IoT Trace Analysis

Lucas Colantuono
INSA Lyon
lucas.colantuono@insa-lyon.fr

Lucas Kummer
INSA Lyon
lucas.kummer@insa-lyon.fr

Shanan Lynch
INSA Lyon
shanan.lynch@insa-lyon.fr

Samuel Sedlmeir
INSA Lyon
S.Sedlmeir@campus.lmu.de

Contents

I Introduction 1

II Related work 1

III Data consistency 2
III-A Geographical consistency 2
III-B Gap filtering 3

IV Our Metric 3

V Future work / Open issues 3

VI Summary 3

References 4

***Abstract*—**

I. Introduction

II. Related work

There are several different articles, that analyse the same or similar datasets and compute different models and metrics.

The first article written by chinese researchers analysed the Taxi GPS traces from Shanghai in order to develop a model that helps understanding the needs of vehicular ad hoc networks for example. Therefore the authors computed several metrics, e.g. the turn probability at all the road crossing, implemented a map-matching algorithm and considered macro- and microscopic travel patterns. Their so called ”META” model shows a higher accuracy than all other models at this time. [1]

A different approach is presented by American researchers in their article. They haven’t used GPS traces, but collected traces from users on a campus via WiFi routers, at which the users’ smartphones automatically registered at. Thereby they are able to develop a model, which makes it possible for them to detect a hierarchy of the different access points and cluster them. These cluster are being analysed in terms of the size distribution as well as inter- and intra-cluster travel patterns. [2]

Another article about the taxi data set from Shanghai aimed to find an algorithm, that allows fair route sharing for every competing taxi driver. One of its goals is to not lose efficiency and a driving cost per customer as low as possible, so that it offers advantages for the driver and the customer equally. This requires a complex evaluation algorithm that considers several priniciples which are respected by the assignment mechanism. [3]

Instead of other public transports, taxi drivers plan their own routes once they drop off a passenger. Some articles are about recommender routes, to save the time for taxi drivers and potential passenger.

In general, there are three ways to develop recommender systems. The first one is content based which suggests items that are similar to those a given user has liked in the past. The second way is based on recommendations according to the tastes of other users that are similar to the target user. The third one is an hybrid solution.

T-Finder provides taxi drivers with detailed information not only the route, also with parking places. [4] The goal of the work described in this article is to propose an approach to detect parking places based on a large number of GPS trajectories generated by taxis, where the parking places stand for the locations where taxi drivers usually wait for passengers with their taxis parked. T-Finder enhance the passenger recommender by estimating the waiting time on a specified nearby road segment in addition to calculating the probability of finding a vacant taxi.

LCP [5] use the same method as T-Finder [4] to extract the representative small areas from the trajectories. LCP is the route recommendation algorithm that they have developed. This development is to find an energy-efficient mobile recommender system by exploiting the energy-efficient driving patterns extracted from the location traces of Taxi drivers. This system has the ability to recommend a sequence of potential pick-up points for a driver in a way such that the potential travel distance before having customer is minimized.

Identifying user interests and providing personalized suggestions, the different recommender systems address and collect information to improve time of passengers and also time of taxi drivers. Recommender systems [6] presents an overview of the field of recommender systems and describes the current generation of recommendation methods and various ways to extend the capabilities of recommender systems as well. This article concluded that recommender systems made significant progress over the last decade when numerous content-based, collaborative and hybrid methods were proposed and several ”industrial-strength” systems have been developed.

Recommender systems by mining large-scale [7], [8] [8] develop an smart recomendation system based on extracted patterns and data from studies focused of large-scale trace data collected by a probe taxi. Large-scale trace data provide us with opportunities to extract useful transportation patterns and utilize these patterns to improve efficiencies.

III. Data consistency

The first task is to check and improve the consistency of the provided data, so that we are able to compute our chosen metric accurately.

At the beginning we have to filter all impossible values of the variables we’re operating on. For example, our R script removes all entries with a negative speed value or timestamps outside of the interval in which these data have been collected. Apparently the trace files only contain consistent data sets as our script has not detected and removed a single error.

Furthermore, there are two more types of errors that have to be considered: Checking the geographical consistency and filtering gaps in the traces.

A. Geographical consistency

In the next step we analysed several statistical metrics concerning the speed variable which showed that there is a wide spectrum of different values, with some outliers bigger than 200 km/h, which does not make any sense in a city centre and is probably an error. That’s why we decided to cut off all the entries with a speed higher than three times the interquartile range. As you can see in figure 1 the speed distribution is skewed right with a single maximum. That justifies our decision, because we don’t lose relevant data entries, when we choose our cut-off speed value. On the other hand side, just performing a coordinate cut is not sufficient: it changes the speed distribution slightly, e.g. we lost some data entries on the street towards one of the airports by removing all the points with a speed higher than 80 km/h. However, it does not remove all error entries.

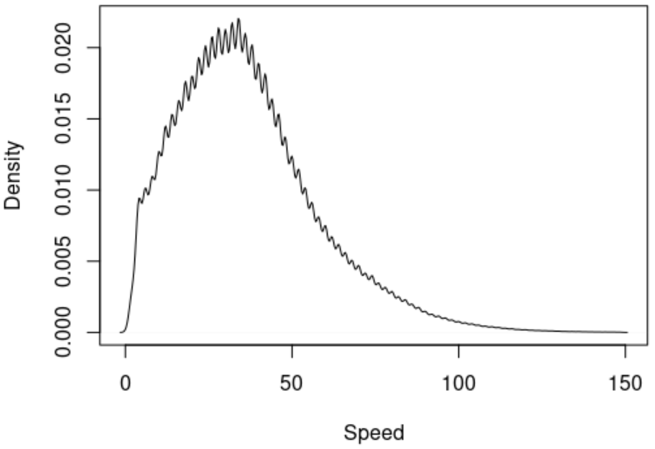

Figure 1. Adapted speed distribution without zero values

The data has still to be filtered by the coordinates. As we decided to focus on the city centre, we only keep those entries in our dataset that have a longitude between 121.38 and 121.57 and a latitude between 31.15 and 31.32. In figure 2 you can see our chosen borders on a map with a random sample of 100.000 points. After restricting the data to this area, we observed a change in the distribution of the speed values as well: In general, the mean slightly decreases whilst the median is increasing strongly. As we have a right skewed distribution, we eliminate higher values with this procedure, which is the result of the fact that some streets outside don't have a low speed limit and make it necessary to perform this cut-off.

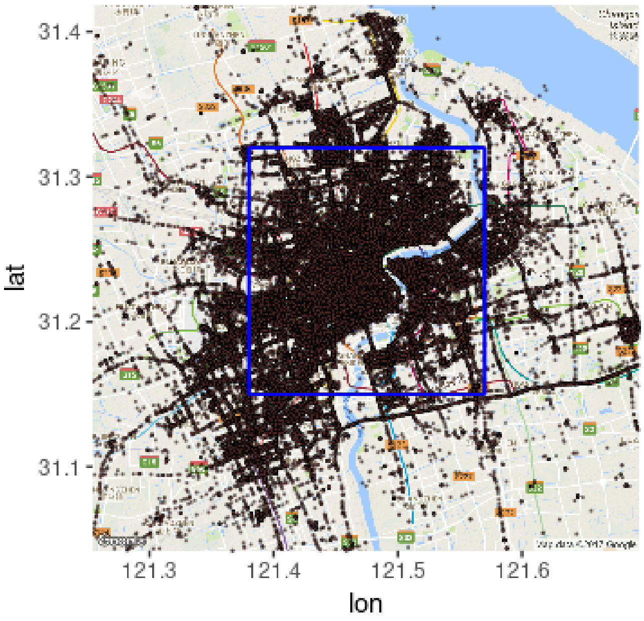

Figure 2. Borders of the city center

B. *Gap filtering*

After this procedure we are already able to plot a schematic map of Shanghai, although our dataset still contains data to be filtered. We search for those taxi traces with a time gap between two data entries. This gap could either be a result of a data transmission error or an indicator for the beginning of a new taxi trip.

To define these trips we had to choose a time gap that was appropriate for this data. So to define a trip we said that if the absolute value of a time gap between two data was bigger than 600 seconds that it would be separated and be a new trip.

The reason we chose this time range was so that for future analyses of these trips we felt that this time range would keep our results consistent and more accurate. Each of these trips would be given random colours to discern one from another if we decided to plot all of these trips like in the example given. As you can see there are many trips that have been plotted and to the eye these trips seem to be valid as well.

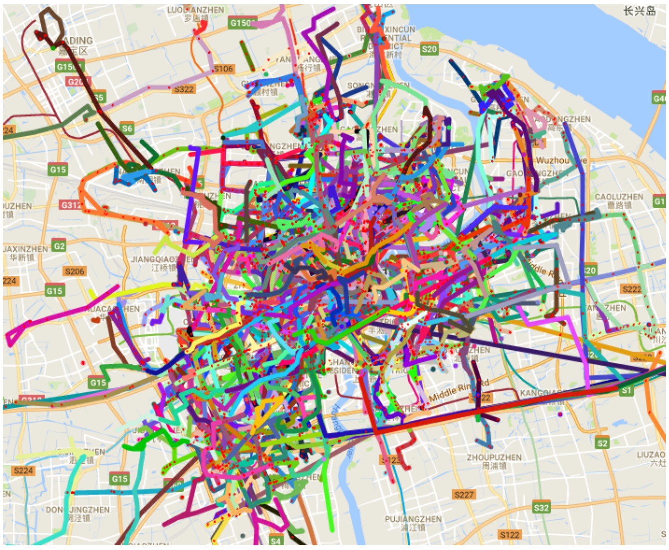

Figure 3. All calculated trips

IV. Our Metric

V. Future work / Open issues

VI. Summary

References

[1] H. Huang, D. Zhang, Y. Zhu, M. Li, and M.-Y. Wu, “A metropolitan taxi mobility model from real gps traces,” *Journal of Universal Computer Science*, 2012.

[2] R. Jain, D. Lelescu, and M. Balakrishnan, “Model t: An empirical model for user registration patterns in a campus wireless lan,” *MobiCom’05*, 2005.

[3] S. Qian, J. Cao, F. L. Mouël, I. Sahel, and M. Li, “Scram: A sharing considered route assignment mechanism for fair taxi route recommendations,” *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD’2015)*, 2015.

[4] N. J. Yuana, Y. Zheng, L. Zhang, and X. Xie, “T-finder: A recommender system for finding passengers and vacant taxis,” *Knowledge and Data Engineering*, 2013.

[5] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, “An energy-efficient mobile recommender system,” *InSIGKDD*, 2010.

[6] G. Adomavicius and A. Tuzhilin, “Towards the next generation of recommender systems: A survey of the state-of-the art and possible extensions,” *TKDE*, 2005.

[7] S. Qian, Y. Zhu, and M. Li, “Smart recommendation by mining large-scale gps traces,” *InWCNC*, 2012.

[8] X. Liu, Y. Wang, J. Biagioni, and S. Jiao, “Mining large-scale, sparse gps traces for map inference: Comparison of approaches,” *ACM*, 2012.