Advanced Knowledge Discovery on Movement Data with the GeoPKDD system

M. Nanni R. Trasarti C. Renso F. Giannotti
KDD-Lab
ISTI-CNR, Pisa, Italy
name.surname@isti.cnr.it

D. Pedreschi
Dipartimento di Informatica
University of Pisa
pedre@di.unipi.it

ABSTRACT

The growing availability of mobile devices produces an enormous quantity of personal tracks which calls for advanced analysis methods capable of extracting knowledge out of massive trajectories datasets. In this paper we present an experiment on a real world scenario that demonstrates the strong analytical power of massive, raw trajectory data made available as a by-product of telecom services, in unveiling the complexity of urban mobility. The experiment has been made possible by the GeoPKDD system, an integrated platform for complex analysis of mobility data. The system combines spatio-temporal querying capabilities with data mining and semantic technologies, thus providing a full support for the Mobility Knowledge Discovery process.

1. INTRODUCTION

Research on moving-object data analysis has been recently fostered by the widespread diffusion of new techniques and systems for monitoring, collecting and storing location-aware data, generated by a wealth of technological infrastructures, such as GPS positioning and wireless networks [6]. These have made available massive repositories of spatio-temporal data recording human mobile activities, such as location data from mobile phones, GPS tracks from mobile devices, etc.: is it possible to discover from these data useful and timely knowledge about human mobility? This is a scenario of great potential opportunities and risks: on one side, mining this data can produce useful knowledge, supporting sustainable mobility and intelligent transportation systems; on the other side, individual privacy is at risk, as the mobility data contain sensitive personal information. The GeoPKDD project [1], since 2005, investigated this direction of research; the lesson learned is that the answer to the question above is yes, but with a big caveat: there is a long way to go from raw data of individual trajectories up to high-level collective mobility knowledge, capable of supporting the decisions of mobility and transportation managers. Such analysts reason about semantically rich concepts, such as systematic vs. occasional movement behavior, purpose of a trip, home-work commuting patterns, etc.; accordingly, the mainstream analytical tools of transportation engineering, such as origin/destination matrices, are based on semantically rich data collected by means of field surveys and interviews. Clearly, the price to pay for this richness is hard: mass surveys are very expensive, so that their periodicity is very broad – every 5 years is a standard in current practice – and obsolescence is rapid; poor data quality is also a plague: people tend to respond elusively (especially about their non-routine activity), inaccurately, or not to respond at all. On the other extreme, automatically sensed mobility data record individual trajectories at mass level, collected on a continuous basis: real mobile activities, faithfully sampled as they occur, in real time. Clearly, the price to pay here is exactly the lack of semantics in raw data: How to bridge this deficiency?

The system illustrated in this paper is capable, for the first time, of demonstrating the striking analytical power of massive trajectory data in unveiling the complexity of urban mobility. The GeoPKDD platform coherently integrates solid analytical methods with a semantic-based query, mining and reasoning system, capable of mastering the complexity inherent to what we call the geographic privacy-aware knowledge discovery process. To our knowledge, no other integrated analytical solution exists, capable of supporting the whole knowledge discovery process in the context of mobility data. GeoPKDD system offers traffic management functionalities that are complementary to other existing commercial traffic management tool, such as traffic simulators (e.g., [2]).

The novel contribution of this paper is to present the GeoPKDD system as the proper integration of previously developed tools for analysis of spatio-temporal data. The system comes with a new query language capable of expressing, in a uniform way analysis, mining and reasoning queries.

2. THE GEOPKDD SYSTEM

A system able to master the complexity of the knowledege discovery process over mobility data needs to support at least four functionalities: (i) trajectory data need to be created, stored and queried through spatio temporal primitives; (ii) trajectory models and patterns representing collective behaviour have to be extracted using trajectory mining algorithms; (iii) such patterns and models have to be represented and stored in order to be re-used or combined; (iv) new mining algorithms may be added.

The first preliminary choice for the implementation of the GeoPKDD system has been to demand the storing and man-

EDBT 2010, March 22–26, 2010, Lausanne, Switzerland.

agement of trajectories to an existing MOD. Hermes [10] has been selected as best candidate, particularly for being based on an object-relational data model. The second choice was adopting the idea (Weka-like) of having a library of algorithms for trajectory mining that might grow during time. The third idea was to design a query language that might allow the analyst to progressively combine mining and querying; namely focusing on a spatio temporal area, selecting a mining algorithm, using it for mining patterns on that area, changing the area, doing again some mining, storing the patterns, quering them, etc. The fourth choice was to require that the ability to dress such data with domain knowledge is tightly integrated with querying and mining.

Figure 1 illustrates the architecture of the GeoPKDD system that assembles together four main components, namely the Data Manager, the Data Mining Query Language Executor, the Library of Trajectory Mining algorithms and the Semantic component.

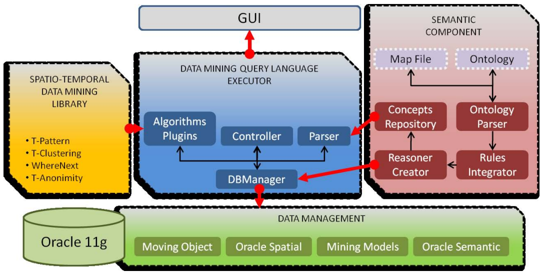

Figure 1: The GeoPKDD Integrated system

The central box of Figure 1 illustrates the main modules of the Data Mining Query Language executor, which is the kernel of the GeoPKDD system and supports the processing of queries expressed with the data mining query language (DMQL) [9]. The introduced query language is an extension of a spatio-temporal SQL with mining calls statements. In of the DMQL component.

The technological solution adopted for data storage of the GeoPKDD Integrated platform is based on Oracle 11g extended with components for moving object data storage and manipulation, spatial objects representation, mining models storage and semantic technology. The moving object support is provided by Hermes [10], which defines a collection of moving object data types, each accompanied with a palette of specialized operations provided as an Oracle data cartridge. Hermes is, in turn, based on Oracle Spatial [8] thus capable of natively representing spatial objects. Oracle 11g has been also extended [9] with data types for representing data mining patterns, thus supporting natively the mined pattern storage and manipulation for all mining algorithms actually available in the system. Furthermore, the Semantic Technology component of Oracle 11g provides primitive support for ontology storage and reasoning. Therefore, Oracle 11g with Hermes becomes the basic data storage of the system for both the data mining and semantic components.

The library of trajectory mining algorithms is plugged into the data mining query language through the Algorithms Manager interface. Some of the plugged algorithms are: Trajectory pattern (T-Pattern) [5], that extracts common itineraries with detailed timings; the Trajectory Clustering (T-Clustering) [11], that groups together similar trajectories, based on a wide choice of different trajectory similarity functions; the WhereNext location prediction algorithm [7], that builds a T-Pattern-based predictive model able to estimate the future location of a trajectory; and, finally, Trajectory anonymization (T-Anonymization) [3], that produces an anonimized version of the input trajectory dataset, following a variant of the standard k-anonimity notion.

The semantic component is derived from [4] and is aimed at enriching both trajectories and mined patterns with domain information encoded in an ontology, thus making an explicit representation of semantic patterns. The results of running the semantic component on a set of trajectories or patterns is that they are, possibly, classified into one or more ontology defined classes, thus producing a kind of ”semantic tagging” of trajectories and patterns.

3. A REAL URBAN MOBILITY SCENARIO

In the demo of the system we will show a set of experiments performed on a real world case study, that demonstrate the capabilities of the GeoPKDD platform and how they can be exploited to extract useful knowledge from raw mobility data.

The analysis capabilities of our system have been applied onto a massive real life GPS dataset, obtained from 17,000 vehicles with on-board GPS receivers under a specific car insurance contract, tracked during one week of ordinary mobile activity in the urban area of the city of Milan, Italy; the dataset contains more than 2 million observations. Using the preprocessing algorithm integrated in the system, we cleaned the GPS dataset and reconstruct the trajectories. This last step yield more than 200,000 trajectories.

By applying our mobility data mining methods to this dataset, we developed a set of novel analytical services for mobility analysis and traffic management, designed and validated in collaboration with Milan Mobility Agency. Three representative examples are discussed below.

Origin-Destination Analysis. The automated construction of Origin/Destination (O/D) matrices from mobility data in a timely, reliable and objective manner, overcoming the limitations of the current survey-based approach. The O/D matrix is a popular tool of transportation engineering, describing users' flows between any pairs of certain geographic areas designated as possible origins and destinations of users' trips; the current practice for estimating O/D flows is through data collected by periodic surveys (every 5 years in Milan, sometimes enriched with road sensor data), with obvious limitations due to high costs of interviews, poor data quality and rapid obsolescence.

Finding Itineraries. Providing insights about how the flow between some given origin and/or destination is distributed along the paths over the road network; e.g., describing the main itineraries towards a specific destination, such as a crucial parking lot.

Detecting Systematic Movements. The detailed analysis and discovery of systematic movement behaviors, i.e., the movements that repeat periodically during the week, with particular emphasis to home-to-work and work-to-home commuting patterns.

Due to the lack of space we briefly present two of the analysis performed with the platform. However, the full set of analysis results will be available during the demonstration.

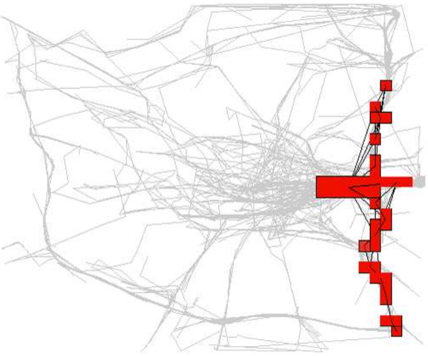

Figure 2: T-patterns to park n. 317

3.1 Finding Itineraries

The analysis tools integrated in the GeoPKDD system allow to analyse the traffic directed to a destination in detail, taking into account the actual routes followed by all vehicles to reach the destination. In particular, we selected all trajectories that ended in one of the parking lots considered above – more precisely, the one close to the Linate airport – and extracted a set of frequent T-Patterns that describe their most common routes with timings. We remind that a T-Pattern describes a sequence of common visited places (modeled through rectangular regions) together with the typical transition times between each pair of consecutive elements of the sequence.

As an example of the expressivity of the DQML adopted by the system, the following mining query was used to obtain the results described in this section:

```
CREATE MODEL mobility_tpattern AS MINE T-PATTERN
FROM (SELECT t.id, t.trajectory
    FROM TrajectoryTable t, RegionTable r
    WHERE r.id=317
    AND t.trajectory.final().intersection(r.area))
WHERE T-PATTERN.support = 0.18 AND
T-PATTERN.side = 0.0045 AND T-PATTERN.time = 60
```

where TrajectoryTable and RegionTable contain, respectively, input trajectories and parking lots, and support, side and time are parameters of the T-Pattern mining tool.

Figure 2 shows an overall view of all the T-patterns obtained with a minimum support threshold set to 18% and time tolerance set to 1 minute. It is clear that the most frequent routes to the parking lot follow the eastern side of the *tangenziale* (the main ring road of the city). That is even clearer on Figure 3, where four most significant patterns are selected and shown in detail, together with the corresponding timings. The starting region of each T-pattern has a black border around it, and consecutive regions in a T-pattern are connected through a line. Beside the visualization of each T-pattern is reported the list of typical transition times between consecutive regions, in the format *"step_number min_time max_time"*, times being expressed in seconds. For instance, the left picture of Figure 3 describes a T-pattern composed of four regions, and therefore three transitions. The first block of transition times is composed of interval $[37.45, 37.89]$ (around half a minute) for the first transition (*step_number* equal to 0), $[35.12, 55.67]$ (from half to one minute) for the second one, and $[125, 125.16]$ (around two minutes) for the last one.

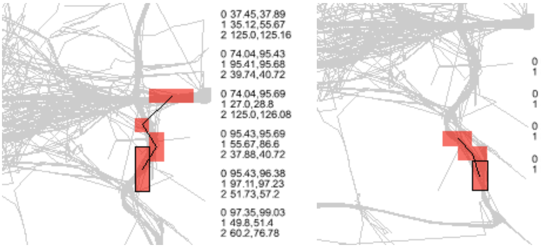

Figure 3: Selected T-patterns to park n. 317

3.2 Detecting Systematic Movements

The objective of this new analysis task is to compute and analyse the systematic movements contained in the dataset. A systematic movement has been defined by the Mobility Agency as *a frequent movement between two stops, in the user's history*. In particular, we distinguished between stops and regions of interests: the first represents the actual information about the trajectory suspension of movement, identified by a timestamp and a location, whilst the latter represents the geographic location where the stop happens.

Given the systematic movements, we exploited the Semantic Component to define a special kind of systematic movements, the Home-Work behaviour. Home-work movements have been defined by the domain expert as *trajectories starting with a systematic movement from a frequent starting point (the home location) ending in a long stop (the work place), possibly followed by other movements and ending again at the home location*. We indicate as *home* a location from which a user frequently starts his/her trajectories, as *long stop* a stop that lasts at least 3 hours and as *frequent move* a move with frequency support at least 3. Since the Home-work movement must finish in the home region, this behavior captures the routine movements of people going to work, possibly moving again for job reasons or for shopping, and then finally going back home.

Figure 4 shows two examples where two vehicles were selected, and all trajectories of each single vehicle are plotted together with the regions that are recognized as its instances of the *Home* and *Work* classes, respectively colored in blue and green. We can observe that: (i) homes and work locations are usually connected to rather fixed routes; (ii) however, several variations can be spotted, as well exemplified in Figure 4(bottom right), where apparently a few intermediate stops are performed in the route between home and work; (iii) as naturally expected, there are also movements that do not belong to the home-to-work, such as in the case in Figure 4(top right), where some trajectories move towards a destination (close to the center) not recognized as *work*.

The definition of systematic and Home-Work movements, can be exploited to focus the analysis on the subset of trajectories which belong to these two classes.

The hourly distribution of systematic movements is shown in Figure 5. Here, we can notice two different emerging behaviors: (i) the systematic movements (red curve) follow the global trend (green curve) but during the central hours of the days are relatively less frequent; (ii) the systematic traffic during the weekend is extremely low. These results provide a valuable insight that supports the hypothesis made by the mobility agency about a possible underestimation of the non-systematic traffic, so far manly based on personal

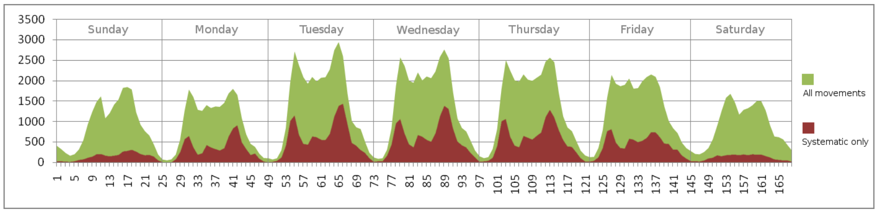

Figure 5: Movement distribution by hour, focused on systematic movements

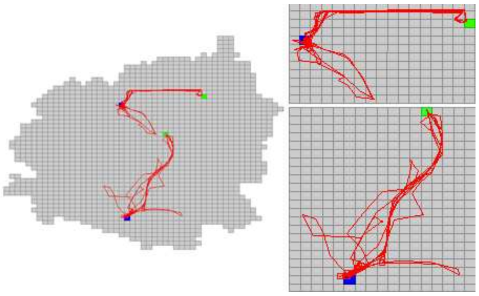

Figure 4: Sample Home-Work locations and movements

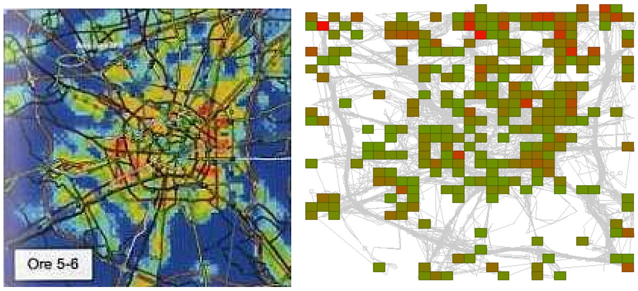

Figure 6: Population distribution between 5 and 6 a.m. compared against the distribution of *homes*

experience and indirect evidence.

Similarly, assuming that the distribution of the population at very early hours of the day reflects that of their residences, we can compare the *homes* distribution with the original figure provided by Milan Agency relative to 5-6 a.m. period, as shown in Figure 6. The two distributions match quite closely, showing essentially the same dense spots on both the figures. This seems to confirm both the value of GPS mobility data for this kind of analysis, and the correctness of the reasoning process followed to infer the location of *homes*.

Running times of queries depend on the task, and range from a few seconds for simple selections to few hours for complex (yet less frequent) operations such as pattern entailment on massive data (measures obtained on a 8-processors Xeon 2GHz, 8GB RAM Windows system).

4. CONCLUSIONS AND ROADMAP

The experiment illustrated in this paper demonstrates the strong analytical power of massive, raw trajectory data, made available as a by-product of telecom services, in unveiling the complexity of urban mobility. The experiment has been made possible by our querying, mining and reasoning system, which provided the analytical framework for mastering this complexity and extracting meaningful mobility knowledge.

We believe that the frontier of this line of research lies primarily in the semantics of *interactions* – both with the surrounding context and between mobile individuals themselves. On one hand, the increasing intelligence, connectivity and context-sensitivity of mobile devices is producing location data tagged with ever richer semantic information: P2P and ubiquitous computing interactions, query logs, sensor data – a whole new picture of social relations intertwined with mobile behavior. On the other hand, these richer data will enable deeper analytics, based either on known interaction schemata studied in natural and social sciences, such as encounters, convergence, leadership, or self-emerging interaction models, extracted by means of statistical and mining approaches. Semantic-based mobility data mining will get us closer to an *archaeology of the present*.

5. REFERENCES

[1] GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery. http://www.geopkdd.eu.

[2] VISUM software. http://www.ptvag.com.

[3] O. Abul, F. Bonchi, and M. Nanni. Never Walk Alone: Uncertainty for anonymity in moving objects databases. In *ICDE'08*, 2008.

[4] M. Baglioni, J. Macedo, C. Renso, R. Trasarti, and M. Wachowicz. Towards semantic intepretation of movement data. In *AGILE Int' Conf. on Geographic Information Science*, 2009.

[5] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD'09*, pages 330–339, 2007.

[6] F. Giannotti and D. Pedreschi, editors. *Mobility, Data Mining and Privacy - Geographic Knowledge Discovery*. Springer, 2008.

[7] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *KDD'09*, 2009.

[8] Oracle. Oracle spatial. http://www.oracle.com/technology/products/spatial.

[9] R. Ortale et al. The daedalus framework: progressive querying and mining of movement data. In *GIS*, page 52, 2008.

[10] N. Pelekis and Y. Theodoridis. Boosting location-based services with a moving object database engine. In *MobiDE*, pages 3–10, 2006.

[11] S. Rinzivillo et al. Visually-driven analysis of movement data by progressive clustering. *Information Visualization*, 7((3/4)):225–239, 2008.