# A Visual Analytics Toolkit for Cluster-Based Classification of Mobility Data

Gennady Andrienko[1], Natalia Andrienko[1], Salvatore Rinzivillo[2], Mirco Nanni[2], and Dino Pedreschi[3]

[1] Fraunhofer IAIS, Sankt Augustin, Germany
[2] ISTI - CNR, Pisa, Italy
[3] Università di Pisa, Pisa, Italy

**Abstract.** In this paper we propose a demo of a Visual Analytics Toolkit to cope with the complexity of analysing a large dataset of moving objects, in a step wise manner. We allow the user to sample a small subset of objects, that can be handled in main memory, and to perform the analysis on this small group by means of a density based clustering algorithm. The GUI is designed in order to exploit and facilitate the human interaction during this phase of the analysis, to select interesting clusters among the candidates. The selected groups are used to build a classifier that can be used to label other objects from the original dataset. The classifier can then be used to efficiently associate all objects in the database to clusters. The tool has been tested using a large set of GPS tracked cars.

## 1 Introduction

The technologies of mobile communications and ubiquitous computing pervade our society, and wireless networks sense the movement of people and vehicles through their location-aware devices, generating large volumes of mobility data of unprecedented quantity, quality and timeliness at a very low cost. However, raw mobility data, such as collections of GPS tracks, are very complex, as they represent rough approximations of complex human activities, and at the same time semantically poor. Therefore it is extremely challenging to develop analysis techniques capable of mastering the complexity of the data and extracting meaningful abstractions, in particular, by discovering and interpreting groups of people or vehicles that exhibit similar mobility behavior. The mentioned problem can be formulated as a trajectory clustering problem: find, for the spatial area and the time interval under analysis, the natural clusters of similar trajectories, together with an intuitive way of presenting the discovered clusters to a human analyst for interpretation, i.e. attaching semantics. A few trajectory clustering techniques have been proposed, but their direct application to realistic mobility datasets, such as the object of study in this paper, is simply unfeasible. Real mobility data are both computationally and analytically complex, and require the involvement of a human analyst with her background knowledge and understanding of the properties of space and time. We present here a visual analytics environment, which enables to progressively find and refine trajectory clusters

N. Mamoulis et al. (Eds.): SSTD 2009, LNCS 5644, pp. 432–435, 2009.

and associated representative prototypes; our experiments demonstrate that natural clusters are found in the data, which characterize movement behaviors at a suitable abstraction level, understandable by mobility managers.

## 2 Analysis Process Description

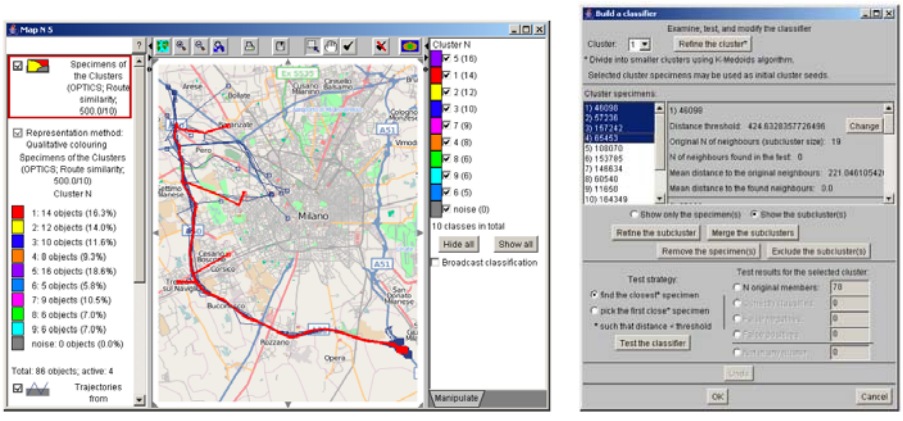
The tool we present here is designed to support a stepwise analysis process for a large trajectory dataset. The analyst can consider a small portion of the dataset, extracting all the interesting clusters from this sample, by means of a density based clustering algorithm, i.e. OPTICS [1]. The visual environment allows the user to validate, refine and revise the found clusters. For each cluster, the system computes and proposes a set of specimens (i.e. representants) that serves as a classifier of the whole cluster: any other new object belongs to that cluster (i.e. has a similar behavior) if it is close to one of these specimens.

Given a trajectory dataset $D$, the analytic process can be formalized as follow:

- Extract a sample $D'$ of trajectories from the database $D$
- Apply OPTICS with a suitable distance function $d$ [2] and get a set of density-based clusters $\{C_1, C_2, \ldots, C_m\}$
- For each cluster $C_i$
  - Select $s$ specimens in $C_i$, with $1 \leq s < |C_i|$, namely $\{c_{i1}^{\epsilon_1}, c_{i2}^{\epsilon_2}, \ldots, c_{is}^{\epsilon_s}\}$, such that the cluster $C_i$ may be described as the set of objects in $D'$ whose distance from one of the objects $c_{ij}^{\epsilon_j}$ is less than the threshold $\epsilon_j$, i.e. $C_i = \{c \in D' | \exists\ j\ s.t.\ d(c, c_{ij}^{\epsilon_j}) < \epsilon_j, j = 1, 2, \ldots, s\}$
- Visually inspect and refine the selected specimens. The set of the specimens for all clusters forms a classifier
- Apply the classifier to the remaining trajectories, attaching each new trajectory to the closest specimens. The trajectories with no close specimen remain unclassified
- Possibly, restart the whole process again for the unclassified trajectories.

## 3 Presentation of the Tool

The user interface of the tools consists of an operational window (Figure 1(b)) and a map window (Figure 1(a)). The two windows are linked, only the selection of the operational window is showed in the map window. The application automatically selects a set of specimens for each cluster and partitions the objects in each group according to the closest specimen. After that, the analyst can manipulate the specimens and, in parallel, her actions are reflected on the corresponding trajectories. The refinement actions available for each cluster are: *(1)* merging two or more sub-clusters, *(2)* removing a sub-cluster, *(3)* splitting a sub-cluster. To show the functionalities of the tool, consider the cluster in Figure 1(a). On the map the specimens and the trajectories are represented with different colors. The analyst can select one subset of specimens as the most representative for the cluster



(a) The map window (b) The operational window

**Fig. 1.** The user interface of the application

(a) The main representative specimen of the Cluster 1 (b) Two specimens of Cluster 1 (c) The specimens are extracted to create a new cluster (Cluster 10)

**Fig. 2.** Selection and merging of two specimens. Split of the cluster.

(Figure 2(a)). If other subclusters are visually inspected and valued as too dissimilar, it is possible to split the cluster. For example, the two specimens in Figure 2(b) are moved in a new cluster (Figure 2(c)) since they describe shorter paths. It is also possible to discard a sub-cluster from the analysis by tagging it as noise.

When the analyst concludes the refinement phase, she can use the resulting classifier to classify all the other trajectories of the original dataset. Figure 3(a) shows 745 trajectories from the database that have been attached to Cluster 1. Figure 3(b) represents these trajectories in a summarized form. For comparison, a summarized representation of Cluster 10 (133 trajectories) is shown in Figure 3(c).


(a) (b) (c)

**Fig. 3.** Here we show complete clusters based on select specimens: (a)745 trajectories of the 1st cluster, (b) their summarized representation, (c) and summarized representation of 133 trajectories in the 2nd cluster

## 4 Conclusions

We have presented a tool to tackle the problem of analyzing a large dataset of trajectories, where the size and the complexity of the data prevent the effective application of traditional data mining methods. The visual tool is based on a step wise user-driven approach, based on the reduction of the problem complexity through sampling and filtering. The proposed methodology is able to find natural clusters in a complex and large dataset, hence it is able to underline meaningful mobility behaviors. In addition to clusters, the method also produces a classifier that can be used for many purposes such as movement prediction, detection of abnormal behaviors etc. The scalability of the tool have been tested on a real dataset of GPS tracked cars (around 200.000 trajectories from 17.000 cars equipped with an on-board GPS receiver, collected during 7 days in the city of Milan, Italy, resulting in more than 2,000,000 irregularly sampled positions).

## 5 Nature of the Demonstration

The demo will present the functionalities provided by the tool. Depending on the availability of newtwork connection, it is possible to test the application both on the online dataset and on a local dump of the data. We will show the steps followed to analyze a real dataset of GPS tracked cars using the visual analytic environment.

## References

1. Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., Andrienko, G.: Visually driven analysis of movement data by progressive clustering. Information Visualization 7(3-4), 225–239 (2008)
2. Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G.L., Theodoridis, Y.: Similarity search in trajectory databases. In: TIME, pp. 129–140 (2007)