BibTeX:

Relating mobility patterns to socio-demographic profiles

Thomas Liebig
TU Dortmund University
Lehrstuhl 8,
Dortmund, Germany
thomas.liebig@tu-dortmund.de

Abstract

Gathering knowledge on mobility patterns and their distribution among the population provides an important input for agent-based mobility simulation. This work proposes a novel approach to link socio-demographic attributes to mobility patterns given a representative data sample. In a first step individual stops are detected. They will be aggregated and linked to points of interest. Among them sequences can be detected and related to socio-demographic features extracted by subgroup discovery. The overall process is applied on a cyclists data set.
Keywords: mobility diary mining, agent-based simulation

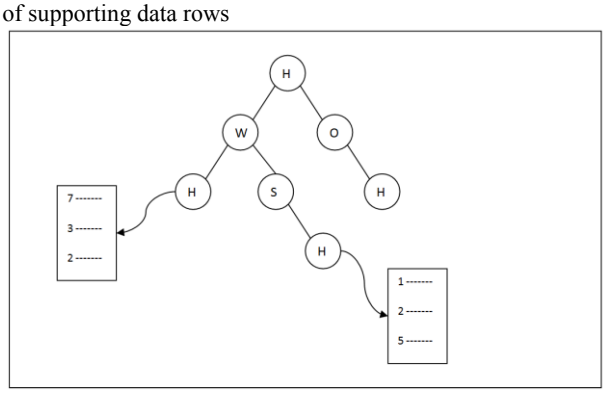
1 Introduction

The need to explore the relationship between persons' activities and their socio-demographic profile appeared in several fields, recently. Foremost in the design of agent-based simulation models (ABM) the task received latest attention [1], since it helps populating the artificial worlds with realistic behaving agents and for mode of transportation decision planning of the simulated agents. But also city authorities have strong interest in knowing how people interact with the urban environment.

Our approach to achieve these valuable insights is to mine the relationship among persons' activities and their socio-demographic profile from their daily movements. This assumes that there is a way to monitor these movements in advance representatively. Having both datasets, i.e. users' movements and associated demographic information, relationships between those two pieces of information can be extracted.

In order to extract such relationships from the raw sensor readings there several steps are required. Given the input data as raw GPS traces, (1) the initial step is data cleaning and preparation, as raw data contains artifacts. Afterwards (2), stay points of the users need to be recovered from these traces. A stay point is a location within a defined radius (for example 300 meters) where a user spends more time than a given threshold (for example 30 minutes). Because of the nature of GPS traces, no two stay points could have the same geographical coordinates. For this reason stay points are clustered together (3) and the averaged coordinates are taken into account. That is why after detecting the stay points a clustering algorithm is used. (4) After the clustering step each cluster is mapped to a possible point of interest, cross mapping the cluster with a POI database (Open Street Maps POIs). (5) Next is to produce a set of stay points which are labelled either as living place (H) or working place (W) or as a Point of interest. (6) The next step is to extract or chunk the stay points set to sequences of labelled movements for example H-W-H or H-W-Shopping-H. (7) Using such sequences for several users, an extended Frequent Pattern Tree [2] (FP-tree) data structure can be populated.

Figure 1: Extended Frequent-Pattern Tree. Attached to the vertices which hold atoms of the movement patterns are lists of supporting data rows



At each node of the tree there is an associated list of data row identifiers (respectively user identifiers, linking the pattern to a subset of the users). Having this structure in place, (8) subgroup-discovery will be applied on the socio-demographic data of the subsets of the users at certain nodes in the FP-tree (Figure 1). This allows us to produce correlations between the users' movement patterns and their socio-demographic data.

2 Preprocessing

As the provided dataset already contains the intermediate results of step (6), we will focus mostly on the remaining steps. For convenience, we sketch the process how to get to this step from raw GPS data within the following sections.

2.1 Stay point Detection

For the stay point detection algorithm it is possible to apply the algorithm used in [3]. The input for the algorithm is a sequence of GPS traces. With each point having the longitude, latitude and a timestamp. The GPS points are connected according to the timestamp sequence. A stay point is defined as a geographical region where a user stays for a while. The geographical region is usually a circle with a specified

diameter as a parameter. The temporal duration of a stay point is also required as a parameter to the stay point detection algorithm.

2.2 Stay point Clustering and Labeling

As mentioned in the introduction section, a clustering algorithm is required to cluster the stay points. In the next step these clusters will be matched to points of interest (POIs) or become labelled as the user's home or work place. A possible algorithm for clustering is the DBSCAN algorithm, described in [4].

2.3 Sequence Generation

Using the previously generated stay points, the raw GPS data needs to be reduced to the sequence of visited locations, e.g. home-work-home.

3 Relating Mobility Patterns to Socio-Demographic profiles

The data provided by the CDC2013 already contains sequences of visited locations for 79 cyclists. Though the creation of these so-called travel diaries did not follow the process described in previous sections, but resulted by manual user annotation, we may use it directly for subsequent analysis.

For further processing, we create a binary matrix holding a single line for each user. The different travel purposes of the users become the columns. Every travel purpose which was ever indicated in a user's travel diary receives the value TRUE whereas the others are FALSE.

First, we focus on the frequent sets among this matrix. These are the sets of items (travel purposes) which co-occur together for a given percentage of the users. We mine them using FP-Growth algorithm, we set the required minimum support threshold to 0.25. This indicates that one fourth of all data rows, respectively users must fulfill the resulting patterns. Every vertex of the used data structure holds candidates of the supporting data rows. This allows easy lookup of user attributes in subsequent analysis. An excerpt of the results is listed in Table 1.

Table1: Frequent Sets with support and their description, for description of the trip purposes see Table 2.

Trip Purpose	Support	Description
9,1	69	To work and to home
9,10	37	To home and to Other
1,10	30	To work and to Other
9,1,10	30	To work and to home and to other
9,4	29	To Home and To Eat
9,1,4	23	To Work and to Eat and to Home,

While these frequent sets already enable the sampling of movement patterns for a population of agents in a multi-agent system, previous knowledge on the inhabitants (usually provided by census) may not be incorporated. Thus it is necessary to determine for every frequent set the subgroup of the population which mostly uses this pattern.

Table 2: Description of Trip purposes (CDC2013)

Trip purpose	Description
1	To work
2	To visit (friends, etc);
3	To work related task;
4	To Food shopping;
5	To Non-food shopping;
6	To School (Student);
7	To Entertainment;
8	To Eat (Lunch, etc);
9	To Home;
10	Other (any other not mentioned)]

The CDC2013 dataset provides attributes to the cyclists consisting of gender, age, salary, employment and experience of cycling. Despite the last one, the attributes are also contained in typical census data. Thus, we remove the experience of cycling attribute from the dataset and use the remaining attributes. The analysis is called subgroup analysis [5]. For each previously computed frequent set the supporting users are identified according to the initial data matrix The extended FP-tree supports this step. For the retrieved users an additional label attribute is set to TRUE and for the other users to FALSE. Subgroup discovery among the socio-demographic attributes searches for a description of the data rows with a TRUE label. An excerpt of the subgroups (with highest accuracy) is listed in Table 3.

Table 3: Subgroups with their conclusion and Frequent Sets

Pattern	Subgroup	Conclusion
9,1	27-30 years=false and FullTimeEducation=false and Employment_Other=false and SelfEmployed=false and Income Low=false	True
9,10	23-26 years=false and 35-38 years=false and 55-58 years=false and EmployedPartTime=false and Employment Other=false	False
1,10 and 9,1,10	27-30 years=false and 47-50 years=false and 51-54 years=false and SelfEmployed=false and Income Medium=false	True
9,4	Single=false and Health Fair=false and Employment Other=false and SelfEmployed=false and Income Low=false	False
9,1,4	43-46 years=false and Married=false and Health-Very Good=false and FullTimeEducation=false	True

Note, that some of the subgroups do not have positive conclusions but negative ones. These are describing the

condition for not belonging to the group where the pattern holds.

4 Discussion

In this work we outlined a method to link mobility patterns and socio-demographic attributes. Our approach bases on an extended FP-Growth algorithm and a subsequent subgroup analysis.

The proposed method was successfully applied to the CDC2013 data set, more sophisticated performance analysis are required. In a multi-agent-based traffic simulation the results can be directly applied for mode of transportation decisions for the agents. For simulators focusing on cyclists our approach supports decisions for travel purposes.

Incorporation spatio-temporal subgroups are subject to future analysis.

Acknowledgements

This work was supported by the European FP7-INSIGHT project. We thank the organizers of CDC2013. Additionally, we thank Meena Shehata for fruitful discussion.

References

[1] T. Bellemans, S. Bothe, S. Cho, F. Giannotti,D. Janssens, L. Knapen, C. Körner, M. May, M. Nanni, D. Pedreschi, H. Stange, R. Trasarti, A.-U.-H. Yasar, and G. Wets. An agent-based model to evaluate carpooling at large manufacturing plants. Procedia Computer Science, 10(0):1221-1227, 2012.

[2] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Min. Knowl. Discov., 8(1):53-87, 2004.

[3] Y. Ye, Y. Zheng, Y. Chen, J. Feng and X. Xie. Mining Individual Life Pattern Based on Location History. Mobile Data Management. 1-10. 2009

[4] M. Ester and H. Kriegel and J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. Second International Conference on Knowledge Discovery and Data Mining. 226-231. Press. 1996.

[5] S. Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In Proc. 1st European Symposion on Principles of Data Mining and Knowledge Discovery, 78-87, 1997.