Mobility, Data Mining and Privacy: The GeoPKDD Paradigm

Fosca Giannotti
KDD Lab - ISTI - CNR
Pisa, Italy

Fabio Pinelli
KDD Lab - ISTI - CNR
Pisa, Italy

Salvatore Rinzivillo
KDD Lab - ISTI - CNR
Pisa, Italy

Roberto Trasarti
KDD Lab - ISTI - CNR
Pisa, Italy

May 28, 2010

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 are 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. This paper gives a short overview of the analytical tools developed within the European Project GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery), a project funded by European Commission under the FET program of the IST FP6 framework.

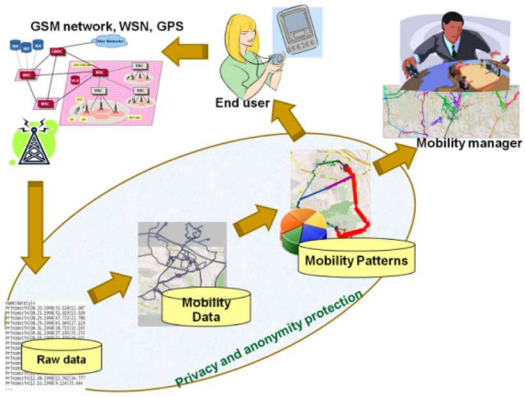

Figure 1: The GeoPKDD process

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. 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. The GeoPKDD project [1], 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.

The GeoPKDD system, originally presented in [5], allows to handle the whole knowledge discovery process from mobility data, in particular it provides tools for reconstructing a trajectory from raw logs, storing and querying trajectory data, classifying trajectories according to means of transportation (pedestrian, private vehicle, public transportation vehicle, extracting spatio-temporal pattern and models as useful abstractions of mobility data, find an optimal trade-off between privacy protection and quality of the analysis.

2 Experimenting on real GSM data

To demonstrate the power of our framework we have tested it on different real scenarios and different data sources. Here we present a set of experiments on a dataset of real GSM data logs. The observations are collected by the *Telecom Italia Lab*, the research laboratory of the main telecommunication company in Italy, using an estimation of the position of the devices by means of triangulation. The dataset contains the points recorded along a whole day (in particular the 21st May, 2009). The first step was the importing of the observations in the data management system which is based on Oracle 11g database. The trajectories are built starting from this raw data using the trajectory reconstruction algorithm, and cleaning them from errors and outliers obtaining a set of useful data (Fig.2).

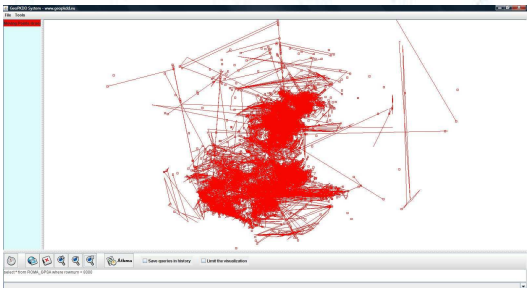

Figure 2: The reconstructed moving points dataset

The resulting set is composed by more than four million trajectories in the urban area of Rome.

2.1 Statistical Analysis. A set of statistical analysis can be easily computed having the data storage system integrated with a set of spatio-temporal primitives, that allow to efficiently compute spatial and temporal measures, like temporal gaps and spatial distances between consecutive points.

2.1.1 Distribution of movements during the day. For this analysis we partition the day in hours (0-24) and we will intersect the trajectory dataset with this periods counting the presence of the trajectories in them. The result is shown in Fig.3.

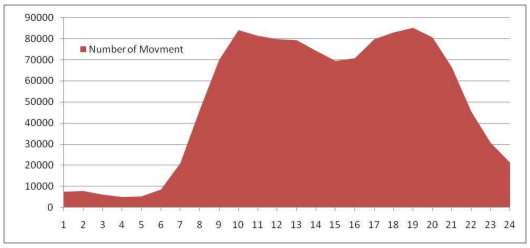

Figure 3: The Time distribution of movements

The analysis shows two major peeks during the day corresponding to the periods of time where the people are going or coming from work. Another aspect of the mobility is highlighted: the period between the peeks has a very high number of movements which gives to the mobility agent a clue about the sustainability of the traffic during such hours.

2.1.2 Density of movements in space. The distribution of movements can be analyzed not only in time but also in space. For example, by dividing the territory in a grid of 50×50 cells, we can compute the density within each cell. In this case we can take advantage from the previous analysis obtaining a spatio-temporal density distribution which can be navigated in both dimensions. The global result is shown in Fig.4.

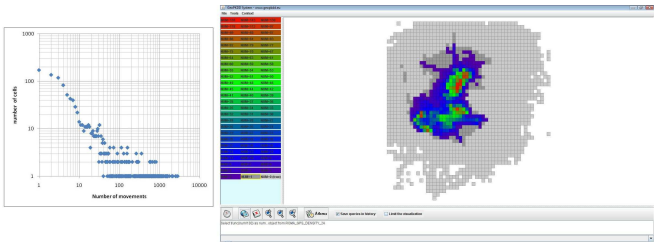

Figure 4: The density distribution plot (Left) and the density distribution on the grid (Right)

Using the temporal dimension we can focus the view only in a specific period, say from 6 am to 12 am, obtaining the Fig.5. As we can see in the morning the mobility is greater between two dense points in the south part of the city giving to the mobility manager the idea of where the peek of detected in the first analysis is focused.

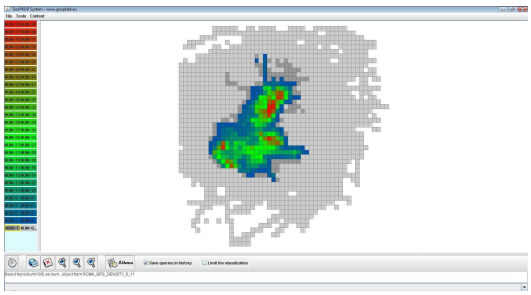

Figure 5: The density distribution in the morningof movements

To better understand the traffic flows of the traffic we proceed with the next statistical analysis called O-D Matrix.

2.1.3 O-D Matrix. Usually this analysis is obtained by the mobility agencies using a survey obtained from direct or by telephonic interview to the citizens. The quality of the data obtained in this way is very poor and has very high costs. Instead thanks to our system this can be done with a very low cost and high precision avoiding incomplete and incorrect data. For this example we introduce a bigger grid 20×20 which simulates the districts of the city. Having more information like real districts or regions of interest given by the mobility agency we can use them to perform this analysis. Joining the matrix with the used grid we can browse it by selecting a region as origin or destination. In Fig.6 we show how the people move from and to a cells.

The next section will show a further step: the statistical analysis using data mining algorithms such as *T-Clustering*

Figure 6: The OD Matrix focused on the yellow cell. The destinations (Left) and the origins (Right)

[3] and *T-Pattern* [2] and how they can interact with the previous analysis thanks to the unifying system.

2.2 Data mining analysis. The understanding of the mobility in a city is a very complex process, here we present an example of how the presented framework helps the analyst in the discovery process. An example of analysis is *looking for common behavior of people who move toward a common destination*. To perform this task the first step is to find this communities of people: the Fig.7 shows two clusters obtained applying the *T-Clustering* using the *common destination* distance function. Once the analyst identifies the clusters of intersect, it is possible to refine the analysis by investigating other regularities in their behavior, for example by applying the *T-Pattern* algorithm on the trajectories of one of the selected cluster (say, for example, the red cluster in Fig.7). The resulting T-Patterns are shown in Fig.8.

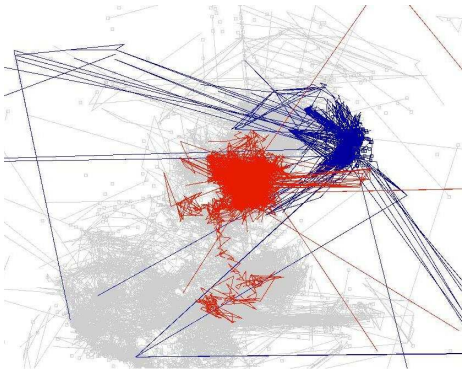

Figure 7: Two Clusters discovered using common destination distance function

To give semantic to the *T-patterns* extracted we can intersect the regions of the *T-Patterns* with a set of interesting places (specified by the mobility agency), and discover that the T-Pattern in Fig.8(b) represents the behavior of people coming in a common area. This simple process show the capabilities of the system allowing the user to perform iterative querying mixing together different data mining

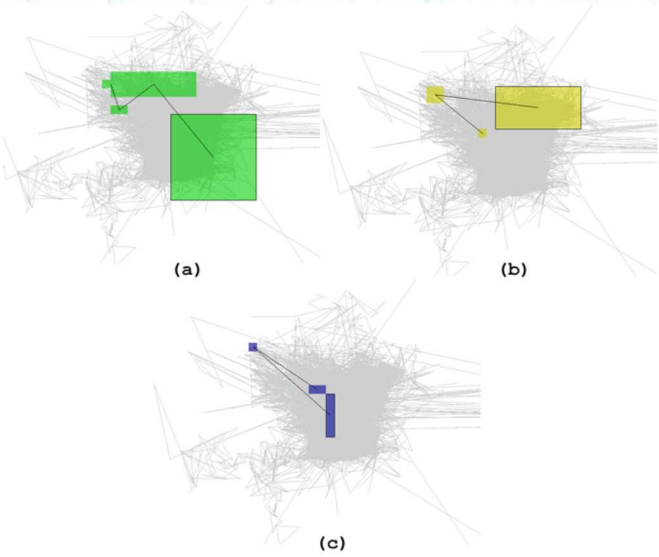

Figure 8: The T-Patterns discovered on the trajectories in a cluster

algorithms obtaining a deep understanding of the data.

The approach of combining the extracted patterns with the information available in the system can be also exploited to learn another type of model: the *Location Prediction* [4]. The location prediction algorithm aggregate the local patterns found in the previous step to produce a global model for the considered dataset: the model gives a high level description of the mobility allowing also to predict the possible destinations of an individual by observing his/her movements in the recent past.

3 Conclusions

The analysis capabilities of our system have been applied onto a massive real life GSM dataset and we demonstrated how the various methods and systems developed in the project support the creation of novel analytical services for mobility management, such as: i)Analysis of the movements in space and time ii) the automated construction of origin/destination matrices from mobility data in a timely, reliable and objective manner. It allows to analyze users's flows between geographical areas, overcoming the limitations of the current survey-based approach; iii) discovery of mobility patterns with different data mining tools which can be combined to go deep on the data understanding iv) the detailed analysis and discovery of systematic movement behaviors, i.e., the movements that repeat periodically during the week, with particular emphasis to commuting patterns like home-to-work and work-to-home.

References

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

[2] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In KDD, pages 330-339, 2007.

[3] G.Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni and Dino Pedreschi. A Visual Analytics Toolkit for Cluster-Based Classification of Mobility Data. SSTD, pag. 432-435(2009)

[4] Anna Monreale, Fabio Pinelli, Roberto Trasarti, Fosca Giannotti: WhereNext: a Location Predictor on Trajectory Pattern Mining.KDD 2009. 15th ACM SIGKDD Conference on Knoledge Discovery and Data Mining

[5] Riccardo Ortale, E. Ritacco, Nikos Pelekis, Roberto Trasarti, Gianni Costa, Fosca Giannotti, Giuseppe Manco, Chiara Renso, Yannis Theodoridis: The DAEDALUS Framework: Progressive Querying and Mining of Movement Data. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008).