Mobility, Data Mining and Privacy – the GeoPKDD project

Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Chiara Renso, Salvatore Rinzivillo, Roberto Trasarti

KDDLab
ISTI CNR, Via Moruzzi, 1 Pisa, Italy

ABSTRACT
The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles, generating large volumes of mobility data. Miniaturization, wearability, pervasiveness is producing traces of our mobile activity, with increasing positioning accuracy and semantic richness: Location data from mobile phones: (GSM cell positions), GPS tracks from mobile devices receiving geo-positions from satellites, etc. The objective of the GeoPKDD project is to discover useful knowledge about human movement behavior from mobility data, while preserving the privacy of the people under observation. While pursuing this ambitious objective, the GeoPKDD project has started a new exciting multidisciplinary research area, at the crossroads of **mobility, data mining,** and **privacy.**

Keywords
Mobility data, Data Mining, Warehousing, Privacy, Anonymity.

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 [1]. These have made available massive repositories of spatio-temporal data recording human mobile activities, that call for suitable analytical methods, capable of enabling the development of innovative, location-aware applications. This is a scenario of great 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 [2], since 2005, investigates how to discover useful knowledge about human movement behavior from mobility data, while preserving the privacy of the people under observation. GeoPKDD aims at improving decision-making in many mobility-related tasks, especially in metropolitan areas:

- Monitoring and planning traffic and public transportation systems
- Localizing new facilities and public services
- Forecasting/simulating traffic-related phenomena
- Geomarketing and location-based advertising
- Innovative infomobility services
- Detecting changes in collective movement behavior.

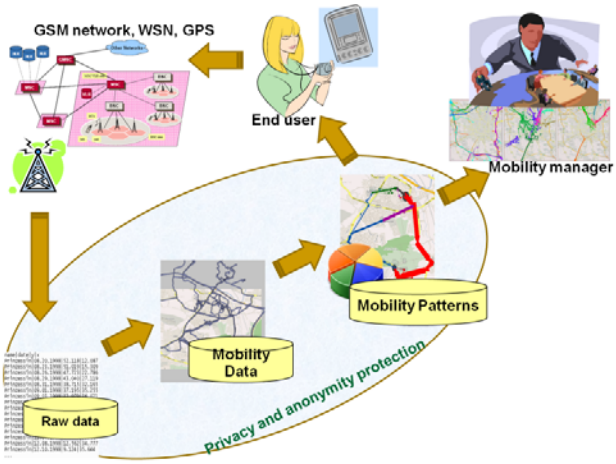

Fig. 1 The GeoPKDD process

2. TRAJECTORY MINING METHODS
The first set of achievements of GeoPKDD consists in a toolkit of breakthrough analytical methods for mining from massive trajectory dataset. A *trajectory* is a sequence of time-stamped locations, sampled from the itinerary of a moving object.

2.1 Trajectory Warehouse
The T-Warehouse is a spatio-temporal data cube representing various aggregated measures of the moving objects, such as presence and speed. Our T-OLAP engine supports exploratory analysis, drilling up and down the space and time dimensions.

2.2 Trajectory Patterns
A T-pattern is a sequence of locations that are frequently visited in the specified order with similar transition times; thus, a T-pattern reveals a frequently followed itinerary. Our mining algorithms automatically discover T-patterns in trajectory data.

2.3 Trajectory Clustering
A T-cluster is a set of similar trajectories, according to a repertoire of trajectory similarity functions; thus, a T-cluster reveals a group of objects sharing a systematic movement behaviour, e.g., home-

work-home commuting. Our density-based clustering algorithms discover T-clusters in trajectory data.

2.4 Trajectory Anonymity

A k-anonymous trajectory dataset is one where the itinerary of each person is indistinguishable from that of other k-1 persons – anonymity viewed as *hiding in the crowd*. Our T-anonymity methods transform a trajectory dataset into a new, k-anonymous dataset, such that the key analytical properties are preserved, together with users' privacy.

3. MASTERING THE GeoPKDD PROCESS

The second set of achievements of GeoPKDD consists in two analytical platforms, supporting the interactive, iterative, combined usage of the various tools to the purpose of discovering knowledge. GeoPKDD developed two prototype platforms: a semantic-based query & reasoning systems, and a visual analytic environment.

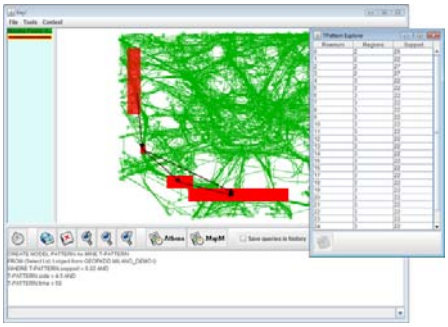

Fig.2 Result of a DMQL query in the semantic-based query & reasoning system

3.1 Semantic-based query & reasoning system

This system allows the user to describe the entire knowledge discovery process using a set of primitives, based onto a **Data Mining Query Language**:

- The *spatio-temporal query primitives* allow reconstructing trajectories from raw data, selecting and pre-processing trajectory data w.r.t. geographic background knowledge, creating anonymous versions of the trajectory datasets.
- The *trajectory mining primivitives* (see Sect. 2) allow extracting and validating mobility patterns and models (Fig.2).

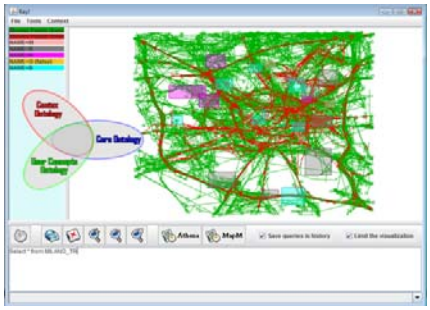

Fig. 3 Semantic-based query & reasoning system

- A *reasoning component* allows to specify domain-driven ontologies, inferring types of trajectories and patterns (Fig.3).

3.2 Visual Analytics

The aim of this system is to navigate through mobility data and patterns and visually drive the analytical process. Some key features:

- *Visualization of T-patterns* to support the navigation of the extracted patterns in the spatial and temporal dimensions (Fig. 4).

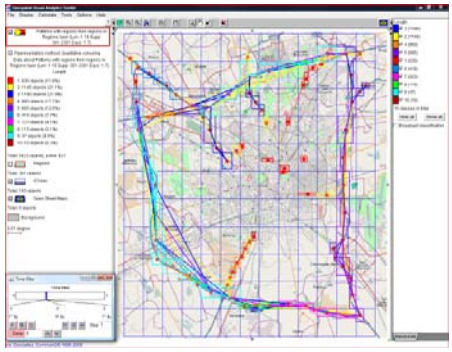

Fig. 4 Discovering T-patterns by visual analytics

- *Progressive refinement of T-clusters:* A user-driven exploration and evaluation of the discovered T-clusters, based on a step-wise iterative method.
- *Visual exploration of the T-Warehouse* to browse aggregated measures of moving objects (In Fig. 5, the triangle base represents the presence measure, while the triangle height is the speed).

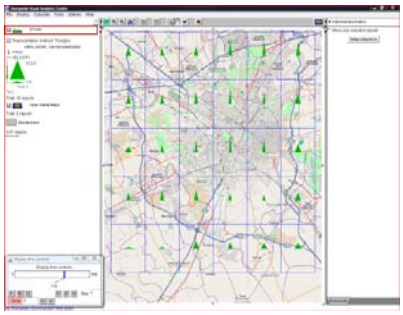

Fig. 5 T-OLAP using visual analytics

4. GeoPKDD CONSORTIUM

GeoPKDD has been carried out in 8 research laboratories: **KDD LAB, ISTI CNR and Univ. Pisa** (IT, coordinator), **Hasselt University** (B), **EPFL Lausanne** (CH), **Fraunhofer IAIS** (D), **CTI and Univ. Pireaeus, Athens** (GR), **Wageningen UR** (NL), **Universidad Politecnica de Madrid** (ES), **Sabanci University, Istanbul** (TK), and the telecommunication company **WIND** (I).

The project coordinator is **Fosca Giannotti** (ISTI-CNR), fosca.giannotti@isti.cnr.it

5. REFERENCES

[1] F. Giannotti and D. Pedreschi (Eds.) *Mobility, Data Mining and Privacy*. Springer, 2008

[2] GeoPKDD website, **`http://www.geopkdd.eu`**