Fusion of Semantic Data for Mobility Information Prospects and Opportunities

Assaf Biderman

SENSEable City Lab
Massachusetts Institute of Technology
MIT
USA
abider@mit.edu

Francisco C. Pereira, Ana Alves

Centro de Informática e Sistemas da
Universidade de Coimbra
CISUC
Portugal
camara@dei.uc.pt, ana@dei.uc.pt

1 Introduction

In this position paper, we argue for the opportunities in combining data at the semantic level with other types of data for extracting information related to mobility. Specifically, we are interested in combining statistical data which describes mobility in the city (e.g. the number of vehicles in a specific street per unit of time) and natural language-based information about activities in the city (e.g. RSS feeds about events and user generated content about an individual's activity). We intend to explore the opportunities and difficulties in performing this work and discuss its possible applications. The paper is organized in three sections: we start by familiarizing the reader with the main concepts involved; we propose our core ideas and point to a number of applications; we then end by illustrating the main challenges facing this work.

2 Assumptions

There is a wide variety of research opportunities in developing methods and applications that rely on fusion between statistical mobility information and semantically represented data sources. By mobility information we refer to any kind of data that directly represents the movement of people or goods. This includes GPS and cell phone traces, cell tower aggregate calling statistics (e.g. Erlang measures, number of calls, handover data, etc.), road sensor data (inductive loops, cameras, and toll collection), public transport ticket sale information, or even parking space availability. We refer to semantics as the study of the meaning in language – a definition borrowed from computational linguistics, which is different from the computer science definition in which we speak of semantics of computer programs or functions. The definition we use refers to interpreting words, symbols, and their constructions (e.g. phrases) in terms of concepts, and these concepts are represented by reference words and symbols that are well defined within a context of communication. For example, the semantics information about "Starbucks Coffee" could include the words "coffee," "espresso," and also "coffeehouse chain," "from Seattle," or "good in the morning".

Data fusion is the integration of two or more distinct sources of data into a single representation. These include data sources that differ in format, type, level of detail, or level of representation. The result of data fusion is united information. For example, the combination of cell phone call handover data with data about the signaling of phones on a cell network to produce estimates of the speed of movement of mobile phones [1]). Aside from literature on modeling, the majority of literature on data fusion is centered on multi-sensor data fusion (e.g. [2,4]). More specifically, it focuses on the integration of distinct low-level signals into a united result. For example, applying Kalman filters to estimate the position of an object using information from a GPS receiver and an accelerometer. Substantial research efforts have been directed in recent years

towards investigating such problems, and commercial products for data fusion of certain types of low-level signals are widely available. There has also been considerable work dedicated to data fusion of information with higher levels of representation, such as the semantic level. For example, various methods have been developed to combine different classifiers for document categorization [5].

We see a potential contribution in the intersection of these two worlds – the integration of semantic data with data of a lower level of representation, such as signal from a sensor. While it may appear an odd combination, our intention in this paper is to describe its potential value.

3 Fusion of statistical Mobility Data with Semantics Data

Typical statistical data about mobility includes information about time, position, and some measure of intensity (e.g. number of vehicles). These three dimensions of information can be combined, individually or collectively, with data that is represented semantically. First, we can associate semantic tags with locations (e.g. position X, Y is associated with Starbucks, Coffee, and John's work), thus adding a semantic representation to a location. We can also analyze more complex semantic information such as geo-tagged user–generated content (e.g. twitter, or other micro-blog entries) which includes information related to mobility, such as "I am stuck in traffic," or "I am at a great street fair." When we add information about intensity (e.g. congestion area, high levels of cell-phone usage, etc.) and analyze the correlation between data types over time, we can produce measures that describe demand for mobility. For example, the concepts "shopping area," "downtown," and "business district" could be possible associations to high intensity of traffic. If we add a time dimension, we can begin to study specific events in correlation with mobility in the city. For example, a sudden congestion can be associated with tags such as "soccer match," "championship finals," or "terrorist attack". Temporal alignment can be made possible with access to carefully selected information sources such as RSS feeds, which have temporal tags and are generally categorized by information type (e.g. traffic, sports, news headlines, etc.). The integration of these data types can ultimately be applicable for making predictions. For example, if a system finds correlation between specific crowded events and congestion, it can trigger a warning when RSS feeds inform that such events are approaching and predict the levels of congestion in the time surrounding those events. This type of data fusion could also by applied to other traffic simulation platforms (e.g. DynaMIT[6]) as an external probabilistic indicator. Another application could be automatic rating of events based on their predictable effect on the city (event X attracts a crowd effect of size N and produces congestion of magnitude S at streets A, B, C).

4 Challenges

Several main challenges are associated with the type of data fusion proposed above. First, natural language processing requires significant precision and computational efficiency. Natural language is often ambiguous to humans and, therefore, is particularly challenging for use in automated computer applications. Obtaining high-quality semantic data is another challenge. Large amounts of information about events, businesses, and commercial activities are presented on the internet. Also, users provide increasing amounts of information about their individual activities and make it available to the public. However, it is hard to determine the quality of such data, or to guarantee stable quantities of data for continuous analysis. Probably, the main challenge lies in determining the real context associated with a data entry of user-generated content: what is being described? Where does it take place? When? Who provided the data? Etc. Determining these involves, amongst other methods, applying information extraction techniques such as regular expressions; annotation tables, as in GATE ; or wrapper generation, as in [7]). For mobility analysis, the main challenge is to extract relevant patterns. Congestion is the most straightforward one, but other patterns may be present in the data such as: what is the regular flow of people every Thursday night between place X to place Y?

References

1. Fontaine, M., Smith, B.: Investigation of the performance of wireless location. J. of Transportation Engineering 133(3) (2007) 157{165

2. Mitchell, H.B.: Multi-Sensor Data Fusion: An Introduction. Springer (2007)

3. Wald, L.: Data Fusion Deffinition and Architectures. Presses de l'Ecole des Mines de Paris (2003)

4. Klein, L.A.: Sensor and Data Fusion: A tool for information assessment and decision making. SPIE Press Monograph (2004)

5. Hull, D., Pedersen, J., Shultze, H.: Method combination for document filtering. In: Proc. of the Annual International ACM SIGIR. (1996)

6. M Ben-Akiva, et al: Dynamit: a simulation-based system for traffic prediction. In: Proceedings of DACCORD Short-term forecasting workshop. (1998)

7. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications -Text Retrieval, Extraction and Categorization. John Benjamins Pub. Comp.