Veracity assessment framework for discovering social activities in urban big datasets

Soror Sahri, Université de Paris
Philipp Brandt, SciencesPo

Motivations

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, economic performance, etc. 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. can work with novel datasets.

Proposed solution

We propose a veracity assessment model with approaches that correlate the data veracity to their various business queries without repairing data. Our proposed model will tie quality of big datasets and quality of their query resultsets, and then end the disconnect between data source and data use. All of data inconsistencies, data inaccuracies, and data incompletenesses will be considered with propagating a fitness for use score of data.

Dataset 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.

Problem statement

The disconnect between data source and data use is one of the prime reasons behind the data quality issues. Moreover, there is still no consensus on what the notion of veracity is, and consequently no standard approach for measurement of veracity.

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.

Related Work

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]. Indeed, 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 is ensured by consistent query answering without modifying sources [3]. This was only investigated for relational data and remains challenging for big datasets.

#VERACITY

#Data Quality

#Big Data

Results

The expected results of this work will be a veracity model based on our model proposed in [4], with a veracity score calculus and veracity assessment approaches to the identified social scientist workloads, such as those proposed in [5]. The evaluation of the effectiveness and the efficiency, of the veracity assessment approaches on each workload, 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.

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, 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 2021: 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.

This work is supported by diiP, IdEx Université de Paris, ANR-18-IDEX-0001