Mining Mobility Behavior from Trajectory Data

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

KDD LAB – ISTI CNR and Univ. Pisa, Italy
{name}.{surname}@isti.cnr.it

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: location data from mobile phones, GPS tracks from mobile devices receiving geo-positions from satellites, etc. The GeoPKDD project, a large European research initiative, has studied how to discover useful knowledge about human movement behavior from mobility data, while preserving the privacy of the people under observation. A new exciting multi-disciplinary research area has thus started, at the crossroads of mobility, data mining, and privacy.

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, as well as discovering deeper knowledge about human mobility. 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 info mobility services
- Detecting changes in collective movement behavior.

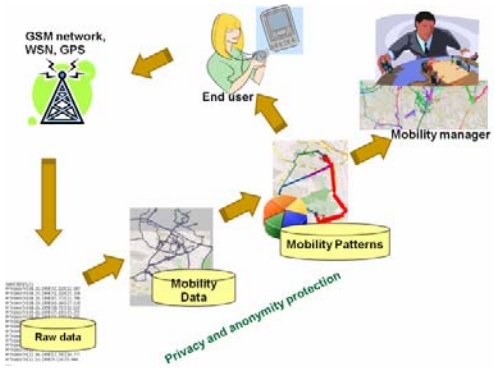
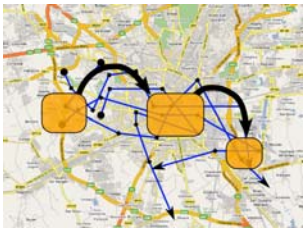


Fig. 1 The GeoPKDD process

Figure 1 represents the overall *Geographic Privacy-aware Knowledge Discovery process*; observe how the mobility data collected by the wireless networks are centralized into a trajectory database management system and warehouse – by a *trajectory* we mean a sequence of time-stamped locations, sampled from the itinerary of a moving object.

2. Mobility Data Mining

GeoPKDD created a toolkit of breakthrough analytical methods for mining from massive trajectory datasets, yielding a variety of mobility patterns and models.



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 [4].

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 mobile behavior, e.g., home-work-home commuting. Our density-based clustering algorithms discover T-clusters in trajectory data [5].

In order to master the complexity of the geographic knowledge discovery process, the various analytical and mining methods are integrated within a **semantic-based query, mining & reasoning system**, called Daedalus [3]:

- The *spatio-temporal query primitives* allow reconstructing trajectories from raw data, selecting and pre-processing trajectory data w.r.t. geographic background knowledge.
- The *trajectory mining primitives*, such as T-pattern and T-clustering, allow extracting and validating mobility patterns and models.
- A *reasoning component* allows to specify domain-driven ontologies, inferring types of trajectories and patterns [4].

3. Discovering Mobile Behavior from Vehicular GPS Tracks

We put to work the findings of GeoPKDD 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; the dataset contains more than 2 million observations, yielding 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:

- The automated construction of **origin/destination 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.
- 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.
- 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.

The novelty of this approach w.r.t. current practice is twofold: the use of real mobility data vs survey data, and the discovery of emergent patterns and models vs the application of preconceived traffic models from transportation engineering and operations research.

3.1. Vehicular GPS data vs survey data

An important aspect to be considered in this experiment is that both the sample population and the data collection method are different from those used by the Mobility Agency. First, the GPS dataset covers only vehicular movements, whereas surveys include all transportation means, including pedestrians. Second, the automatic collection procedure applied for GPS data ensures that all movements are correctly captured, whereas surveys leave space to omissions or distortions. Finally, GPS data provides no explicit semantic information about the purpose of movements, the final destination, and profiling of the citizens involved, whereas surveys explicitly collect them. Significant differences hold also for the mere size of the sample: ca. 17.000 vehicles for the GPS dataset, against ca. 45.000 vehicles and ca. 210.000 physical persons covered by surveys – although the number of GPS-equipped cars is continuously increasing. Concerning the periodicity of the sample, the difference is striking: near real time for GPS tracks vs years for the surveys.

In order to assess the coherence and statistical significance of the GPS dataset against survey data, the former was used to recompute a set of statistics published in the most recent periodic mobility report in Milano, which was based on survey data collected from 2005 and 2006. The results achieved confirmed the validity of GPS data, and more importantly showed that these data contain information about occasional mobility, which is known **not** to be covered by survey data. Using the preprocessing primitives integrated in the Daedalus system, the input dataset of trajectories was reduced to a set of trajectory segments, that describe the movement between two subsequent stops. Then, the movements of a single day of the week

(Monday, April 2nd, 2007) were selected, and the number of vehicles on the move at each hour of the day was counted. The result is shown in Fig. 2 (right), together with the corresponding results obtained by the Mobility Agency Fig. 2 (left).

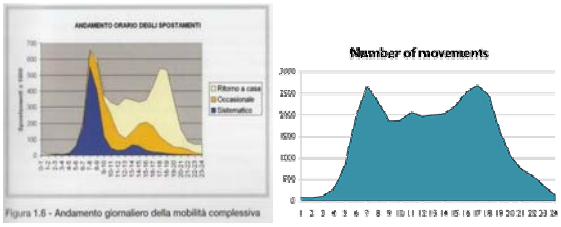


Fig. 2 Hourly distribution of movements in Milano: (left) estimated through surveys, and (right) estimated by means of GPS-based analysis

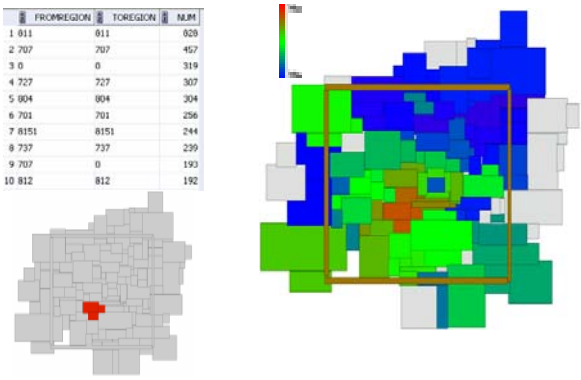
The two distributions match significantly, the most relevant difference being a slight overestimation provided by our analysis in the central and (to a minor extent) the later part of the day. Actually, the assessment with the Mobility Agency revealed not only that the results are coherent, but also that the survey distribution were known to underestimate the movements where the mismatch occurs. The explanation of this phenomenon is that GPS data also capture non systematic movements, while survey data don't, as interviewed people tend not to report their occasional mobility, such as going to the dentist or visiting a friend. Also, GPS data contain mobile activity of people that do not live in the greater metropolitan area, which is the subject of the survey.

3.2. Automated construction of the O/D Matrix

The basic step is to define the areas to be considered in the O/D matrix. The current level of detail adopted by Mobility Agency (612 small districts, Fig. 3) is considered too fine, since typical decisions are taken by reasoning on larger areas; however, the areas used in the analysis should be built by merging existing zones, in order to make the results comparable with current ones. For this reason, a smaller set of 63 regions was extracted by aggregating original districts, and approximated by simple rectangles. Finally, an extra region was introduced, that covers the border of the sensed area, in order to identify the movements that arrive from or go towards outside the metropolitan area.

Using the primitives provided by Daedalus, each movement was associated with the region of its origin and the region of its destination. Then, the number of movements that take place between each pair of regions is computed.

Fig. 3 Milano districts



	FROMREGION	TOREGION	NUM
1	811	811	620
2	707	707	457
3	0	0	319
4	727	727	307
5	804	804	304
6	701	701	256
7	8151	8151	244
8	737	737	239
9	707	0	193
10	812	812	192

Fig. 4 (top-left) Top fragment of O/D matrix for Sunday. (bottom-left) The selected origin to analyze. (right) Density of destinations from the fixed origin.

Fig. 4(top-left) shows the top 10 most frequent pairs of regions. As expected, most of them represent movement that start and end in the same regions. The third one, in particular, represents movements that arrive from outside the area covered by GPS data (the "border region", that has code 0) and end outside. By querying the trajectory dataset w.r.t. the regions, it is possible to identify the trajectories that contribute to specific projections of the O/D matrix. Fig. 4(right) illustrates the intensity of movements starting from a specific origin of interest, shown in Fig. 4(bottom-left). The intensity is represented on a color scale from blue to red. As a result, we can learn how people move from that particular area of the city: e.g., there is a strong flow towards outside the city, especially through the south-west gate (dark green), or the north-west gate (light green). Interestingly, in the north-east part, within the large blue area, there is a small dark green area: this can be an evidence of something interesting, that deserves further analysis to be unveiled.

3.3. Discovering systematic movements

A more challenging analysis task is to discover and analyze the *systematic* movements hidden in the dataset. Compared to previous analyses, here we show how combining querying and mining we can achieve a higher-level *semantic* characterization of movement. A *systematic* movement is defined by the Mobility Agency as a movement between two *stops,* which occurs *frequently* in the individual user's history. We need a finer granularity of space in this analysis, and therefore we use a grid to compute the stops and their associated regions of interest. The next step is to exploit the T-Pattern algorithm to compute systematic movements of each user as frequent sequences of regions of interest; this mining is performed separately on each user, with appropriate frequency thresholds. On top of these systematic movement patterns, we can now build a special O/D matrix (left part of Fig. 5), the *systematic O/D matrix*, which can be used to analyze the systematic traffic flows in the city. In the right part of Fig. 5 we show a view of the matrix by selecting a particular region as origin (the blue square) and highlighting all the destination regions in the matrix (the red squares). This analysis is useful in understanding how the systematic traffic affects the city, and how public transportation could be improved to better meet the needs of systematic mobility.

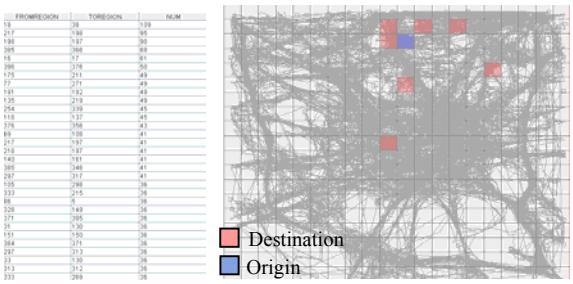


Fig. 5 (left) O/D matrix for systematic movements. (right) The destinations for a particular origin.

4. Concluding remarks

The experiment sketched in this paper demonstrates the analytical power of mobility data, made available as a by-product of telecom services, in unveiling the complexity of urban mobility. The prospects of providing travelers and planners with multimodal information about mobile behavior will, for the first time, integrate people's behavior information with transportation information. If such integration can be achieved based on the patterns observed from the real trajectories of people – and their relations between space, time, and activities – we will have the possibility of presenting to travelers and planners with comparable information on travel options across modalities. In order to make this vision real, we need solid analytical methods and systems, capable of mastering this complexity, summarizing trajectory data and extracting meaningful and simple knowledge. Our GeoPKDD project provided the framework and the basic building blocks, which include *prediction* and *anonymity*. The *WhereNext* method predicts the next location of a moving object given its past trajectory, based on the collective behavior modeled by T-Patterns [7]. Anonymity is based on the idea of *k*-anonymous trajectory dataset, 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 *k*-anonymous dataset, such that the key analytical properties are preserved, together with users' privacy [8].

Where is the frontier of this line of research? In a word, *interactions*. The increasing intelligence and context-sensitivity of mobile devices is producing location data tagged with ever greater semantic information, recording interactions with the surrounding context and with the other mobile peers – a whole new picture of social relations intertwined with mobile behavior. These richer data will enable deeper analytics, such as classification methods for inferring, e.g., the transportation means or the purpose associated to a trajectory. And mining these data, possibly combined with complex network analytics, will get us closer to an *archaeology of the present*.

References

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

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

[3] M. Baglioni, J. Macedo, C. Renso, R. Trasarti and M. Wachowicz. Towards semantic interpretation of movement behavior. AGILE 2009.

[4] R. Ortale, E. Ritacco, N. Pelekis, R. Trasarti, F. Giannotti, G. Manco, C. Renso, Y. Theodoridis. The DAEDALUS Framework: Progressive Querying and Mining of Movement Data. ACM GIS 2008.

[5] Mirco Nanni, Dino Pedreschi. Time-focused density-based clustering of trajectories of moving. J. of Intelligent Information Systems, 2006.

[6] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli. Trajectory Pattern Mining. In ACM Int. Conf. on Knowledge Discovery and Data Mining, SIGKDD 2007.

[7] R. Trasarti, A. Monreale, F. Pinelli, F. Giannotti. WhereNext: a Location Predictor based on Trajectory Pattern Mining. In ACM Int. Conf. on Knowledge Discovery and Data Mining, SIGKDD 2009.

[8] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects database. In 24th IEEE Int. Conf. on Data Eng., ICDE 2008.