Mob-Warehouse: A semantic approach for mobility analysis with a Trajectory Data Warehouse

Ricardo Wagner[1], José Antonio Fernandes de Macedo[1], Alessandra Raffaetà[2], Chiara Renso[3], Alessandro Roncato[2], and Roberto Trasarti[3]

[1] Universidade Federal do Ceará, Brazil
[2] Ca' Foscari University of Venice, Italy
[3] ISTI-CNR, Italy

Abstract. The effective analysis and understanding of huge amount of mobility data have been a hot research topic in the last few years. In this paper, we introduce *Mob-Warehouse*, a Trajectory Data Warehouse which goes a step further to the state of the art on mobility analysis since it models trajectories enriched with semantics. The unit of movement is the (spatio-temporal) point endowed with several semantic dimensions including the activity, the transportation means and the mobility patterns. This model allows us to answer the classical Why, Who, When, Where, What, How questions providing an aggregated view of different aspects of the user movements, no longer limited to space and time. We briefly present an experiment of *Mob-Warehouse* on a real dataset.

1 Introduction

With the incredible availability of mobile devices equipped with geographical localization services, it has become economical and technically feasible to collect a huge amount of moving object traces in real life. Many interesting applications may take advantage of such data performing analysis on the moving objects. For example, in a traffic management system, traffic jams may be determined by mining movement patterns of groups of cars. Similarly, in a zoology application, the analysis of a group of bird trajectories can help explaining their migration patterns. However, trajectory data as they are collected by mobile devices are usually represented as a set of triples (latitude, longitude, timestamp), lacking other non spatio-temporal information like why objects move and other contextual semantic information. These non spatio-temporal aspects are key factors for the real success and deployment of the trajectory analysis results in any application scenario.

Semantic enrichment of mobility data refers to the process of integrating domain knowledge with trajectory data. In other words, the spatio-temporal points forming the trajectory need to be linked to application domain data giving rise to what is called *semantic trajectories*. How can semantic trajectories be useful for application analysis? Let us assume that we have a trajectory dataset

about people moving around a city, and we want to answer the following query: *Which is the average distance traveled by people using public transportation to visit at least one cultural attraction?* In this scenario, we should integrate each trajectory relative to some user with city's points of interest and their categories (like "cultural attraction") where the user stops, and with the transportation means used by the moving user. Then, we should compute the average distance for all trajectories satisfying the required conditions. Other queries may need not only domain knowledge integrated into trajectory data, but also the capability to combine complex analysis methods with the trajectory dataset. For example, the query *Which are the car trajectories belonging to a traffic congestion where the average speed inside this congestion is less than 30 km/h?* Here we use the notion of traffic congestion which is not native in spatio-temporal data, but can be computed with trajectory mining algorithms. Again, we need to combine different non spatio-temporal aspects like traffic jams with pure spatio-temporal components like the speed of the car, and then aggregate to compute the average velocity. The combination and aggregation of these different aspects on large trajectory datasets can be handled by a Trajectory Data Warehouse (TDW).

Traditional DW techniques and current analytic tools do not satisfy the requisites for a TDW, since the representation and aggregation of trajectories requires the capability of modeling and aggregating varied and interrelated data types, such as geometries, context information and spatio-temporal objects. Consequently, despite several proposals, there is still no consensus about how trajectories should be modeled multidimensionally and organized in different levels of aggregation in a DW, and about what functionality such DWs should support [11]. For example, in [7] a data model for storing measures related to trajectories is presented, focusing on the efficient approximation of aggregates. The same data model has been adopted in [6], to evaluate design solutions that integrate moving object databases (MOD) and DW. In [9] such a framework has been used to examine traffic data, in combination with tools for the visual analysis of spatio-temporal data. Other proposals extend spatial Data Warehouses to include in the model a temporal dimension for dynamic spatial data (e.g, [3]). In [5] two modeling approaches are proposed, both based on design patterns, for devising trajectory data schemas for relational and multidimensional environment. The main limitation of the mentioned approaches is the fact that they do not deal with *semantic trajectories*, but simply with sequences of spatio-temporal points.

Substantial research has been conducted on providing methods and prototype systems for enriching trajectories with domain knowledge like points of interest visited by the users, transportation means, user activities and annotations [8]. The model CONSTAnT [1] proposes a conceptual model for semantic trajectories, but with no specific reference to a data warehouse model. Nevertheless, CONSTAnT model inspired our current work in combining spatio-temporal and semantic aspects in a general concept of semantic trajectory.

Given this context, we propose a comprehensive TDW model for mobility called *Mob-Warehouse*, enriching trajectory data with domain knowledge. In

particular, the model is based on the so called 5W1H (Who, Where, When, What, Why, How) framework [13]. This is a well-known approach for getting the complete story on a subject, often mentioned in journalism, research, and police investigations. In our case, we intend to use this framework for specifying contextual information on trajectories and analyze the different aspects of the "mobility story" that the user "is writing" with his/her tracks. It will also guide the specification of the ETL process, which integrates trajectory data and domain application data. We developed a prototype implementation of the model and we experiment it in a case study consisting of a large dataset of car trajectories.

This paper is organized as follows. Section 2 introduces some basic concepts like trajectories and semantic trajectories. Section 3 briefly describes the 5W1H model and presents the conceptual model of *Mob-Warehouse*. Section 4 discusses the case study and shows some interesting queries that Mob-Warehouse allows to answer. Section 5 draws some conclusions.

2 Preliminaries

Several works in the literature address the analysis of trajectory data. Even the definition of a trajectory can have several variants. A trajectory can be defined as a representation of the spatio-temporal evolution of a moving object. However, since trajectories are usually collected by means of position-enabled devices, the notion of trajectory has to deal with the concept of *sampling*. Here we call *raw trajectory* the discrete representation of a trajectory as a sequence of spatio-temporal points or *samples* as collected by the device.

Definition 1 (Raw Trajectory). *A trajectory T is an ordered list of spatio-temporal points or* samples $p_1, p_2, p_3, \ldots, p_n$. *Each* $p_i = (id, x_i, y_i, t_i)$ *where id is the identifier of the trajectory,* x_i, y_i *are the geographical coordinates of the sampled point and* t_i *is the timestamp in which the point has been collected, with* $t_1 < t_2 < t_3 < \ldots < t_n$.

From a raw trajectory it is possible to infer a number of properties of a trajectory like the speed, the acceleration and the traveled distance. However, some other aspects are missing like the places visited by the object, or the performed activities. The concept of semantic trajectory has been proposed as a way to overcome the lack of semantics characterizing raw trajectories. A well known definition of semantic trajectory relies on the "stop and move" approach: a trajectory is segmented into parts where the object is stopped (the "stop") and the parts where the object is changing his/her position (the "move") [10]. This approach evolved to the more general definition of *episodes* to represent segments of a trajectory complying to some predicate representing the semantics of that segment, like the transportation mean, the goal or activity [8]. A further evolution towards this direction brought to the definition of a conceptual model for semantic trajectories as proposed in [1] where several contextual aspects contribute to create the concept of semantic trajectory. Formally:

Definition 2 (Semantic Trajectory). *A semantic trajectory is a trajectory that has been enhanced with annotations and/or one or several complementary segmentations.*

3 Mob-Warehouse

This section introduces the Mob-Warehouse model which is organized around the notion of semantic trajectory where different aspects contribute to describe the context.

As already discussed, raw trajectories are, by nature, semantically poor and they have to be enriched with domain knowledge in order to achieve better understanding of moving objects behavior. In the literature, to the best of our knowledge, no systematic and comprehensible method for accomplishing this task exists. Thus, we need a basic framework for describing object's movement, which provides a minimal set of information that is expressive enough for helping analyzing object behavior.

3.1 The 5W1H model

In this research, we propose the use of a narrative method as a conduit for systematically explaining the context involving a moving object trajectory. To this end, we resort to the six narrative questions, which were first mentioned in the poem Six Honest Serving Men [4], coined by the mnemonic 5W1H. The basic idea of this approach is to apply six narrative questions of *Who, What, When, Where, Why* and *How* to provide a consistent amount of understanding of the context of a circumstance. The 5W1H framework has been recurrently used by journalists as a guide for narrating a fact.

Each narrative question of the 5W1H model is mapped to a specific trajectory feature. In this way, we describe an object (*Who*) moving by a transportation means and/or having a certain behavior (*How*), performing an activity (*What*), for a certain reason (*Why*), at a given time (*When*) and place (*Where*).

Using this narrative approach, we may increase the level of semantic information into our model allowing to perform more meaningful queries about moving object habits. Below, we discuss the correlation of each question with trajectory features:

Who: This addresses the identification of a moving object, which is easily answered in case all objects are identified by the tracking system.

Where: This concerns the place where the trajectory point is located. Having the georeferenced location of each trajectory point, we may associate the latitude and longitude with a set of points of interest.

When: This question refers to the time extent related with trajectory points. This question is necessary to associate sampled points with specific calendar events, week periods, at different levels of details.

What: This question refers to what a moving object is doing, what task it is trying to perform or achieve. Clearly, answering this question is a challenge

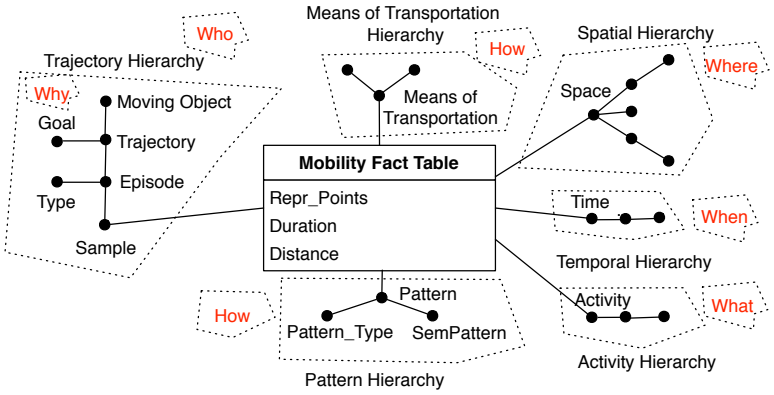

Fig. 1. TDW Conceptual model

since we should infer from each trajectory point the corresponding activity. This process could be facilitated using the information provided by other questions.

Why: This represents the motivation for traveling. This is an issue since this is so deeply rooted into the moving entity intent.

How: This question identifies both the way the object is moving, like the transportation means, and the behavior like belonging to a traffic congestion. Answering to this question could be challenging: when trajectories use multi-modal transportation means, identifying the transportation means could be far to be trivial. Moreover, mining algorithms have to be used to extract meaningful behavior: trajectories can be associated with one or more patterns depending on the fact the trajectory *entails* or satisfies the behavior.

3.2 The TDW conceptual model

Following the 5W1H Model we define a TDW conceptual model with six dimensions, as illustrated in Fig. 1. First, two dimensions represent space and time and they correspond respectively to the *Where* and *When* questions of the 5W1H Model. With respect to [6, 9] here the spatial domain can be structured according to the application requirements, providing the user with a greater flexibility. The spatial hierarchy is no longer restricted to consist of simple regular grids, the user can define hierarchies of regions with ad-hoc shapes. While a hierarchy of regular grids can be used to analyze objects that can move freely in the space, hierarchies with ad-hoc shapes are useful for objects whose movements are constrained, such as objects that can only move along a road network (e.g., cars). Moreover, Voronoi tessellation can be employed in order to build hierarchies of regions based on the actual distribution of the points forming the trajectories. This kind of partitioning turns out to be particularly suited for highlighting the directions of the trajectory movement.

In [6, 9] a third dimension named *object group* represents features of the objects under analysis. Here we call this dimension *Trajectory*, as it becomes a

central component of our model and it is used to represent the trajectory of the moving objects. At the base granularity it represents a single sample (id, x, y, t) belonging to the trajectory identified by *id*. The hierarchy having *Sample* as a root mixes together semantic and geometric features. A sample belongs to an *episode*, which can be classified according to its *Type* (e.g., a *stop* or a *move*) and it is grouped into a *Trajectory*. Each *Trajectory* is associated not only with the *Moving object* but also to a *Goal*, which is the main objective of such a trajectory. This dimension allows one to model *Who* is performing the action (the moving object) and the attribute *goal* answers the question *Why*. A fourth dimension, called *Activity*, states the activity the object is doing in a certain sample. This allows one to describe in a very detailed manner *What* is going on at the different samples of a trajectory. We can build a hierarchy of activities which classifies properly the variety of things an object can perform. Usually this hierarchy is application dependent hence in the general model it is not specified and should be instantiated case by case depending on the application requirements. Then the dimension *Means of Transportation* represents which transportation means the object is using for the movement. The last dimension, called *Pattern*, collects the patterns mined from the data under analysis. In this way we can directly relate trajectories to the patterns they belong to. An example of hierarchy on this dimension could be: each pattern is associated with its *Type*, such as cluster, frequent pattern, flock, and with a *Semantic Pattern*, which expresses the interpretation of such a pattern. For instance, in [2] to compute the movements of commuters in the city of Milan a clustering algorithm is applied to the trajectories of the moving objects. We can store these mined clusters in Mob-Warehouse as follows. Each cluster c_i is represented as a row in the dimension table *Patterns*: the identifier of the cluster is the primary key, the attribute *Semantic Pattern* can state *north-east commuters*, to point out that the semantic interpretation of cluster c_i is a group of people moving from North to East whereas the attribute *Pattern_type* assumes *cluster* as value. The latter two dimensions express the concept of *How* the movement is performed.

The fact table stores measures concerning the samples of the trajectories. Differently from [6, 9] where at the minimum granularity data were already aggregated, in this paper we want to record the detailed information to give the user the flexibility to analyze the behavior according to various points of view: at the minimum granularity we store information related to a single sample of a trajectory, specifying the kind of activity is doing, the means of transportation is using, the patterns, the space and time it belongs to. Then by aggregating according to the described hierarchies we can recover also properties concerning the whole trajectory or groups of trajectories satisfying certain conditions.

In the fact table we store measures related to a given sample $s = \langle id, x, y, t \rangle$

- *Repr_Points* is a spatio-temporal measure containing the spatial and temporal component of the sample, i.e., (id, x, y, t);
- *Duration* is the time spent to reach the sample from the previous point of the same trajectory in the same granule. It is set to 0, if this is the first point of the trajectory in such a granule;

– *Distance* is the traveled distance from the previous point to the sample of the trajectory in the same granule. It is set to 0, if this is the first point of the trajectory in such a granule.

As far as the aggregate functions are concerned, for the measures *Duration* and *Distance*, we use the *distributive* function *sum*: super-aggregates are computed by summing up the sub-aggregates at finer granularities. On the other hand, the aggregate function for the measure *Repr_points* can be defined in different ways according to the application requirements. The simplest way is to use the *union* operator to join together the points satisfying given conditions. Differently one can return a bounding box enclosing all the points or compress the points removing the ones which are spatio-temporally similar. For our experiments in Section 4 we will assume the union operator.

4 Experiments

Here we present a real case study using a trajectory dataset of people traveling by car in Milan (Italy), during one week in April 2007. The dataset contains track of 16,946 cars and 48,906 trajectories for a total of 1,806,293 points. We start describing the ETL process designed for building the Mob-Warehouse. Next, we present some queries that help analyzing people behavior in Milan city.

ETL Process. Although the description of the ETL process is not the main focus of this paper, we discuss briefly here the steps performed to populate Mob-Warehouse. The ETL process includes a semantic enrichment step, whose goal is to associate semantic information from the application domain with the trajectory data. Clearly, the semantic enrichment step is application dependent and may be very complex. Here, we adopted a quite simple semantic enrichment approach sufficient to illustrate the underlying idea and to show that the model is powerful enough to perform interesting analyses.

An important preprocessing task consists in transforming the GPS samples (id, x, y, t) into trajectories. The first step of such process is called *trajectory segmentation*, where the main goal is to split the samples into groups of related elements. We assume that trajectories end at three o'clock in the morning. The next step is the *stop identification*, where each sample must be classified as a move or a stop. This is done by means of a speed-based approach [12]. In our experiment, when the distance and the time difference between two consecutive samples are below the thresholds of 100 meters and 20 minutes, the second point is labeled as a stop. Otherwise, it is considered as a move. Hence a trajectory can be viewed as an alternation of a stop and a move.

The **spatial dimension** (space_dim) was populated using data extracted from the Open Street Map project[1] covering the city of Milan. Each kind of spatial object like streets, neighborhoods, districts, municipalities, states, countries and points of interest (POI) are represented as geometries. While streets are represented with linestrings, the other spatial objects are represented with points,

[1] http://www.openstreetmap.org

polygons or multipolygons. Moreover, the POI is also described by its category in a specific text attribute (poi_category). The **temporal dimension** (time_dim) is populated by using the temporal component of trajectory samples, which are copied to the temporal hierarchy lowest level. Higher levels (day of the week, month, year) are built by aggregating lower levels. The **trajectory dimension** (traj_dim) is populated, at its lowest granularity with trajectory samples. Higher levels of the trajectory hierarchy are populated with the episodes, which could be a *begin*, an *end*, a *stop*, or a *move*. We distinguished two special kinds of stops, namely *Home* and *Work*. To identify them, we compute the most frequent locations and we assigned the first most frequent to *Home* and the second most frequent to *Work*. The **activity dimension** (activity_dim) is populated with 12 predefined activities derived from a transportation research survey performed in the context of the project DataSIM[2]. Clearly, the approach is parametric with respect to the list of possible activities and can be changed depending on the application requirements. The **transportation means** dimension is not instantiated since we use only data describing car movements. The **pattern dimension** is populated with the mined patterns (clustering, frequent patterns, and flock detection) from the Milan trajectory dataset. For instance, in [2] a clustering algorithm is applied to extract groups of trajectories ending in a similar place. The identifier of each cluster is stored into the pattern dimension and it is associated with the samples of the trajectories belonging to such a cluster. Moreover, the attribute *SemPattern* specifies a semantic interpretation of the cluster, when it is known.

Queries Evaluation. Mob-Warehouse can be queried for analyzing people behavior in Milan city. We next discuss two examples. The first query aims at understanding what people usually do after leaving home, by finding *the most frequent activity after home*. Considering the 5W1H, such query uses the *What* perspective. For answering this query, we created a view, called *stopsAtHome*, to represent when a person is at home, i.e. his/her stops at home.

```
SELECT ac.category, COUNT(*)
FROM points_fact pf, traj_dim tr, activity_dim ac, stopsAtHome stops
  WHERE (join conditions) AND tr.trajectory = stops.trajectory
  AND tr.episode = stops.episode + 2 AND ac.category <> 'HOME'
GROUP BY ac.category ORDER BY 2 DESC;
```

where `join conditions` consists of equalities between the foreign keys in the fact tables and the corresponding primary keys in the dimension tables. The attribute `episode` contains a progressive number identifying the segments composing a trajectory. Hence the condition `tr.episode = stops.episode + 2` states that we look for the first stop after 'HOME'.

The result of this query is reported in the table below. We notice that the prevalent activity is *Working*. The second most frequent is *Leisure*, i.e., going for leisure activities like jogging in a park. At the third position we find *Services* like

[2] http://www.datasim-fp7.com

visiting a doctor or a bank. The low number of education activities is probably due to few POIs related to education.

Variants of the same query may be used to find activities after work or before home/work by only modifying the view.

Activity	After Home
Working	6401
Leisure activities	1866
Services	1403
Shopping	593
Sports	523
Eating	473
Social activities	434
Education	63
Religion	40

The second query aims at exploring the *When*, *Why*, *Where* and *Who* perspectives by *finding the users and their time spent traveling from home to work*. Below, we show the query used to answer this question. We use, in this query, a view, called *firstWorkStop*, to select the timestamp of the first sample, after leaving home, having as activity Work and classified as a stop.

```
SELECT traj_dim.trajectory, SUM(points_fact.duration) AS time
   FROM points_fact, traj_dim, activity_dim, time_dim
   WHERE (join conditions) AND
   ((traj_dim.episode_type = 'BEGIN' AND activity_dim.category = 'HOME')
       OR (traj_dim.episode_type <> 'BEGIN'))
   AND time_dim.minute <= firstWorkStop(traj_dim.trajectory)
GROUP BY traj_dim.trajectory;
```

The result of this query is a list of users with the duration of their travel from home to work. The effective values are not worth to be presented. We just mention that there is a high variability in the durations (from less than one hour to several hours). This deserves a further deeper analysis to properly understand what happens in travels with longer durations.

5 Conclusion and Future Work

We presented Mob-Warehouse, a semantic-enhanced warehouse for trajectories. The main contribution is the introduction of a model where the spatio-temporal component of trajectory data is properly integrated with context related information like transportation means, performed activities and mobility patterns. The key idea is to consider the spatio-temporal point sampled by a device as the lowest granularity to which semantic information is linked. We followed the 5W1H model to describe the context where objects move. This allows us to express a wide range of queries, involving the questions Who, What, Why, When, Where and How. Due to lack of space we only briefly sketched some experiments

performed with Mob-Warehouse on a real dataset of trajectories. Future works include the development of a more sophisticated ETL process to enrich raw trajectory data with contextual information and the definition of more complex aggregate functions for the measure *Repr_Points*, like *representative trajectories*.

6 Acknowledgments

We acknowledge European Projects SEEK Marie Curie Action N. 295179, DATASIM FET N. 270833, the national research project PON TETRis (no. PON01 00451) for partially supporting this work.

References

1. 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. *Transactions in GIS, Online First*, 2013.
2. F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB J.*, 20(5):695–719, 2011.
3. 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.
4. Kipling. Six Honest Serving Men Poem. `http://www.kipling.org.uk/poems_serving.htm`. Accessed: 2013-06-05.
5. B. Leal, J. A. F. de Macêdo, V. C. Times, M. A. Casanova, V. M. P. Vidal, and M. T. M. de Carvalho. From Conceptual Modeling to Logical Representation of Trajectories in DBMS-OR and DW Systems. *JIDM*, 2(3):463–478, 2011.
6. G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà, and Y. Theodoridis. Building Real World Trajectory Warehouses. In *Proc. of MobiDE*, pages 8–15, 2008.
7. S. Orlando, R. Orsini, A. Raffaetà, A. Roncato, and C. Silvestri. Trajectory Data Warehouses: Design and Implementation Issues. *Journal of Computing Science and Engineering*, 1(2):240–261, 2007.
8. C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. A. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys*, 45(4), 2013. To appear.
9. A. Raffaetà, L. Leonardi, G. Marketos, G. Andrienko, N. Andrienko, E. Frentzos, N. Giatrakos, S. Orlando, N. Pelekis, A. Roncato, and C. Silvestri. Visual Mobility Analysis using T-Warehouse. *J. of Data Warehousing and Mining*, 7(1):1–23, 2011.
10. S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1):126–146, 2008.
11. A. A. Vaisman and E. Zimányi. What Is Spatio-Temporal Data Warehousing? In *Proc. of DaWaK*, volume 5691 of *LNCS*, pages 9–23. Springer, 2009.
12. Z. Yan. Towards Semantic Trajectory Data Analysis: A Conceptual and Computational Approach. In *VLDB PhD Workshop*, 2009.
13. L. Yang, Z. Hu, J. Long, and T. Guo. 5W1H-based Conceptual Modeling Framework for Domain Ontology and Its Application on STPO. In *Proc. of SKG*, pages 203–206. IEEE, 2011.