Causal Graphs to Predict and Explain Air Pollution Patterns Using Citizen-contributed Odor Reports and Environmental Sensing Data

Advisors

- Angela Bonifati (Lyon 1 Univ., angela.bonifati@univ-lyon1.fr)
- Andrea Mauri (Lyon 1 Univ., andrea.mauri@univ-lyon1.fr)

This project will be done in collaboration with Dr. Yen-Chia Hsu (http://yenchiah.me), Assistant Professor in the Multimedia Analytics Group (https://multix.io) in the Informatics Institute at the University of Amsterdam.

Context

Air pollution is a problem in many urban areas that impacts many local people's living quality. For example, the steel industry in Pittsburgh (Pennsylvania, USA) has created many fugitive emissions that are harmful to residents' health. Nowadays, many air quality sensing stations are installed in cities to monitor the flow of air pollutants. However, it is still challenging to precisely understand the impact of pollution sources in a local region, as the local geography and climate conditions can be complicated, and there can be many pollution sources in the local region.

Objective

In this project, you will work with the Smell Pittsburgh dataset. Smell Pittsburgh is a mobile application for citizens to report bad smells in the city. The dataset contains nearly 70000 smell reports (contributed by local citizens in Pittsburgh) from Oct 2016 to Jan 2022. The smell reports have geographical locations, severity rating (from 1 to 5), description of the smell, and accompanying symptoms. The dataset also contains environmental sensing data from 12 monitoring stations involving different air pollutants, such as particulate matter, sulfur dioxide, ozone, carbon monoxide, wind information, etc.
You can explore the smell reports and particulate matter sensor reading data on the following page.
A preliminary data analysis and the background of the Smell Pittsburgh work can be found at "https://smellpgh.org/analysis" and also in the following paper [1].

In this project, we are interested in using methods based on causal graphs to infer the relation between the different factors (such as, wind direction, smell, air quality, air pollutants, etc..) to develop automatic or semi-automatic data analysis pipeline to discover and explain multiple underlying pollution patterns in the dataset (i.e., how an air pollutant moves in the urban environment and what influences its movement). While the local community already has hypotheses in mind about air pollution patterns, verifying its validity it's a difficult task.

Tasks

- Familiarize yourself with the data (and the work behind it [1]), the preliminary analysis, and the concepts of causal/probabilistic graphs [2,4]. Also, understand what kind of information is needed by the community [1,3]
- Design a graph data model and a set of queries to provide that information.
- Develop algorithms to transform the Smell Pittsburgh dataset in the data model you defined in Neo4J (or Memgraph).
- Evaluate the data model by running the queries.

Requirements

- Good programming skills (Python)
- Familiarity with graph structure and graph database (Neo4j, Memgraph)
- Familiarity with multiple types of time-series data (qualitative reports from citizens, and environmental sensing data).
- An interest in environmental science is a plus.

Reference

[1] Yen-Chia Hsu, Jennifer Cross, Paul Dille, Michael Tasota, Beatrice Dias, Randy Sargent, Ting-Hao Huang, and Illah Nourbakhsh. 2020. Smell Pittsburgh: Engaging Community Citizen Science for Air Quality. ACM Transactions on Interactive Intelligent Systems.

[2] Sakr, S., Bonifati, A., Voigt, H., Iosup, A., Ammar, K., Angles, R., ... & Yoneki, E. (2021). The future is big graphs: a community view on graph processing systems. Communications of the ACM, 64(9), 62-71.

[3] https://plumepgh.org/?date=2023-06-11

[4] Pearl, J. (2000). Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 19(2), 3.