DEPARTAMENTO DE ELETRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA
UNIVERSIDADE DE AVEIRO

EXPLORAÇÃO DE DADOS / DATA MINING

(Part I) Homework assignment: experimental work

The aim of this assignment is to perform an experimental work, to validate an idea ('research' question), in opposition to the previous assignment that focused mainly on the automation of tasks in conventional business information systems.

Datsets

- T-Drive is a smart driving direction service based on GPS trajectories of taxis in Beijing. The T-Drive trajectory dataset[1] is a sample that contains the trajectories of 10 357 taxis captured during one week. The sampling rate is variable (the range is 5 to 90 seconds, in most cases). Despite the volume of data captured, one of the main issues when using this dataset for data analytics is data sparsity.

- Open Street Map (OSM) is a world map free to use under an open license[2]. Geofabrik[3] is server that allows downloading OSM data by continent and country. This server is normally updated every day. In this assignment, we will provide a clipping of the map of china centered at the city of Beijing. We will also provide a map with the road intersections and the decomposition of the streets into segments (each segment gives a path between two intersections).

- GPS data accuracy depends on several factors, including satellite geometry, signal blockage, atmospheric conditions, and receiver design features/quality. This means that most GPS coordinates in the T-Drive dataset do not intersect any road. So, if you want to create statistics such as, the number of cars in a road segment, it is necessary to implement a map-matching procedure. In this assignment, you will also have a data sample with the GPS location adjusted to the road segments.

Formulate a question (hypothesis)

You must formulate a question that you think it would be interesting to answer using the data described above, e.g.:

- How many taxis will be circulating on a given road segment within 10 minutes?

- What is the fastest route between a source and a destination based on historical data (rather than using the speed-constraint-based approach)?

- Should we aggregate the data to decrease sparsity?

- Did weather conditions influence the flow of traffic? (To answer this question, you need to look for weather data in Beijin during a specific week in 2008)

[1] https://www.microsoft.com/en-us/research/publication/t-drive-trajectory-data-sample/

[2] https://www.openstreetmap.org/#map=6/39.602/-7.839

[3] http://download.geofabrik.de/

Validate your hypothesis

- Design a multidimensional data model for road traffic data analytics focused on your ‘research’ question.

- Create Data Mart and implement the procedures to study your data. No need to implement a nice user interface, unless you consider that this is the focus of your work (solution). In most cases, text outputs and charts will be fine to demonstrate your achievements.

- Results do not need to be complete or definitive, but it is important that you are able to formulate a problem and to implement a methodology to study that problem and to discuss the results.

Next steps

Tasks to be completed until October 24, 2019:

- ‘Research’ question (carefully written).
- Choose a paper or your own methodology to develop your work. Use ACM DL (https://dl.acm.org/), IEEE xplore (https://ieeexplore.ieee.org/) or Science Direct (https://www.sciencedirect.com/).
- Draft version of the multidimensional data model.

Deliverables

- You must prepare a detailed presentation (Powerpoint or similar) explaining your ‘research’ question (you can add notes to your slides), the design of your data model, the methodology and the procedures implemented to solve the problem and an in-depth discussion of obtained results.
- Include all code and packages, and a readme file with the instructions to run your work.

The workload is expected to be 15 to 20 working hours per student. After the paper submission, there will be a presentation to demonstrate the results.

Related publications

Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. Driving with knowledge from the physical world. In The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD’11, New York, NY, USA, 2011. ACM.

Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. T-drive: driving directions based on taxi trajectories. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS ’10, pages 99-108, New York, NY, USA,2010. ACM.

See also:

- https://www.microsoft.com/en-us/research/project/t-drive-driving-directions-based-on-taxi-traces/#!publications
- https://dl.acm.org/citation.cfm?id=1869807