The *Baquara*^{2} Knowledge-based Framework for Semantic Enrichment and Analysis of Movement Data

Renato Fileto[a], Cleto May[a], Chiara Renso[d], Nikos Pelekis[b], Douglas Klein[a], Yannis Theodoridis[c]

[a] *INE/CTC, Federal University of Santa Catarina, Florianópolis-SC, Brazil*
[b] *Department of Statistics and Insurance Science, University of Piraeus, Greece*
[c] *Department of Informatics, University of Piraeus, Greece*
[d] *IST/CNR, Pisa, Italy*

Abstract

The analysis of moving objects behaviors frequently requires more than just spatio-temporal data. Thus, despite recent progresses in trajectories handling, there is still a gap between movement data and formal semantics. This gap hinders movement data analyses benefiting of available knowledge, with well-defined and widely agreed semantics. This article describes the *Baquara*^{2} framework to help narrow this gap by exploiting knowledge bases to semantically enrich and analyze movement data. It allows structuring and abstracting movement data in an arbitrarily deep hierarchy of progressively detailed movement segments that generalize concepts such as trajectories, stops, and moves. A general customizable process convert textual annotations of movement data into semantic annotations that point to concepts and objects described in ontologies and Linked Open Data (LOD) collections. This process helps to produce semantically enriched movement data compliant with an ontology that enables queries for movement analyzes based on application and domain specific knowledge. The proposed framework has been used in experiments to semantically enrich movement data collected from social media with geo-referenced LOD. The obtained results enable powerful queries that illustrate *Baquara*^{2} capabilities.

Keywords: trajectories of moving objects. social media posts. ontologies, linked data, semantic enrichment, movement data analysis.

1. Introduction

Nowadays, large amounts of data about movements of objects in the geographic space can be gathered by using a variety of devices (e.g., smart phones equipped with GPS or just connected to a GSM network, vehicles equipped with RFID) and information systems (e.g., social media that can detect changing locations of their users). We call *movement data* any collection of spatio-temporal data representing sampled or inferred positions of moving objects.

This definition encompasses, among other things, moving objects' trajectories, and sequences of social media users' posts.

A *trajectory* is a temporally ordered sequence of spatio-temporal positions occupied by a moving object. Nowadays, it is possible to get accurate trajectories, by using state-of-the art sensors and fine sampling rates (e.g., every second, every 3 meters). However, it is hard to gather large volumes of annotated trajectories, because annotating them is a laborious task [1, 2, 3, 4]. Nevertheless, social media posts usually have plenty of associated textual contents (e.g., text, hash tags). These contents can be regarded as textual annotations that may provide hints to explain movements. We call a *user's spatio-temporal trail* (or simply a *trail*) one temporally ordered sequence of posts of a user in a particular system (e.g., Twitter, FourSquare, Facebook, Instagram), with the spatio-temporal coordinates of each post along with its contents. Differently from trajectories, social media users' trails are usually sparse, due to the asynchronous nature of the users' posts, and usually are less precise than trajectories, due to limitations of their gathering processes or access restrictions. However, they can be useful in several kinds of spatio-temporal information analysis [5, 6, 7, 8, 9], and can be combined (e.g., spatio-temporally fused) with trajectories to exploit the best traits of each one of these categories of movement data [10].

Lots of information can be extracted from freely annotated movement data such as trajectories and trails, and combinations of them, with a myriad of useful applications [11]. However, freely annotated movement data lack well-defined semantics to support information analysis. For instance, the tag `"Rio"` in social media user's post may refer to a city, a state, or even a restaurant or nightclub, among other possibilities. Therefore, to realize potential applications it is necessary to develop appropriate methods to turn freely annotated movement raw data into semantically rich movement data. The recent progress in movement data handling [12, 13, 14, 15, 16] includes plenty of significant contributions for structuring and analyzing movement data based on its spatio-temporal components. Notwithstanding, it is recognized by the scientific community that semantic issues, including the exploitation of textual and contextual information that can be associated to movement data, must be addressed yet to better understand and exploit movement data [17, 18, 1, 19, 20, 21].

This article describes *Baquara*^{2}, a conceptual framework that includes an upper ontology and a general semantic enrichment process to semantically enrich movement data and support knowledge-based queries for movement analysis. The *Baquara*^{2} upper ontology has a rich repertoire of constructs to semantically describe progressively detailed movement segments organized in an arbitrarily deep hierarchy for movement analysis. The general concept of movement segment generalizes concepts such as trajectories, trails, and episodes (e.g., stops and moves). The *Baquara*^{2} semantic enrichment process converts movement data annotated with text into semantically annotated movement data. The produced semantic annotations reference concepts (classes) and objects (instances of concepts) of movement analysis facets, such as space, time, and goals. Some of these facets can be built from extracts of description logics compatible knowledge-bases, such as LOD collections and their associated ontologies.

The proposed ontology and semantic enrichment process are customizable to for use with a variety of freely annotated movement data, domain and application ontologies, and LOD collections. The adaptation points of the *Baquara*^{2} conceptual framework are the domain knowledge used to semantically describe movements and the tasks of the semantic enrichment process. The domain ontologies and LOD collections used to turn free textual annotations into semantic annotations can be selected according to the spatio-temporal scope of the movement to be analyzed, and the application domain (traffic analysis, tourism, emergency planning, etc.). Several methods can be used to find connections of movement data with linked data, including spatio-temporal and lexical matching (e.g., tag Rio associated with an stop occurring at a bar called Rio), and produce semantic annotations for different movement analysis aspects (places and events effectively visited or intended to visit, goals of movements).

Baquara^2 enables queries referring to concepts (classes) and/or objects (instances of concepts) used to semantically enrich the movement data, such as:

Query 1 *Select the social media user's trails with at least one stop to visit a mountain called Corcovado in the city of Rio, followed by one stop in a marketplace, where he/she does at least one finer stop in a restaurant.*

Query 2 *Determine the percentage of European's trails in Brazil that make at least a stop in a nature reserve, where he/she does at least one finer stop in a tourist shop.*

These queries can be expressed in languages like SPARQL[1], and its extensions with spatial operators, such as GeoSPARQL [22] and ST-SPARQL [23], among other alternatives. The viability of the proposal have been investigated in case studies using movement data extracted from Flickr[2] and Twitter[3], and semantically enriched with labled geo-referenced places of several subclasses taken from DBPedia[4] and LinkedGeoData[5].

The rest of this article is organized as follows. Section 2 defines general constructs for structuring and abstracting movement data. Section 3 describes the *Baquara*^{2} upper ontology, that provides conceptual support to annotate and query movement data according to a variety of semantic description facets. Section 4 describes a basic semantic enrichment process, and the *Baquara*^{2} general architecture to semantically enrich and analyze movement data using ontologies and LOD. Section 5 reports experiments that apply our proposal to semantically enrich and analyze movement data collected from social media with geo-referenced LOD. Finally, Section 6 discusses related work, and Section 7 summarizes our contributions and future work.

[1] http://www.w3.org/TR/sparql11-query

[2] https://www.flickr.com

[3] https://twitter.com

[4] http://dbpedia.org

[5] http://linkedgeodata.org

2. General Structures and Abstractions for Movement Data

This section defines general concepts to structure and abstract movement data in several refinement levels. A moving object's positions sequence (MOPS) represents the known movement history of an object (e.g., a car monitored by GPS, a social media user) during a certain period of time as a temporally ordered sequence spatio-temporal positions occupied by this object. A movement segment (MS) is an abstraction that refers to any continuous subsequence of a MOPS. These concepts generalize notions like social media user's trails (time-ordered sequences of posts), semantic trajectories, and episodes (e.g., stops, moves). A MOPS can be successively segmented in several levels of detail. The MSs referring to successively smaller segments of a MOPS can be organized in hierarchy with many refinement levels for information analyses purposes. Annotations can be associated with a MOPS, MS, or known spatio-temporal position of a moving object to help describe its movements. Annotations of movement data and movement patterns are formally defined in Section 3.2.

The first step for doing this kind of movement information analysis is to collect and time-order positions of each moving object in a MOPS (Definition 1).

Definition 1. A **moving object's positions sequence (MOPS)** is a tuple:

$$mops = (idMO, PS, A)$$

where:

$idMO$ is the unique identifier of a moving object;

$PS = \langle p_1, \ldots, p_n \rangle$ is a time ordered sequence of spatio-temporal positions of the moving object identified by $idMO$;

A is a set of annotations associated with the whole $mops$.

Each position p_i of PS ($i, n \in \mathbb{N}; 1 \leq i \leq n; n \geq 1$) is a tuple of the form:

$$p_i = (i, geom, t, A_i)$$

where:

i is the temporal order of the position p_i in PS;

$geom$ is a geometry that represents the moving object identified by $idMO$ during the time t;

$t = [begin, end]$, with $begin$ and end being instants of time and $begin \leq end$, is the time interval ($begin < end$) or instant ($begin = end$) when the position of the moving object identified by $idMO$ is represented in space by the geometry $geom$;

A_i is a set of annotations associated with p_i.

An $idMO$ identifies a moving object in a given data source. A moving object can sometimes be decomposed in a relevant moving entity (e.g., instance of person, animal, or vehicle) and a data gathering device (e.g., cell phone, GPS navigator) used to tack the moving entity positions. The entity id (possibly fake, for privacy reasons) of the same real world entity, can be different in different data sources. For example, though the same person can post in different social media systems (e.g., Twitter, Flickr, Facebook), her id can be different in each system. The same can happen to a device. In addition, a moving entity can hold several devices, and the same device may be held by different users at different times. Consequently, the movement data of the same real world entity and/or device taken from different sources is separated in distinct MOPS (each one with a different idMO). Identifying if movement positions coming from different data sources pertain to the same moving entity and/or device is beyond the scope of this work. This problem has been addressed as data fusion in [10].

A position p_i of a MOPS represents its location and shape in a given time. If the moving object shape can be neglected due to its small size compared to the space where it moves (e.g., a person or a car moving in a city), the geometry $p_i.geom$ can be a point. Otherwise, it can be a polygon or multi-polygon (e.g., representing a moving storm). The positions of a MOPS must be totally ordered by their respective times, and their times cannot overlap, i.e., the following constraint must apply to the positions sequence PS of any MOPS:

$$\forall p_i, p_{i+1} \in PS : p_i.t[end] < p_{i+1}.t[begin]$$

A MOPS can be segmented for information analysis purposes by using a variety of methods proposed in the literature to produce subsequences such as trajectories and episodes [1, 18, 16]. Although a structured trajectory can be defined as a time-ordered sequence of episodes [1], both trajectories and episodes refer to subsequences of movement positions satisfying particular predicates. For instance, the segmentation of a MOPS into trajectories can be determined according to constraints on time (e.g., a trajectory per day), space (e.g., segment according to some geographic boundaries or after reaching certain traveled distances), or both (e.g., segment every time there is a sampling gap or stop lasting longer than a given threshold) [24]. Episodes, such as *stops*/*moves*, on the other hand, can be determined by predicates like “speed below/above a certain threshold” [25] or “moving object inside a given region for a time period longer/shorter than a certain threshold” [26, 27]. In this work, we propose a generalized concept for constructs such as trajectories and episodes called movement segment (MS).

An MS (Definition 2) is an abstraction for a portion of the movement history of a moving object identified by *idMO*. Possible values for an MS’ `type` include: `TRAIL` (time ordered sequence of social media user’s posts), `TRAJ` (trajectory), `STOP` episode, and `MOVE` episode. An MS is associated to a subsequence of spatio-temporal positions $\langle p_i, \ldots, p_j \rangle$ of a positions sequence $mops.PS = \langle p_1, \ldots, p_n \rangle$. However, it can abstract these positions. The geometry *ms.geom* of the MS *ms* is an approximation of the movement portion that *ms* refers to. For example, a stop can be represented in the space by the centroid of the spatial coordinates of its constituent points, while a move can be represented by line segments. The time span *ms.ts* temporally fits all the positions associated to *ms*.

Definition 2. A **movement segment (MS)** of a moving object's positions sequence $mops = (idMO, PS, A)$ is a tuple of the form:

$$ms = (idMO, idMS, type, geom, p_i, p_j, ts, father, level, prev, next, ord, A_{ms})$$

where:

$idMO$ is a moving object unique identifier;

$idMS$ is the unique identifier of ms;

$type$ is the type of ms;

$geom$ is the geometry used to abstractly represent ms in the space;

p_i, p_j are respectively the initial and final positions of the corresponding subsequence $\langle p_i, \ldots, p_j \rangle$ of the time ordered positions sequence $\langle p_1, \ldots, p_n \rangle$ of the moving object identified by $idMO$ $(i, j, n \in N; n \geq 1; 1 \leq i \leq j \leq n)$;

ts $= [b, e]$ $(b = p_i.t[begin], e = p_j.t[end])$ is time span of ms;

$father$ is the shortest movement segment of $idMO$ such that $ms.father \neq NULL \quad \rightarrow \quad ms.ts \subset ms.father.ts$;

$level$ is the distance of ms to its ancestry root, i.e., $ms.level = 0$ if $ms.father = NULL$, otherwise $ms.level = 1 + ms.father.level$;

$prev$ is the chronologically closest previous sibling of ms;

$next$ is the chronologically closest next sibling of ms;

ord is the distance of ms to its sibling that happened first in time plus 1, i.e., $ms.order = 1$ if $ms.prev = NULL$, otherwise $ms.order = 1 + ms.prev.order$;

A_{ms} is a possibly empty set of annotations associated to ms.

Sibling movement segments are those that have the same father, i.e., two movement segments ms and ms' are siblings if $ms.father = ms'.father$. The set of predecessors of a movement segment ms is given by the transitive closure $predecessors(ms) = prev(ms) \cup predecessors(ms.prev)$, with the stop condition $predecessors(NULL) = \emptyset$, and the set of successors of ms by the transitive closure $successors(ms) = next(ms) \cup successors(ms.next)$, with the stop condition $successors(NULL) = \emptyset$. The set of siblings of ms is $siblings(ms) = predecessors(ms) \cup successors(ms)$.

The $children$ of a movement segment ms are the movement segments having ms as $father$, i.e., $children(ms) = \{ms' \mid ms'.father = ms.idMS\}$. The set of descendants of ms is given by the transitive closure $descendants(ms) = children(ms) \cup \{\cup_{ms' \in children(ms)} ms'\}$. The set of ancestors of ms is given by the transitive closure $ancestors(ms) = ms.father \cup ancestors(ms.father)$ with $ancestors(NULL) = \emptyset$. The set of movement segments in the lineage of ms is $lineage(ms) = ancestors(ms) \cup descendants(ms)$.

Notice that, according to Definition 2, for any movement segment ms its time span $ms.ts$ must be contained in that of $ms.father$ if $ms.father \neq NULL$. In addition, sibling movement segments do not overlap in time, i.e.:

$$\forall ms', ms'' \in children(ms) : ms'.ts \cap ms''.ts = \emptyset$$

These restrictions ensure that the time span of an MS covers the time spans of all its descendants, and that sibling movement segments are always organized in a consistent total ordering in time, in the sense that no segment begins before its previous one finishes. Thus, movement segments can be arranged in a tree-like hierarchy to support information analysis at different levels of detail, determined by their time span, as stated by Definition 3.

Definition 3. An **movement segments hierarchy (MSH)** for a MOPS denoted by *mops* is a tree denoted by *msh* such that:

1. each node of the *msh* is a movement segment of *mops*;
2. the father of the unique root node of *msh* is $NULL$.

MSHs can support semantic analysis of movement data in different levels of detail. For example, at the root of the hierarchy a MOPS *mops* can be regarded as a sequence of semantic trajectories [1]. Each semantic trajectory can be refined in the next level by a sequence of episodes, each one referring to subsegments of a semantic trajectory that satisfy some kind of predicate (e.g., stops inside cities and moves between them). Then stops in each city can be further refined in finer stops (e.g., in places like an airport, a university, a shopping mall) and moves between such smaller stops that refine bigger stops in cities. Finally, stops in relatively big places of cities (e.g., a university or a marketplace), can be further segmented in lower level stops in smaller places (e.g., particular departments of the university or shops of the marketplace) and moves between such finest stops.

Figure 1 illustrates a hierarchy of movement segments. At the top level the corresponding MOPS is segmented and abstracted in sequences of trajectories inside big countries or world regions, such as Brazil, the US, and the EU. These trajectories are further segmented in stops and moves progressively detailed in the lower levels of the hierarchy. Stops are represented by circles, and moves by dashed lines between them, with arrows indicating the movement direction. Many of these movement segments have associated annotations, which begin with the sign @. In this example, the annotations indicate the places where trajectories or stops occur, and the transportation means of some moves. The portion of the second hierarchical level presented in Figure 1 details `Trajectory 3` by showing stops in some Brazilian cities and moves between them. The third hierarchical level details `stop 3.1` in `Rio`, with stops and moves in smaller places that are inside `Rio` , including `stop 3.1.1` at the `GIG airport`, followed by stops at `Sun hotel`, `SC market`, `Corcovado`, `Ipanema Beach`, and so on. Finally, in the lowest level some details of the movement inside `SC Market` are presented, including `stop 3.1.3.1` at `BB ATM`, stops in some shops, and so on.

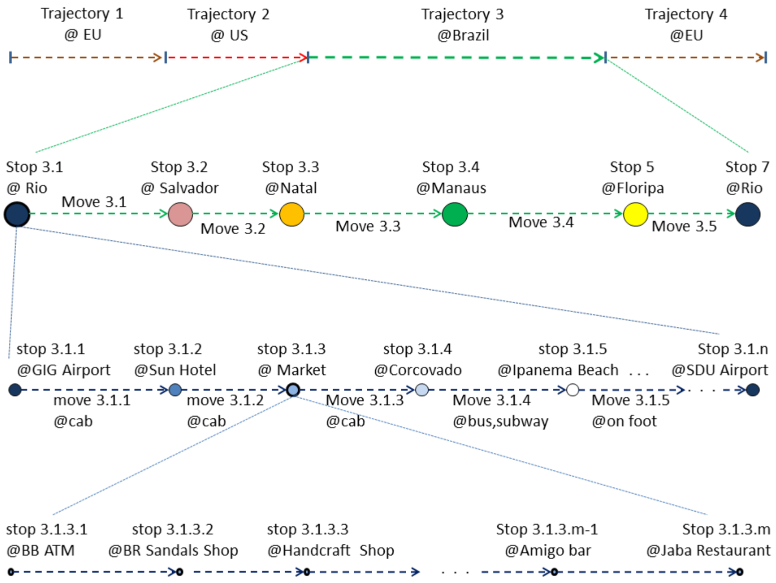


Figure 1: A movement segments hierarchy example

3. The *Baquara*^{2} ontology

The conceptual modeling core of our approach for semantic enrichment and analysis of movement data is the *Baquara*^{2} upper ontology. It has been designed to serve as a conceptual framework for describing movement segments in several application domains, ranging from urban transportation to animal ecology. Such adaptation can be done by specializing some of its pre-defined classes, and by creating new relationships among them.

Figure 2 shows the high level concepts (classes) of *Baquara*^{2} ontology, and the major semantic relationships between them. Each labeled rectangle represents a concept. Nesting denotes subsumption (`IS_A` relationship), i.e., each nested concept is a subclass of its enclosing concept. For instance `Episode IS_A MovementSegment`. The plus sign on the top left corner of a rectangle indicates that the respective concept can be further specialized, according to application domains and analyses needs. A dashed line between concepts denotes a semantic relationship, such as composition (`PART_OF`) or a specific relationship (e.g., between an `Event` and a `Place` where it occurs). The main constituents of the *Baquara*^{2} ontology are described in the following.

3.1. Movement Analysis Facets

Movement segments and other abstractions for describing movements like some movement patterns (Section 3.3) can be semantically annotated by linking them to concepts (classes) and/or objects (instances of classes) of semantic facets for movement description and analysis (Definition 4).

Definition 4. A **semantic facet** is a graph $G(V, E)$, where:

V is a set of resources, each one referring to a concept (class) or object (instance of a class);

E is a set of semantic relationships between resources of V.

Each facet has an intentional level (TBox) and an extensional level (ABox) based on description logics [28]. The first has at least one conceptualization hierarchy (classes organized according to their `IS_A` relationships), and the latter at least one objects hierarchy (instances that can be organized by some partial ordering relationship, such as `PART_OF`).

Baquara facets describe information and knowledge about relevant themes for movement analysis. The *Baquara*[2] pre-defined facets cover those of the CONSTAnT model [20], namely: `Space` (e.g., `Places of Interest (POIs)`), `Time`, `Goal` (e.g., `Eat`, `Watch a game`), `Behavior` (e.g., `Flock` [29], `Chasing` [30], `Avoidance` [31]), `TransportationMeans`, `EnvironmentCondition` (e.g., `Windy`), `Activity` (e.g., `Running`), and `MovingObject`. Baquara also includes the facet `Event` (e.g., `CulturalEvent`, `SportEvent`), and allows adding new specific facets for movement analysis in particular application domains.

A `MovingObject` is described in Baquara as an association between one `MovingEntity` (e.g., `Person`, `Animal`) and a `MovementMonitoringMeans`, that can be specialized in a `MovementMonitoringSystem` (e.g., `SocialMedia`), or a `MovementMonitoringDevice` (e.g., `CellPhone`), as discussed in Section 2.

Baquara[2] employs the OGCs Geospatial Features Model[6] and the W3Cs Time Ontology[7] as foundations to describe space (places and their relationships) and time (instants, periods of time, and their relationships), respectively. Their concepts and instances are used to describe the spatio-temporal scope of movement, as well as the places, times and events of interest for movement analysis. A place is a spatial feature relevant for movement data description and/or movement analysis. It can be anything with a geometry, and at least a name. A places geometry, represented in some coordinate system, can be simple (point, line, or region) or complex (set of points, lines, and/or regions). Specializations of place relevant for the tourism domain may, for example, include `Country`, `City`, `Airport`, `Hotel`, and `Cafe`. A time can be an instant or a period of time.

An event is any circumstance relevant for movement analysis in an given domain. It has at least one label, occurs in a time (instant or period), and may have relationship(s) with some place(s). For instance, `SportEvent` and `CulturalEvent` are subclasses of `Event` relevant for the tourism domain. A `CulturalEvent` can be specialized to `MusicFestival`, `DanceFestival`, etc. Conversely, `Carnival` can be regarded as a subclass of `TraditionalParty`. Instances of event (and its subclasses in any abstraction level) can be related to specific instances of place and time, as indicated by the dashed line linking these

[6] http://www.opengeospatial.org/standards/sfa

[7] http://www.w3.org/TR/owl-time

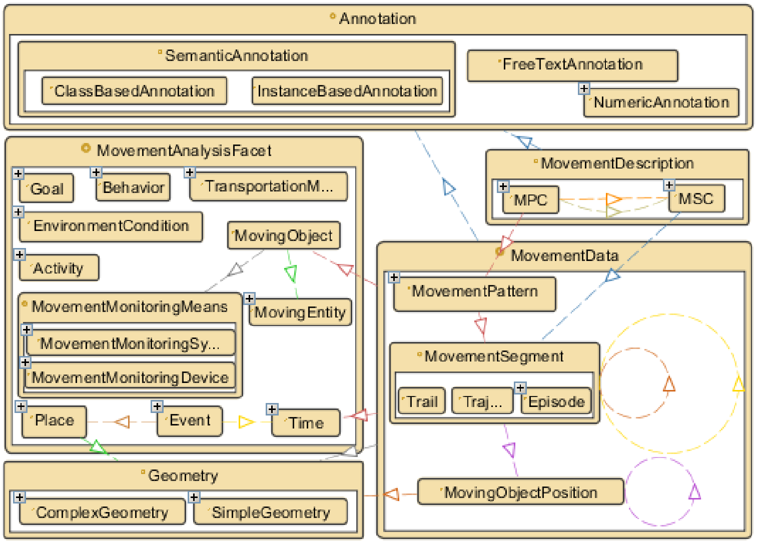


Figure 2: The backbone of the *Baquara*[2] ontology

classes in the left bottom portion of Figure 2. For instance, the city `Rio` can be semantically related to events occurring there, such as `Pan American Games 2007` and `Word Cup Final 2014`. On the other hand, events like `Christmas holidays` may not be associated with any particular place, because they occur in many places. Remember that a facet represents descriptive information for movement analysis in the extensional level (e.g., `Maracanã`, `July 13 2014 5:00 PM - 8:00 PM`, `watch World Cup 2014 Final`), and the intentional level (e.g., `Stadium`, `Sunday evening`, `watch soccer game`).

3.2. Movement Annotations

A set of annotations can be associated with movement segments to describe what is going on (e.g., place or event of interest, goal, environmental conditions), as stated by Definition 5.

Definition 5. An **annotation** is a triple of the form:

$$annot = (target, property, value)$$

where:

target is the annotated thing;

property is a descriptive property defined for instances of *target*;

value is a typed literal (e.g., text, number) or a reference to a resource (object or concept) described in a knowledge base.

The *target* is a reference to a movement object positions sequence (MOPS), a movement segment (MS), or a specific moving object position (defined in Section 2). The *target* can also be a movement pattern or a movement abstract description (defined in Section 3.3). The property indicates a relation of the *target* with the annotation value. For example, the property *occursAt* indicates that a movement segment (e.g., a stop) occurs at a particular place denoted by the value of this property.

An annotation can be free or semantic. The value of a free annotation is a literal, such as a string, free text, or a number. Thus, it may not have precise semantics. The value of a semantic annotation, on the other hand, must be a reference to resource, i.e., a concept (class) or an object (instance of a class) described in a description logics compatible knowledge base, to better describe its semantics [32]. In Baquara each semantic annotation is a reference to an object or a concept of a movement analysis facet described in Section 3.1.

3.3. Movement Patterns and Abstract Movement Descriptions

The *Baquara*^{2} upper ontology also provides constructs to express movement patterns, and abstract movement descriptions (movement segment categories, and movement pattern categories). A **movement pattern (MP)** is as a collection of movement segments that satisfies some predicate, based on: spatio-temporal constraints of movement segments and/or semantic, ordering, and timing constraints on related segments. Examples of spatio-temporal constraints include moving clusters and meetings [30]). Semantic constraints are expressed by the type of the movement segments and their associated semantic annotations. Ordering constraints refer to the exact or relative order of movement segments among their siblings or lineage. Timing constraints refer to the duration of some movement segment(s) or the elapsed time between them.

A **Movement Segment Category (MSC)** is an abstract description for movement segments (MS). One MSC can be represented with by a tuple analogous to the one proposed to represent an MS in Definition 2, but without any geometry (*geom*) or time span (*ts*), minimum and maximum allowed duration instead of a single exact duration (*dur*), the possibility of relative instead of absolute sibling order, and the possibility of any relatives instead of just immediate relatives (*predecessors* instead of just *prev*, *successors* instead of just *next*, *ancestors* instead of $father$, and many *children*) always referring to other MSCs instead of MSs.

A **Movement Pattern Category (MPC)** is an abstract description for movement segments (MS) is an abstract description for movement patterns (MP). One MPC can be expressed by a reference to an MSC and its related MSCs, along with possible (partial) ordering and/or timming restrictions. For example, any collection of segments $S = ms_i$ such that: (i) ms is a `Stop` of an `EuropeanPerson` in `Market`; (i) ms_i takes part in a meeting pattern; (ii) ms_i is preceded by at least a sibling `Stop` in an `Airport` and another one in a `Hotel`; (iii) ms_i is followed by a a sibling `Stop` in a `TouristicPlace` called `Corcovado`; and (iv) ms_i is detailed in a number of shorter stops, being at least one of them in a `Bar` for at most 1 hour, which is immediately followed by another short `Stop` in a `Restaurant`. Notice that *stop*3.1.3 presented in Figure 1 satisfies all these semantic and ordering constraints, and maybe the spatio-temporal constraint (take part in a meeting pattern) and the time constraint of the short stop in a `Bar` for at most 1 hour as well.

Abstract movement descriptions (MSCs and MPCs) semantically describe movement by relaying on annotations, without referring to any concrete movement segment of any existent moving object. However, many concrete movement segments and movement patterns can semantically match such a description. More formal definitions for restricted MSCs and MPCs, with detailed data structures for representing them can be found in [33], along with examples of their use for movement analysis. Generalized formal definitions for these abstract movement descriptions and the investigation of semantic consistency rules among them, and the semantic matching concrete movement segments and movement patterns with their respective categories are out of the scope of this article, and left to future work.

4. Semantic Enrichment and Analysis of Movement Data

This section describes a customizable process to semantically enrich movement data by using a description logics compatible knowledge base (KB), which can be built with (portions of) domain ontologies and LOD collections. The semantic enrichment process turns textually annotated movement data (either raw or structured) as described in Section 2, into semantically annotated movement data compliant with the *Baquara*^{2} upper ontology described in Section 3. The movement data structuring in movement segments hierarchies can take place before, during, or after this semantic enrichment process. The general architecture for semantic enrichment and analyses of movement data allows data processing in the KB and/or conventional and spatio-temporal database, to allow flexibility for the semantic enrichment process, as well as for querying, reasoning, data warehousing, and data mining with the semantically enriched movement data.

4.1. A General and Basic Process for Semantic Enrichment

Figure 3 illustrates the inputs, outputs, stages, and major tasks of the proposed process to semantically enrich movement data by using ontologies and LOD. The inputs are: (i) movement data (movement segments or individual positions of moving objects) annotated with text (e.g., sequences of social media posts with their textual contents, annotated trajectories); and (ii) resources (concepts and instances) of domain ontologies and LOD collections (e.g., DBpedia, LinkedGeoData) organized in a KB with a variety of semantic facets (e.g., space, time, goal) to describe and analyze the movement data (Section 3.1). The outputs are semantically enriched movement data, i.e., movement data with semantic annotations that refer to resources of the KB used for semantic enrichment. Such a resource can be a concept (e.g., bar, restaurant) or an object (instance of such a concept).

The proposed semantic enrichment process is didactically organized in two stages: Data Pre-processing and Linking. Each of these stages have a sequence of

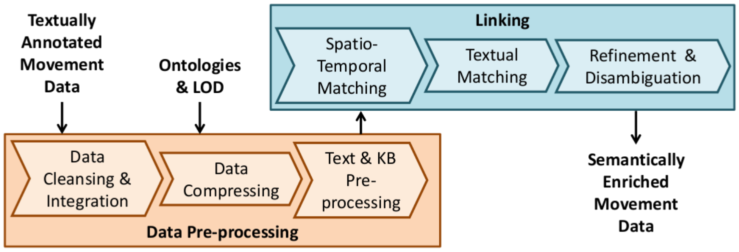

Textually Annotated Movement Data
Ontologies & LOD
Linking
Spatio-Temporal Matching
Textual Matching
Refinement & Disambiguation
Data Cleansing & Integration
Data Compressing
Text & KB Pre-processing
Data Pre-processing
Semantically Enriched Movement Data

Figure 3: General process to semantically enrich textually annotated movement data

tasks that are customizable in terms of the techniques used to accomplish them, parameters tunning, and even the existence of some tasks and their relative order. Such customization can be done according to the characteristics of the data and knowledge provided as inputs, and the data enrichment and analysis purposes. The Data Pre-processing stage typically include the tasks: Data Cleansing & Integration that filters out invalid data (e.g., outliers) and sometimes integrates data obtained from different sources (e.g., social media posts of different systems, social media posts with trajectories [10], LOD of different collections); Data & Compressing that can compact data for speeding up the following tasks, by using a variety of techniques such as those described in [18, 1, 16]; and Text & KB Pre-processing that prepares the textual data and KB resources for the linking stage, by applying techniques such as textual contents filtering, classification, stemming, and named entities recognition [34, 35, 36, 37, 38].

Then, the Linking stage does the entity linking [39, 38, 40], i.e., the connection of movement data with the KB resources describing the entities mentioned in their text annotations. The specific tasks and techniques employed for solving this problem vary with the nature of the movement description facet. For example, Spatio-Temporal Matching is a crucial task to link movement data to facets like space and events, by taking into account the geographic extensions of places that are visited or where events and their operational times. In this case, Textual Matching may help to refine the matchings, before applying the final task of Refinement & Disambiguation.

Algorithm 1 is a basic solution for linking textually annotated movement segments with one or more moving object positions to KB resources referring to places or events. It exploits spatial proximity and textual similarity [41, 42, 43] to match movement segments with geo-referenced KB resources, and generate semantic annotations for the former linked to the latter. It employs the thresholds $\tau_s \in [0, \infty]$ and $\tau_s \in [0, 1]$ ($\tau_s, \tau_s \in R$) to filter the resources most likely to match each movement segment according to some geographic distance function, and some text similarity function, respectively. The spatial join of line 3) returns the pairs of movement segment (s) and resource (r) that are closer than τ_s along with the distance between them. Then, for each segment $s \in S$ the algorithm looks for the pairs (s,r) that are the closest in space, and among them those whose textual similarity is the highest (lines 4 to 17). Finally,

Algorithm 1: Link movement segments to co-located resources

```
    input : $S = \{s_0, ..., s_n\}$; // Pre-processed movement segments
            $R = \{r_0, ..., r_m\}$; // Pre-processed resources set
            $\tau_s \in \mathbb{R}^+$; // Spatial distance threshold in meters
            $\tau_t \in \mathbb{R}^+$; // Textual similarity threshold
    output: $SA$; // Semantic annotations of movement segments in $S$
 1  begin
 2      $SA \leftarrow \emptyset$;   // Semantic Annotations ($SA$) set initially empty
 3      $SJ \leftarrow (\Pi_{s \leftarrow S.*, r \leftarrow R.*, geoDist}(S \bowtie_{(geoDist \leftarrow dist(s.geom, r.geom)) \leq \tau_s} R))$;
 4      foreach $s \in S$ do
 5          $k \leftarrow 0$;         // Initialize best matching measures for $s$
 6          $minDist \leftarrow \tau_s$;
 7          $maxSim \leftarrow \tau_t$;
 8          foreach $(s, r, geoDist) \in SJ$ do
 9              if $geoDist \leq minDist$ then
10                  $textSim \leftarrow textualSimilarity(s.ppText,\ r.ppText)$;
11                  if $textSim \geq maxSim$ then
12                      if $geoDist < minDist \ \vee \ textSim > maxSim$ then
13                          $k \leftarrow 0$;    // Better matching resource $r$ found
14                          $minDist \leftarrow geoDist$;
15                          $maxSim \leftarrow textSim$;
16                      $k{+}{+}$;            // Increment number of matchings
17                      $bestMatching[k] \leftarrow r$;           // Add matching $r$
18          while $k > 0$ do
19              $k{-}{-}$;     // Create semantic annotations for segment $s$
20              $SA \leftarrow SA \cup (s, visits, bestMatching[k])$;
21      return $SA$;
```

the set *bestMatching* of resources satisfying these conditions with respect to segment s is used to create the semantic annotations for s (lines 18 to 20).

In this version, Algorithm 1 generates semantic annotations of segments with the property *visits*. The investigation of methods to generate annotations for other properties (e.g., *comesFrom* and *goesTo* for moves, *hasGoal* for stops and moves) is theme for future work. Nevertheless, notice that a number of optimizations and customizations can be easily done in this algorithm. For example, if tuples resulting of the spatial join of line 3 are grouped according to s, and the tuples for each s are put in ascending order of *geoDist*, then the second loop can be interrupted when $geoDist > minDist$. In addition, line 9 can be pushed down to be executed after the two line bellow it, to semantically annotate each movement segment s with the textually most similar resource(s) that are the closest to s. Recently proposed spatio-textual similarity joins [44, 45] can also be considered to speed-up the data processing. In addition, a variety of entity linking techniques [34, 39, 38, 40] can be employed to better refine and disambiguate generated links from movement segments to KB resources.

4.2. General Systems Architecture for Semantic Enrichment and Analysis

Figure 4 illustrates a general system architecture to realize the proposed approach for semantic enrichment and analyses of movement data. Firstly, the *Baquara*^{2} ontology must be loaded in a KB handled by a Knowledge Management System. Secondly, domain specific knowledge (e.g., ontologies and LOD) with the same spatio-temporal scope as the movement data to be enriched and analyzed must be selected, and customized if necessary. The domain knowledge selection and customization can be done, for example, by using SPARQL endpoints or REST APIs of a variety of LOD collections available on the Web. Thirdly, the Semantic Enrichment Process described in Section 4.1 must be executed to generate the semantic annotations for the movement data using the collected domain specific knowledge.

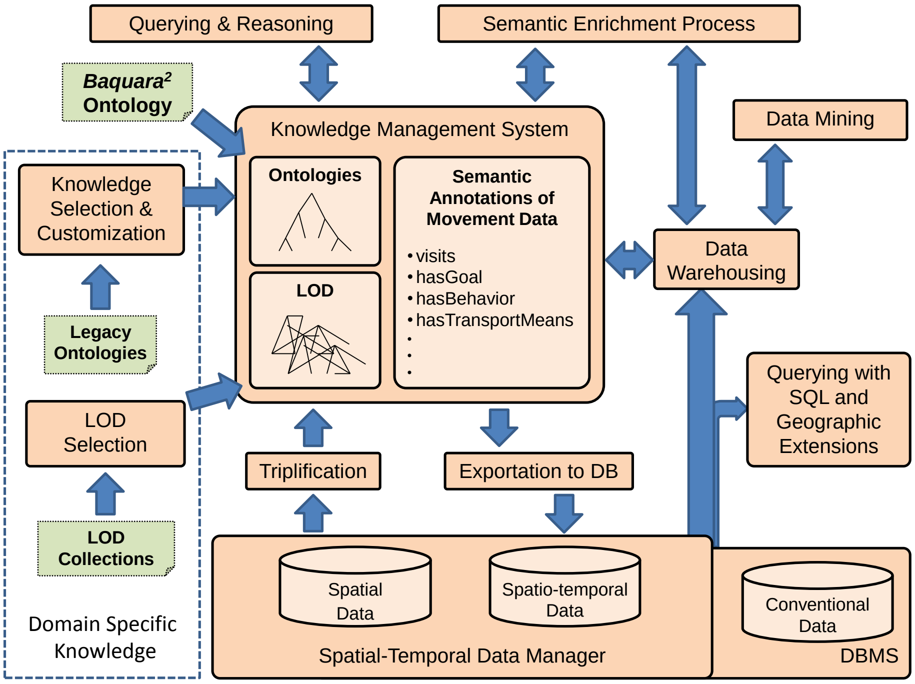


Figure 4: General architecture for semantic enrichment and analysis of movement data

The resulting semantically enriched movement data can be represented as a collection of RDF/RDFS triples, and maintained for Querying & Reasoning in the KB. This KB can be queried by using languages such as SPARQL and geoSPARQL. It can also be analyzed or further enriched by using a variety of reasoners that can be connected to the KB. The general architecture of Figure 4 also includes a conventional (e.g., relational) Database Management System (DBMS) and a Spatio-Temporal Data Manager. Triplification can be used to convert spatio-temporal data (e.g., movement data) from these systems into RDF triples, for allowing their processing in the KB whenever necessary or convenient. Conversely, Exportation to DB can be used to convert RDF triples of the KB into conventional, spatial, and spatio-temporal data managed by systems that efficiently support Querying with SQL and Geographic Extensions, Data Warehousing, and Data Mining.

5. Experiments

The viability of the proposed approach to semantically enrich and analyze movement data has been investigated in two case studies, using data extracted from Flickr and Twitter, respectively, and LOD from DBpedia and LinkedGeoData. Specializations of the semantic enrichment process described in Section 4.1 have been implemented on PostGIS[8], and applied to derive semantic annotations for stops and individual positions of moving objects. The resulting semantically annotated movement data supports queries in SQL with geographic extensions on PostGIS. They can also be converted into RDF triples, and stored in a KB compliant with the *Baquara*[2] ontology, to be queried with SPARQL and geoSPARQL, among other possibilities.

5.1. Flickr

Sample Flickr data for experiments were extracted from CoPhIR[9], by filtering tuples with spatio-temporal points inside Brazil, and eliminating moving objects positions sequences (MOPS) having any subsequence with speed higher than 500 km/h. The resulting raw data collection, that consists in 14,504 positions of 564 distinct users, was segmented in 2,143 user's trails (by just breaking each MOPS in a daily basis, as more sophisticated methods for trajectories reconstructing is beyond the scope of this paper). The positions are associated with 12,443 distinct tags. The total number of textual annotations is 117,146. Thus, each spatio-temporal sample point is associated to 8.08 tags in average. The user's daily trails were further segmented in 971 stops, each one corresponding to a period of at least 30 minutes without moving more than 500 meters. These stops are associated to 6,278 distinct tags, in a total of 45,768 (stop,tag) pairs, i.e., around 47 different tag values associated to each stop, in average. Figure 5 illustrates the distribution of the obtained trails across Brazil (left side), and the tags associated to a particular stop in the city of Rio.

The Flickr users' positions have been semantically enriched with 97,242 resources of DBpedia and LinkedGeoData associated to geographic coordinates

[8] http://postgis.net/

[9] http://cophir.isti.cnr.it

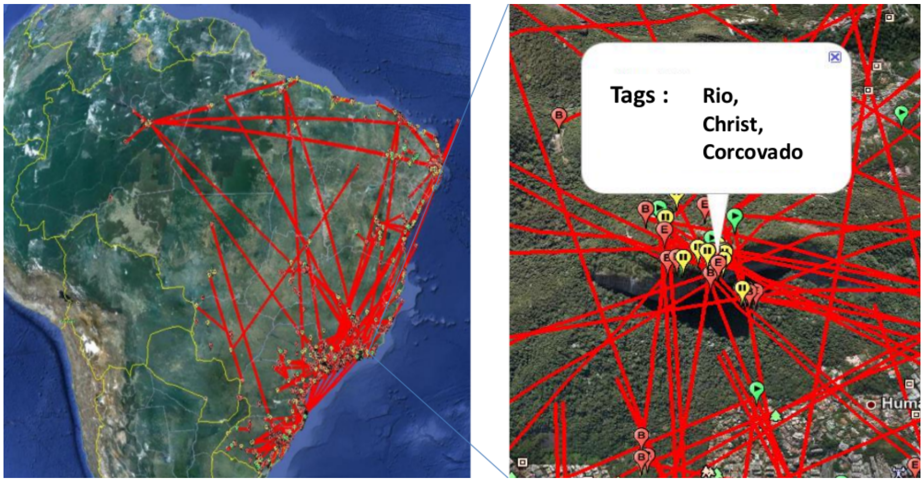


Figure 5: CoPhIR Flickr trails across Brazil (left), and close to Corcovado in Rio (right)

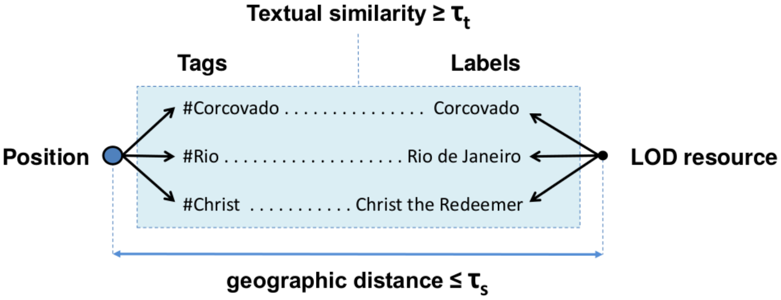


Figure 6: Linking a Flickr user tagged position to a geo-referenced labeled LOD resource

inside an MBR fitting the Brazilian territory, and having at least a label. The linking of these resources with Flickr positions was done with a variation of Algorithm 1 that employs Euclidean distance as geographic distance with a threshold $\tau_s = 1$ km to match the positions. For the pairs position-resource within 1 kilometer from each other, the set of tags associated to the position is then compared with the set of labels of the resource, by using soft-TFIDF [43] to compose Jaro-Winkler [42] similarities of pairs (tag,label) that are above the textual threshold τ_t in a unique similarity measure for the respective pair (position,resource). Figure 6 illustrates the linking criteria with the textual similarities considered between tags and labels of a pair (position,resource) referring to the place called `Corcovado` in the city of `Rio`. Finally, Figure 7 illustrates a situation in which the similarity between tags and labels is crucial to link the moving object's position to the correct resource. This position is in a densely populated area, and its spatial coordinates are not precise enough to decide what is the best matching among the many resources in the surroundings.

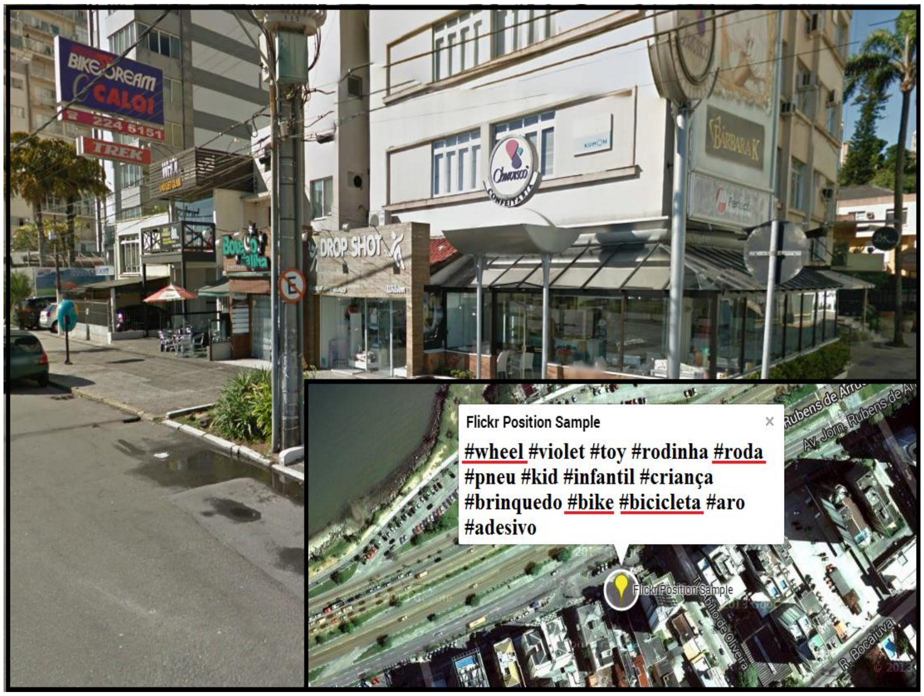

BIKE DREAM
CALOI
224 5151
TREK
DROP SHOT
Flickr Position Sample
#wheel #violet #toy #rodinha #roda
#pneu #kid #infantil #criança
#brinquedo #bike #bicicleta #aro
#adesivo
Flickr Position Sample
Av. Jorn. Rubens de Arruda
R. Bocaiuva

Figure 7: Example of position-resource link enabled by Soft-TF-IDF textual similarity

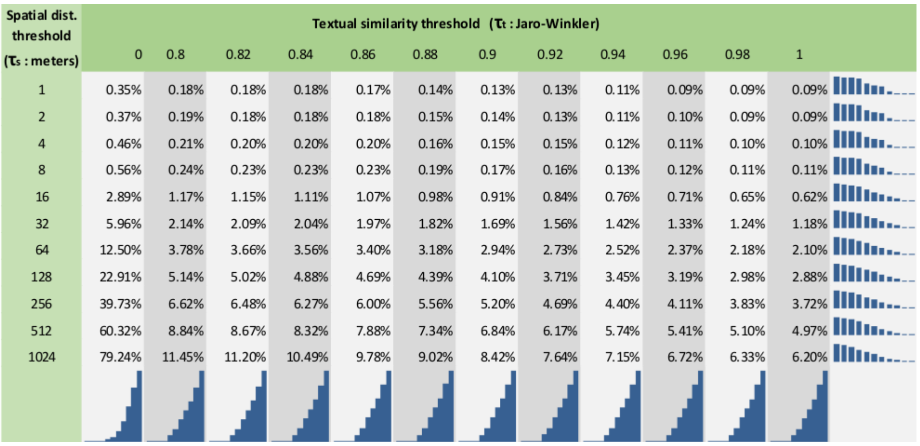
5.2. Twitter

Another variation of Algorithm 1 has been run on 57,099,806 tweets collected via the Twitter API[10] during the World Cup 2014 period (6 June 2014 to 7 July 2014). These tweets have geographic coordinates inside an MBR fitting the Brazilian territory. They were filtered to take the 1,183,354 ones originated from FourSquare, to increase the probability of matching their locations with the 97,242 geo-referenced resources selected of DBPedia and LinkedGeoData in July 21 2014. The linking algorithm associates a tweet with a LinkedGeoData resource (e.g., an instance of shop) when the coordinates of the resource are up to τ_s meters away from the tweet, and the LinkedGeoData resource has at least a label value whose Jaro-Winkler similarity [43] with the name of the tweet location value is higher than τ_t.

The first experiments with Twitter data just filtered the matchings satisfying the spatial and textual thresholds (τ_s and τ_t, respectively), i.e., without doing the refinements of lines 4 to 17 of Algorithm 1. Figure 8 shows the distribution of the percentage of the tweets associated with at least one LOD resource for distinct values of the thresholds τ_s and τ_t. The percentage of tweets with at least one resource in a radius of τ_s meters (column $\tau_t = 0$) jumps from 0.35%

[10] https://dev.twitter.com/docs/api

to 79, 24% by increasing τ_s from 1 to 1024 meters. Nevertheless, values of the textual threshold τ_t between 0.8 and 1 (range that properly filters similarities in our observations) eliminate a considerable percentage of matchings. Figure 9 presents the average number of associated LOD resources per tweet for the same scales of values for τ_s and τ_t as those of Figure 8. Notice that $\tau_t \geq 0.8$ results in an average number of associations per tweet equal to 1 (in bold) or close to 1, meaning no or few ambiguities (the same movement segment associated to distinct LOD resources), respectively. However, for values of τ_s close to 1 km, almost half of the associations constitute ambiguities. On the other hand, filtering with low values for τ_s and high values for τ_t eliminates ambiguities, but may also eliminate a considerable number o valid matchings.

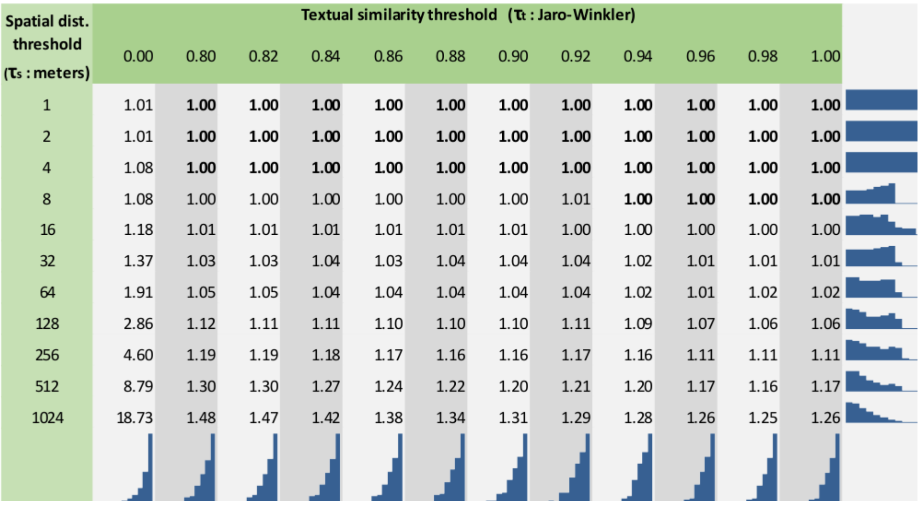


Spatial dist. threshold	Textual similarity threshold (τt : Jaro-Winkler)											
(τs : meters)	0	0.8	0.82	0.84	0.86	0.88	0.9	0.92	0.94	0.96	0.98	1
1	0.35%	0.18%	0.18%	0.18%	0.17%	0.14%	0.13%	0.13%	0.11%	0.09%	0.09%	0.09%
2	0.37%	0.19%	0.18%	0.18%	0.18%	0.15%	0.14%	0.13%	0.11%	0.10%	0.09%	0.09%
4	0.46%	0.21%	0.20%	0.20%	0.20%	0.16%	0.15%	0.15%	0.12%	0.11%	0.10%	0.10%
8	0.56%	0.24%	0.23%	0.23%	0.23%	0.19%	0.17%	0.16%	0.13%	0.12%	0.11%	0.11%
16	2.89%	1.17%	1.15%	1.11%	1.07%	0.98%	0.91%	0.84%	0.76%	0.71%	0.65%	0.62%
32	5.96%	2.14%	2.09%	2.04%	1.97%	1.82%	1.69%	1.56%	1.42%	1.33%	1.24%	1.18%
64	12.50%	3.78%	3.66%	3.56%	3.40%	3.18%	2.94%	2.73%	2.52%	2.37%	2.18%	2.10%
128	22.91%	5.14%	5.02%	4.88%	4.69%	4.39%	4.10%	3.71%	3.45%	3.19%	2.98%	2.88%
256	39.73%	6.62%	6.48%	6.27%	6.00%	5.56%	5.20%	4.69%	4.40%	4.11%	3.83%	3.72%
512	60.32%	8.84%	8.67%	8.32%	7.88%	7.34%	6.84%	6.17%	5.74%	5.41%	5.10%	4.97%
1024	79.24%	11.45%	11.20%	10.49%	9.78%	9.02%	8.42%	7.64%	7.15%	6.72%	6.33%	6.20%

Figure 8: Percentage of tweets associated with at least one LOD resource

Figure 10 shows that the vast majority of the associated tweets are linked to just 1 LOD resource for $\tau_s = 16$ meters and $\tau_t = 0.9$ (left side), and that there are ambiguities in almost half of the associated tweets for $\tau_s = 1024$ meters and $\tau_t = 0.8$. It suggests that the textual similarity can play just a limited role on disambiguation. Notwithstanding, textual similarity can be crucial to make correct links, as in the scenario illustrated by Figure 11, in which the best matching resource is not the geographically closest to the tweet.

The experiments have been repeated with the same Twitter data and LOD as input, but performing the refinements of lines 4 to 17 of Algorithm 1 to link each movement segment ms only to the resource(s) that are the closest to ms in the geographic space, and whose location name is most similar to at least a label of that resource(s). Figure 12 presents the average number of associated resources per tweet for different values of τ_s and τ_t. In these results, the number of ambiguities do not vary monotonically with variations in τ_s and τ_t, because the refinement phase optimize the results for minimum τ_s and maximum τ_t, and the number of matching resources can vary for such optimized values. Notice that the average number of associated resources per tweet has decreased to values



Spatial dist. threshold (τ_s : meters)	Textual similarity threshold (τ_t : Jaro-Winkler)											
	0.00	0.80	0.82	0.84	0.86	0.88	0.90	0.92	0.94	0.96	0.98	1.00
1	1.01	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
2	1.01	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
4	1.08	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
8	1.08	1.00	1.00	1.00	1.00	1.00	1.00	1.01	**1.00**	**1.00**	**1.00**	**1.00**
16	1.18	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00
32	1.37	1.03	1.03	1.04	1.03	1.04	1.04	1.04	1.02	1.01	1.01	1.01
64	1.91	1.05	1.05	1.04	1.04	1.04	1.04	1.04	1.02	1.01	1.02	1.02
128	2.86	1.12	1.11	1.11	1.10	1.10	1.10	1.11	1.09	1.07	1.06	1.06
256	4.60	1.19	1.19	1.18	1.17	1.16	1.16	1.17	1.16	1.11	1.11	1.11
512	8.79	1.30	1.30	1.27	1.24	1.22	1.20	1.21	1.20	1.17	1.16	1.17
1024	18.73	1.48	1.47	1.42	1.38	1.34	1.31	1.29	1.28	1.26	1.25	1.26

Figure 9: Average number of associated LOD resources per tweet

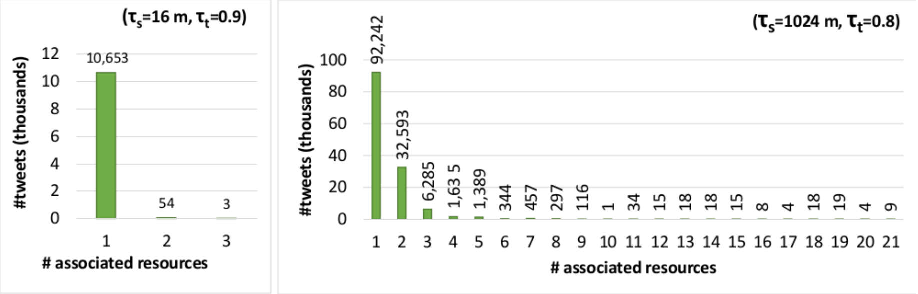


Figure 10: Distribution of the number of associated resources to tweets

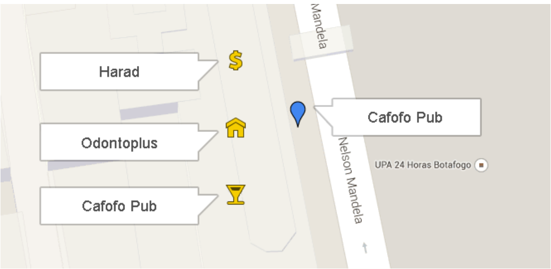
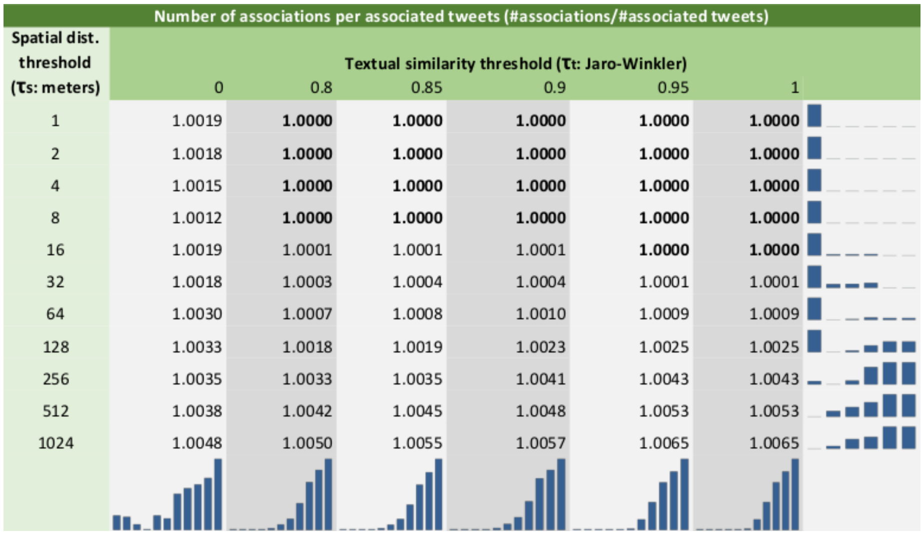


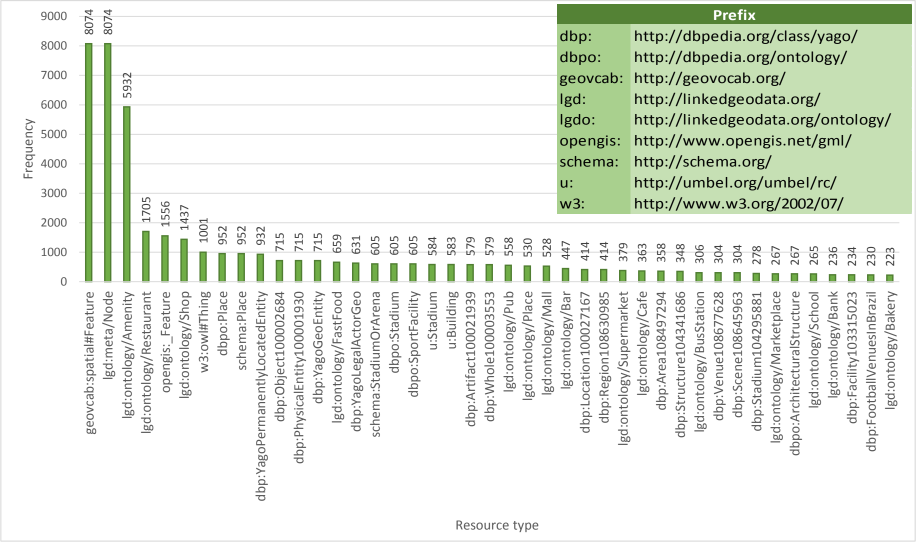
Figure 11: Example of a tweet position linked to most similarly labeled place

equal or closer to 1 than those in Figure 9, even for high values of τ_s. In fact, the total number of ambiguities have been reduced to less than 0.5% for $\tau_s = 1$ km and $\tau_t = 0.8$, and is around 0.01% for $\tau_s = 16$ m and $\tau_t = 0.9$. In addition, some slight gains in the execution time have been observed in experiments that use the refinement portion of Algorithm 1, compared to just filtering by τ_s and τ_t. It happens because the textual similarity is only calculated in Algorithm 1 when a closer resource than the previously matched one(s) is found in line 9.



Number of associations per associated tweets (#associations/#associated tweets)						
Spatial dist. threshold (τs: meters)	Textual similarity threshold (τt: Jaro-Winkler)					
	0	0.8	0.85	0.9	0.95	1
1	1.0019	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
2	1.0018	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
4	1.0015	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
8	1.0012	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
16	1.0019	1.0001	1.0001	1.0001	**1.0000**	**1.0000**
32	1.0018	1.0003	1.0004	1.0004	1.0001	1.0001
64	1.0030	1.0007	1.0008	1.0010	1.0009	1.0009
128	1.0033	1.0018	1.0019	1.0023	1.0025	1.0025
256	1.0035	1.0033	1.0035	1.0041	1.0043	1.0043
512	1.0038	1.0042	1.0045	1.0048	1.0053	1.0053
1024	1.0048	1.0050	1.0055	1.0057	1.0065	1.0065

Figure 12: Average number of associated LOD resources per tweet (refined)



Prefix	
dbp:	http://dbpedia.org/class/yago/
dbpo:	http://dbpedia.org/ontology/
geovcab:	http://geovocab.org/
lgd:	http://linkedgeodata.org/
lgdo:	http://linkedgeodata.org/ontology/
opengis:	http://www.opengis.net/gml/
schema:	http://schema.org/
u:	http://umbel.org/umbel/rc/
w3:	http://www.w3.org/2002/07/

Figure 13: Top matched LOD types with tweets collected during Brazil World Cup 2014

Finally, Figure 13 presents the distribution of the top matched LOD resource types in the generated associations with the collected tweets whose geographic coordinates are inside the MBR fitting the Brazilian territory during World Cup 2014, in experiments that used the whole Algorithm 1, with $\tau_s = 16$ meters, and $\tau_t = 0.9$. These matchings help to bring forth interesting results for a variety of queries, such as the ones presented in the following.

5.3. Example Queries

Baquara^2 ontology compliant knowledge bases (KBs) built with movement data extracted from sources such as Flickr and Twitter, and semantically enriched with LOD of collections such as DBPedia and LinkedGeoData can be stored in an RDF triples repository such as that of Virtuoso[11], which supports SPARQL and GeoSPARQL. Of course, besides Algorithm 1, further processing is necessary for producing hierarchies of semantically annotated movement segments as described in Section 2. A variety of methods and tools can help perform this task [46, 25, 26, 18, 1, 16, 27].

The following SPARQL queries complement the geoSPARQL ones presented in [47], by providing examples involving semantic constraints in multiple levels of movement segments hierarchies. Consider that the prefix `bq` refers to the URI of the *Baquara*^{2} ontology[12], and that LOD from different sources have been pre-processed to consolidate the properties of resources linked by the `sameAs` property. The property `hasDuration` is available for movement segments in the *Baquara*^{2} ontology to avoid having to calculate the movement segment duration in minutes in queries by using its time span.

Query 1 *Select the social media user's trails with at least one stop to visit a mountain called Corcovado in the city of Rio, followed by one stop in a marketplace, where he/she does at least one finer stop in a restaurant.*

```
SELECT ?t WHERE {
 ?t a bq:Trail.
 ?sc bq:father ?t;  a bq:Stop;  bq:ord ?sc_ord;
     bq:visits ?corcovado.
 ?corcovado a <http://schema.org/Mountain>;
            rdfs:label "Corcovado";
            <http://dbpedia.org/property/location> ?rio.
 ?rio a <http://dbpedia.org/ontology/City>;
      rdfs:label "Rio de Janeiro".
 ?sm bq:father ?t;  a bq:Stop;  bq:ord ?sm_ord;
     bq:visits ?mp.
 ?mp a <http://linkedgeodata.org/ontology/Marketplace>.
 ?sr  bq:father ?sm;  a bq:Stop;  bq:visits ?r.
 ?r a <http://linkedgeodata.org/ontology/Restaurant>.
 FILTER(?sc_ord < ?sm_ord)}
```

[11] http://virtuoso.openlinksw.com

[12] Prefix bq:<`http://www.seek-project.eu/Baquara02`>

Query 2 *Determine the percentage of European's trails in Brazil that make at least a stop in a nature reserve, where he/she does at least one finer stop in a tourist shop.*

```
SELECT COUNT(DISTINCT ?ts) / COUNT(DISTINCT ?t) WHERE {
 ?ts a bq:Trail;  bq:isOfMOclass bq:European;  bq:visits ?b.
 ?t a bq:Trail;  bq:isOfMOclass bq:European;  bq:visits ?b.
 ?b a <http://dbpedia.org/ontology/Country>;
     rdfs:label "Brazil".
 ?sr bq:father ?ts;  a bq:Stop; bq:visits ?r.
 ?r a <http://linkedgeodata.org/ontology/NatureReserve>.
 ?hs bq:father ?r;  a bq:Stop;  bq:visits ?s;
 ?s <http://linkedgeodata.org/ontology/TouristShop>.}
```

Query 3 *Select the `Cities` with the largest number of `trails` inside them, with at least one stop in a `Marketplace`, that is preceded by a `stop` in a `Hotel`, and followed by a `stop` in a `Nightclub`, that lasts at least 2 hours. The stop in the `Marketplace` must be detailed in at least one stop in a `Pub`.*

```
SELECT ?cityLabel, COUNT(DISTINCT ?t) AS ?nts WHERE {
 ?t a bq:Trail;  bq:visits ?city.
 ?city a <http://linkedgeodata.org/ontology/City>;
     rdfs:label ?cityLabel.
 ?sm bq:father ?t;  a bq:Stop;  bq:ord ?sm_ord;
     bq:visits ?mp.
 ?mp a <http://linkedgeodata.org/ontology/Marketplace>.
 ?sh bq:father ?t;  a bq:Stop;  bq:ord ?sh_ord;
      bq:visits ?h.
 ?h a <http://linkedgeodata.org/ontology/Hotel>.
 ?snc bq:father ?t;  a bq:Stop;  bq:ord ?snc_ord;
      bq:hasDuration ?snc_dur;   bq:visits ?nc.
 ?nc a <http://linkedgeodata.org/ontology/Nightclub>.
 ?sp bq:father ?sm;  a bq:Stop;   bq:visits ?p.
 ?p a <http://linkedgeodata.org/ontology/Pub>.
 FILTER((?snc_dur >= 120) && (?snc_ord > ?sm_ord) &&
        (?sh_ord < ?sm_ord))}
GROUP BY ?cityLabel  ORDER BY DESC(?nts)
```

6. Related Work

The present article extends and improves a previous work [47], that introduced the Baquara ontology, and firstly exploited links between movement data and LOD. First, it generalizes movement segments hierarchies to allow arbitrary levels of refinement. Secondly, it details on the *Baquara*^{2} ontology, and on the linking process, that had been just sketched in [47]. Thirdly, it provides further experimental results with bigger amounts and variety of social media data for assessing the effectiveness of the proposed linking algorithm.

A core contribution of this work is the conceptual model conveyed by the *Baquara*^{2} ontology, which enables knowledge-based semantic description and analysis of movements in several abstraction levels. A pioneering work on conceptual modeling of spatio-temporal objects is MADS (Modeling Application Data with Spatio-temporal features) [48]. MADS extends the basic ER model with spatio-temporal constructs. The key MADS premise is that spatial and temporal concepts are orthogonal. MADS uses the object-relationship paradigm, including the features of the ODMG (Object Database Management Group) data model, and provides spatial and temporal data types, attributes, and relationships. It offers a wide range of conceptual constructs to model the spatio-temporal world. A more recent contribution with focus on conceptual modeling of trajectories (spatio-temporal objects changing their geographical positions but not their shapes) comes from Spaccapietra *et al.* [12]. This model represents semantic trajectories as stops and moves, i.e., trajectory segments in which the object is stationary or moving, respectively. It has been the first attempt to embed semantics in the movement representation, but it lacks generality since other relevant semantic aspects are not explicitly taken into account. An extension of the "Stop-Move" model towards overcoming these limitations comes from the CONSTAnT conceptual model [20], which defines several semantic dimensions for movement analysis (e.g., goal, behavior).

Although the conceptual modeling of trajectories have seen a "convergence" mainly to the "Stop-Move" model [12], ontologies for movement data did not find so far an agreed approach. Focusing only on the works most related to our approach, we recall, for example, Yan *et al.* [1]. The ontology proposed in that work includes three modules: the Geometric Trajectory Ontology describes the spatio-temporal features of a trajectory; the Geographic Ontology describes the geographic objects; and the Domain Application Ontology describes the thematic objects of the application. These ontology modules are integrated into a unique ontology that supports conjunctive queries in a traffic application.

The work presented in [49] introduces a design pattern for semantic trajectories to enable the publishing as Linked Data. They describe the geo-ontology design pattern in OWL expressing the basic features of a semantic trajectory like the spatio-temporal information as a sequence of fixes and the semantic information like Points of Interest visited and device information.

The proposal of [11] exploits a movement ontology for querying and mining trajectory data enriched with geographic and application information. Here the ontology has been used to infer application-dependent behavior from raw and mined trajectory data. A pioneering work on movement patterns is instead the one of Dodge *et al.* [30] where authors propose a taxonomy of movement patterns distinguishing generic patterns (e.g. moving clusters, co-location) that represents any form of movement behavior and can be extracted applying generic data mining algorithms from behavioral patterns (e.g. flock, leadership) where the movement has a clear semantics and can be considered higher level movement patterns. A recent survey on semantic trajectory modeling and analysis is reported in [18].

However, these works do not address the automatic enrichment of trajectories with semantically precise information about specific places (e.g., restaurants, hotels, touristic spots), events (e.g., sport events, cultural events), and other relevant entities of the open dynamic world in which trajectories occur. In this article, movement segments are linked to specific concepts and/or instances via ontological relationships that can describe their precise semantics. Such semantic enrichment requires lots of continuously updated information, with well-defined and widely agreed semantics.

The conjugated use of textual and spatial information in social networks is another growing research theme. The recent work [50] proposes a method to exploit spatial proximity and users' common interests for querying location-based social networks. The problem is clearly NP-complete and the authors propose two efficient algorithms that explore the search space using two distinct criteria, that give good results in terms of performance compared to the state of the art.

Entity linking of social media data (e.g., tweets) is also a growing research theme, because in this context the task is particularly challenging: the text is noisy, short, and informal [36, 35]. Entity linking has been largely explored on the Web, mainly relying on the context around the entity [51, 52, 53]. However, these methods cannot be applied to tweets due to the insufficient context information. The interesting article [54] proposes a graph-based framework to collectively link all the named entity mentions in all the tweets posted by a user, with the assumption that each user has an underlying topic of interest distribution over various named entities.

In the field of semantic extraction from text, the challenge is to find patterns or associations in the text, in particular the co-occurrence of terms. An example of these kind of approach is [55], that proposes a generic framework for mining semantic associations in text. They base their idea on a co-occurrence graph and a set of primitives to mine four kinds of semantic associations, namely: topical anchors, semantic siblings, topical markers, and topic expansions.

7. Conclusions and Future Work

Vast amounts of linked open data (LOD) about real world entities and events have been fed and continuously updated on the Web. However, their potential to leverage movement understanding has not been exploited yet. This article proposes the *Baquara*^{2} knowledge-based framework as a bridge between movement analysis and knowledge bases, by allowing movement data and associated knowledge to be connected and queried together. It unleashes the use of growing collections of ontologies and LOD available in the Web to help semantically enrich and analyze a wide variety of movement data. The major contributions of this article are: (i) an ontology for semantic trajectories enrichment with linked data; (ii) an automated method to derive semantic annotations from textually annotated movement data; (iii) experimental results showing the viability of the proposal in case studies with real data available in the Web. Though this paper focus on knowledge management in engines such as triple stores, the latter constitute just an alternative means to handle movement data and knowledge. The semantically enriched movement data produced by using *Baquara*^{2} can be stored and efficiently processed in alternative kinds of spatio-temporal database management systems.

In our future work we plan to: (i) further evaluate the performance and the effectiveness of the proposed in several application domains; (ii) collect data and judgments of volunteers to serve as ground true for analyzing the quality of the semantic annotations generated by the proposed methods, with measures such as precision and recall; (iii) develop efficient and effective methods to derive precise semantic annotations from different movement data and LOD collections; and (iv) investigate the use of ontologies and LOD for movement data warehousing and data mining.

Acknowledgements

This work has been supported by the European Unions IRSES-SEEK (grant 295179) and FP7-DATASIM (grant 270833) projects. The Brazilian researchers have also been supported by CNPq (grant 478634/2011-0), CAPES, FAPESC, and FEESC. Nikos Pelekis research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales.

References

[1] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, K. Aberer, Semantic trajectories: Mobility data computation and annotation, ACM TIST 4 (3).

[2] S. Rinzivillo, F. de Lucca Siqueira, L. Gabrielli, C. Renso, V. Bogorny, Where Have You Been Today? Annotating Trajectories with DayTag, in: SSTD, Vol. 8098 of LNCS, Springer, 2013, pp. 467–471.

[3] A. Doulamis, N. Pelekis, Y. Theodoridis, Easytracker: An android application for capturing mobility behavior, 2012 16th Panhellenic Conference on Informatics 0 (2012) 357–362.

[4] G. Broll, H. Cao, P. Ebben, P. Holleis, K. Jacobs, J. Koolwaaij, M. Luther, B. Souville, Tripzoom: An app to improve your mobility behavior.

[5] Z. Cheng, J. Caverlee, K. Lee, D. Z. Sui, Exploring millions of footprints in location sharing services, in: L. A. Adamic, R. A. Baeza-Yates, S. Counts (Eds.), ICWSM, The AAAI Press, 2011.

[6] Z. Yin, L. Cao, J. Han, J. Luo, T. S. Huang, Diversified trajectory pattern ranking in geo-tagged social media, in: SDM, SIAM / Omnipress, 2011, pp. 980–991.

[7] M. Azmandian, K. Singh, B. Gelsey, Y.-H. Chang, R. T. Maheswaran, Following human mobility using tweets, in: L. Cao, Y. Zeng, A. L. Symeonidis, V. Gorodetsky, P. S. Yu, M. P. Singh (Eds.), ADMI, Vol. 7607 of LNCS, Springer, 2012, pp. 139–149.

[8] L. Gabrielli, S. Rinzivillo, F. Ronzano, D. Villatoro, From tweets to semantic trajectories: Mining anomalous urban mobility patterns, in: J. Nin, D. Villatoro (Eds.), CitiSens, Vol. 8313 of LNCS, Springer, 2013, pp. 26–35.

[9] S. Kumar, F. Morstatter, H. Liu, Twitter Data Analytics, Springer Briefs in Computer Science, Springer, 2014.

[10] R. G. B. Nabo, R. Fileto, C. Renso, M. Nanni, Annotating Trajectories by Fusing them with Social Media Users' Posts, in: Brazilian Symposium on Geoinformatics, GeoInfo, Campos do Jordão, SP, Brazil (to appear), 2014.

[11] C. Renso, M. Baglioni, J. A. F. de Macêdo, R. Trasarti, M. Wachowicz, How you move reveals who you are: understanding human behavior by analyzing trajectory data, Knowl. Inf. Syst. 37 (2) (2013) 331–362.

[12] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. F. de Macêdo, F. Porto, C. Vangenot, A conceptual view on trajectories, Data Knowl. Eng. 65 (1) (2008) 126–146.

[13] M. Nanni, R. Trasarti, C. Renso, F. Giannotti, D. Pedreschi, Advanced knowledge discovery on movement data with the geopkdd system, in: EDBT, Vol. 426 of ACM International Conference Proceeding Series, ACM, 2010, pp. 693–696.

[14] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, R. Trasarti, Unveiling the complexity of human mobility by querying and mining massive trajectory data, VLDB J. 20 (5) (2011) 695–719.

[15] B. Furletti, L. Gabrielli, C. Renso, S. Rinzivillo, Analysis of GSM calls data for understanding user mobility behavior, in: BigData Conference, IEEE, 2013, pp. 550–555.

[16] N. Pelekis, Y. Theodoridis, Mobility Data Management and Exploration, Springer, 2014.

[17] S. Spaccapietra, C. Parent, Adding meaning to your steps (keynote paper), in: M. A. Jeusfeld, L. M. L. Delcambre, T. W. Ling (Eds.), ER, Vol. 6998 of LNCS, Springer, 2011, pp. 13–31.

[18] C. Parent, S. Spaccapietra, C. Renso, G. L. Andrienko, N. V. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. A. F. de Macêdo, N. Pelekis, Y. Theodoridis, Z. Yan, Semantic trajectories modeling and analysis, ACM Comput. Surv. 45 (4), article 42.

[19] A. S. Furtado, R. Fileto, C. Renso, Assessing the attractiveness of places with movement data, JIDM 4 (2) (2013) 124–133.

[20] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, L. O. Alvares, CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects, T. GIS 18 (1) (2014) 66–88.

[21] N. Pelekis, Y. Theodoridis, D. Janssens, On the management and analysis of our lifesteps, SIGKDD Explorations 15 (1) (2013) 23–32.

[22] R. Battle, D. Kolas, Enabling the geospatial Semantic Web with Parliament and GeoSPARQL, Semantic Web 3 (4) (2012) 355–370.

[23] K. Kyzirakos, M. Karpathiotakis, M. Koubarakis, Strabon: A semantic geospatial dbms, in: ISWC, Vol. 7649 of LNCS, 2012.

[24] G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà, Y. Theodoridis, Building Real World Trajectory Warehouses, in: MobiDE, ACM, 2008, pp. 8–15.

[25] J. A. M. R. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, V. Bogorny, DB-SMoT: A direction-based spatio-temporal clustering method, in: IEEE Conf. of Intelligent Systems, IEEE, 2010, pp. 114–119.

[26] V. Bogorny, H. Avancini, B. C. de Paula, C. R. Kuplich, L. O. Alvares, Weka-STPM: a Software Architecture and Prototype for Semantic Trajectory Data Mining and Visualization, T. GIS 15 (2) (2011) 227–248.

[27] F. Moreno, A. Pineda, R. Fileto, V. Bogorny, SMoT+: Extending the SMoT Algorithm for Discovering Stops in Nested Sites, Computing and Informatics 33 (2) (2014) 327–342.

[28] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, 2003.

[29] M. Wachowicz, R. Ong, C. Renso, M. Nanni, Finding moving flock patterns among pedestrians through collective coherence, International Journal of Geographical Information Science 25 (11) (2011) 1849–1864.

[30] S. Dodge, R. Weibel, A.-K. Lautenschütz, Towards a Taxonomy of Movement Patterns, Information Visualization 7 (3) (2008) 240–252.

[31] L. O. Alvares, A. M. Loy, C. Renso, V. Bogorny, An algorithm to identify avoidance behavior in moving object trajectories, Journal of the Brazilian Computer Society 17 (3) (2011) 193–203.

[32] A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff, Semantic Annotation, Indexing, and Retrieval, Web Semantics: Science, Services and Agents on the World Wide Web 2 (1) (2004) 49–79.

[33] R. Fileto, A. Raffaetà, A. Roncato, J. A. P. Sacenti, C. May, D. Klein, A semantic model for movement data warehouses, in: 16th DOLAP, Shanghai, China, November 7 (to appear), 2014.

[34] W. Zhang, J. Su, C. L. Tan, W. T. Wang, Entity linking leveraging: Automatically generated annotation, in: 23rd Intl. Conf. on Computational Linguistics, COLING, 2010, pp. 1290–1298.

[35] X. Liu, S. Zhang, F. Wei, M. Zhou, Recognizing named entities in tweets, in: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 of HLT, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 359–367.

[36] A. Ritter, S. Clark, Mausam, O. Etzioni, Named entity recognition in tweets: An experimental study, in: Conf. on Empirical Methods in Natural Language Processing, EMNLP, 2011, pp. 1524–1534.

[37] J. Nothman, N. Ringland, W. Radford, T. Murphy, J. R. Curran, Learning multilingual named entity recognition from wikipedia, Artificial Intelligence 194 (0) (2013) 151 – 175.

[38] A. Gattani, D. S. Lamba, N. Garera, M. Tiwari, X. Chai, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, A. Doan, Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach, Proc. VLDB Endow. 6 (11) (2013) 1126–1137.

[39] X. Han, L. Sun, J. Zhao, Collective entity linking in web text: A graph-based method, in: Proc 34th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR, 2011, pp. 765–774.

[40] Z. Guo, D. Barbosa, Entity linking with a unified semantic representation, in: 23rd Intl. Conference on World Wide Web, WWW Companion, 2014, pp. 1305–1310.

[41] G. Navarro, A guided tour to approximate string matching, ACM Comput. Surv. 33 (1) (2001) 31–88.

[42] W. W. Cohen, P. D. Ravikumar, S. E. Fienberg, A comparison of string distance metrics for name-matching tasks, in: S. Kambhampati, C. A. Knoblock (Eds.), IIWeb, 2003, pp. 73–78.

[43] E. Moreau, F. Yvon, O. Cappé, Robust Similarity Measures for Named Entities Matching, in: 22nd Intl. Conf. on Computational Linguistics - Volume 1, COLING, 2008, pp. 593–600.

[44] P. Bouros, S. Ge, N. Mamoulis, Spatio-textual similarity joins, Proc. VLDB Endow. 6 (1) (2012) 1–12.

[45] S. Liu, G. Li, J. Feng, Star-join: Spatio-textual similarity join, in: 21st ACM Intl. Conf. on Information and Knowledge Management, CIKM '12, 2012, pp. 2194–2198.

[46] A. T. Palma, V. Bogorny, B. Kuijpers, L. O. Alvares, A clustering-based approach for discovering interesting places in trajectories, in: R. L. Wainwright, H. Haddad (Eds.), SAC, ACM, 2008, pp. 863–868.

[47] R. Fileto, M. Krüger, N. Pelekis, Y. Theodoridis, C. Renso, Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data, in: W. Ng, V. C. Storey, J. Trujillo (Eds.), ER, Vol. 8217 of LNCS, Springer, 2013, pp. 342–355.

[48] C. Parent, S. Spaccapietra, E. Zimányi, Conceptual modeling for traditional and spatio-temporal applications - the MADS approach, Springer, 2006.

[49] Y. Hu, K. Janowicz, D. Carral, S. Scheider, W. Kuhn, G. Berg-Cross, P. Hitzler, M. Dean, D. Kolas, A geo-ontology design pattern for semantic trajectories, in: T. Tenbrink, J. Stell, A. Galton, Z. Wood (Eds.), Spatial Information Theory, Vol. 8116 of LNCS, Springer, 2013, pp. 438–456.

[50] Y. Li, D. Wu, J. Xu, B. Choi, W. Su, Spatial-aware interest group queries in location-based social networks, Data & Knowledge Engineering 92 (0) (2014) 20–38.

[51] S. Cucerzan, Large-scale named entity disambiguation based on wikipedia data, in: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), June 28-30, Prague, Czech Republic, 2007, pp. 708–716.

[52] R. C. Bunescu, M. Pasca, Using encyclopedic knowledge for named entity disambiguation, in: EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy, 2006.

[53] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of wikipedia entities in web text, in: 15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, KDD, ACM, New York, NY, USA, 2009, pp. 457–466.

[54] W. Shen, J. Wang, P. Luo, M. Wang, Linking named entities in tweets with knowledge base via user interest modeling, in: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, KDD, 2013, pp. 68–76.

[55] A. R. Rachakonda, S. Srinivasa, S. Kulkarni, M. Srinivasan, A generic framework and methodology for extracting semantics from co-occurrences, Data & Knowledge Engineering 92 (0) (2014) 39–9.