Veracity assessment framework for discovering social activities in urban big datasets

- Master internship -

Philipp Brandt, SciencesPo
Soror Sahri, Université de Paris, LIPADE, diNo

Motivation

Digital technologies provide access datasets that have been unfamiliar to social scientists, including behavioral traces (e.g., point of sales, geolocation data, social media scrapings, CCTV recordings), machine-readable texts, and code and data repositories. These secondary data sources produced without research goals in mind require new technical skills and computing capacities to manage their scale and content. A particular recent trend for social scientists is to understand the potential of big data in complementing traditional research methods and their value in making decisions. Several major issues have to be closely investigated around big data in social sciences, including political polarization, viral information diffusion, and economic performance. The veracity and value characteristics of big data are the main concerns for social scientists [1].

This master internship will focus on urban data, particularly the NYC taxi dataset, to develop technical procedures that help social scientists deal with this and similar urban datasets. Social scientists have used the NYC dataset in the past and yet left many dimensions unexplored. Most problematically, they have not yet provided a technology that allows for fast, flexible data access and a strategy for ensuring the quality of the data. Once such an infrastructure is in place, the NYC taxi dataset can lead to better understanding of core questions in the social sciences, such as economic decision-making and labor mobility, as well as a strategy for how social scientists can work with novel datasets.

Proposed work and implementation

Background and state-of-the-art

The main contribution of our work to the state-of-the art, a veracity assessment model with approaches that correlate the data veracity to their various business queries without repairing data. Existing work studying the quality of urban datasets investigate quality methods focusing only on data cleansing by detecting errors and repairing them, particularly for urban datasets [2]. Data quality, and then veracity, is mostly focusing on data cleansing that undertake a data repair action to remove errors from data sources. However, in real-world applications, cleansing the data is costly, and may lead to a loss of potentially useful data. In some work, data quality mainly measured based on the consistency metric, is ensured by consistent query answering without modifying sources [3]. This was only investigated for relational data and remains challenging for big datasets.

In this work, we would study data quality issues in urban big datasets by considering all of data inconsistencies, data inaccuracies, and data incompletenesses. The social science problems with data are relevant to define appropriate metrics to characterize and measure veracity depending on the

application domain, and investigate veracity approaches without repairing data. The availability of veracity metrics is critical for social scientists, who find themselves increasingly exposed to data formats and scales that they have not worked with previously and therefore lack the technical skills and expertise.

Data description

The NYC taxi contains over a billion of individual taxi trips in the city of New York from January 2009. Each individual trip record contains precise location coordinates for where the trip started and ended (spatial data), timestamps for when the trip started and ended (temporal data), plus a few other attributes including fare amount, payment method, and distance traveled. A subset of this data for 2013 also includes anonymized driver and medallion (vehicle) identifiers.

Internship work plan

The intern will start by studying the NYC taxi dataset, evaluate its veracity metrics (ex. the accuracy of spatial data, the completeness of data and the consistency), and identify interesting workloads, in connection with the social scientist needs from SciencesPo. For each workload (or query), the focus will be put on whether veracity metrics of the whole data impact the veracity of query answers or not. Based on our model proposed in [4], the intern will adapt the veracity score calculus and the veracity assessment approaches to the identified social scientist workloads, such as those proposed in [5]. Finally, the intern will evaluate the effectiveness and efficiency of the veracity assessment approaches on each workload. This step will be performed in connection with the social scientists to make fundamental progress in understanding labor supply decisions and labor mobility with implications for the rising gig economy.

Desired background

We are looking for a student in Master 2 or engineering school in computer science. The ideal candidate would have excellent programming skills, good knowledge in data management, and an interest in handling large amount of data.

How to apply

Apply by emailing your detailed CV (including course marks) to Soror Sahri:
soror.sahri@parisdescartes.fr

This internship is supervised by Prof. Philipp Brandt (SciencesPo) and Prof. Soror Sahri (Université de Paris). The internship will last 6 months, and is fully funded. The selected intern will become a member of diNo team of LIPADE at Université de Paris.

References

[1] Abiteboul, S., Dong, X.L., Etzioni, O., Srivastava, D., Weikum, G., Stoyanovich, J., Suchanek, F.M.: The elephant in the room: getting value from big data. In: Proceedings of the 18th International Workshop on Web and Databases, Melbourne, VIC, Australia, May 31, 2015. 1-5.
[2] Freire, F., Bessa, A., Chirigati, F., Vo, H., Zhao, K.: Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips. IEEE Data Eng. Bull. 39(2): 63-77 (2016)
[3] Bertossi, L.: Consistent query answering in databases. SIGMOD Rec. 35(2), 6876 (2006).
[4] R. Moussa, S. Sahri. Customized Eager-Lazy Data Cleansing for Satisfactory Big Data Veracity. IDEAS: 25th International Database Engineering & Applications Symposium. July 2021, pp 157-165.
[5] Brandt, Philipp, and Stefan Timmermans. "Abductive Logic of Inquiry for Quantitative Research in the Digital Age." Sociological Science 8 (2021): 191-210.