Semantics-Aware Warehousing of Symbolic Trajectories

Goce Trajcevski*
Dept. of EECS
Northwestern University
Evanston, IL, USA
goce@eecs.northwestern.edu

Ivana Donevska
Dept. of CS
Indiana University Purdue University
Fort Wayne, IN, USA
ivanadonevska@yahoo.com

Alejandro Vaisman
Instituto Tecnológico de Buenos Aires
Buenos Aires, Argentina
avaisman@itba.edu.ar

Besim Avci
Dept. of EECS
Northwestern University
Evanston, IL, USA

Tian Zhang
Dept. of EECS
Northwestern University
Evanston, IL, USA

Di Tian
Dept. of EECS
Northwestern University
Evanston, IL, USA

besim,t-zhang,d-tian@eecs.northwestern.edu

ABSTRACT

We address the problem of extending the querying capabilities of Trajectories Data Warehouses (TDW) for symbolic trajectories, by introducing *Semantic Relatedness* (SR) as part of the formal model. This enables capturing the similarity between different annotations describing Points of Interest (POI), locations and activities. We formally define the inclusion of the relationship between different terms used as descriptors in symbolic trajectories and present the *Semantic Relatedness in Trajectories Data Warehouse* (SR-TDW) model. We introduce newly enabled queries in the SR-TDW model and illustrate the impacts of the added functionality. Our experiments demonstrate the benefits of the proposed approaches in terms of enriching the answer-sets for the common OLAP-based queries, and the sensitivity in terms of the various measures of semantic similarity.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Warehousing, Semantics, Queries

Keywords

Trajectory Warehouses, Semantic Relatedness

*Research supported by the NSF-CNS 0910952, NSF-III 1213038, and ONR N00014-14-10215.

IWGS'15, November 03-06, 2015 Bellevue, WA, USA

DOI: http://dx.doi.org/10.1145/2833165.2833174. .

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 in plethora of GIS applications. An O(Peta-Bytes) per year are generated from the GPS of smart phone users, with up to 400-fold increase if cell-tower locations are included [14]. It is estimated that by 2020, more than 70% of mobile phones will have GPS capability – compared to 20% in 2010 – with similar trends in cars equipped with dashboard GPS devices.

Efficient storage and retrieval of the (*location, time*) information is essential for various applications – e.g., navigation, traffic management, recommendation systems, disaster mitigation, etc. traditionally, managed by Moving Objects Databases (MOD) [11]. 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 [12, 17, 24]. The sequence of spatiotemporal positions of a moving object, having a certain start and end, is called the object's *raw trajectory* – 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 (POIs) may suffice. For this, we need to identify 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* in 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* [19]. 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.,

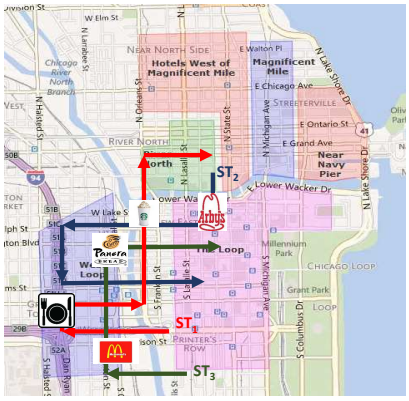


Figure 1: Chicagoland trajectories

street corners, restaurants, etc.).

Trajectory Data Warehouses (TDWs) [8] and Semantic Trajectory Data Warehouse (STDW) [3,21,26] are aimed at aggregating and analyzing trajectory data, e.g., using OLAP and data mining techniques – as exemplified with the query **Q_1** below, over 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 detect 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 Arby's (a fast-food place), followed by a stop at Starbucks; (2) Similarly ST_3 stopped first at McDonald's (a fast-food place), and then at Panera (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. As another 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 take a first step towards extending TDWs and STDWs with the notion of *semantic relatedness* [2, 7, 18], 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.

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 similarities – 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 between different (classes of) terms.
- We introduce novel queries which incorporate the value of the semantic relatedness when determining the answer-set and may enrich (augment) the answer set.
- We present experimental observations evaluating the effectiveness of the novel SR-TDW model when applied on a dataset of semantic trajectories from Chicago, illustrating the impact of 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 Symbolic Trajectories (ST) and TDWs.

2.1 Semantically Enriched Trajectories

Symbolic(synonymously, Semantic or Enriched) *Trajectories* [3, 6, 19] embed contextual and/or situational knowledge into location-in-time data. In a MOD [11] 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 o_{ID}, 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. [6,19]), 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)
(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),
(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, coffee, eat, dessert)
]
ST_2 = [(move, Dearborn St, 60, 60, 11:30, 60, 40, 11:45, walk),
(stop, "Arby's" , 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)]

Note that there is a match between the third and the second *stop* activities in ST_1 and ST_2, respectively (i.e., both involve "Starbucks"). However, it is also worth noticing that the first *stop* activity of ST_2, that is, stopping at "Arby's", involves "fast-food". This activity can be considered, in some sense, semantically similar to stopping by at the "Roditis" restaurant, the first stop in ST_1, since the latter is labeled "restaurant".

2.2 Warehousing Trajectory Data

Due to space limitations, we assume the reader is familiar with the basic notions of traditional OLAP and DWs (cf. [25]), so we omit details in this sense.

Several works have used OLAP techniques for exploration of spatial data – named 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. The next extension was in the context of spatial objects that may vary across time, which spurred 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 [21, 22, 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 [22, 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.

3. SEMANTIC RELATEDNESS AND TRAJECTORY DATA WAREHOUSE

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

3.1 Semantic Relatedness

Intuitively, the notion of *semantic relatedness* quantifies the "semantic proximity" of two concepts or entities not only by the similarity between objects, but also via other features, like their "popularity" or how often those two entities appear together in text-corpora or are referenced by users [7, 20]. Much work has been done in the field of semantic similarity/relatedness and there are various measures and evaluation techniques [2, 20] based on multiple connections (even multiple hierarchies) that 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].

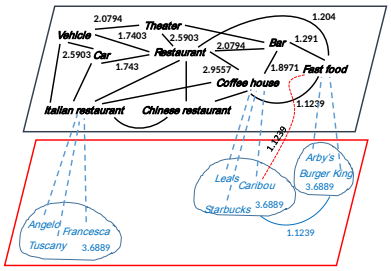


Figure 2: Semantic relatedness

One way to model/represent semantic similarity between terms in a given collection, in addition to a simple matrix of weights, is via graph in which nodes correspond to terms, and edges represent (the strengths of) the respective semantic connections. There are different approaches for assigning weights and targeting a different group of semantic connections [2, 4, 20] – e.g., in [7] a graph is constructed by making five passes over the existing connections, where the first pass inserts the core nodes (nouns extracted from the WordNet repository [23]) which are then connected to their sense, and the weight assigned to the likelihood of transition from one node to the other captures popularity of that sense. Subsequently, weights are 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). There are various measures for capturing the semantic similarity between terms [4,20], and a recent comprehensive comparative survey is available in [18].

While the distinction between the notions of *similarity* and *relatedness* has often been blurred, more recent works

consider them as separate descriptors of relationships between terms. For instance, the measures used in our experiments are referred to as "relatedness" measures in [23], whereas [18] classifies them as "similarity" measures. Existing works [2,18] have also pointed out the need for augmenting the popular WordNet repository with other Internet-based sources/dictionaries [13] such as Wiktionary, along with hybrid-measures. The main reason is that different measures have been shown to exhibit variable quality (in terms of certain criteria) in different dictionaries and corpora. These issues, while relevant to our work, are beyond the scope of this paper.

The concepts that we discussed are illustrated in Figure 2 which shows a portion of a relatedness graph along with the weights between edges, with the values corresponding to the Leacock and Chodorow (LC) measure from WordNet [23]. For example, the weight of the edge between Restaurant and Theatre is based on various relationships between the two terms (e.g., co-occurrences in sentences). We note that the bottom portion is separated to illustrate an ontological type of a relationship *"Is an Instance of"*, as opposed to the traditional *IS-A* (i.e., Italian restaurant *IS-A* restaurant, whereas Tuscany *Is an Instance of* Italian restaurant). The bottom part does not show all the edges between the instances – instead, we put 3.6889 which is the maximum value of relatedness in the `LC` measure. Thus, all of *Angelo, Francesca* and *Tuscany* can be thought of as having pairwise edges with weight 3.6889, and each of them has a relatedness of 3.6889 with Italian Restaurant. Similarly for the respective instances of Coffee House and Fast Food. However, the relatedness value between instances of different classes is assumed to be equal to the one between their classes. Thus, `Lea's` and `Arby's` are assumed to have an edge with weight 1.1239. The relatedness between an instance of a class and another class has the same value as the two classes. Thus, `Caribou` and Fast Food have relatedness of 1.1239 too.

Semantic similarity/relatedness has been extensively studied and a detailed survey is well beyond the scope of this work [2–4, 10, 13, 18, 20]. Here, we use the notion of relatedness to augment the use of 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 a two-fold impact over the query results: (1) allows to obtain answers which, otherwise, would remain hidden; (2) it eliminates certain answers which are not related-enough (modulo given measure and user's preferences). 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 the corresponding nouns and/or verbs from the *tagList* (cf. Section 2.1).

3.2 Extending TDWs

We now proceed with introducing a generic SR-TDW model which extends STDWs with the notion of semantic relatedness. 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

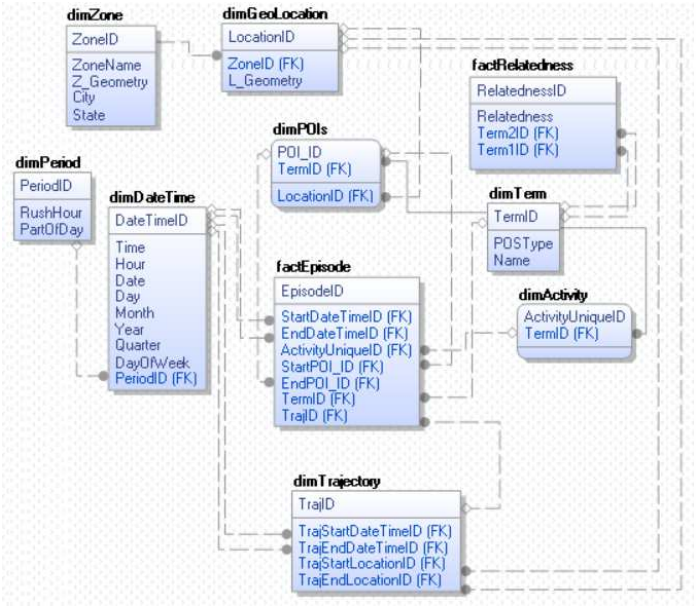


Figure 3: TDW with Semantic Relatedness

the distance traversed in the cell); and (b) the "vector-like" one [21,25], where trajectory segments are represented as geometric types. Extended models incorporate the concept of continuous fields (cf. [22]), which we do not consider in this work. 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. We note that the ETL (Extract, Transform, Load) process is an important component of a DW – and, in particular the SR-TDW [28]. However, that issue is beyond the scope of this paper, and in the sequel we assume properly populated tables. We assume that motion is represented as a finite set of points which are semantically annotated [3,6,19] (cf. Section 2.1). Each trajectory consists of sequential episodes defined with actions that:

- Take 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;

Figure 3 shows the SR-TDW model: it is based on a constellation schema, with 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.

Dimensions dimActivity and dimTerm are both connected to factEpisode, however, they are also connected with each other via the dimension dimTerm which, in a sense, is a bridge-table between them. 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. This provides a two-fold genericity in the design of the SR-TDW: (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 word-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 SR-TDW

We now illustrate the novel aspects in the categories of queries enabled by the SR-TDW model by incorporating the *Relatedness*(A, B, α) predicate, the values of which are readily available from the corresponding fact table (cf. Figure 3). We note that due to a lack of space, we cannot present a detailed analysis of conformance with the taxonomy of analytics-motivated queries pertaining to the traditional TDW settings (cf. [24]).

As discussed in Section 3, there are different measures for semantic similarity/relatedness, and each of them has different ranges of values. For uniformity, in the sequel we use "%" in the queries syntax to indicate the similarity threshold modulo a particular measure.

We start with a variant of query **Q_1** from Section 1, that takes advantage of the notion of semantic relatedness:

Q'_1: *Daily number of trajectories throughout the first week of June 2015, that started at the Loop, first stopped at a location having a semantic relatedness value* $\geq 75\%$ *with a restaurant, and then stopped at a location having a semantic relatedness value* $\geq 75\%$ *with a coffee house, both within 2 miles from West Loop.*

Q'_1 is an example of a geometrically constrained query (Loop and West Loop are names of zones in Chicago) coupled with a sequence-constraint (restaurant visited before coffee house). However, we augment the answer-set with the count of trajectories which were not bound to explicitly stop by at a restaurant and coffee house, but at places having a certain relatedness with those terms.

To process **Q'_1**, one may proceed with selecting the POIs inside or within 2 miles from West Loop, and select the trajectories which started in the Loop during the first week of June in 2014, respectively. The crux of processing **Q'_1** is in retrieving 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"*. We note that the measure, as well as both the spatial and temporal dimensions could have varied in **Q'_1** in the sense of e.g., weekly average throughout June, for the trajectories from the entire Chicagoland.

Examples of other kinds of queries enabled by the semantic relatedness embedded in SR-TDW follow.

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[1] 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).

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

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.*

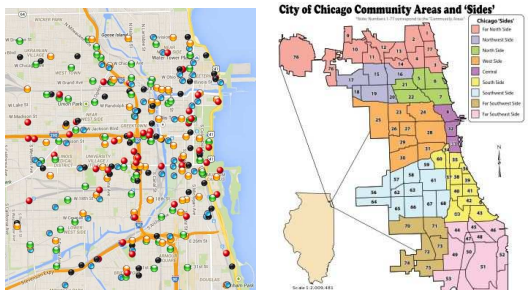
Q_3 aims at detecting an average trip of the trajectories for the individuals who tend to visit semantically "close" `POIs` which are also located within same spatial boundaries (at the level of zone in this case). As an additional condition – e.g., for the purpose of targeted online advertising, the 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.

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. [28]).

5. EXPERIMENTAL EVALUATION

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



(a) Trajectories (b) Zones

Figure 4: Data generation

We generated collections having 500, 1000, 2000 and 3000 trajectories using the Chicago road network, and with drive times of 500, 1000, 2000, 4000, 8000 and 10000 seconds, using the MNTG (Minnesota Traffic Network Generator) tool, publicly available at *http://mntg.cs.umn.edu/tg/index.php* [15]. The routes of the trajectories are within a rectangular boundary 5×10 miles^2 around the downtown area.

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:	LC	Res	WP
Intervals of Values	0-3.6889	0-12	0-1
(*The Gage, Cadillac Palace*)	2.0794	3.9425	0.7778
(*Starbucks, BoA 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 the trajectories generated via MNTG do not have stop-points, we randomly picked trajectories passing on a road-segment along a given POI and "induced" a stay between 5 and 180 minutes, respectively shifting the time-stamps in the subsequent points. 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. Figures 4(a) and 4(b) illustrate 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. In total, we had approximately 4.5GB of data. The experiments were performed on a Windows PC with Intel Quad-core i7-4790 processor (3.6GHz) with 16 GB of RAM.

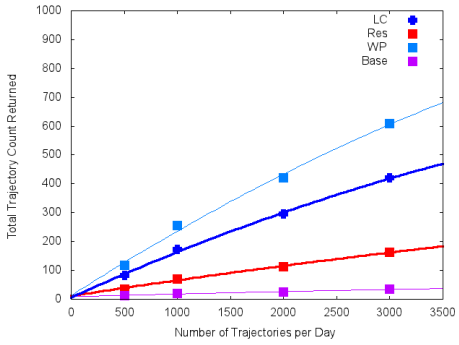


Figure 5: Relatedness and Answer-sets (Q_1')

In total, we have 10,000 pairs of terms in the factRelatedness table. 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, 20, 23]. 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 (cf. Section 3): namely, for each instance 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 equal to the maximum value for a respective measure (e.g., 1 for Wu & Palmer), re-iterating that the distribution of similarity values between pairs of terms exhibits variations among measures. We note that all of our datasets and scripts used for: conversion; uploading the data in the tables; and executing queries – are publicly available (*http://www.eecs.northwestern.edu/~goce/Similarity.html*).

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 Table 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, increases 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` (meaning, no relatedness) values are the ones for 1000 trajectories.

Day:	$\Theta = 50\%$	$\Theta = 75\%$
Monday	[49,20,49,5]	[20,5,37,5]
Tuesday:	[83,69,83,5]	[69,5,81,5]
Wednesday:	[42,17,43,1,]	[17,1,35,1]
Thursday:	[54,21,52,5]	[25,5,51,5]
Friday:	[23,10,23,2]	[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 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 about, when selecting a particular measure, a context-based topic which we plan to investigate in the future.

Although we did not explicitly address the issue of efficiency of queries processing, for an intuitive idea of the trade-offs, our last set of experiments measured the computational overhead induced by including semantic relatedness in the queries. As shown in Table 3, incorporating the relatedness does affect the overall time to process a particular query – a trade-off to be considered as part of business policies. Again we show the averaged values of the execution times for the different ranges of the parameter Θ ($\in \{50\%, 75\%, 90\%\}$) for **Q_1'** and we observe that the execution overheads increase with the size of the input trajectories data. Given the intended analytics use of the SR-TDW, coupling the values of Θ with the more traditional optimization techniques and/or indexing, might balance the richness of the answer-set and the time-efficiency.

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)

Summarizing, our experiments have demonstrated the benefits of adding the semantic awareness in TDWs in terms of enriching the answer-sets based on similarity preferences when query parameters are bound or free over a given domain. The observations were consistent across different measures although, as noted, the selection of measures may have impact on the quantitative values in the answers to particular queries.

6. RELATED WORK

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 [12] 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. [19], 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. With a great level of detail [3] presented a geo-spatial 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 [27, 28] 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 for a given answer set based on semantic similarity enabled by the approaches and measures that we used in this work [2, 4].

7. CONCLUSIONS AND FUTURE WORK

We addressed the problem of adding semantic-awareness to TDWs of symbolic trajectories by augmenting them with semantic relatedness data, for the purpose of increasing their flexibility when generating answers to users' queries. We gave the corresponding constellation schema and described novel queries enabled by the SR-TDW model. Our experiments demonstrated the effectiveness of the proposed methodologies in terms of yielding richer answer-sets, the extent of which, as we discussed, may vary based on the measure used. As part of our future work, we plan a detailed formal classification of the SR-TDW enabled queries along the existing TDW taxonomies [24]. We will also address efficiency-related tasks from the perspectives of the design of warehouse schemata, queries optimization and the impact of the relatedness measures.

8. 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*, 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*, 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 *Proc. of WISE*, 2014.

[6] M. L. Damiani and R. H. Güting. Semantic trajectories and beyond. In *Proc. of IEEE - MDM*, 2014.

[7] I. Donevska. Advancing the semantic relatedness approach by using sense popularity. In *Proc. of IEEE - ICSC*, 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*, 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 *Proc. of WISE*, 2008.

[11] R. H. Güting and M. Schneider. *Moving Objects Databases*. Morgan Kaufmann, 2005.

[12] 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.

[13] G. Liu, R. Wang, J. Buckley, and H. M. Zhou. A wordnet-based semantic similarity measure enhanced by internet-based knowledge. In *Proc. of SEKE*, 2011.

[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 *Proc. of IEEE - ICDE*, 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 *Proc. of DaWaK*, 2007.

[18] A. Panchenko. *Similarity Measures for Semantic Relation Extraction*. Phd thesis, Universite catholique de Louvain, 2013.

[19] 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.

[20] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Computational Linguistics and Intelligent Text Processing*, 2003.

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

[22] C. Renso, S. Spaccapietra, and E. Z. (editors). *Mobility Data: Modeling, Management and Understanding*. Cambridge University Press, 2013.

[23] P. University. About wordnet, 2010. http://wordnet.princeton.edu.

[24] A. A. Vaisman and E. Zimányi. What is spatio-temporal data warehousing? In *Proc. of DaWaK*, 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*, 2013.

[27] R. Wannous, A. Bouju, J. Malki, and C. Vincent. Ontology inference using spatial and trajectory domain rules. In *Proc. of WorldComp*, Las Vegas, USA, 2014.

[28] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM TIST*, 4(3):49, 2013.