked lists of features, extracted from Linked Data sources, as a representation of the entities to be compared. The similarity between two entities is then estimated by comparing their ranked lists of features. Section 4.1 gives an overview of SELECTOR, a two-module framework that takes as input Linked Data entities, ranks the lists of entity features according to their relevance for describing the entities, compares the ranked lists using rank correlation metrics, and outputs the entities similarity. Section 4.2 describes experiments with museums descriptions found in DBpedia. Section 4.3 assesses the proposed approach with experiments in two other domains: dataset descriptions found in a Linked Open Data catalogue; and (iii) computer science conferences available in the Linked Data version of DBLP. Finally, Section 4.4 summarizes some conclusions.

4.1.
SELEcTor framework

Figure 16 gives an overview of the SELECTOR framework, which takes as input Linked Data entities, extracts from Linked Open Data lists of their relevant features, and then compares the ranked features according to some rank correlation metric to generate a similarity score for the entities. The first module is the *ranked features extractor*, which communicates with Linked Open datasets to extract the ranked list of relevant features that describe the entities. The second module, the *entity similarity processor*, takes these lists as input and compares

them using the proper list correlation metric to generate as output a similarity score for the entities. We detail both modules in the following sections.

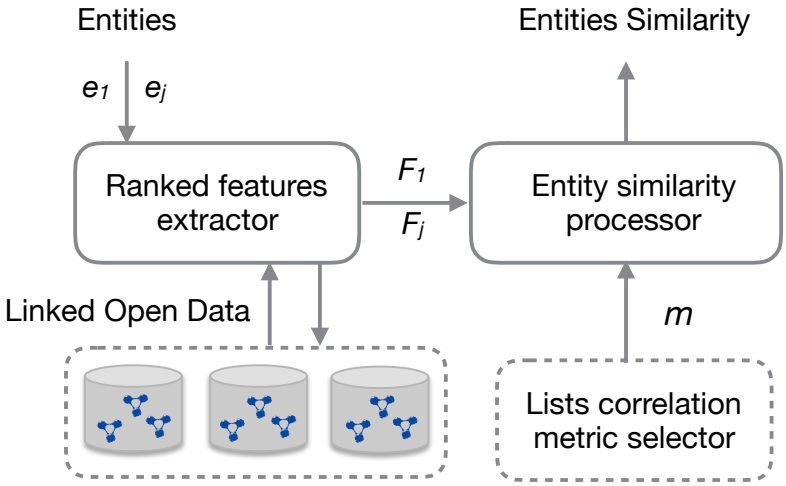

Figure 16: Overview of the SELECTOR framework.

4.1.1. Extracting ranked features

The *ranked features extractor* module is in charge of generating ranked lists of features that describe a certain entity. As it can be seen in Figure 16, this module receives as input two Linked Data entities and outputs their respective ranked features. For each input entity, the module navigates through the nodes of Linked Data graphs that are connected to the entity to extract its features. The module accesses the Linked Data graphs through its respective SPARQL endpoints. It is important to notice that the features' identification is part of an analysis process that can be aided by a domain expert.

Consider the museums scenario. We claim that two museums can be compared on the basis of the art movements of their art pieces. We also investigated other Linked Data features, such as those related to their popularity based on the number of visitors, but we found that, compared with other features available as Linked Data, the art movements better describe the museums. Therefore, we claim that museums with art pieces of similar art movements are similar, whereas two museums with no art movements in common would be completely different.

We call *query-based* and *graph-exploration-based* the strategies to order the features to generate a ranked list.

The *query-based strategy* performs a pre-defined SPARQL query over one or more Linked dataset endpoints to generate the ranked features. In the case of museums, the SPARQL query would match a certain path pattern to get all the art movements (features) that describe the museum (the input entity), according to some criterion of relevance. The criterion of relevance is applied using some group function that aggregates the features, for instance, counting the number of art pieces of each art movement.

We note that, depending on the aggregation function chosen, a tie may occur between two or more features, in the sense that they all have the same feature values. In these cases, the SPARQL query can be re-formulated to untie the elements according to some other criteria.

When applying the *graph-exploration-based strategy*, the module navigates through the RDF graph and then calculates the importance of each node to describe a certain entity. There are several approaches in the literature to measure relationships within a graph, commonly referred as *centrality measures*, such as the Katz score (KATZ, L., 1953) or the SCS score (NUNES *et al.*, 2013) (See Section 2.3). They can be profitably applied in this context to generate a ranked list of features (nodes) that represent a certain entity.

In both strategies (*query-based* and *graph-exploration-based*) the module performs SPARQL queries over the Linked datasets. The main difference is that in the query-based strategy, the query already retrieves the ranked features, i.e., it orders the features and the answer of the query is the ranked features itself, while in the graph-exploration-based strategy, the SPARQL query is used to retrieve an RDF graph which the module will use to measure the importance of each entity.

4.1.2. Computing entity similarity

The *entity similarity processor* takes as input the lists of features, and compares them to measure how similar they are, using a rank correlation metric. The module, therefore, outputs the similarity score for a pair of entities.

The module can choose one of the similarity measures introduced in Section 2.4 to compute the similarity between the entities, according to the nature of the ranked features. If the lists have the same items (i.e., if they are conjoint), the module may choose Kendall, Spearman's p, among others. Otherwise, the module may choose AO (Average Overlap) (FAGIN *et al.*, 2004) or its top-weighted parameterized extension, RBO (WEBBER *et al.*, 2010).

The module output is a similarity score that measures how similar the entities are. When the entities do not have any feature in common, the correlation coefficient is 0, and when they have exactly the same features and in the same order (and also the same weights for weighted ranks), the output is 1.

4.2. Experiments with museums descriptions found in DBpedia

In this section, we instantiate the SELECTOR framework and then evaluate our approach comparing the similarity measure obtained using SELECTOR with a ground truth. We argue that the similarity between Linked Data entities is better captured if we first rank their features, in a way that the more relevant features appear before the less relevant ones, and then compare the ranked lists. We adopt the museums domain as an example of trajectory POI to be compared.

4.2.1. Museums on DBpedia

DBpedia provides entities that represent museums around the world, which are instances of `dbo:Museum` class. An example of triple is `<dbr:Louvre, rdf:type, dbo:Museum>` (`dbr` is the prefix of `http://dbpedia.org/resource/`) stating that Louvre is an entity of type museum.

The museum instances can be linked to other entities through the `dct:subject` property, often used to represent the topic of the entity. Some of these entities are hierarchically related to each other through the `skos:broader` property and in some cases they have a direct link to the `dbc:Museum_by_type` class. We call *categories* all the entities linked to `dbc:Museum_by_type` directly or indirectly through the `skos:broader` property.

Figure 17 shows some categories of Louvre. A category directly related to Louvre is `dbc:Museums_of_Ancient_Greece`. The indirectly related categories are `dbc:History_museum` and `dbc:Civilization_museums`.

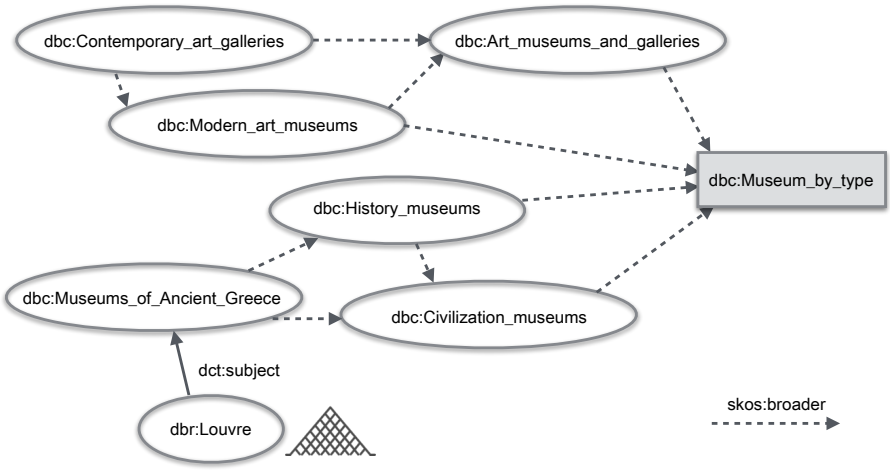

Figure 17: DBpedia concepts describing museum categories.

Figure 18 illustrates other museum properties to be explored on DBpedia. The `dbo:museum` property links a museum to its art pieces, instances of `dbo:Artwork`. In turn, each art piece may be linked to its creator/artist through the `dbo:author` property. Finally, the artists may be related to one or more art movements by the `dbo:movement` property. The RDF graph shown in Figure 18 represents that the *J Paul Getty museum* has as art piece the *Irises* painting by Vincent van Gogh, an impressionist (art movement) artist.

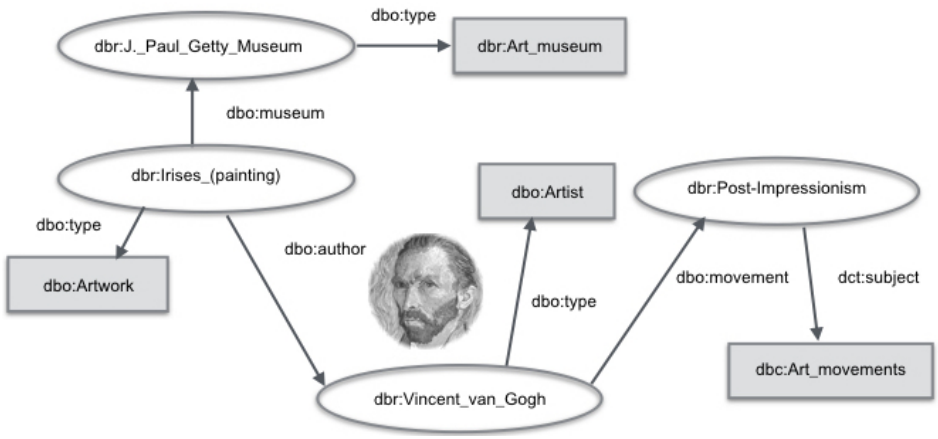

Figure 18: DBpedia links describing J. Paul Getty museum features.

In the experiments, we explored the DBpedia graphs shown in Figure 17 and Figure 18.

4.2.1.1. Ranking the features

Following the first step of the framework, the extraction of ranked features, we applied to the DBpedia graph both the graph-exploration-based and the query-based strategies, described in Section 4.1.1.

When applying the graph-exploration strategy, we explored DBpedia entities and properties as shown in Figure 17. Given a Linked Data entity *e* that represents a museum, the *ranked features extractor* module queries DBpedia via its SPARQL endpoint following only the properties `dct:subject` and `dct:broader`. We navigate the graph from the root entity (the museum) reaching the entities that represent their categories.

We consider such museum categories (the entities having direct or indirect links to the `dbc:Museum_by_type` class) as the features to be ranked by this module. We used the depth-first approach with depth distance 4, as adopted in (NUNES *et al.*, 2013), which means that we have considered the entities from the root until all its 4-hop neighbors.

In order to measure the relevance of each feature with respect to the museum, we have calculated the distance from the museum to all the features *F* (the categories) using the SCS score (see Section 2.4) (NUNES *et al.*, 2013). We then ordered the features according to the score, generating the ranked features list. It important to notice that, even though SCS is a connectivity measure (i.e., the semantic association) used to identify connected entities, we used it in this experiment as an alternative to automatically find the relevant nodes of the entities.

Table 3 shows Louvre's ranked features using the graph-exploration-based strategy, ordered in descending order by SCS score. The feature `dbc:Museums_of_Ancient_Near_East`, at the first position, and the feature `dbc:Museums_of_Ancient_Greece`, at the second position, are the more relevant to describe Louvre (both with SCS score 0.5) while the less relevant is `dbc:Civilization_museums`, with SCS score 0.25.

Table 3: Louvre features by the graph-exploration approach.

Louvre categories	SCS score
dbc:Museums_of_Ancient_Near_East	0.5
dbc:Museums_of_Ancient_Greece	0.5
dbc:History_museums	0.48
dbc:Civilization_museums	0.25

When applying the graph-exploration-based strategy for a group of museums, we found that the ranked features did not represent well the museums for two reasons: (i) some of the most famous museum have few DBpedia categories that represent them, such as Louvre; and (ii) in some cases, the categories found are very generic and thus do not represent well the different museums (such as `dbc:Society_museums`, `dbc:Art_museums_and_Galleries` and `dbc:Civilization_museums`).

We therefore also applied the query-based strategy, exploiting the DBpedia graph paths shown in Figure 17. Instead of exploiting the museum categories, we focused on the art movements of the art pieces that can be reached using the artists, as can be seen in Figure 18.

The *ranked features extractor* module performed the following SPARQL query, which aggregates the art movements and orders them by the art pieces frequency. The `?museum` parameter represents the input entity.

```
SELECT ?artMovement
WHERE {
  ?artWork <dbo:museum> ?museum.
  ?artWork <dbo:author> ?artist.
  ?artist <dbo:movement> ?artMovement.
}
GROUP BY ?art_movement
ORDER BY DESC(count(?artWork))
```

Table 4 shows the SPARQL results for the Getty Museum ranked features. Note that, since DBpedia is constantly updated, the results may vary. In some cases, the SPARQL query may return items with the same feature value, i.e., two or more art movements with the same art pieces frequency. In this cases, one may choose another criteria to untie these features, such as the *out* or the *in* degree of

the feature, which represents respectively the number of RDF links that leave from the entity and the number of links that arrive at the entity.

Table 4: Getty features extraction by the query-based strategy.

J Paul Getty ranked features
dbr:Symbolism_(arts)
dbr:Baroque
dbr:Expressionism
dbr:Romanticism
dbr:High_Renaissance
dbr:Dutch_Golden_Age_painting
dbr:Academic_art
dbr:Post-Impressionism
dbr:Mannerism

According to the query-based strategy and using the SPARQL query shown above, the feature that better describes the Getty museum is the *Symbolism* art movement and the less relevant feature is the *Mannerism* art movement, which means that the museum has more art pieces that belong to the Symbolist art period than to the Romanticism art period.

Comparing the graph-exploration-based strategy with the query-based strategy for the museums scenario, we found that the later strategy is able to extract better features, both in quantity (the ranked lists have more items) and in quality (the art movements are more domain-specific than the museum categories available on DBpedia).

4.2.1.2. Computing the similarity between entities

As explained in Section 4.1.2, the *entity similarity processor* module takes as input the ranked features, representing the entities to compare, and gives as output their similarity score.

Table 5 shows the ranked features that describe the Getty and the Louvre museums. They have been generated in the previous step using the query-based strategy by performing the SPARQL query previously presented.

Table 5: Comparing the ranked features.

J Paul Getty ranked features	Louvre ranked features
dbr:Symbolism_(arts)	dbr:Romanticism
dbr:Baroque	dbr:High_Renaissance
dbr:Expressionism	dbr:Neoclassicism
dbr:Romanticism	dbr:Baroque
dbr:High_Renaissance	dbr:Italian_Renaissance
dbr:Dutch_Golden_Age_painting	dbr:Dutch_Golden_Age_painting
dbr:Academic_art	dbr:The_Renaissance
dbr:Post-Impressionism	dbr:Classicism
dbr:Mannerism	dbr:Realism_(arts)
	dbr:Flemish_Barique_paiting
	dbr:Early_Netherlandish_painting
	dbr:Caravaggisti

In order to compare the two ranked features, the *entity similarity processor* in this case chooses the RBO measure (see Section 2.4) (WEBBER *et al.*, 2010), since it handles non-conjoint lists (the museums are not described by the same art movements) and also allows to weight higher ranks more heavily than lower ranks with the parameter *k* (their top-k art movement is more relevant then the top-k+1, and so on). It is important to notice that the choice of parameter *k* was inspired by the experiments performed in [11], which consider *k* ranging from 0.9 to 0.998.

When computing the similarity between the Getty and the Louvre museums, with the top-weighted parameter *p* equals to 0.95, the RBO score is 0.437. In fact, from a total of 4 common features (the art movements), the first art movement to match in both lists (`dbr:Baroque`) appears in the 4nd position and the last art movement to match (`dbr:Dutch_Golden_Age_Paiting`) appears on the 6th position.

When comparing the Getty Museum with the *Museum Of Modern Art*, in New York, the RBO similarity score with p = 0.95 is 0.117. In fact, they only have 2 art movements in common, the first art movement matches in the 8th position (`dbr:Post-Impressionism`) and the last art movement matches in the 14th positions (`dbr:Expressionism`). Considering our museums dataset shown in

Table 6, the Art Institute of Chicago is that which is least similar to the Getty Museum.

We computed the similarity between all museums of our dataset (shown in Table 6) using RBO. Then, we generated, for each museum, the list of the most similar museums. Table 6 shows the most similar museums to the Getty Museum, with $p = 0.95$ and $p = 0.98$.

Table 6: The most similar museums to the J. Paul Getty Museum.

Getty similars	**RBO score** ***p* = 0.95**	**RBO score** ***p* = 0.98**
dbr:Metropolitan_Museum_of_Art	0.437	0.491
dbr:Louvre	0.404	0.429
dbr:Kunsthistorisches_Museum	0.385	0.419
dbr:Museum_of_Fine_Arts,_Boston	0.360	0.381
dbr:Vatican_Museums	0.351	0.380
dbr:Uffizi	0.261	0.302
dbr:National_Gallery_of_Art	0.247	0.281
dbr:Musée_d'Orsay	0.161	0.195
dbr:Philadelphia_Museum_of_Art	0.161	0.195
dbr:Museum_of_Modern_Art	0.117	0.139
dbr:Art_Institute_of_Chicago	0.103	0.130

4.2.1.3. Evaluation

Constructing the ground truth

Since there is not a specific ground truth containing museums similarity data to validate our approach, we built it using a well-known Web site about art history. SmartHistory[9] is a non-profit organization that makes art history learning content freely available and provides a number of articles discussing the most important masterpieces, ranging from ancient to contemporary art. We chose SmartHistory as the ground truth because it is a rich source of museums data entirely authored by human domain experts, its creation process is totally independent from the DBpedia (or similar) data, and it is not affected by popularity bias.

[9] smarthistory.org

Each SmartHistory article (very often an article is about an art piece) is categorized according to a hierarchical taxonomy, which includes time periods, art movements, and other relevant facets. For each artwork, the hosting museum is also mentioned.

Given the SmartHistory data, we defined the similarity between two museums on the basis of the categories found in the articles mentioning the museums. To avoid sparsity, we limited to the top-2 levels of the category hierarchy. Museum similarity was then computed as the cosine similarity of the museum's categories. Cosine similarity was adopted as it allows to properly weight the richness of a given museum in a specific category, and avoids boosting large museums with several works of art.

Choosing the set of museums

First, we pre-filtered 32 museums in DBpedia with at least 8 art pieces and 5 art movements in order to avoid poorly described museums. Then, we filtered the museums that have also categories in the ground truth, Smart History, resulting in the 12 richest museums in both datasets, shown in Table 7.

Table 7: Chosen museums for the experiment.

Museum	#DBpedia art movements	#SmartHistory art movements
dbr:Metropolitan_Museum_of_Art	28	84
dbr:Louvre	17	37
dbr:Museum_of_Modern_Art	36	17
dbr:National_Gallery_of_Art	29	17
dbr:J._Paul_Getty_Museum	9	14
dbr:Uffizi	11	12
dbr:Museum_of_Fine_Arts,_Boston	7	16
dbr:Musée_d'Orsay	11	10
dbr:Art_Institute_of_Chicago	26	9
dbr:Philadelphia_Museum_of_Art	15	13
dbr:Kunsthistorisches_Museum	9	11
dbr:Vatican_Museums	16	13

We chose as baseline the semantic relatedness measure proposed by Milne and Witten, WLM (WITTEN *et al.*, 2008) (see Section 2.3). Even though WLM is

intended to be a generic approach, we chose it as a baseline, since it also measures the semantic relatedness (a broader concept regarding to the similarity concept) of two Linked Data entities. Furthermore, as far as we know, there is no metric defined specifically to compare two museums available as Linked Data. To compute the WLM similarity, we used Dexter, an Open Source Framework for Entity Linking (CECCARELLI *et al.*, 2013).

Our strategy to evaluate the results is based on the comparison of the lists of similar museums (such as in Table 6) generated by the three different approaches: (i) our approach, namely the SELECTOR framework, (ii) the WLM measure, which represent our baseline; and (iii) the SmartHistory data, from which we have built the ground truth.

Then, for each museum, we have confronted the SELECTOR list with the ground truth list. Table 8 shows an example of the lists of museums similar to the Getty Museum generated by SELECTOR and SmartHistory.

As can be seen in Table 8, according to SELECTOR, *MET* (Metropolitan Museum of Art, in New York) is the most similar museum to The J. Paul Getty Museum, the second most similar is the Louvre and the third one is the Kunsthistorisches Museum, an art museum in Vienna, and so on.

Table 8: Comparing SELECTOR with the ground truth.

SELEcTor: List of museums similar to the Getty Museum	**SmartHistory: List of museums similar to the Getty Museum**
dbr:Metropolitan_Museum_of_Art	dbr:Metropolitan_Museum_of_Art
dbr:Louvre	dbr:Vatican_Museums
dbr:Kunsthistorisches_Museum	dbr:Louvre
dbr:Museum_of_Fine_Arts,_Boston	dbr:National_Gallery_of_Art
dbr:Vatican_Museums	dbr:Art_Institute_of_Chicago
dbr:Uffizi	dbr:Museum_of_Fine_Arts,_Boston
dbr:National_Gallery_of_Art	dbr:Musée_d'Orsay
dbr:Musée_d'Orsay	dbr:Philadelphia_Museum_of_Art
dbr:Philadelphia_Museum_of_Art	dbr:Kunsthistorisches_Museum
dbr:Museum_of_Modern_Art	dbr:Uffizi
dbr:Art_Institute_of_Chicago	dbr:Museum_of_Modern_Art

We also confronted the lists of similar museums generated by WLM (our baseline) with the SmartHistory lists. Table 9 shows an example of the similar lists generated by WLM and SmartHistory, again for the Getty Museum.

Table 9: Comparing WLM with the ground truth.

WLM: List of museums similar to the Getty Museum	**SmartHistory: List of museums similar to the Getty Museum**
dbr:National_Gallery_of_Art	dbr:Metropolitan_Museum_of_Art
dbr:Musée_d'Orsay	dbr:Vatican_Museums
dbr:Philadelphia_Museum_of_Art	dbr:Louvre
dbr:Museum_of_Fine_Arts,_Boston	dbr:National_Gallery_of_Art
dbr:Kunsthistorisches_Museum	dbr:Art_Institute_of_Chicago
dbr:Art_Institute_of_Chicago	dbr:Museum_of_Fine_Arts,_Boston
dbr:Metropolitan_Museum_of_Art	dbr:Musée_d'Orsay
dbr:Uffizi	dbr:Philadelphia_Museum_of_Art
dbr:Museum_of_Modern_Art	dbr:Kunsthistorisches_Museum
dbr:Vatican_Museums	dbr:Uffizi
dbr:Louvre	dbr:Museum_of_Modern_Art

According to the baseline WLM (Table 9), the most similar museum to the Getty Museum is the *National Gallery of Art*, an art museum in Washington D.C. The second most similar is *Musée d'Orsey*, in Paris, and the third most similar is the *Philadelphia Museum of Art*, and so on.

Analyzing the results, it can be noticed that the geographic proximity between two museums influences the similarity score between them, according to WLM, as expected. This is because this measure considers all links found in their respective Wikipedia articles, including some geographic-related links. An example is the link <dbc:Modern_art_museums_in_the_United_States>, connected to the museum through the property <dct:subject>, that can be found both in the *Metropolitan* page as in the *National Gallery of Art* page (dbc is a prefix for http://dbpedia.org/page/Category and dct is a prefix for http://purl.org/dc/ terms/). This explains why, according to WLM, some museums – for instance The Louvre and the Vatican Museum – are in the last positions on the similarity list, while in the SELECTOR lists of similar museums they appears in the first-half of the list. Analogously, WLM considered as similar

some museums that SELECTOR does not consider – for instance, the Philadelphia Museum of Art, which is also located in the United States.

Lastly, we calculated the accuracy of SELECTOR lists and the WLM lists comparing both with the ground truth. We compared the lists using NDCG (Normalized Discounted Cumulative Gain) (JÄRVELIN *et al.*, 2002), a well-known metric used in Information Retrieval to measure ranking. The measure accumulates the gain from the top of the list to the bottom, penalizing lower ranks. It may be parameterized to take into account the top k first elements of the lists quality (See Section 2.4).

Figure 19 shows the results considering only the Getty Museum, with *k* from 3 to 8. The X axis represents *k*, which ranges from 3 to 8, while the Y axis represents the NDCG@k score itself. Considering the top 3 items, the SELECTOR accuracy score is 0.886, while the WLM score is 0.697. Considering the top 4 items, the SELECTOR score decreases to 0.876, but is still higher than the WLM score, 0.714. The highest SELECTOR accuracy is 0.924, achieved when $k = 8$.

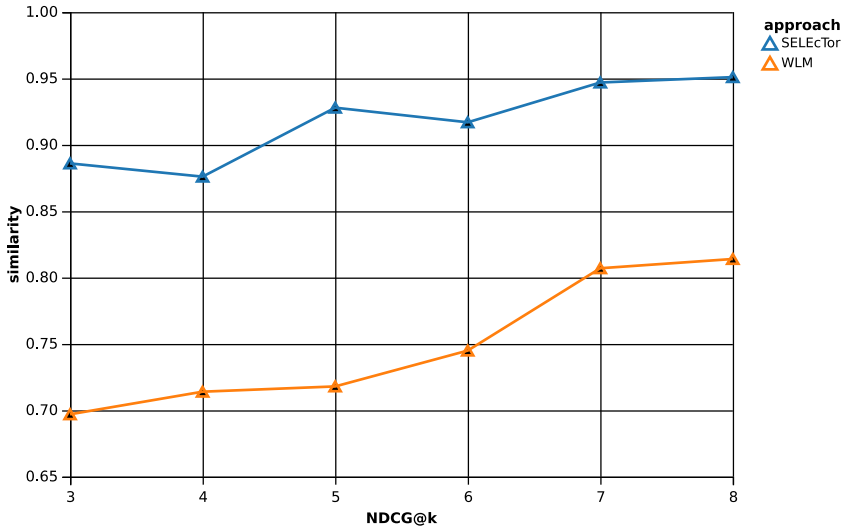

Figure 19: NDGC results for Getty museum.

Finally, we compared the similarity lists of all museums using the same idea. Figure 20 shows the results. Again, the X axis represents *k*, which ranges from 3 to 8, while the Y axis represents the NDCG@k score itself. Considering all museums, SELECTOR performs significantly better than WLM, for any *k*, with the highest accuracy being 0.924 (when $k = 6$).

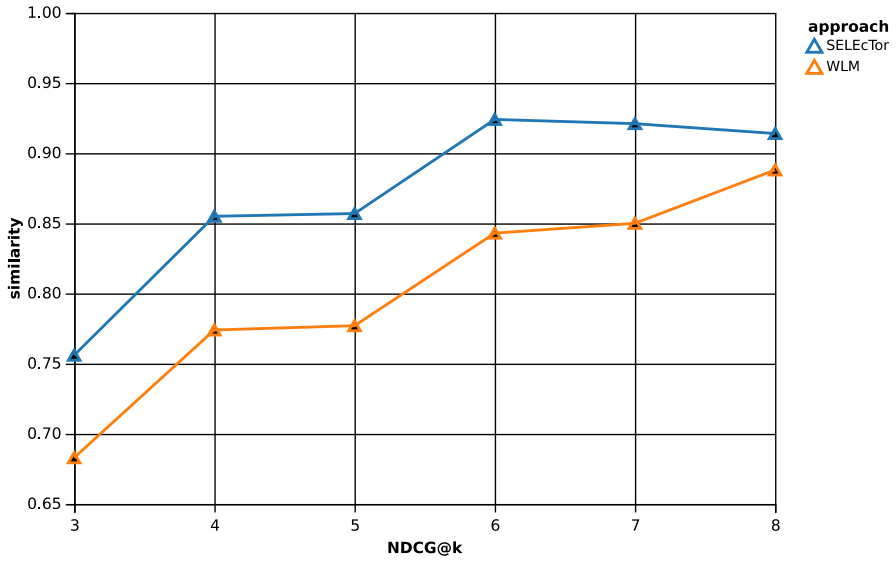

Figure 20: The average NDGC top k items.

Since SELECTOR filters the entities that better describe a museum, it can be consider more selective than WLM, in the sense that it focuses on the specific features that describe the museum - such as its art movements.

In this experiment, we compare RBO results with the semantic relatedness measure WLM (*Wikipedia Link-based Measure*). Although this research focuses on the semantic *similarity* between Linked Data entities, we used the semantic *relatedness* measure WLM as a baseline, since it is a well-known metric to compare different Linked Data entities.

Although in this experiment we use SELECTOR to compare museums (which can be trajectory POIs), the approach can be tuned or generalized to other application scenarios. Indeed, the next section applies SELECTOR to two different scenarios: comparing datasets and comparing computer science conferences.

4.3. Experiments with other domains

4.3.1. Comparing LOD datasets

Datasets in the Mannheim Linked Data Catalog

The data used in this experiment was obtained by a crawling process of the LOD cloud, detailed in (CARABALLO *et al.*, 2016) and conducted in April 2014, that started with the description of the datasets contained in the Mannheim Linked Data Catalog[10].

Figure 21 illustrates the overview of the experiment. The upper part of the figure comprises the Mannheim Linked Data Catalog extraction process. We extracted 390 datasets from the Mannheim catalogue, filtering out only the datasets that have SPARQL endpoints available.

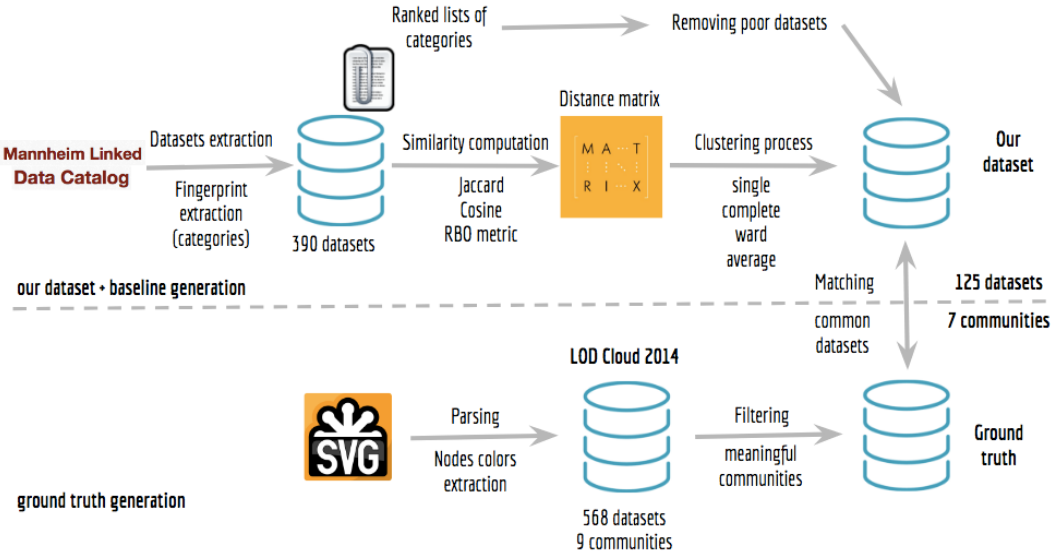

Figure 21: Datasets selection.

Extracting and ranking features

There are several ways to extract relevant features from datasets. In this experiment, we adopted a profiling techniques described in (CARABALLO *et al.*, 2016), which basically generates *profiles* or *fingerprints* for textual resources,

[10] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/

extracted from the datasets (see Figure 21). The method, detailed in (CARABALLO *et al.*, 2016), has five steps:

1. Extract entities from a given textual resource.
2. Link the extracted entities to the English Wikipedia articles.
3. Extract the English Wikipedia categories for the articles.
4. Follow the path from each extracted category to its top-level category and compute a vector with scores for the top-level categories thus obtained.
5. Perform a linear aggregation in all dimensions of the vectors to generate the final profile, represented as a histogram for the 23 top-level categories of the English Wikipedia, shown in Table 10 (`dbc` is a prefix for `https://en.wikipedia.org/wiki/Category`).

The Wikipedia top-level categories shown in Table 10 represent the features that describe the datasets. For each of the 390 datasets, we then aggregate the categories by counting the number of entities extracted from textual resources of each category.

Finally, we avoided poorly described datasets, i.e., datasets described by few categories (see Figure 21). We considered a minimum of 15 categories (out of 23), which ruled out 10% of the datasets.

Table 10: Wikipedia Top-level categories.

Wikipedia Top-level categories
dbc:Agriculture
dbc:Applied_sciences
dbc:Arts
dbc:Belief
dbc:Business
dbc:Chronology
dbc:Culture
dbc:Education
dbc:Enviroment
dbc:Geography
dbc:Health
dbc:History
dbc:Humanities

dbc:Language
dbc:Law
dbc:Life
dbc:Mathematics
dbc:Nature
dbc:People
dbc:Politics
dbc:Science
dbc:Society
dbc:Technology

Computing entity similarity

We generate ranked lists using the frequency of each category found for the dataset. Table 11 shows the frequency of the categories for two of the datasets available at the catalogue: the *eu-agencies-bodies* dataset[11], a dataset about agencies and decentralized bodies in the EU; and the *rkb-explorer-citeseer*[12] dataset, a semantic research repository with co-reference information from the research index CiteSeer[13]. Table 11 shows the top 5 categories of each dataset in boldface.

Table 11: Top-level categories frequency.

Top-level category	***eu-agencies-bodies* dataset**	***rkb-explorer-citeseer* dataset**
dbc:Agriculture	133	1
dbc:Applied_sciences	57	28
dbc:Arts	85	75
dbc:Belief	290	930
dbc:Business	**687**	468
dbc:Chronology	163	**1377**
dbc:Culture	427	773
dbc:Education	508	**1458**
dbc:Enviroment	411	91
dbc:Geography	240	14
dbc:Health	38	3
dbc:History	20	33

[11] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/sk/dataset/eu-agencies-bodies
[12] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/sk/dataset/rkb-explorer-citeseer
[13] http://citeseer.ist.psu.edu/index

dbc:Humanities	460	861
dbc:Language	84	367
dbc:Law	15	2
dbc:Life	291	401
dbc:Mathematics	248	**2142**
dbc:Nature	**1650**	893
dbc:People	1	-
dbc:Politics	118	26
dbc:Science	**1529**	**2391**
dbc:Society	**979**	**1650**
dbc:Technology	**987**	948

As can be noticed in Table 11, the second dataset does not contain `dbc:People` entities. Therefore, the lists representing the datasets do not always have all the 23 top-level categories, i.e., their lists of categories are not always conjoint.

The *entity similarity processor* may choose different similarity measures to compare the ranked lists and generate a similarity score. In the example shown in Table 11, if the similarity measure chosen is the *cosine distance*, the score is 0.784. Choosing *RBO* as the similarity measure (see Section 3), the score is 0.887 (with p = 0.98) and 0.940 (with p = 0.99). In turn, the *Jaccard distance* gives 0.956 as the similarity score.

Although the Jaccard distance gives the highest similarity score (since the lists have 22 out of the 23 categories in common), the two datasets appear in different communities in the LOD Cloud (considered as the ground truth for the experiment): the *eu-agencies-bodies* is in the *Publication* community and the *rkb-explorer-citeseer* is in the *Government* community. In fact, the relevance of the categories representing the datasets is very different. In these cases, the Jaccard distance is not a reasonable option, since it is not unlikely that two datasets have several categories in common (in total, there are 23 categories), which makes the Jaccard similarity score usually high.

A ground truth for the datasets domain

The Linked Open Data cloud[14] diagram describes datasets that have been published as Linked Data based on metadata collected and curated by contributors to the Data Hub.

We constructed the ground truth from a fragment of the August 2014 version diagram. Each circle represents a dataset and the circle size indicates the number of edges connected to each dataset. The circle color indicates the dataset community. In this version of the diagram, there is a total of 568 datasets, classified into 9 communities: Government (136 datasets), Publications (133 datasets), Social Networking (89 datasets), Life Sciences (63 datasets), User generated content (42 datasets), Cross-domain (40 datasets), Geographic (24 datasets), Media (21 datasets) and Linguistics (20 datasets).

We extracted the 568 datasets from the 2014 LOD Cloud (see Figure 21), parsing the SVG file to read the circles color to identify the datasets communities (see Figure 22). Then, we selected 7 meaningful communities, discarding the Cross-Domain and the Linguistic communities, since they mix different dataset domains.

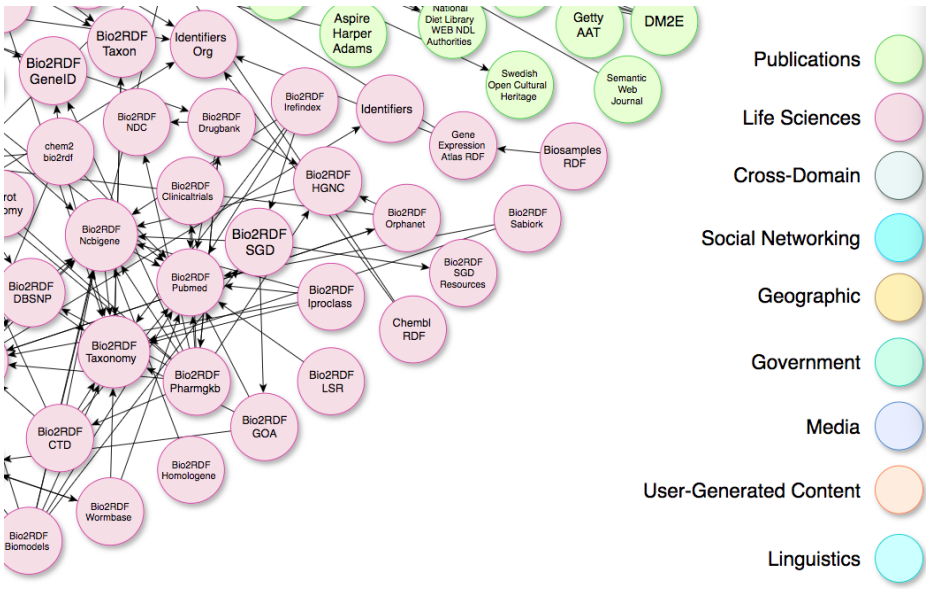

Figure 22: LOD cloud diagram fragment.

Evaluating the results

[14] http://lod-cloud.net/

For the final evaluation, we considered only the 125 datasets present in both the LOD ground truth (with 568 datasets) and in our set of datasets (with 390 datasets) (see Figure 21).

We consider as baselines two well-known similarity measures: Jaccard distance and Cosine distance. We compared the datasets with each other using the three different similarity measures: Jaccard distance, Cosine distance (the two baselines) and RBO. We generated a distance matrix representing the distance between all dataset pairs. From such distances, we generated 7 clusters, using the Hierarchical Agglomerative algorithm with different linkage criteria: Single, Complete, Ward and Average.

To evaluate the proposed entity similarity metric for the datasets domain, we clustered the datasets using the proposed entity similarity metric and compared the clusters thus obtained with the ground truth communities (or clusters). We assumed that the most similar datasets should belong to the same category.

There is a wide variety of clustering algorithms (hierarchical agglomerative, centroid-based, among others). Besides the similarity measure used to compare the items (which can be, for instance, *Jaccard*, *Cosine* or *RBO*), such algorithms also depend on a method for grouping the items, called *linkage criteria*, listed as follows:

1. *Single Linkage* minimizes the minimum distance criterion between items in pairs of clusters (see Figure 23(a))
2. *Complete Linkage* minimizes the maximum distance between items in clusters (see Figure 23(b)).
3. *Average Linkage* minimizes the average distance between items in clusters (see Figure 23(c)).
4. *Ward Linkage* minimizes the sum of squared differences between items.

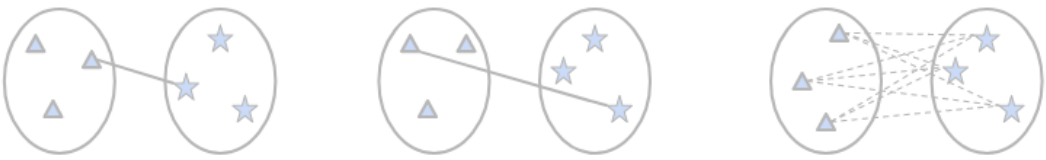

Figure 23: (a) Single Linkage; (b) Complete Linkage; (c) Average Group.

As adopted in (GARCÍA *et al.*, 2017), we chose the hierarchical agglomerative clustering algorithm (MURTAGH *et al.*, 2014), which starts with each dataset as a single cluster and then merges pairs of clusters, using similarity measures, until achieving the desired number of clusters.

To evaluate the clustering performance, we used the Adjusted Rand Index (ARI) (see Section 2.4) (YEUNG *et al.*, 2001). Table 12 shows the ARI values by considering two types of parameters: (a) the similarity measure used to compare the datasets before clustering them (represented by the lines of Table 12); and (b) the clustering linkage metric used to merge the clusters (represented by the columns of Table 12).

Table 12: Adjusted Rand Index of the clustering algorithms.

	Single	**Complete**	**Average**	**Ward**
Jaccard	0.018	0.170	0.242	0.142
Cosine	0.161	0.277	0.284	0.267
RBO, *p* = 0.98	0.008	0.281	0.302	0.298
RBO, *p* = 0.99	0.008	0.205	0.149	0.273

The worst performance, measure by the ARI index, was obtained using the Jaccard distance. This was expected, since Jaccard considers only the presence or the absence of an item in the lists. The cosine distance performed better than Jaccard, since it considers the frequency of the categories that describe the datasets. The best performances (0.302 and 0.298) were obtained using *RBO* (with $p = 0.98$) as similarity measure, and the *Average* and *Ward* as clustering linkage metrics, respectively.

Figure 24 shows the confusion matrix (a confusion matrix compares two classification models) for the best performing RBO case (with ARI index 0.302, for $p = 0.98$ and *Average* as the clustering linkage metric) as compared with the ground truth classification. The lines of the confusion matrix correspond to the communities (or clusters) of the ground truth. They represent, from 0 to 6, respectively, Government, Geographic, Publications, Life Sciences, User Generated content, Social networking and Media. The columns, in turn, represent the clusters found in the experiment, using the best performing RBO case.

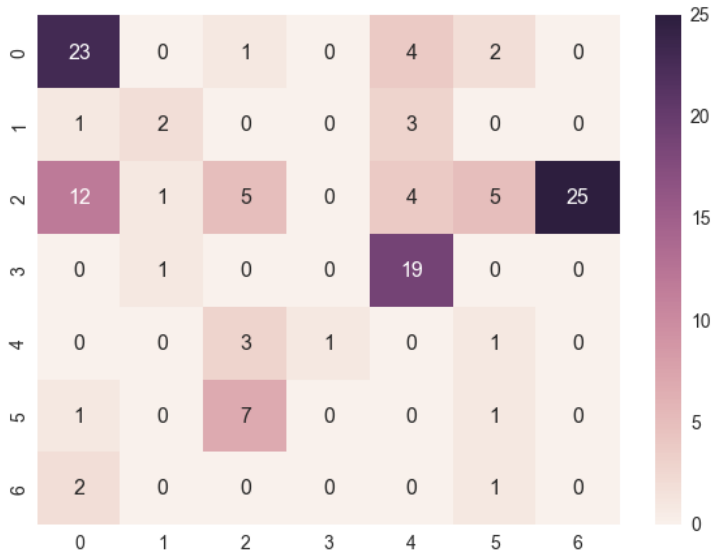

Figure 24: Confusion matrix for the best performance case of the dataset experiment.

Analyzing the quality of the generated clusters, we note that Government (line 0), Publications (line 2), Life Sciences (line 3) and Social networking (line 5) were clearly recognized as communities. For the Government community, from a total of 30 datasets in the ground truth, 23 were assigned to the same cluster. For the Publications community (line 2), 25 out of 52 were assigned to the same cluster. For the Life Sciences (line 3), 19 from 20 datasets were in the same clusters. The other communities (Geographic, User Generated content, Social networking and Media) were not recognized. A possible reason for this lies in their low density in the ground truth. In fact, the Social Networking community has 9 datasets, the Geographic community has 6 datasets, the User Generated content community has 5 datasets**,** and the Media community has only 3 datasets.

This experiment demonstrates that the best performing algorithm is that which consider the entity features as ranked lists, in our case, the RBO metric.

4.3.2. Comparing DBLP Computer Science conferences

Computer Science Conferences in DBLP

The DBLP repository[15] stores Computer Science bibliographic data for more than 4,500 conferences and 1,500 journals. DBPL is a joint service of the University of Trier and the Schloss Dagstuhl. Table 13 shows DBLP statistics in August 2017. In this experiment, we extracted computer science conferences from the DBLP repository to instantiate the SELECTOR framework.

Table 13: DBLP statistics in August 2017.

Entity type	Number of entities
Publications	3,859,721
Authors	1,946,939
Conferences	5,163
Journals	1,544

Extracting and ranking features

We considered the keywords extracted from the papers published in a conference as features to describe the conference. Basically, we extracted the stem-words from the keywords, in order to cope with different variations of the same root term. For instance, the *retriev* stem-word matches both with the *retrieval* and with the *retrieving* keywords; analogously, the *relev* stem-word matches both with *relevant* and with the *relevance* keywords. This strategy is detailed in (GARCÍA *et al.*, 2016).

Table 14 the top 16 stem-words (out of 1,847 stem-words) extracted from papers from the SIGIR conference[16], the International ACM Conference on Research and Development in Information Retrieval.

Table 14: SIGIR top stem-words.

stem-word	frequency
retriev	808
search	600
inform	551
queri	475

[15] http://dblp.uni-trier.de/

[16] http://dblp.uni-trier.de/db/conf/sigir/

model	467
web	317
base	261
document	255
text	244
evalu	239
rank	232
language	220
relev	216
learn	203
user	189
cluster	179

A ground truth for the conferences domain

We chose as ground truth for the experiment the list of academic Computer Science conferences defined in Wikipedia[17], with 248 conferences grouped into 13 groups. Although the 13 groups are subdivided into smaller groups, we considered only the 13 more general groups available, listed in Table 15, together with the number of conferences of each group.

Table 15: Computer science conference groups

Group	**# conferences**
Artificial intelligence	38
Computer networking	35
Languages and software	27
Algorithms and theory	27
Computer architecture	25
Concurrent, distributed and parallel computing	24
Data Management	21
Security and privacy	14
Computer graphics	9
Human–computer interaction	9
Operating systems	8
Education	6
Computational biology	5

[17] https://en.wikipedia.org/wiki/List_of_computer_science_conferences

Computing entity similarity

Our strategy to compare the conferences was analogous to the strategy adopted for the dataset experiment (see Section 4.3.1).

First, we compared all 248 conferences using similarity measures to generate a similarity matrix, with 61,256 cells, from which 30,380 cells (the lower triangular part of the matrix) are filled with the similarity between the pairs of distinct conferences. Then, we clustered the conferences using the hierarchical agglomerative clustering algorithm.

Evaluating the results

We consider again as baselines the well-known similarity measures: Jaccard distance and Cosine distance. To evaluate the clustering process, we again adopted the Adjusted Rand Index (ARI). Table 16 shows the ARI values for two parameters: (a) the similarity measure used to compare the conferences before clustering them; and (b) the clustering linkage metric used to merge the clusters when executing the hierarchical agglomerative clustering.

Table 16: ARI for the clustering algorithms comparing conferences.

	Single	**Complete**	**Average**	**Ward**
Jaccard	0.343	0.599	0.612	0.586
Cosine	0.343	0.589	0.630	0.713
RBO, *p* = 0.97	0.464	0.598	**0.794**	0.602
RBO, *p* = 0.98	0.562	0.661	0.742	0.727
RBO, *p* = 0.99	0.361	0.670	0.754	0.727

The worst performance was again obtained using the Jaccard distance. The cosine distance had better results when combined with Ward as linkage criteria (ARI=0.713). Note that, in general, the RBO had the best performances. The best overall performance, with ARI=0.794, was obtained using RBO as similarity measure, with p=0.97, and *Average* as clustering linkage metric.

We performed the experiments using Python with Jupyter in a Macbook air 1,6 GHz Intel Core i5 4 GB 1600MHz DDR3. Using the Jaccard distance, it took 27 seconds to construct the 248×248 similarity matrix (for the 248 conferences). Using RBO, it took around 210 seconds. Using the Cosine distance,

it took around 25 hours. By contrast, regarding the previous two experiments, since the number of entities to be compared was considerable smaller (12 museums and 125 datasets), the computational cost of the similarity measures was negligible. Figure 25 shows the confusion matrix for the best performance case (ARI = 0.794). The lines correspond to the ground truth clusters and the columns refer to the clusters generated by the best performing case.

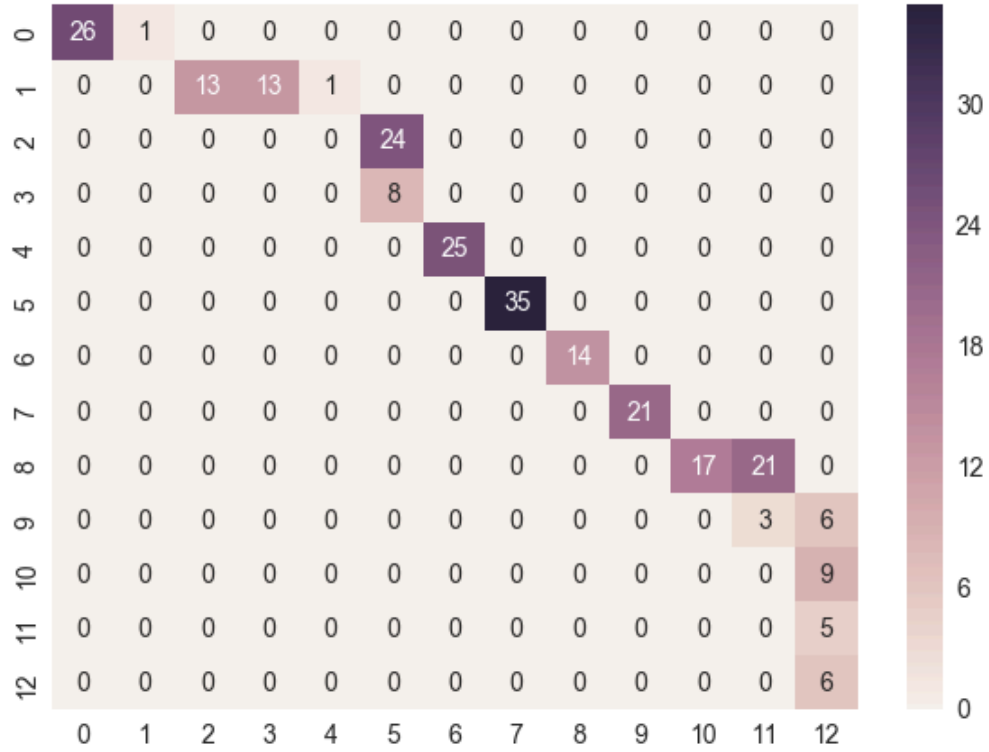

Figure 25: Confusion matrix for the best performance case for the conferences experiment.

By analyzing the quality of the generated clusters, we notice that four conference groups were entirely identified by the clustering process: The *Computer architecture* group (line 4), with 25 conferences, the *Computer networking* group (line 5), with 35 conferences, the *Security and privacy* group (line 6), with 14 conferences, and *Data Management,* with 21 conferences (line 7). The *Algorithms and theory* group (line 0) had 26 out of its 27 conferences assigned to cluster 0 and one conference to cluster 1. From the 27 conferences of the *Languages and software* group (line 1), 13 were assigned to cluster 2, 13 to cluster 3, and one conference to cluster 4. All 24 conferences of the *Concurrent, distributed and parallel computing* group (line 2) went to cluster 4, together with

8 other conferences from the *Operating systems* group (line 3). The biggest group, *Artificial intelligence* (line 8, 38 conferences), had 17 conferences in cluster 10, and 21 in cluster 11. The *Computer graphics* group (line 9) had 3 conferences assigned to cluster 11 and the other 6 to cluster 12. Finally, the last three groups, *Human–computer interaction*, with 9 conferences (line 10), *Computational biology*, with 5 conferences (line 11), and *Education*, with 6 conferences (line 12), were merged into only one cluster, cluster 12.

This last experiment also demonstrates that the best performing algorithm is that which consider the entity features as ranked lists, in our case, the RBO metric. Also, it also demonstrates that the cosine distance would be a reasonable option, only if the number of entities is fairly small.

4.4. Conclusions

This chapter addressed the problem of estimating semantic entity similarity using entity features available as Linked Data, on proposing SELECTOR, a two-module framework to estimate entity similarity. We assessed the accuracy of the proposed similarity metrics by instantiating the framework, in three different domains, and by carrying out detailed experiments.

In the first experiment, we compared museums represented in DBpedia. First, we chosen as features DBpedia categories to describe museums, but we found that they are very generic to describe museums. We then found that the art movements of the museums' artworks are high quality features. In the second experiment, we compared datasets represented in a Linked Data repository, using their Wikipedia top-level categories as features. In the last experiment, we compared computer science conferences, also provided as Linked Data in the DBLP repository, using the keywords extracted from their publications as features. We achieved better results with ranked lists of features than chosen baselines in all experiments.

We used WLM as a baseline for the first experiment, but not for the second experiment (comparing datasets) or for the third experiment (comparing conferences), because WLM uses the Wikipedia structure (or its RDF version) to compute the entity similarity, and the entities used in the second and third

experiments do not have an entry in Wikipedia (and hence in DBpedia). Instead, we compared RBO both with Jaccard and cosine, other widely used similarity metrics. n the first experiment (with museums on DBpedia), we used NDCG to evaluate the results, since we compare lists of similar museums generated by our approach with the lists generated by the ground truth. However, in the second and in the third experiment, we chose clustering techniques to evaluate the results.

Comparing with other approaches in the literature that compute similarity between Linked Data entities, our main contribution is that we take advantage of Linked Data in the similarity computation, by ranking the relevant features of the entities to be compared.

5
Comparing Semantic Trajectories

This chapter focuses on **RQ3**. *How can we compare semantic trajectories by considering their semantic dimension, extracted from Linked Data?* We first introduce a case study of a real trajectory repository with data collected from tourists in Italian cities, provided by TRIPBUILDER (BRILHANTE *et al.*, 2013) (Section 5.1). We then present how to describe and analyze a single trajectory using its POIs (Section 5.2). We next present different strategies to compare two semantic trajectories, focusing on the semantic facet (Section 5.3). Finally, we analyze a group of semantic trajectories from a trajectories repository (Section 5.4).

5.1.
Introduction and running example

We proposed in Section 4.2 an approach to compare different Linked Data entities, such as museums, that can represent trajectory POIs. This idea can be useful to compare different trajectories by comparing their POIs. However, the trajectory dataset we used in this part of the thesis does not contain trajectories, which are rich enough with respect to their entities, like museums, even if they have a reasonable number of POIs. Therefore, although the approach we propose to compare entities could be applied to compare different trajectories, we did not use it in this part of the thesis. Instead, we considered a trajectory as an *entity* to be compared with other *entities* (other trajectories).

We illustrate the trajectory knowledge discovery process with a running example of trajectories of tourists covering three Italian cities: Pisa, Florence and Rome. We use the TRIPBUILDER trajectory dataset (BRILHANTE *et al.*, 2013),

taken from user-generated content obtained from Flickr, combined with Wikipedia. The collection and pre-processing steps are detailed in (BRILHANTE *et al.*, 2013) and summarized in Section 3.2.1.

Table 17: Statistics about the trajectories dataset (BRILHANTE *et al.*, 2013)

City	# POIs	# Users	# Photos	# Traj.
Pisa	112	1825	18,170	3,430
Florence	891	7049	102,888	16,522
Rome	490	13772	234,616	35,522

Table 17 shows the statistics related to each of the cities: Pisa, Florence and Rome. The columns are detailed as follows:

1. POIs: Points of interest of the city, collected from Wikipedia, with the attributes *userId*, *name*, *latitude* and *longitude*.
2. Users: Users from Flickr.
3. Photos: Geo-referred photos taken from Flickr, with *userId*, *photoId*, *dateTaken*, *dateUpload*, *latitude* and *longitude*.
4. Traj: Trajectories preprocessed by: assigning a cluster of each photo, creating the users' history and splitting into trajectories. Contains the attributes *userId* and *trajectory*.

As expected, the city with fewer pictures collected was Pisa (the smaller of the three cities) and the city with more pictures was Rome (the bigger of the cities). However, Florence has more POIs than Rome.

The TRIPBUILDER dataset also contains two other data: (a) the *compound POIs* and (b) the list of *categories* related to each POI. For instance, the open-air sculpture gallery of antique and Renaissance art *Loggia della Signoria* comprises the following POIs, representing its sculptures: *Perseo con la testa di Medusa*; *Loggia della Signoria*; *Patroclo e Menelao*; *Ratto delle Sabine_(Giambologna)*; *Ercole e il Centauro Nesso*; *Ratto di Polissena*; *Ercole e Caco*.

Besides, this POI is related, in Wikipedia, to the following three categories: *scultureafirenze* (sculptures in Florence); *loggedifirenze* (Loggia in Florence)*; operedigiambologna* (works of Giambologna).

5.2. Representing a single trajectory

In this section, we discuss two different ways to represent a single trajectory. First, we consider a trajectory as a *set* of POIs (Points-of-interest), i.e., we see a trajectory as a set of POIs in which the elements are the visited during the journey, ignoring any information regarding the sequence of the POIs. It is important to notice, however, that we consider a trajectory with its POIs as a *semantic* trajectory, since it is enriched with contextual data (Wikipedia in this case), which is not part of the raw trajectory.

When representing a single trajectory, we may consider, besides its POIs, the categories to which they belong, which may describe the interests in general of the tourist – for instance, churches, museums, gardens, among others. Besides that, when comparing two trajectories (to be discussed in the next section), we can find a similarity between them by considering their categories, even when the trajectories do not share any POI in common. We consider, inspired by (FURTADO *et al.*, 2016), as an alternative way to represent trajectories, the POI categories. However, we go a step ahead in a way that each category has a different weight according to its relevance.

Consider the raw trajectory *T1* in Florence, shown in Figure 26. The tourist trajectory is present in the TRIPBUILDER trajectory dataset (BRILHANTE *et al.*, 2013) and the tourist stopped at 17 POIs in the city.

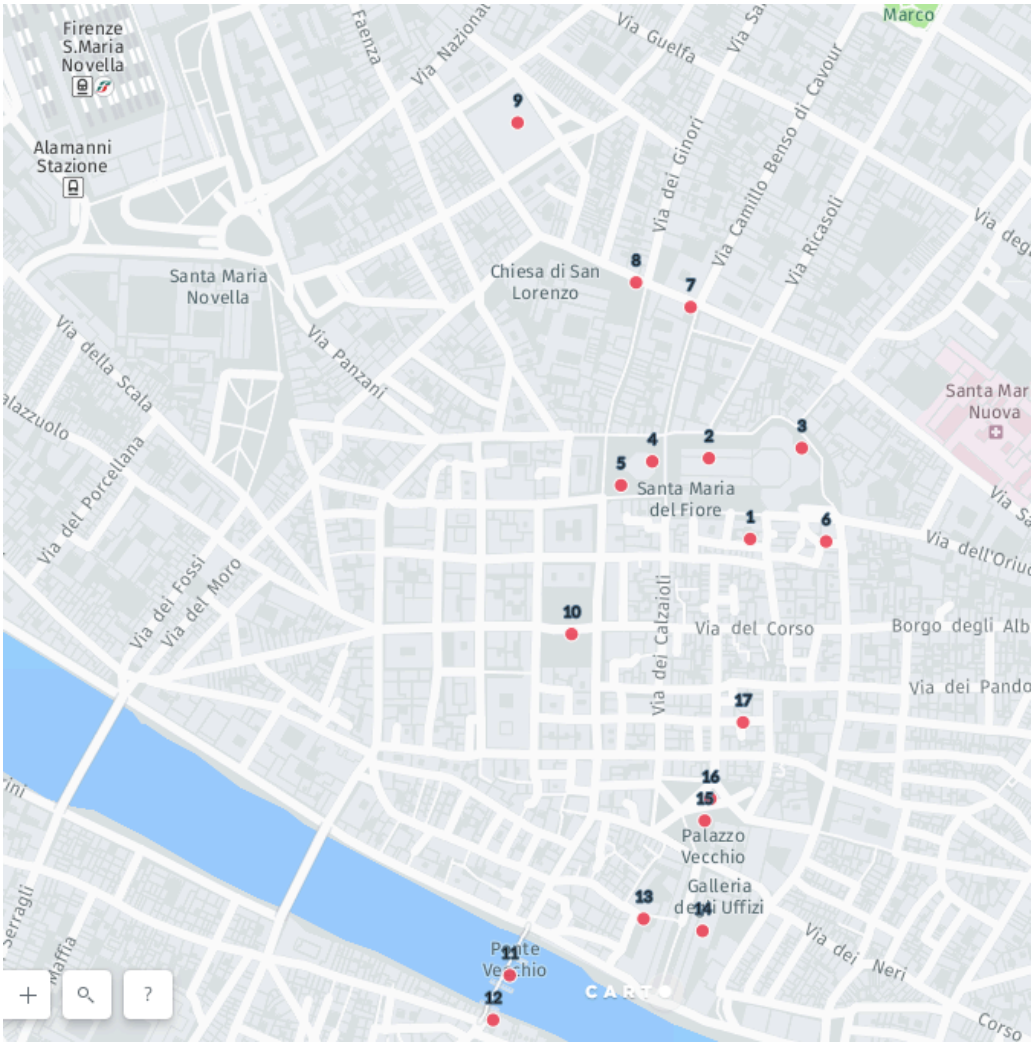

Figure 26: A raw trajectory T1 in Florence with 17 geo-referenced points.

Now, consider the raw trajectory shown in Figure 26, but enriched with metadata about its POIs. Table 18 shows the semantic trajectory *ST1* represented as the set of its POI labels (each table line represents a POI). Note that some POIs are compound, i.e., they comprise two or more other POIs.

Table 18: A semantic trajectory *ST1* in Florence as a set of POIs.

Trajectory *ST1*
Palazzo_del_Capitolo_dei_Canonici, Piazza_del_Capitolo
Chiesa_di_Santa_Reparata, Cattedrale_di_Santa_Maria_del_Fiore
Cupola_del_Brunelleschi
Porta_del_Paradiso, Battistero_di_San_Giovanni_(Firenze)
Piazza_San_Giovanni_(Firenze)
Piazza_delle_Pallottole
Chiesa_di_San_Giovannino_degli_Scolopi

Monumento_a_Giovanni_delle_Bande_Nere
Mercato_Centrale_(Firenze)
Piazza_della_Repubblica_(Firenze), Mercato_Vecchio
Monumento_a_Giovanni_delle_Bande_Nere
Ponte_Vecchio
Torre_dei_Mannelli
Torre_dei_Pulci
Galleria_degli_Uffizi
Fontana_del_Nettuno_(Firenze), Piazza_della_Signoria
Statua_equestre_di_Cosimo_I_de_Medici
Tabernacolo_della_Quarquonia, Ospizio_della_Quarconia, Torri_dei_Galigai

Finally, consider the semantic trajectory *ST1* represented by its POIs together with their Wikipedia categories. Table 19 shows the semantic trajectory including the POIs categories.

Table 19: A semantic trajectory as a set of POIs with their categories.

Trajectory *ST1*
Palazzo_del_Capitolo_dei_Canonici, Piazza_del_Capitolo palazzidifirenze, chiesedifirenze, piazzedifirenze
Chiesa_di_Santa_Reparata, Cattedrale_di_Santa_Maria_del_Fiore chiesedifirenze, monumentidifirenze, cattedralidellaprovinciadifirenze, architetturedifirenze, duomo
Cupola_del_Brunelleschi cattedralidellaprovinciadifirenze, architetturedifirenze, cupole
Porta_del_Paradiso, Battistero_di_San_Giovanni_(Firenze) scultureafirenze, architetturedifirenze, battisteridellatoscana, chiesedifirenze
Piazza_San_Giovanni_(Firenze) piazzedifirenze
Piazza_delle_Pallottole piazzedifirenze
Chiesa_di_San_Giovannino_degli_Scolopi chiesedifirenze
Monumento_a_Giovanni_delle_Bande_Nere monumentidifirenze, scultureafirenze, fontanedifirenze
Mercato_Centrale_(Firenze) architetturedifirenze, mercatidifirenze
Piazza_della_Repubblica_(Firenze), Mercato_Vecchio piazzedifirenze, architetturedifirenze, mercatidifirenze
Monumento_a_Giovanni_delle_Bande_Nere monumentidifirenze, scultureafirenze, fontanedifirenze
Ponte_Vecchio pontidifirenze
Torre_dei_Mannelli torridifirenze
Torre_dei_Pulci

torridifirenze
Galleria_degli_Uffizi museidifirenze, pinacotecheditalia, uffizi
Fontana_del_Nettuno_(Firenze), Piazza_della_Signoria operedigiambologna, piazzedifirenze, fontanedifirenze, scultureafirenze
Statua_equestre_di_Cosimo_I_de_Medici monumentidifirenze, scultureafirenze, operedigiambologna
Tabernacolo_della_Quarquonia, Ospizio_della_Quarconia, Torri_dei_Galigai salecinematografichedifirenze, tabernacolidifirenze, ospedalidifirenze, teatridifirenze, torridifirenze

Considering the POIs categories shown in Table 19, we may want to aggregate those categories considering the frequency they occur in the trajectory. An option would be to represent the trajectory using the idea of ranked features introduced in Section 4.1.1. However, the ranked lists generated for objects (i.e., trajectories) in this domain tend to have many POI categories with the same frequency (as can be seen in Table 20). In this case, since one of the analysis tasks is to compare trajectories, a better alternative would be to represent a trajectory as the frequency vector of the categories of its POIs. Besides that, the Wikipedia categories found in the TRIPBUILDER POIs are, usually, generic categories – such as architecture, squares and churches, representing general preferences of tourists. Table 20 shows a frequency vector that represents the semantic trajectory *ST1*.

Table 20: A semantic trajectory *ST1* represented as the frequency vector of the categories of its POIs.

POI category	**frequency**
architetturedifirenze	5
piazzedifirenze	5
chiesedifirenze	4
scultureafirenze	4
monumentidifirenze	3
torridifirenze	3
cattedralidellaprovinciadifirenze	2
fontanedifirenze	2
mercatidifirenze	2
operedigiambologna	2
battisteridellatoscana	1
cupole	1
duomo	1
museidifirenze	1

ospedalidifirenze	1
palazzidifirenze	1
pinacotecheditalia	1
pontidifirenze	1
salecinematografichedifirenze	1
tabernacolidifirenze	1
teatridifirenze	1
uffizi	1

The frequency vector shown in Table 20 may describe the interests in general of the tourist – in this case, architecture (*architetturedifirenze*), squares (*piazzedifirenze*), churches (*chiesedifirenze*) and sculptures (*scultureafirenze*) are the top 4 interests of the tourist.

5.3. Computing the Semantic Similarity between trajectories

Measuring the similarity between trajectories is one of the most important tasks when dealing with trajectory data, since it serves as the foundation of several types of analyses such as similarity search, clustering, and classification (WANG *et al.*, 2013).

As a starting point to compare trajectories, we propose to use their sets of POIs and consider the two trajectory representations described in the last section: (i) a trajectory as a set of POIs; or (ii) a trajectory as a feature vector of the POI categories.

Consider again the semantic trajectory *ST1* presented in the last section, and another trajectory in Florence, *ST2*, both present in the TRIPBUILDER trajectory dataset (BRILHANTE *et al.*, 2013). Table 21 shows those trajectories.

Table 21: Comparing two trajectories represented as sets of POIs.

Trajectory *ST1*	**Trajectory *ST2***
Palazzo_del_Capitolo_dei_Canonici, Piazza_del_Capitolo palazzidifirenze, chiesedifirenze, piazzedifirenze *Chiesa_di_Santa_Reparata, Cattedrale_di_Santa_Maria_del_Fiore* chiesedifirenze, monumentidifirenze, cattedralidellaprovinciadifirenze, architetturedifirenze, duomo *Cupola_del_Brunelleschi*	*Stazione_di_Firenze_Santa_Maria_Novella* architetturedifirenze, stazioniferroviariedifirenze *Piazza_Bambine_e_Bambini_di_Beslan* piazzedifirenze *Padiglione_Spadolini* architetturedifirenze *Fortezza_da_Basso* fortezzedellatoscana, firenzefiera,

cattedralidellaprovinciadifirenze, architetturedifirenze, cupole *Porta_del_Paradiso, Battistero_di_San_Giovanni_(Firenze)* scultureafirenze, architetturedifirenze, battisteridellatoscana, chiesedifirenze *Piazza_San_Giovanni_(Firenze)* piazzedifirenze *Piazza_delle_Pallottole* piazzedifirenze *Chiesa_di_San_Giovannino_degli_Scolopi* chiesedifirenze *Monumento_a_Giovanni_delle_Bande_Nere* monumentidifirenze, scultureafirenze, fontanedifirenze *Mercato_Centrale_(Firenze)* architetturedifirenze, mercatidifirenze *Piazza_della_Repubblica_(Firenze), Mercato_Vecchio* piazzedifirenze, architetturedifirenze, mercatidifirenze *Monumento_a_Giovanni_delle_Bande_Nere* monumentidifirenze, scultureafirenze, fontanedifirenze *Ponte_Vecchio* pontidifirenze *Torre_dei_Mannelli* torridifirenze *Torre_dei_Pulci* torridifirenze *Galleria_degli_Uffizi* museidifirenze, pinacotecheditalia, uffizi *Fontana_del_Nettuno_(Firenze), Piazza_della_Signoria* operedigiambologna, piazzedifirenze, fontanedifirenze, scultureafirenze *Statua_equestre_di_Cosimo_I_de_Medici* monumentidifirenze, scultureafirenze, operedigiambologna *Tabernacolo_della_Quarquonia, Ospizio_della_Quarconia, Torri_dei_Galigai* salecinematografichedifirenze, tabernacolidifirenze, ospedalidifirenze, teatridifirenze, torridifirenze	architetturedifirenze *Giardini_della_Fortezza* giardinidifirenze *Porta_a_Faenza* architetturedifirenze *Piazza_Santa_Maria_Novella* piazzedifirenze *Monumento_funebre_dellantipapa_Giovanni_XXIII* scultureafirenze, monumentidifirenze *Palazzo_Del_Bembo* palazzidifirenze *Piazza_San_Giovanni_(Firenze)* scultureafirenze, monumentidifirenze *Loggia_del_Bigallo* museidifirenze, loggedifirenze *Torre_dei_Caponsacchi, Caffè_Le_Giubbe_Rosse* caffèstoricidifirenze, architetturedifirenze, torridifirenze *Bottegone_(Firenze)* esercizistoricidifirenze *Fontana_del_Nettuno_(Firenze), Piazza_della_Signoria* operedigiambologna, piazzedifirenze, fontanedifirenze, scultureafirenze *Perseo_con_la_testa_di_Medusa, Loggia_della_Signoria, Patroclo_e_Menelao, Ratto_delle_Sabine_(Giambologna), Ercole_e_il_Centauro_Nesso, Ratto_di_Polissena, Ercole_e_Caco* scultureafirenze, loggedifirenze, operedigiambologna *Torre_dei_Pulci* torridifirenze

When comparing the semantic trajectories shown in Table 21, considering only the presence or absence of the POIs, we may use the *Jaccard index* since it takes into consideration the proportion of POIs the trajectories share, i.e., the intersection of POIs over the union of POIs between the two trajectories (See section 2.3).

In our example, the Jaccard index returns, in a range from 0 to 1, the score 0.1 (the trajectories have only three common POIs: (i) *Fontana_del_Nettuno_(Firenze)*; *Piazza_della_Signoria*, (ii) *Torre_dei_Pulci* and (iii) *Piazza_San_Giovanni _(Firenze)*). The Jaccard index is, as expected, very low, since the trajectories have few POIs in common and, intuitively, are not very similar considering the tourist preferences (the first tourist has a behavior of a more *art/museum lover* tourist then the second tourist).

It is important to mention that on comparing two trajectories geographically far from each other (for instance, in two different cities), the Jaccard index between their POIs is always 0, since they do not have any POI in common. This fact also motivated us to consider the POIs categories in the comparison (the second alternative). Table 22 presents the frequency vectors of the POIs categories of *ST1* and *ST2*.

Table 22: Two semantic trajectories as frequency vectors of POI categories

POI category	**frequency in *ST1***	**frequency in *ST2***
architetturedifirenze	5	5
piazzedifirenze	5	4
chiesedifirenze	4	-
scultureafirenze	4	3
monumentidifirenze	3	1
torridifirenze	3	2
cattedralidellaprovinciadifirenze	2	-
fontanedifirenze	2	1
mercatidifirenze	2	-
operedigiambologna	2	2
battisteridellatoscana	1	-
cupole	1	-
duomo	1	-
museidifirenze	1	1
ospedalidifirenze	1	-
palazzidifirenze	1	1
pinacotecheditalia	1	-
pontidifirenze	1	-
salecinematografichedifirenze	1	-
tabernacolidifirenze	1	-
teatridifirenze	1	-
uffizi	1	-

stazioniferroviariedifirenze	-	1
fortezzedellatoscana	-	1
firenzefiera	-	1
giardinidifirenze	-	1
loggedifirenze	-	1
caffèstoricidifirenze	-	1
esercizistoricidifirenze	-	1
operedigiambologna		

When comparing the semantic trajectories as frequency vectors shown in Table 22, we may use the cosine index (See section 2.3). In this case, the cosine index is 0.771. Note that, as expected, the score similarity for the second approach (with the frequency vectors) is higher than the score of the first approach (with the sets of POIs), since it widens the range of possibilities to find similarities not found considering only the POIs.

We may conclude that, therefore, if we want to compare two trajectories located in the same city, considering only their Points-of-Interest in common, the Jaccard index between their POIs might be a reasonable solution. Otherwise, if the trajectories are not taken in the same region or if we want to consider other aspects of the POIs – such as the places categories, the better strategy is to represent the frequencies of the POIs categories as a frequency vector and then compare them using for instance, the cosine index.

5.4. Analyzing groups of trajectories

A repository containing semantic trajectories enables a variety of analysis to understand the behavior of moving objects. For instance, by analyzing similar animal trajectories, it is possible to determine their migration patterns. Considering a city traffic monitoring system, it is possible to locate popular vehicle routes (WANG *et al.*, 2013).

In the context of tourism, it is possible to find popular tourist trajectories, to find the most popular POIs (i.e., the most visited POIs), and to analyze the most popular POI categories, to mention only a few examples.

Table 23 shows the most popular POIs found in the city of Florence, present in the TRIPBUILDER trajectory dataset (BRILHANTE *et al.*, 2013). The total number of trajectories in the city is 16522 and the total number of POIs, 891).

Table 23: Most popular POIs in Florence.

POI cluster	frequency
(1) Loggia_della_Signoria	1797
(2) Piazza_della_Signoria	1667
(3) Battistero_di_San_Giovanni	1647
(4) Ponte_Vecchio	1498
(5) Campanile_di_Giotto	935
(6) Porta_della_Mandorla	900
(7) Piazzale_Michelangelo	888
(8) Torre_dei_Pulci	716
(9) Palazzo_Vecchio	701
(10) Cattedrale_di_Santa_Maria_del_Fiore	677
(11) Corridoio Vassariano	648
(12) Cupola_del_Brunelleschi	622
(13) Basilica_di_Santa_Croce	601

Note that some POIs are located very close to each-other, and can be regarded as part of the same complex of monuments or buildings. For instance, POI (1), *Loggia della Signoria* is a building on a corner of the *Piazza della Signoria,* POI (2), where the town hall of Florence, *Palazzo Vecchio*, POI (9), is also located. In addition to these three POIs, other four POIs are geographically very close to each other, and located in the most visited square in Florence: *Battistero di San Giovanni* (3), *Campanile di Giotto* (5), *Porta della Mandorla* (6) and *Cattedrale_di_Santa_Maria_del_Fiore* (10).

In order to assess the *popularity* feature of those POIs, we consider as ground truth the TripAdvisor[18] Web site, the most used trip Web site, according to the *comscore* report of June 2017[19]. Table 24 shows the top places to visit in

[18] https://www.tripadvisor.com

[19] http://www.comscore.com/Insights/Rankings

Florence recommended by a Trip Advisor article[20], and the corresponding POIs in the TRIPBUILDER trajectory dataset.

Note that 8 out of the 13 popular POIs in the TRIPBUILDER dataset are also in the Trip Advisor recommendations list. We also compared the set of places recommended by Trip Advisor with the whole TRIPBUILDER dataset, using the Jaccard index, that ranges from 0 to 1, which returned as an average score of 0.032. The score was very low because we consider all trajectories in the repository, which contains a lot of *poor trajectories* – i.e, with few POIs.

Table 24: Matching popular POIs with Trip Advisor recommendations.

Place to visit by Trip Advisor	POI matched in TripBuilder
(1) Church of Santa Maria Novella	-
(2) Mercato Centrale	-
(3) Duomo – Cattedrale di Santa Maria del Fiore	10
(4) Campanile di Giotto	5
(5) Baptistery of San Giovanni (Battistero)	3
(6) Sasso di Dante (restaurant)	-
(7) Piazza della Signoria	1, 2 and 9
(8) Palazzo Pitti	-
(9) Uffizi Gallery	-
(10) Ponte Vecchio	4
(11) Piazzale Michelangelo	7

We also considered another popular tourist guide: the Lonely Planet[21]. Table 25 shows the places in Florence recommended in their Web site article[22]. The Lonely Planet suggestions differ from those in Trip Advisor. We found 4 distinct TRIPBUILDER POIs out of the 13 in their suggestions. We again compared the set of places recommended by Lonely Planet with the TRIPBUILDER dataset using the Jaccard index, which returned, as expected, a lower average score 0.010 (as compared with 0.032 for Trip Advisor). These results intuitively indicates that the Trip Advisor recommendations better reflect the places tourists actually visit in

[20] https://www.tripadvisor.com/Guide-g187895-k265-Florence_Tuscany.html

[21] https://www.lonelyplanet.com/

[22] https://www.lonelyplanet.com/italy/florence/attractions/a/poi-sig/360059

the City of Florence than the Lonely Planet recommendations (in the sense that they have more POIs in common).

Table 25: Matching popular POIs with Lonely Planet recommendations.

Place to visit by Lonely Planet	POI matched in TripBuilder
(1) Galleria degli Uffizi	-
(2) Basilica di Santa Maria Novella	-
(3) Palazzo Vecchio	9
(4) Museo di San Marco	-
(5) Duomo	10
(6) Museo dell'Opera del Duomo	10
(7) Cupola del Brunelleschi	10
(8) Galleria dell'Accademia	-
(9) Museo Novecento	10
(10) Museo delle Cappelle Medicee	4
(11) Basilica di Santa Croce	13

We also analyzed the POI categories found in the TRIPBUILDER dataset. Figure 27 shows a word cloud with the main categories related to the POIs found in Florence, comprising the 67 most frequent categories, out of the 123 (we omitted the others to facilitate visualization).

Figure 27 shows that the most frequent category is *PalazzidiFirenze* (Buildings in Florence), found in 8,908 trajectory POIs, followed by *ArchitetturediFirenze* (Architecture in Florence), that appears 8,167 times, and *SculptureaFirenze* (Sculptures in Florence), with frequency 7,775.

Figure 27: POIs' categories in Florence.

5.5. Conclusions

In this section, we discussed, using a running example of trajectories in Florence, two different ways to represent a single trajectory. First, as a *set* of POIs, visited during the journey, and then as a *frequency vector* with the relevance of each POI category. We also presented an approach to compare a pair of trajectories by using these two different representations, and focusing on the *semantic* aspects of the trajectories, ignoring the geo-spatial dimensions. We concluded that, when comparing two trajectories geographically far from each other, a better alternative to find some similarity between them is to compare the semantic aspects – such as their POIs categories. Otherwise, if the trajectories are in the same region, we might use both approaches (or combine them): compare their set of POIs or compare the frequency vectors of their POIs categories. Finally, we analyzed a group of trajectories, considering their POI categories, and compared them with the places recommended by two of the most famous trip guides: Trip Advisor and Lonely Planet. Contrasting with other approaches that compare trajectories, our main contribution is that we exploit the semantic dimension (that we extract from the Linked Open Data) in the comparison, while most of the literature so far focus on the spatial and temporal dimensions.

6
Conclusions and Future Works

6.1.
Conclusions

In the last years, a growing number of devices that track moving objects have been generating an impressive amount of mobility data, which support a deep understanding of movement behavior. Another unprecedented global space that is also growing fast is the Web of Data, thanks to the emergence of the Linked Data Initiative.

In this thesis, we presented novel approaches to study how the Linked Data Initiative can support the trajectory enrichment process to generate semantic trajectories, and how to compare these enriched trajectories.

First, we presented a conceptual framework aiming at guiding the whole trajectory enrichment process to generate a repository containing semantic trajectories, taking advantage of Linked Data to answer the research question **RQ1** detailed in Section 1.2. First, we represented semantic trajectories according to the Linked Data principles. Then, we used the Web of Data as the main source of contextual information to enrich movement data. We highlighted the different steps and how the availability of such repository improves the ability to formulate application analysis questions, thanks to the richness of the linked contextual data. This part of the thesis was guided by a running example of a semantic trajectory of a tourist in Florence, Italy.

Inspired by a number of interesting analysis that might be done over a semantic repository – i.e., a repository containing semantic trajectories, we investigated the similarity between POIs (Points-of-Interest) and other Linked Data entities. This work concerns the research question **RQ2**. In particular, we proposed SELECTOR, a two-module framework that takes as input Linked Data entities, ranks the lists of entity features according to their relevance for

describing the entities, and compares the ranked lists using rank correlation metrics. We performed experiments with our framework in three different domains. We first compared Linked Data entities that represents museums on DBpedia, and found that the art movements of the museums' artworks are high quality features. Then, in the second experiment, we compared datasets represented in a Linked Data repository, using their Wikipedia top-level categories as features. Finally, we compared computer science conferences, also provided as Linked Data in the DBLP repository, using the keywords extracted from their publications as features. We achieved better results than chosen baselines in all experiments.

Lastly, to answer the research question **RQ3**, we discussed different approaches to represent semantic trajectories: (i) as set of POIs; and (ii) as frequency vectors of their POIs categories. We then compared trajectories considering those two types of representations, considering that the trajectories to be compared may be geographically far from each other. Finally, we analyzed a group of trajectories, considering their POI categories, and comparing a group of trajectories with the places recommended by two known trip guides: Trip Advisor and Lonely Planet.

It is important to notice that, when running the experiments, we faced several challenges, mainly related to the data quality. In some cases, the dataset was considered *poor* in terms of the features to describe the entities. For instance, in the first experiments with museums on DBpedia (see Section 4.2), the categories describing the museums were very generic to distinguish a museum from another. Besides, although the trajectories found in the TRIPBUILDER dataset (see Section 5.1) comprise a lot of stops representing POIs, it not contains trajectories *rich* enough (with several POIs with features available as Linked Data, such as museums).

6.2. Future Works

We visualize as future work in the research fields covered by this thesis the following topics.

The ideas discussed in this thesis could be applied to several systems that aim at understanding urban mobility behaviors to recommend new places or routes to tourists, such as in recommender systems in the tourism domain, like Trip advisor[23] and Booking[24].

In this thesis, more specifically, we discussed how semantic trajectories might be compared considering their sets of POIs, ignoring any information regarding the POIs sequences. Therefore, a complete framework might also combine our strategy with the representation of trajectories as *sequences* of POIs, in order to understand the tourist behavior to find out when and how the order of POIs visited matters when comparing trajectories.

We might also compare groups of real trajectories geographically far from each other, for instance, located in different cities, by taking advantage of the second approach to represent trajectories – as frequency vectors of their POI categories. In these cases, the semantic aspects of the trajectory POIs (for instance the places categories) are fundamental to find similarities between trajectories. Other research directions include the similarity of trajectories enriched with aspects other than POIs with their categories, including social media data, such as the reviews or the opinions and other contextual data.

Another possibility of applying the ideas of this thesis in other domains would be to compare different *students trajectories* in universities, by considering the courses as *stops* and their academic life as whole trajectories. This approach would have to consider aspects such as the dependency between courses and would be useful to plan or recommend to the students the next courses to take, considering the academic trajectory of similar students.

[23] https://www.tripadvisor.com.br/

[24] https://www.booking.com

Bibliography

ALVES, Ana et al. Semantic enrichment of places: Ontology learning from web. **International Journal of Knowledge-based and Intelligent Engineering Systems**, v. 13, n. 1, p. 19-30, 2009.

AUER, Sören et al. Dbpedia: A nucleus for a web of open data. **The semantic web**, p. 722-735, 2007.

BACCIU, Clara et al. Accommodations in Tuscany as Linked Data. In: **International Conference on Language Resources and Evaluation** (**LREC**). 2014. p. 3542-3545.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data-the story so far. **Semantic services, interoperability and web applications: emerging concepts**, p. 205-227, 2009.

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

BRILHANTE, Igo et al. Where shall we go today?: planning touristic tours with tripbuilder. In: **Proceedings of the 22nd ACM international conference on Information & Knowledge Management**. ACM, 2013. p. 757-762.

BRILHANTE, Igo Ramalho et al. On planning sightseeing tours with TripBuilder. **Information Processing & Management**, v. 51, n. 2, p. 1-15, 2015.

BUDANITSKY, Alexander; HIRST, Graeme. Evaluating wordnet-based measures of lexical semantic relatedness. **Computational Linguistics**, v. 32, n. 1, p. 13-47, 2006.

BUCHIN, Maike; DODGE, Somayeh; SPECKMANN, Bettina. Similarity of trajectories taking into account geographic context. **Journal of Spatial Information Science**, v. 2014, n. 9, p. 101-124, 2014.

CARABALLO, Alexander Arturo Mera et al. Automatic Creation and Analysis of a Linked Data Cloud Diagram. In: **International Conference on Web**

Information Systems Engineering. Springer International Publishing, 2016. p. 417-432.

CASANOVA, Marco A. et al. On materialized sameAs linksets. In: **International Conference on Database and Expert Systems Applications**. Springer, Cham, 2014. p. 377-384.

CECCARELLI, Diego et al. Learning relatedness measures for entity linking. In: **Proceedings of the 22nd ACM international conference on Information & Knowledge Management**. ACM, 2013. p. 139-148.

CHAVES, Marcirio; FREITAS, Larissa; VIEIRA, Renata. Hontology: a multilingual ontology for the accommodation sector in the tourism industry. 2012.

DOU, Dejing; WANG, Hao; LIU, Haishan. Semantic data mining: A survey of ontology-based approaches. In: **Semantic Computing (ICSC), 2015 IEEE International Conference on**. IEEE, 2015. p. 244-251.

ENDRES-NIGGEMEYER, Brigitte. Semantic Mashups: Intelligent Reuse of Web Resources. **Springer Publishing Company**, Incorporated, 2013.

FAGIN, Ronald; KUMAR, Ravi; SIVAKUMAR, Dakshinamurthi. Comparing top k lists. **SIAM Journal on discrete mathematics**, v. 17, n. 1, p. 134-160, 2003.

FAGIN, Ronald et al. Comparing and aggregating rankings with ties. In: **Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems**. ACM, 2004. p. 47-58.

FILETO, Renato et al. A semantic model for movement data warehouses. In: **Proceedings of the 17th international workshop on data warehousing and OLAP**. ACM, 2014. p. 47-56.

FILETO, Renato et al. Baquara: A holistic ontological framework for movement analysis using linked data. In: **International Conference on Conceptual Modeling**. Springer, Berlin, Heidelberg, 2013. p. 342-355.

FURTADO, Andre Salvaro et al. Multidimensional similarity measuring for semantic trajectories. **Transactions in GIS**, v. 20, n. 2, p. 280-298, 2016.

GARCÍA, Grettel Monteagudo et al. Comparing and recommending conferences.

In: **Proceedings of the 5th BraSNAM—Brazilian Workshop on Social Network Analysis and Mining,** Porto Alegre, Brazil. 2016.

GARCÍA, Grettel Monteagudo et al. Techniques for comparing and recommending conferences. **Journal of the Brazilian Computer Society**, v. 23, n. 1, p. 4, 2017.

GRIESER, Karl et al. Using ontological and document similarity to estimate museum exhibit relatedness. **Journal on Computing and Cultural Heritage (JOCCH)**, v. 3, n. 3, p. 10, 2011.

GUC, Baris et al. Semantic annotation of GPS trajectories. In: **11th AGILE international conference on geographic information science**. 2008. p. 1-9.

HAJMOOSAEI, Abdolreza; SKORIC, Petra. Museum ontology-based metadata. In: **Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on**. IEEE, 2016. p. 100-103.

HEATH, Tom; BIZER, Christian. Linked data: Evolving the web into a global data space. **Synthesis lectures on the semantic web: theory and technology**, v. 1, n. 1, p. 1-136, 2011.

HOGAN, Aidan; DECKER, Stefan; HARTH, Andreas. Reconrank: A scalable ranking method for semantic web data with context. 2006.

HU, Yingjie et al. A geo-ontology design pattern for semantic trajectories. In: **International Conference on Spatial Information Theory**. Springer, Cham, 2013. p. 438-456.

HULPUŞ, Ioana; PRANGNAWARAT, Narumol; HAYES, Conor. Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In: **International Semantic Web Conference**. Springer, Cham, 2015. p. 442-457.

IOFFE, Sergey. Improved consistent sampling, weighted minhash and l1 sketching. In: **Data Mining (ICDM), 2010. IEEE 10th International Conference on**. IEEE, 2010. p. 246-255.

JÄRVELIN, Kalervo; KEKÄLÄINEN, Jaana. Cumulated gain-based evaluation of IR techniques. **ACM Transactions on Information Systems (TOIS)**, v. 20, n. 4, p. 422-446, 2002.

KATZ, Leo. A new status index derived from sociometric analysis. **Psychometrika**, v. 18, n. 1, p. 39-43, 1953.

LEME, Luiz André P. Paes et al. Identifying candidate datasets for data interlinking. In: **International Conference on Web Engineering**. Springer, Berlin, Heidelberg, 2013. p. 354-366.

LIU, Yu et al. Understanding intra-urban trip patterns from taxi trajectory data. **Journal of geographical systems**, v. 14, n. 4, p. 463-483, 2012.

MIRIZZI, Roberto et al. Ranking the linked data: the case of dbpedia. In: **International Conference on Web Engineering**. Springer, Berlin, Heidelberg, 2010. p. 337-354.

MURTAGH, Fionn; LEGENDRE, Pierre. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. **Journal of Classification**, v. 31, n. 3, p. 274-295, 2014.

NGOMO, Axel-Cyrille Ngonga; AUER, Sören. Limes-a time-efficient approach for large-scale link discovery on the web of data. In: **International Joint Conference on Artificial Intelligence (IJCAI)**. 2011. p. 2312-2317.

NUNES, Bernardo Pereira et al. Combining a co-occurrence-based and a semantic measure for entity linking. In: **Extended Semantic Web Conference**. Springer, Berlin, Heidelberg, 2013. p. 548-562.

PARENT, Christine et al. Semantic trajectories modeling and analysis. **ACM Computing Surveys (CSUR)**, v. 45, n. 4, p. 42, 2013.

PASSANT, Alexandre. Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. In: **AAAI spring symposium: linked data meets artificial intelligence**. 2010. p. 123.

PELEKIS, Nikos; THEODORIDIS, Yannis. Semantic aspects on mobility data. In: **Mobility Data Management and Exploration**. Springer New York, 2014. p. 189-209.

POLLERES, Axel et al. RDFS and OWL reasoning for linked data. In: **Reasoning Web. Semantic Technologies for Intelligent Data Access**. Springer Berlin Heidelberg, 2013. p. 91-149.

QUERCIA, Daniele; SCHIFANELLA, Rossano; AIELLO, Luca Maria. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In: **Proceedings of the 25th ACM conference on Hypertext and social media**. ACM, 2014. p. 116-125.

RANACHER, Peter; TZAVELLA, Katerina. How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. **Cartography and geographic information science**, v. 41, n. 3, p. 286-307, 2014.

RENSO, Chiara et al. How you move reveals who you are: understanding human behavior by analyzing trajectory data. **Knowledge and information systems**, p. 1-32, 2013.

ROA-VALVERDE, Antonio J.; SICILIA, Miguel-Angel. A survey of approaches for ranking on the web of data. **Information Retrieval**, v. 17, n. 4, p. 295-325, 2014.

ROCHA, Jose Antonio MR et al. DB-SMoT: A direction-based spatio-temporal clustering method. In: **Intelligent systems (IS), 2010 5th IEEE international conference**. IEEE, 2010. p. 114-119.

RUBACK, Lívia et al. Enriching mobility data with linked open data. In: **Proceedings of the 20th International Database Engineering & Applications Symposium**. ACM, 2016. p. 173-182.

RUBACK, Lívia et al. SELEcTor: discovering similar entities on LinkEd DaTa by ranking their features. In: **Semantic Computing (ICSC), 2017 IEEE 11th International Conference on**. IEEE, 2017. p. 117-124.

RUBACK, Lívia et al. Computing Entity Semantic Similarity using Ranked Lists of Features. 2017a. Journal under revision, not published.

SANKARARAMAN, Swaminathan et al. Computing similarity between a pair of trajectories. **arXiv preprint arXiv:1303.1585**, 2013.

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

TOOHEY, Kevin; DUCKHAM, Matt. Trajectory similarity measures. **SIGSPATIAL Special**, v. 7, n. 1, p. 43-50, 2015.

VIDAL, Vânia MP et al. Specification and incremental maintenance of linked data mashup views. In: **International Conference on Advanced Information Systems Engineering**. Springer, Cham, 2015. p. 214-229.

VINH, Nguyen Xuan; EPPS, Julien; BAILEY, James. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. **Journal of Machine Learning Research**, v. 11, n. Oct, p. 2837-2854, 2010.

VLACHOS, Michail; KOLLIOS, George; GUNOPULOS, Dimitrios. Discovering similar multidimensional trajectories. In: **Data Engineering, 2002. Proceedings. 18th International Conference on**. IEEE, 2002. p. 673-684.

VOLZ, Julius et al. Silk-A Link Discovery Framework for the Web of Data. **Linked Data on the Web (LDOW)**, v. 538, 2009.

WANG, Haozhou et al. An effectiveness study on trajectory similarity measures. In: **Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137**. Australian Computer Society, Inc., 2013. p. 13-22.

WEBBER, William; MOFFAT, Alistair; ZOBEL, Justin. A similarity measure for indefinite rankings. **ACM Transactions on Information Systems (TOIS)**, v. 28, n. 4, p. 20, 2010.

WITTEN, Ian H.; MILNE, David N. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. 2008.

YAN, Zhixian et al. Trajectory ontologies and queries. **Transactions in GIS**, v. 12, n. s1, p. 75-91, 2008.

YEUNG, Ka Yee; RUZZO, Walter L. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. **Bioinformatics**, v. 17, n. 9, p. 763-774, 2001.

YILMAZ, Emine; KANOULAS, Evangelos; ASLAM, Javed A. A simple and efficient sampling method for estimating AP and NDCG. In: **Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval**. ACM, 2008. p. 603-610.

ZHANG, Yinuo et al. Event Recommendation in Social Networks with Linked

Data Enablement. In: **ICEIS**. 2013. p. 371-379.

ZHENG, Vincent W. et al. Collaborative location and activity recommendations with gps history data. In: **Proceedings of the 19th international conference on World Wide Web**. ACM, 2010. p. 1029-1038.

ZHENG, Yu et al. Understanding transportation modes based on GPS data for web applications. **ACM Transactions on the Web (TWEB)**, v. 4, n. 1, p. 1, 2010a.