Chapter 69
The Annotation of Global Positioning System (GPS) Data with Activity Purposes Using Multiple Machine Learning Algorithms

Sofie Reumers
Hasselt University, Belgium

Davy Janssens
Hasselt University, Belgium

Feng Liu
Hasselt University, Belgium

Geert Wets
Hasselt University, Belgium

ABSTRACT

The aim of this chapter is to evaluate whether GPS data can be annotated or semantically enriched with different activity categories, allowing GPS data to be used in the future in simulation systems. The data in the study stems from a paper-and-pencil activity-travel diary survey and a corresponding survey in which GPS-enabled Personal Digital Assistants (PDAs) were used. A set of new approaches, which are all independent of additional sensor data and map information, thus significantly reducing additional costs and making the set of techniques relatively easily transferable to other regions, are proposed. Furthermore, this chapter makes a detailed comparison of different machine learning algorithms to semantically enrich GPS data with activity type information.

INTRODUCTION

Current simulation systems in social sciences and the transportation area are based on either traditional surveys or on full (activity) diaries to model the individual behavior of the agent in the system. Collecting these data either in paper-and-pencil format or by means of computer-aided technology such as small hand-held computers is a demanding and burdensome task for respondents. The reason for this is that data about the principal choice dimensions underlying the simulation model has to be collected. Indeed,

DOI: 10.4018/978-1-5225-1759-7.ch069

more advanced simulation models in the transportation field, such as the activity-based transportation model, fully reflect spatial and temporal constraints and opportunities. They model interactions among agents, capture time use and allocation behavior, and consider activity participation along a continuous time dimension (Pendyala, 2005). The spatial and temporal dimensions require special attention because previous research suggests that there is a bias in reporting these dimensions and one may get a fake idea of precision: while a location and time record may be available in the data, the level of precision remains unknown.

In this study, it is our objective to evaluate whether GPS data can be annotated or semantically enriched with different activity categories, such that this source of data can be used in the future in simulation systems. The basic objective of the semantic enrichment process is to introduce a domain-dependent characterization of movement data and patterns, namely a method for activity characterization of GPS movement data.

There is a rich literature, originating from the artificial intelligence field, on developing methods that try to automatically infer people's behavior. The problem with most of these approaches is twofold; (1) the lack of validation and (2) the spatial transferability. The first problem stems from the fact that one must have both diary and GPS data for the same set of respondents. However, if available, both data can be used: the GPS traces for annotation and the diary data as an independent validation source, i.e. as ground truth. The data used in this study stems from a mixed-mode survey design, in which two types of data collection methods were used, namely a paper-and-pencil activity-travel diary survey and a corresponding survey in which GPS-enabled PDAs were used. The second problem stems from the fact that geographic information (e.g. points of interest) is often used during the annotation process, which limits transferability. In our study, we propose a set of new approaches, which are all independent of additional sensor data and map information, thus significantly reducing additional data collection costs and making the set of techniques relatively easily transferable to other regions.

A group of state-of-the-art machine learning methods, including Multiclass Support Vector Machine (SVM), Multinomial Logistic Regression (MNL), Decision Tree (DT), and Random Forest (RF) are adopted in this study. The chapter makes a detailed comparison of the different learning algorithms at the level of the training data, the validation data and the test data.

The chapter is divided into seven sections. The first section thoroughly reviews the literature on the semantic enrichment of travel data. In the second section of this chapter, the data collection method, a description of the data and the data pre-processing step are briefly discussed. The research methodology is presented in the third section. The following section presents the prediction results. Finally, the fifth and sixth section, respectively, offer a discussion on the prediction results, and suggest some conclusions and future research efforts.

LITERATURE

This section begins by reviewing the literature concerning traditional travel survey methods, and current technological advancements in travel data collection with a particular focus on GPS-based surveys. In the second part of the literature review, we take a closer look at studies in which machine learning algorithms and artificial intelligence techniques are applied to automatically infer activity and travel attributes from raw GPS data.