WRAST: Warehousing Relatedness-Aware Semantic Trajectories

Goce Trajcevski[1⋆], Ivana Donevska[2], Alejandro Vaisman[3], Besim Avci[1], Tian Zhang[1], and Di Tian[1]

[1] Dept. of EECS, Northwestern University, Evanston (IL), USA
goce,besim,t-zhang,d-tian@eecs.northwestern.edu.com
[2] Dept. of CS, Indiana University – Purdue University, Fort Wayne (IN), USA
ivana@semantichouse.com
[3] Instituto Tecnológico de Buenos Aires, Buenos Aires, Argentina
avaisman@itba.edu.ar

Abstract. This work introduces methodologies for extending the modelling and querying capabilities of Trajectories Data Warehouses (TDW) in the context of semantic trajectories. Specifically, we incorporate the notion of *Semantic Relatedness* (SR) as part of the formal model of a TDW, which enables capturing the similarity between different annotations describing Points of Interest (POI), locations and activities used in specifying semantic trajectories. We formally define the functionality capturing the relatedness between different terms used as descriptors in semantic trajectories and present the *Semantic Relatedness in Trajectories Data Warehouse* (SR-TDW) model. We also present the newly enabled (categories of) queries in the SR-TDW model and illustrate them with specific examples. Our experimental observations demonstrate the benefits of the proposed approaches in terms of enriched answer-sets of the common OLAP-based queries and illustrate the sensitivity in terms of the relatedness measure.

1 Introduction and Motivation

The omnipresence of computing and sensing devices, and advances in networking and communications enabled the generation of huge volumes of location-in-time data – O(Peta-Bytes) per year – from the GPS of smart phone users, with up to 400-fold increase if cell-tower locations are included [14]. Efficient storage and retrieval of such information is essential for various applications – e.g., navigation, traffic management, disaster mitigation, etc. It is estimated that by 2020, more than 70% of mobile phones will have GPS capability, up from 20% in 2010 (similar trends apply to cars equipped with dashboard GPS devices) and smart routing [11] using data produced by such devices is expected to be around $500 billion.

The field of Moving Objects Databases (MOD) [12] has traditionally tackled the problem of storing and querying moving data produced by entities carrying

⋆ Research supported by NSF grants – CNS 0910952 and III 1213038.

location-aware devices. Recent research has extended moving objects analysis with an OLAP (Online Analytical Processing) kind of functionality for aggregating application-demanded knowledge, enabling decision-support tasks related to mobile data. Data Warehousing (DW) models and tools [25] have been augmented with capabilities for processing complex queries in Spatial OLAP (SOLAP) and Spatio-Temporal (ST-OLAP) settings [13, 17, 24].

The sequence of spatiotemporal positions of a moving object, having a certain start and end, is called the object's *raw trajectory*. These trajectories are useful for querying MOD data (e.g., "When is the next train to London expected to arrive?"). Mobility analysis, however, often does not require the full raw trajectory, and replacing raw data by certain places of interest (or street and crossing names) may suffice. For this, we need to identify places of interest (POIs) where an object stopped for a certain amount of time – or, the other way around, i.e, a POI may be discovered through the analysis of the time spent at a certain position. Thus, trajectories can be segmented into a sequence of *episodes* characterized as a sequence of *stops* at POIs, and *moves* occurring between two stops. This sequence, having a given start and end, is called a *semantic trajectory*. Episodes can be further annotated with contextual information, leading to the notion of *semantically-annotated trajectories* [6,18]. Figure 1 shows three semantically-annotated trajectories, ST_1, ST_2, and ST_3, along with some POIs (restaurants, fast food places, etc), where the trajectories stopped. The trajectory lines link the different kinds of POIs (e.g., street corners, restaurants, etc.).

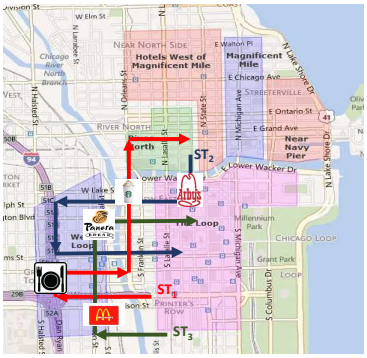


Fig. 1. Chicagoland trajectories

Trajectory Data Warehouses (TDWs) [8] and Semantic Trajectory Data Warehouse (STDW) [3, 20, 26] are aimed at aggregating and analyzing trajectory data, e.g., using OLAP and data mining techniques, as exemplified with the query **Q_1** below to a TDW containing ST_1, ST_2, and ST_3 in Figure 1:

Q_1: *Daily number of trajectories in the first week of June, that started in the Loop, first stopped at a restaurant, and then at a coffee house, both within 2 miles from West Loop.*

Typical proposals that extend trajectories with annotations [5] and account for spatial data [26], would return ST_1 as the only trajectory satisfying both the semantic and spatial conditions in Q_1, returning "1" as a result of the `COUNT` aggregate function. However, a careful observation of Figure 1 reveals that: (1) ST_2 may also be an acceptable answer, since it did stop at a fast-food place, followed by a stop at Starbucks; (2) Similarly ST_3 stopped first at a fast-food place, and then at a pastry, and thus it may also be an acceptable answer. Both ST_2 and ST_3 could satisfy **Q_1** depending on the application and/or user requirements, which must state to what extent

we can consider a fast-food place analogous to a restaurant, a pastry similar to a coffee-shop, and so on. For example, we may consider that ST_3 is "closer" to ST_2 than ST_1 or viceversa, depending on the similarity model adopted. To account for this problem, in this paper we extend TDWs and STDWs with the notion of *semantic relatedness* [2, 7], which enables retrieving concepts of interest and computing aggregates with a predefined correlation value instead of a strict term matching. We call this novel model SR-TDW (Semantic Relatedness in Trajectory Data Warehouses). In our example, given a threshold Θ, if the similarity measure for the attributes correlated to the ones in ST_1 in both ST_2 and ST_3, is $\geq \Theta$, we would obtain "3" as an outcome of the `COUNT` value.

Essentially, semantic relatedness quantifies the knowledge of "how close" are two terms used in the annotation of the respective attributes of the participating trajectories, examples of which abound. Consider for instance a collection of trajectories segmented according to "stop" and "move" episodes. Each "move" episode could be annotated with its associated mean of transportation: the transportation mode of one episode may be a "car", whereas an episode in the same or in another trajectory may be "vehicle". Both are, intuitively, more related to each other than the term "bicycle". Similarly, the tags used in activities description (cf. [6]) may vary from "restaurant", through "fast-food", to "eatery", and all are semantically closer to each other than the term "bar". Note that, even though the notion of relatedness may comprise the concept of generalization (like in the car-vehicle case), it is clearly more general, e.g., there is no generalization between the concepts of restaurant and bar, although both may be considered as an specialization of the concept of "food house". The above example can be straightforwardly extended to various domains and, to the best of our knowledge, TDWs have not fully exploited the concept of semantic relatedness, an issue at the core of our motivation, for which our main contributions are:

— We present the SR-TDW model, which augments the TDW models both by capturing extended information about semantic annotations of trajectories, and the relatedness among different (classes of) terms.

— We introduce novel queries/operators which incorporate the value of the semantic relatedness when determining the candidates for an answer-set.

— We present experimental observations evaluating the benefits of the novel SR-TDW model when applied on a dataset of semantic trajectories from Chicago, illustrating the impact of the proposed approach and the different measures for semantic relatedness on the answer-sets.

In the remainder of this paper, Section 2 introduces the basic terminology and background about the formalisms used. Section 3 introduces the main modelling results – the notion of semantic relatedness and how it is incorporated in the SR-TDW model. In Section 4 we present examples of queries and aggregation with semantic relatedness. Section 5 presents our experimental observations, Section 6 compares our work with relevant literature, and Section 7 concludes the paper and outlines directions for future work.

2 Preliminaries

We now introduce the basics of Semantic Trajectories (ST) and TDWs.

2.1 Semantically Enriched Trajectories

Semantic (synonymously, Symbolic or Enriched) *Trajectories* [3, 6, 18] embed contextual and/or situational knowledge into location-in-time data. In a MOD [12] a trajectory is modelled as a structure of the form $Tr_i = [o_{ID}, (x_{i1}, y_{i1}, t_{i1}), \ldots, (x_{ik}, y_{ik}, t_{ik})]$, where x_{ij} and y_{ij} $(1 \leq j \leq k)$ are the coordinates of the location $(l_{ij} = (x_{ij}, y_{ij}))$ of the object with a unique identifier oID, obtained at time instant t_{ij}. In-between two consecutive updates, objects are assumed to move in accordance with some kind of an interpolation. STs attempt also to describe the kinds of activities associated with a particular location and time – e.g., "stop", "move", "walk", "eat", etc. Formally (cf. [6,18]), a semantic trajectory ST_i is a sequence of so-called, semantic episodes $se_{i,m}$ as follows:
$ST_i = [se_{i1}, se_{i2}, se_{i3}, \ldots se_{im}]$, where the j-th semantic episode of the i-th semantic trajectory is a tuple of the form:
$se_{ij} = (da_{ij}, sp_{ij}, x_{ij}^{in}, y_{ij}^{in}, t_{ij}^{in}, x_{ij}^{out}, y_{ij}^{out}, t_{ij}^{out}, tagList_{ij})$ where:

- da_{ij} = defining annotation; typically expressing an activity (verb) such as stop or move.
- sp_{ij} = semantic location/position of the activity, like a POI (e.g., a museum, restaurant, zoo), home, work, etc.
- t_{ij}^{in} and t_{ij}^{out} = entry/exit times of a semantic position.
- $x_{ij}^{in}, y_{ij}^{in}, x_{ij}^{out}, y_{ij}^{out}$ = entry/exit coordinates of a semantic position.
- tagList$_{ij}$ = any additional semantic information, like transportation mode, additional activity description (e.g., eat), etc.

As an example, assume that there is a coordinate center (0,0) located at the bottom-left corner in Figure 1 and the axes are 100 units in length each. Then, the semantic trajectories ST_1 and ST_2 in Figure 1 can be specified as:

ST_1 = [(drive, Adams_St, 50, 10, 10:45, 10, 10, 11:00, drive, car, VW_Passat)
(stop, "Roditis", 10, 10, 11:00, 10, 10, 11:45, restaurant, eat, salad),
(walk, parking_lot, 10, 10, 11:45, 11, 10, 11:50, car, VW_Passat),
(drive, Randolph_St, 11, 10, 11:55, 25, 10, 12:00, car),
(stop, traffic_light, 25, 10, 12:00, 25, 10, 12:03, car),
...
(stop,"Starbucks", 25, 40, 12:25, 25, 40, 1:30, restaurant, eat, dessert, coffee)]

ST_2 = [(move, Dearborn St, 60, 60, 11:30, 60, 40, 11:45, walk),
(stop, "Arbys" , 60, 40, 11:45, 60, 40, 12:30, fast-food, eat, beef),
(move, Dearborn St, 60, 40, 12:30, 60, 35, 13:00, walk),
(move, Chicago Ave, 50, 35, 13:00, 25, 35, 13:25, ride, bus_14),
(stop, "Starbucks", 25, 35, 13:25, 25, 35, 13:50, coffee, desert),
...
(move, Jackson St, 10, 20, 14:15, 50, 20, 14:40, ride, bus_151)]

While there is a match between the third and the second *stop* activities in ST_1 and ST_2, respectively (i.e., both involve "Starbucks"), the first *stop* activity of ST_2 involves "fast-food". However, stopping at "Arby's" (ST_2) is, in some sense, semantically related to stopping by at the "Roditis" restaurant (ST_1).

2.2 Warehousing Trajectory Data

Many works have tackled the problem of using OLAP techniques for exploration of spatial data. This has been called SOLAP (for Spatial OLAP) [1]. The basic idea of the solutions proposed is to add spatial data type support to conventional DW dimensions and measures, yielding the concept of Spatial DW. When spatial objects vary across time, we are in the field of spatiotemporal data warehousing (STDW) [24]. Trajectory Data Warehouses (TDW) [17, 24] are a particular case of STDW, where trajectories (raw or semantic ones) are part of the DW, either as dimensions or measures. Typically, trajectories are facts which are segmented into episodes according to associated dimensions, which can be traditional (i.e., containing alphanumerical data) or spatial [20, 21, 25]. Another, simpler approach, consists in dividing the space into a 2- or 3-dimensional grid (i.e., the dimensions are the x,y,z spatial coordinates). We may also have additional dimensions representing the moving objects' profile, the time dimension, etc. The measures in this approach are a collection of pre-aggregated values of the trajectories. For example, a measure could be the number of trajectories in a cell of the grid in a certain time interval. That means, trajectories themselves are lost. Details can be found in [21, 25]. Finally, some recent work also make use of the emerging semantic trajectories paradigm, to model so-called semantic TDWs [8, 26].

In this paper we consider semantic episodes as the basic building blocks for the SR-TDW model, equivalently, a trajectory segment. Each such fact-episode is linked to the spatial and temporal hierarchies, and to other dimensions such as POIs and their geo-coordinates along with other semantic-based information. Due to space limitations, we assume the reader is familiar with the notion of traditional OLAP and DWs, so we omit details in this sense.

3 Semantic Relatedness and Trajectories Warehousing

We now introduce the concept of semantic relatedness, apply it to symbolic trajectories, and define the SR-TDW model.

3.1 Semantic Relatedness

The notion of *semantic relatedness* quantifies the "semantic proximity" of two concepts or entities not only via similarity between objects, but also incorporating other features, like their "popularity" or how often those two entities appear together or are referenced by users [7, 19]. As discussed in [2, 19] there are various measures and evaluation techniques for semantic relatedness and multiple

connections (even multiple hierarchies) can exist between entities – e.g., common contexts and synonyms, like (*car*, *automobile*); hypernymy relationships, e.g., (*car*, *vehicle*) (that means, an *isA* or subcategory relationship); meronymy (is-part-of) relationship, like in (*finger*, *hand*); or other functional association not based on lexical relationships, like in (*penguin*, *Antartica*) [10].

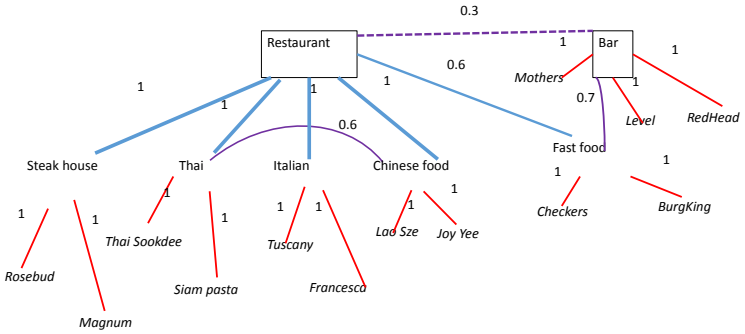


Fig. 2. Semantic relatedness between terms

Broadly speaking, the calculation of semantic relatedness is based on a graph in which nodes correspond to terms, and edges represent (strengths of) semantic connections. There are different approaches for assigning weights and targeting a different group of semantic connections [2, 4, 19]. For example, the approach in [7] makes five passes over the existing connections, where the first pass inserts the core nodes, which are nouns extracted from WordNet (http://wordnet.princeton.edu/wordnet/). Nouns are then connected to their sense, and the probability of the transition from one node to the other is the popularity of that sense. Weight is then given to synonymy, hyponymy/hypernymy relationships, and to words appearing in similar contexts (based on the number of occurrences of a particular meaning in a given context). For a completed graph and a given edge (n_1, n_2) between nodes n_1 and n_2, let P_t denote a non-cyclic path from node "A" to node "C". The relatedness between "A" and "C" is calculated as: $R(C|A) = \sum_{P_t} P_{P_t}(C|A)$ – sum of all the acyclic paths P_t from node A to node C, each one representing the likelihood of the relatedness between A and C based on a given context.

The value of a particular path-similarity between two nodes is calculated as the product of the weights of all edges along the path that connects them $P_{P_t}(C|A) = \prod_{(n_1,n_2)} P(n_2, n_1)$, with edge-weight $0 < P(n_2, n_1) \leq 1$. As the edges are multiplied along the path, the similarity value gradually decreases, which is the desired behaviour: as the number of edges separating the two nodes increases, the similarity value decreases.

These concepts are illustrated in Figure 2 where, for example, the relatedness between *Thai Soodkee* and *Siam Pasta* is 1; the relatedness between *Siam Pasta* and *Joy Yee* is $1 + 0.6 = 1.6$; and the relatedness between *Rosebud* and *Checkers* is $(0.6 + 0.3*0.7 + 0.6*0.6) = 1.16$.

In this paper we use the notion of relatedness in the semantic trajectory setting, to augment the traditional geo-spatial and activity-based attributes such as POIs, walk, etc., with an explicit representation of their relatedness. This semantic enhancement, which, to the best of our knowledge has not been fully exploited in TDW setting, has an impact over the query results, allowing to obtain answers which, otherwise, would remain hidden. We provide a generic framework for comparing specific POIs, as well as other contextual relatedness linking the nouns (e.g., in *da*'s and *sp*'s from a particular semantic trajectory) with both nouns and verbs from the *tagList* (cf. Section 2.1).

3.2 Extending Trajectories Data Warehouses with Semantic Relatedness

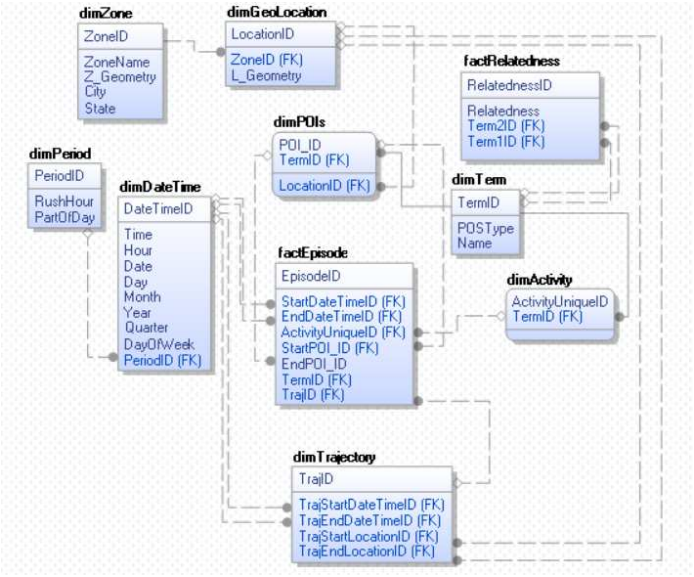


Fig. 3. TDW with Semantic Relatedness

We now proceed with extending STDW with the notion of semantic relatedness, yielding the SR-TDW model. As we outlined in Section 1, when it comes to implementing the advanced capabilities for analytical solutions based on trajectory warehousing, there are two foundational approaches: (a) the "raster-like" one [17] where the 2D geographic space is decomposed into cells of a grid, and,

for each trajectory, only aggregated data within a cell are kept (e.g., the maximum speed of the trajectory in the cell, or the distance traversed in the cell); and (b) the "vector-like" one [20, 25], where trajectory segments are represented as geometric types. Extended models incorporate the concept of continuous fields (cf. [21]), which we do not consider in this work. In this paper we follow the "vector-like" trajectory data warehouse model, and we extend its traditional functionalities beyond the currently available geo-spatial properties. More specifically, we augment the use of semantics by incorporating the concept of semantic relatedness as a new fact-table which, essentially, stores instances of the predicate *Relatedness*(A, B, α), where A and B denote two terms, and α is the numerical value of their relatedness.

Note: The ETL (Extract, Transform, Load) process is an important component of a DW – and, in particular the SR-TDW [29]. However, that issue is beyond the scope of this paper. Thus, in the sequel we assume that the values of the parameters in each episode and relatedness triplet, as well as the dimension tables are populated correctly into the SR-TDW, from the respective ST database along with the other GIS-based facts pertaining to a given geographical region.

We assume that motion is represented as a finite set of points which are semantically annotated [3, 6, 18] (cf. Section 2.1). Each trajectory consists of sequential episodes defined with actions that are:

– taking place at a given geo-location with a timestamp related to a POI; or

– have a duration and are taking place in-between two geo-locations;

Our SR-TDW model is illustrated in Figure 3. We can see that it is based on a constellation schema, where there are two fact tables – one pertaining to semantic episodes and one to relatedness – sharing dimension tables, which we explain next. Trajectory episodes (stored in the fact table factEpisode) are defined by dimensions dimPOIs, dimActivity, dimDateTime, and dimTrajectory. Thus, a tuple in factEpisode corresponds to a certain trajectory episode occurring in a time interval, between two (possibly coinciding) POIs, and with a certain activity occurring throughout that interval. Measures in factEpisode (not shown in the figure) may, for instance, quantify some activity within each episode, or be precomputed from the associated trajectory (e.g., the length and/or velocity within the episode). A more detailed discussion on this issue can be found in [25]. Note that, in addition to being linked with each of its episodes in the fact table, dimTrajectory has attributes recording its start and end times.

Dimension dimActivity is connected to factEpisode, although it is also a part of the hierarchy extending the dimension dimTerm. Similarly for dimension dimPOIs. The rationale is that in a semantic trajectory one needs a coupling between the *da* (defining annotation) specifying the main activity and *sp* (semantic position) – which can range between nouns and verbs – essentially being respective specializations of dimTerm consisting of *Part of Speech* (POS). This provides a two-fold genericity in the design of the SR-TDW in the following sense: (1) For different couplings between nouns and verbs (e.g., (*noun*, *noun*), (*noun*, *verb*), (*verb*, *noun*)) one can lookup the value of their relatedness from the factRelatedness fact table; and (2) Such lookup is enabled among broader POS types,

e.g., adverbs, adjectives, etc. which, in turn, enables one to also incorporate the various additional descriptors of a given ST – namely, the ones available in the *tagList* (cf. Section 2.1). We note that the "ISA" kind of relationship is not introduced from the perspective of the (values in the) respective entries from factRelatedness, but from a standpoint of the warehouse design. The factRelatedness fact table contains triplets of the form (*Term1ID*, *Term2ID*, *Relatedness*) which, as mentioned, list the values of the coefficients of relatedness for POS' couplings. This enables comparisons of similarities between items such as "*restaurant*" and "*eat*", as well as specific instances – e.g., "*Magnum*" and "*eat*". It also enables retrieving the relatedness between terms such as *"move"* and *"bicycle"*, or a pairwise relatedness between *"stop"*, *"eat"* and *"salad"*. We assume the availability of the typical aggregate operators (`COUNT, MAX`, etc.) for relatedness.

POIs are also organized into a geographic hierarchy, and are described by two level attributes indicating the POI's name and type (types follow the ones in Figure 2) – proceeding further with dimGeoLocation and dimZone. Dimensions dimGeoLocation and dimZone are assumed to have the corresponding geometric attributes (i.e., `L_Geometry` and `Z_Geometry`) capturing the respective geometric features such as coordinates, polygonal boundary of a zone, etc., along with the traditional operators for evaluating spatial predicates (e.g., `INTERSECT, UNION`, etc) [1,25]. `LocationID` is a unique key of a given geo-location such as an address within a city. Note that dimension dimZone is not further normalized towards the city and state hierarchy, although in certain practical scenario that may be the case. Lastly, as shown in Figure 3, the temporal and time period dimensions allow supporting timestamps and temporal intervals.

4 Querying Trajectory Warehouses with Semantic Relatedness

We now illustrate the novel categories of queries enabled by the SR-TDW model, with extensions pertaining to *Relatedness*(A, B, α) predicate, the values of which are readily available from the corresponding fact table. We begin with the variant of **Q_1** from Section 1, incorporating the notion of semantic relatedness:
Q'_1: *Daily number of trajectories throughout the first week of June, that started in the Loop, first stopped in a location having a semantic relatedness value* ≥ 0.75 *with a restaurant, and then stopped in a location having a semantic relatedness value* $\geq 75\%$ *with a coffee house, both within 2 miles from West Loop.*
Given the schema in Section 3.2, the corresponding SQL query statement is:

```
WITH AtORNearWestLoop(POI_ID, TermID) AS
(SELECT POI_ID, TermID
FROM dimPOIs, dimZone,
     dimGeoLocation L1
WHERE L1.LocationID = dimPOIs.LocationID
AND dimZone.ZoneName = 'West Loop'
AND (L1.ZoneID = dimZone.ZoneID OR
DISTANCE(L1.L_Geometry, dimZone.Z_Geometry) < 2))
```

```
WITH StartAtLoopWJ(TrajID) AS
(SELECT TrajID
FROM dimTrajectory Tr, dimGeoLocation L1,
     dimDateTime DatT, dimZone Z1
WHERE Tr.StartDateTimeID = DatT.DateTimeID
Tr.TrajStartLocationID = L1.Location_ID AND
L1.Zone_ID = Z1.Zone_ID AND Z1.Name = 'Loop' AND
DatT.Date BETWEEN '2014/06/01' AND '2014/06/07'
```

```
WITH AtWestLoopRest(TrajID, TimeID) AS
(SELECT TrajID, EpisodeID, POI_ID
```

```
WITH AtWestLoopCoffee(TrajID, TimeID) AS
(SELECT TrajID, EpisodeID, POI_ID
```

```
FROM StartATLoopWJ SLTr, AtORNearWestLoop WL1
factEpisode FE1, dimPOIs, dimTerm DT1, dimTerm DT1
dimDateTime DatT, factRelatedness FR1
WHERE SLTr.TrajID = FE1.TrajID
      AND FE1.StartPOI_ID = FE1.EndPOI_ID
          AND FE1.StartPOI_ID = WL1.POI_ID
          AND FE1.TimeID = DatT.Time_ID
          AND WL1.POI_ID = dimPOIs.POI_ID
          AND ((dimPOIs.TermID = DT1.TermID
               AND DT1.name = 'restaurant')
               OR
               (dimPOIs.TermID = DT2.TermID
                AND DT1.TermID = FR1.Term1ID
                AND DT2.TermID = FR1.Term2ID
                AND FR1.Relatedness > 0.75)))
```

```
FROM StartATLoopWJ SLTr, AtORNearWestLoop WL2
factEpisode FE2, dimPOIs, dimTerm DT1, dimTermDT2
dimDateTime DatT, factRelatedness FR2
WHERE SLTr.TrajID = FE2.TrajID
AND FE2.StartPOI_ID = FE2.EndPOI_ID
AND FE2.StartPOI_ID = WL2.POI_ID
AND FE2.TimeID = DatT.Time_ID
AND WL2.POI_ID = dimPOIs.POI_ID
AND ((dimPOIs.TermID = DT1.TermID
     AND DT2.name = 'coffee house')
     OR
     dimPOIs.TermID = DT2.TermID
      AND DT1.TermID = FR2.Term1ID
      AND DT2.TermID = FR2.Term2ID
      AND FR2.Relatedness > 0.75)))
```

```
SELECT DatT1.Date, COUNT(*)
FROM StartAtLoopWJ SLTr, AtWestLoopRest WLR, AtWestLoopCoffee WLC,
   dimDateTime DatT1, dimDateTime DatT2
WHERE SLTr.TrajID = WLR.TrajID AND SLTr.TrajID = WLC.TrajID
    AND WLR.TimeID = DatT1.TimeID
    AND WLC.TimeID = DatT2.TimeID
    AND DatT1.Date = DatT2.Date
    AND DatT1.Time < DatT2.Time
GROUP BY DatT1.Date
```

The first pair of WITH clauses select the POIs inside or within 2 miles from West Loop and the trajectories which started in the Loop during the first week of June in 2014, respectively. The crux of processing **Q'_1** is in the next pair of WITH clauses, which retrieve all the places at or near West Loop, having semantic relatedness $> 75\%$ with the term *"restaurant"* as well as the term *"coffee house"*. Clearly, this is an overhead which involves accessing extra tables to generate the respective POIs. However, this provides an enrichment to the answer-set, as opposed to having only *"restaurant"* and *"coffee house"*. The main SQL query references the previous two tables and ensures the sequence of the visit by the candidate-trajectories.

Examples of other categories of queries enabled by the semantic relatedness embedded in SR-TDW follow. Due to a lack of space, instead of presenting their full SQL-based syntax, we describe their main features and discuss approaches for processing them in the context of SR-TDW.

Q_2: *Weekly average semantic relatedness of any two downtown locations visited by the same trajectory within 1 hour from each other, throughout the month of January 2015.*

This query exemplifies an analytics-motivated scenario where one may be interested in quantifying the relatedness among the places that a particular individual would visit sequentially within 1 hour (e.g., from *theater* to a *restaurant*; from *ATM* to a *bar*; etc.). In some sense, queries like **Q_2** may be used as another kind of context for exploring a strength of semantic proximity between terms – e.g., the "semantic strength" of the relationship between *ATM* and *bar* may be detected to be greater than the average, in the sense of sequentiality of visits within temporal bounds. In addition, one may reason about the variations in the relatedness values based on the temporal hierarchy.

To process **Q_2**, we first need to identify the pairs (fE1, fE2) of factEpisode's, such that: (1) they belong to a same trajectory (`fE1.TrajID = fE2.TrajID`);

(2) the two instances of the factEpisode's are of a type *"stop"* at POIs; the location of the POIs are within the *"downtown"* zone; and the value of the *Time* attribute of the respective `EndDateTimeID` of the first stop-episode is no earlier than 1 hour from the `Time` of the respective `StartDateTimeID` of the second stop-episode. Note that, depending on the dataset (i.e., if there are many "historic trajectories"), one would probably first eliminate all the episodes that are not from the month of January. Subsequently, this temporary result can be projected upon the respective `StartPOI_ID` attributes[4] for each of the fE1 and fE2, join the result of this projection with the corresponding pairs of values in the factRelatedness table (via respective matching values `POI_ID` in dimPOIs and `TermID` in dimTerm). The value of the `AVG(...)` aggregate is then applied to the `Relatedness` column of this temporary table, grouped by the `Week`.

Q_3: *Average duration of the trajectories who have visited sequentially at least two POIs within the same geographic zone, and with semantic relatedness greater than the maximum relatedness between a restaurant and any other POI in that zone.*

Q_3 aims at detecting an average trip of the trajectories for the individuals who tend to visit semantically "close" `POIs` which are also located within same spatial boundaries (at the level of zone in this case). As an additional condition – e.g., for the purpose of targeted online advertising, the semantic proximity of the POIs is required to be greater than the highest one between a restaurant in that zone and any other POI.

To calculate the answer-set for **Q_3**, the main observation is that we first need to obtain the average of all the tuples from the factRelatedness table, for which one of the TermID1 or TermID2 is bound to *"restaurant"*, denote it `MAX-RestSR`. In addition, we select the TrajID, duration, and the semantic episodes having a *"stop"* at some POIs, filtering out the ones with ≤ 1 such episodes. We can execute a *Θ-join* over the last temporary table, retaining only those pairs of tuples for a given TrajID for which the stops at POIs are consecutive (i.e., there does *not* exist any other factEpisode with a stop-kind of POI at a time that is in-between the ones for the pair with itself) *and* their locations are in the same zone. Finally, we filter out all the tuples for which the pair of POIs has semantic relatedness $<$ `MAX-RestSR`, and report the average duration of the rest of them.

Q_4: *Number of triplets of locations, each being visited by more than 1000 trajectories throughout the month of March 2015, and having at least one pairwise-relatedness value smaller than the average relatedness involving any coffee house.*

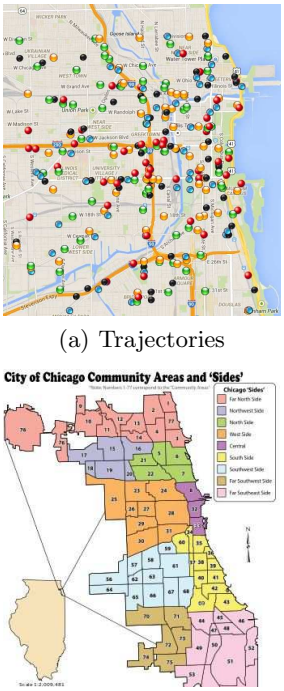
The peculiarity of **Q_4** stems from the fact that it retrieves triplets of locations, whereas table factRelatedness has only pairs of terms, along with the corresponding semantic relatedness value. To find the POIs visited by $>$ 1000 trajectories throughout March 2015, we firstly select the trajectories for which the respective `Date` of `TrajBeginDateTimeID` key is no later than '2014/03/31', *or* the respective `Date` of `TrajEndDateTimeID` key is not earlier than '2014/03/01'. Using the remaining `TrajID` values we can join factEpisode and dimPOIs, and then group them by `POI_ID, Traj_ID`, with a subsequent con-

[4] Since each episode is of a *"stop"* type, the *StartPOI_ID* and *EndPOI_ID* coincide.

dition of `HAVING COUNT(*) >= 1000`. Retaining the time values in this temporary result, we can construct the triplets of such POIs enforcing that each triplet is sorted by the `StartDateTimeID` of the corresponding semantic episode. This will alleviate the problem of a permutation of the same triplet occurring multiple times – which can not be eliminated via simple `SELECT DISTINCT...`. Then, for each pair of a given triplet, we need to check whether it satisfies the condition of having the relatedness value smaller than the average relatedness for all the pairs from factRelatedness having one of the terms being `'coffee house'`.

The issue of efficiency of processing is beyond the scope of this work. We note however that, depending on the actual instance of the SR-TDW, an alternative plausible strategy would be to generate the triplets of `TermIDs` from factRelatedness for which at least one pair has relatedness smaller than the average relatedness for all the pairs having one of the terms being `'coffee house'`. This can be joined with the triplets of `POIs` visited by > 100 trajectories – however, the join condition should account for the possibility of a permutation in the representations.

We close this section with a reminder that, while the features of the SR-TDW model were illustrated using scenarios involving eateries and coffee places from Chicagoland, the applicability is more general (cf. [29]).



(a) Trajectories

(b) Zones

Fig. 4. Data generation

5 Experimental Evaluation

We now present the details of our experimental evaluation, firstly discussing the dataset and queries, followed by the quantitative observations.

We generated collections having 500, 1000, 2000, 3000 and 4000 trajectories using Chicago roadnetwork, and for each cardinality of trajectories, we further generated sets with drive times of 400, 1000, 1000, 4000, 8000 and 10000 seconds. The routes of the trajectories are within a rectangular boundary $5 \cdot 10$ miles^2 around the downtown area, using the MNTG (Minnesota Traffic Network Generator) tool, publicly available at *http://mntg.cs.umn.edu/tg/index.php* [15]. As mentioned, the ETL phase is beyond the scope of this paper, however, for the purpose of conducting the experiments – given that the maps used in MNTG are based on the Open Street Map (OMS – http://www.openstreetmap.org), we used sources based on OMS (*http://poidirectory.com/poifiles/united_states/*) to introduce actual POIs from the underlying map – including restaurants, coffee houses, fast food places, bars and theaters.

Measures:	Leacock and Chodorow	Resnik	Wu and Palmer
Intervals of Values	0—3.6889	0—12	0—1
(*The Gage, Cadillac Palace*)	2.0794	3.9425	0.7778
(*Starbucks, Bank of America Theatre*)	2.0926	5.3823	0.8421
(*Quartino, Urban Counter*)	1.204	0.6144	0.3529
(*Urban Counter, Starbucks*)	1.1239	0.6444	0.3529
(*coffehouse, restaurant*)	2.9957	8.3	0.9474
(*Starbucks, The Purple Pig*)	2.9957	8.3	0.9474

Table 1. Semantic Measures

Since MNTG allows a generation of trajectories for at most 1000 time-units (time-unit = 2 sec.), we repeated the process and appended the outcomes, in order to have the datasets of duration described above. Also, the trajectories generated via MNTG do not have stop-points, therefore, we randomly picked trajectories passing on a road-segment along a given POI and "induced" a stay between 5–180 minutes, respectively shifting the time-stamps in the subsequent points, yielding up to 3000 trajectories. We repeated the above procedure in order to generate a week-worth of trajectories data, varying the timings and the POIs. Lastly, we relied on the map of Chicagoland neighborhoods (http://en.wikipedia.org/wiki/Community_areas_in_Chicago) to generate the boundaries of the respective zones. Figure 4 illustrates the data sources' settings used in our experiments. The corresponding semantic trajectories were inserted as UDTs in Microsoft SQL Server 2012, which enables direct manipulation of *(latitude, longitude)* values in the `ST_Geography` – an added convenience when translating the trajectories and POIs data.

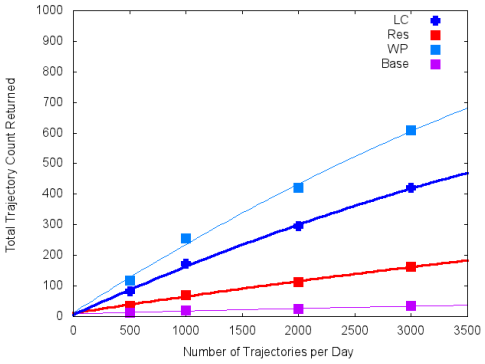


Fig. 5. SR and Answer-sets

In total, there are 10,000 pairs of terms in the factRelatedness table and – to provide an extra degree of context – we used three different sources for the values stored in the "Relatedness" attribute of the factRelatedness table (cf. Figure 3), based on three different measures: Leacock & Chodorow (LC); Resnik (Res); and Wu & Palmer (WP) [4,19]. As recognized in the literature, different measures have different numeric values and distributions, and we illustrate these effects with sample-values shown in Table 1. As can be seen, the largest range of values is associated with the `Resnik` measure, whereas the smallest range is associated with `Wu` & `Palmer`. Looking at the last two rows, we see that in all the measures, the values for the pair (*coffehouse,*

restaurant) coincide with the ones for (*Starbucks, The Purple Pig*) – which illustrates how we added actual POIs to the concepts available at WordNet: namely, for each POI from Chicagoland, we obtained its type and then added it as a new "link" to the term matching its type, and with a weight of 1.[5] However, there is another interesting observation – namely, the distribution of similarity values among pairs of terms exhibits variations among measures.

Our first set of experimental observations illustrates the dependency of the size of the answer-set on the size of the trajectories data, averaged over 3 different values of the semantic relatedness Θ for each of the three measures. Specifically, we used $\Theta \in \{50\%, 75\%, 90\%\}$ of the interval of values in each of the three measures from 1 in **Q'_1** from Section 4 and averaged the size of the output. What is apparent from Figure 5 is that, as expected, regardless of the measure, the difference between the size of the answer-sets with relatedness and without one, is increasing proportionally with the number of trajectories. Table 2 shows actual samples of values of the `COUNT` aggregate distributed per day of week for two values of Θ (50% and 75%) obtained as part of our experiments. The quadruples in each cell show the values when `LC`, `Res`, `WP` and `Base` (No Relatedness) values are the ones for 1000 trajectories.

Day:	Mon.	Tue.	Wed.	Thu.	Fri.
$\Theta = 50\%$	[49,20,49,5]	[83,69,83,5]	[42,17,43,1,]	[54,21,52,5]	[23,10,23,2]
$\Theta = 75\%$	[20,5,37,5]	[69,5,81,5]	[17,1,35,1,]	[25,5,51,5]	[10,2,15,2]

Table 2. Examples of COUNT values

Two observations from Table 2 reveal the impact of the relatedness: (1) As expected, the smaller the threshold value, the larger the increase of the size of the answer-sets; (2) Unlike `LC` and `WP`, the `Res` measure has a sharp decline in the increase of the dataset with the increase of Θ. The reason for it is that the most of the values in `Res` are distributed close to the middle of the range, in a much denser manner than the ones in `LC` and `WP`. This, in turn, has a practical consequence that one needs to be cautious when selecting a particular measure – a context-based topic which we plan to investigate in the future.

The second part of our experimental observations is aimed at illustrating another perspective of the impact of Θ values for different measures. Figure 6 shows the extreme discrepancies in the sizes of the answer-sets for each measure. Thus, for instance, at $\Theta = 50\%$, the largest answer-set for the `LC` measure was 109 on the 2nd day of the week, at which day the `Base` variant of the query had a value of 5 for the count of the trajectories – hence, the discrepancy of 104. Again we observe that `Res` has a sharp decline with the increase of Θ, followed by `LC`, whereas `WP` retains the capability of generating substantially larger answer-sets even with $\Theta = 90\%$.

[5] We note that all of our: datasets and scripts used for conversion; scripts for uploading the tables; the database instance(s), queries and the scripts for executing them – are publicly available (http://www.eecs.northwestern.edu/~goce/research)

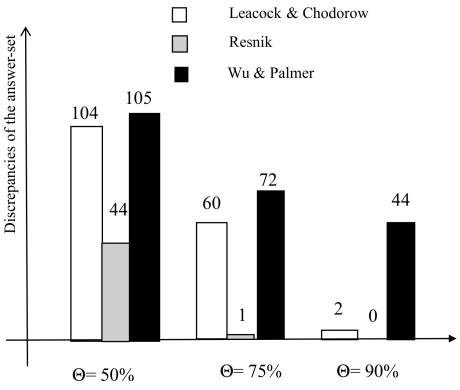


Fig. 6. Θ and Extreme Discrepancies

Our last set of experiments measured the computational overheads induced by allowing semantic relatedness as part of the queries processing. As expected – and illustrated in Table 3 – incorporating the relatedness does affect the overall time to complete processing a particular query processing. However, it is a trade-off that one has to consider as part of the business policies related to a particular query. Once again we show the averaged values of the execution times for the different ranges of the parameter Θ ($\in \{50\%, 75\%, 90\%\}$) and we observe that the execution overheads increase with the size of the input trajectories data. Given the intended use of the analytics enabled with the SR-TDW, a careful selection of Θ might strike a balance between the richness of the answer-set and the time-efficiency of the execution.

Dataset Size:	500	1000	2000	3000
With Semantic Similarity	108	204	390	820
Without Semantic Similarity	49	99	182	296

Table 3. Execution Times (seconds)

6 Related Work

A cohesive collection of works tackling various aspects of mobility data was presented in [21]. The paradigm of semantic trajectories [6, 18] generated novel challenges addressed by the database community. In [23], the traditional settings of Nearest-Neighbor query for spatial data were augmented by allowing keywords associated with the locations. A modified distance function – extending the Euclidian spatial one – was introduced in order to incorporate the matching between the list of keywords associated with location data, along with a novel indexing structure (IR^2 tree) to speed up query processing. Further, Chen et al. [5] proposed to evaluate the distance (respectively, similarity) between two sequences of visited locations, not only based on geographic distances but also in terms of (minimum) matching of the keywords associated with such locations. The process of combining the raw (*location*, *time*) data with segmentation and annotation was addressed in [29], where a platform for semantic enrichment of trajectories was presented. While capitalizing on the definitions of semantic trajectories, our work differs in two main aspects: (1) instead of a set-based matching and/or containment between collections of terms, we consider the semantic relatedness

among the annotations/descriptors; (2) we focus on the role and impact of the semantic relatedness in the context of aggregation in SR-TDWs.

Traditional Data Warehouses [25] have demonstrated their applicability with transaction-level data and computing its various aggregates. However, recent expansion of user-needs for data with contexts beyond the standard dimensions – specifically: location/geography, time and semantic description of the activities – have brought various novelties to the DW models. A taxonomy of different spatial, temporal and spatio-temporal DWs is presented in [24] and, building upon those formalisms, several works have addressed problems related to our proposal. A framework for modeling Trajectory Data Warehouse (TDW) was presented in [13] providing key insight about OLAP operations for moving objects. Related problems were investigated in [9] from the perspective of formalizing the process of the design and querying a TDW, and [17] addressing the computation of aggregate functions in TDW. We leveraged upon the TDW model and OLAP operations tackled in these works, augmenting the scope of applicability of these approaches by seamlessly incorporating the notion of semantic relatedness both in the modelling and the querying aspects of TDWs. The work by Parent et al. [18], which incorporates fundamental definitions for the notion of semantic trajectories, was enriched by Wagner et al. [26] via a data model capturing the Why, Who, When, Where, What and How (5W1H) aspects, focusing around a central fact connected to dimensions that source the semantic information on the transaction level. The addition of ontologies to the data models [16] enabled semantically meaningful hierarchies. As a next step in the evolution of the semantic/symbolic data representation, [6] pays particular attention to adding semantic tags, annotations and definitions in the representation for trajectories/moving objects. With a great level of detail [3] represented a geospatial semantic data model which encapsulates most of the semantic annotation, tags, actions and definitions previously mentioned. The work enabled answering questions related to the trajectory behaviour, goal and transportation means. Extending the semantics behind the trajectories, [8] implemented movement segment hierarchies, distinguishing concepts from instances or objects. While introducing ontologies to represent the semantics of the movement segments and their categories, the work does not go beyond these concepts to represent the semantics of the trajectories and their activities. Additional works stemming from the semantic representation of trajectories [22, 27, 29] advanced the semantic trajectories approach with ontologies, cross-scale analysis and a semantic computing platform, respectively. All these approaches introduce a certain level of semantics-based description to augment the raw spatio-temporal data – however, none of them addresses the inferences of semantic meaning enabled by the approaches and measures that we used in this work [2, 4].

7 Concluding Remarks and Future Work

We addressed the problem of augmenting trajectories data warehouses with the concept of semantic relatedness and increasing their similarity-awareness when

answering users' queries. We presented the corresponding constellation schema and described novel queries enabled by the SR-TDW model. Our experiments demonstrated that the proposed methodologies yield richer answer-sets, which vary based on the measure used. As part of our future work, we are planning to devise efficient approaches for similarity among semantic trajectories combining both semantic relatedness and dynamics of motion in the distance functions (cf. [5, 23]). We will also tackle problems related to incorporating moving shapes (e.g., landslide, hurricanes, oil-spills) and different spatio-temporal patterns (e.g., flocks, clusters) [8] in the framework of SR-TDW, along with extending the formalisms and platform in [28, 29] with semantic relatedness. Among our primary objectives is to address efficiency-related tasks, both from the perspective of the design of warehouse schemata (e.g., different constellation-models [20]) and queries optimization and, as part of that process, to deeper investigate the impact of the relatedness measures as well as augmenting the types of measures used [7]. Lastly, we will try to increase the impact of the relatedness by both broadening the terms sources [2] as well as increasing the efficiency by application-based narrowing of the context [3].

References

1. Y. Bédard, S. Rivest, and M. Proulx. Spatial online analytical processing (SOLAP): Concepts, architectures, and solutions from a geomatics engineering perspective. In R. Wrembel and C. Koncilia, editors, *Data Warehouses and OLAP: Concepts, Architectures and Solutions*, chapter 13, pages 298–319. IRM Press, 2007.
2. R. Bill, Y. Liu, B. T. McInnes, G. B. Melton, T. Pedersen, and S. V. S. Pakhomov. Evaluating semantic relatedness and similarity measures with standardized meddra queries. In *AMIA, American Medical Informatics Association Annual Symposium*, 2012.
3. V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares. Constant - A conceptual data model for semantic trajectories of moving objects. *T. GIS*, 18(1):66–88, 2014.
4. A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
5. W. Chen, L. Zhao, J. Xu, K. Zheng, and X. Zhou. Ranking based activity trajectory search. In *Proceedings of WISE*, pages 170–185, 2014.
6. M. L. Damiani and R. H. Güting. Semantic trajectories and beyond. In *Proceedings of IEEE - MDM*, pages 1–3, 2014.
7. I. Donevska. Advancing the semantic relatedness approach by using sense popularity. In *Proceedings of IEEE - ICSC*, pages 246–247, 2014.
8. R. Fileto, A. Raffaetà, A. Roncato, J. A. P. Sacenti, C. May, and D. Klein. A semantic model for movement data warehouses. In *Proceedings of DOLAP*, pages 47–56, 2014.
9. L. I. Gómez, B. Kuijpers, and A. A. Vaisman. A data model and query language for spatio-temporal decision support. *GeoInformatica*, 15(3):455–496, 2011.
10. J. Gracia and E. Mena. Web-based measure of semantic relatedness. In *Proceedings of WISE*, pages 136–150, 2008.
11. C. Guo, M. Ma, B. Yang, C. S. Jensen, and M. Kaul. Ecomark: Evaluating models of vehicular environmental impact. In *Proceedings of GIS*, 2012.

12. R. H. Güting and M. Schneider. *Moving Objects Databases*. Morgan Kaufmann, 2005.
13. L. Leonardi, S. Orlando, A. Raffaetà, A. Roncato, C. Silvestri, G. L. Andrienko, and N. V. Andrienko. A general framework for trajectory data warehousing and visual OLAP. *GeoInformatica*, 18(2):273–312, 2014.
14. Mckinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity, 2011.
15. M. F. Mokbel, L. Alarabi, J. Bao, A. Eldawy, A. Magdy, M. Sarwat, E. Waytas, and S. Yackel. A demonstration of MNTG - A web-based road network traffic generator. In *Proceedings of IEEE - ICDE*, pages 1246–1249, 2014.
16. V. Nebot, R. B. Llavori, J. M. Pérez-Martínez, M. J. Aramburu, and T. B. Pedersen. Multidimensional integrated ontologies: A framework for designing semantic data warehouses. *J. Data Semantics*, 13:1–36, 2009.
17. S. Orlando, R. Orsini, A. Raffaetà, A. Roncato, and C. Silvestri. Spatio-temporal aggregations in trajectory data warehouses. In *Proceedings of DaWaK*, pages 66–77, 2007.
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, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42, 2013.
19. S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing*, pages 241–257, 2003.
20. N. Pelekis and Y. Theodoridis. *Mobility Data Management and Exploration*. Springer, 2014.
21. C. Renso, S. Spaccapietra, and E. Z. (editors). *Mobility Data: Modeling, Management and Understanding*. Cambridge University Press, 2013.
22. A. Soleymani, J. Cachat, K. Robinson, S. Dodge, A. Kalueff, and R. Weibel. Integrating cross-scale analysis in the spatial and temporal domains for classification of behavioral movement. *J. Spatial Information Science*, 8(1):1–25, 2014.
23. Y. Tao and C. Sheng. Fast nearest neighbor search with keywords. *IEEE Trans. Knowl. Data Eng.*, 26(4):878–888, 2014.
24. A. A. Vaisman and E. Zimányi. What is spatio-temporal data warehousing? In *Proceedings of DaWaK*, pages 9–23, 2009.
25. A. A. Vaisman and E. Zimányi. *Data Warehouse Systems - Design and Implementation*. Data-Centric Systems and Applications. Springer, 2014.
26. R. Wagner, J. A. F. de Macêdo, A. Raffaetà, C. Renso, A. Roncato, and R. Trasarti. Mob-warehouse: A semantic approach for mobility analysis with a trajectory data warehouse. In *Advances in Conceptual Modeling - ER Workshops*, pages 127–136, 2013.
27. R. Wannous, A. Bouju, J. Malki, and C. Vincent. Ontology inference using spatial and trajectory domain rules. In *Proceedings of WorldComp*, 2014.
28. F. Wenzel and W. Kießling. Aggregation and analysis of enriched spatial user models from location-based social networks. In *Proceedings of GeoRich@SIGMOD*, page 8, 2014.
29. Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM TIST*, 4(3):49, 2013.