Visual Data Quality Analysis for Taxi GPS Data

Zuchao Wang* Xiaoru Yuan† Tangzhi Ye‡ Youfeng hao§ Siming Chen¶ Jie Liang‖
Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University
Qiusheng Li** Haiyang Wang†† Yadong Wu‡‡
School of Computer Science and Technology, Southwest University of Science and Technology

Abstract

We present a novel visual analysis method to systematically discover data quality problems in raw taxi GPS data. It combines semi-supervised active learning and interactive visual exploration. It helps analysts interactively discover unknown data quality problems, and automatically extract known problems. We report analysis results on Beijing taxi GPS data.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI)

1 Introduction

Most research in taxi GPS data cleaning focus on proposing algorithms to cleanse specific kinds of quality problems, yet the problem of discovering quality problems remains largely unanswered. In practice, analysts derive the lists of existing quality problems by trial and error. However, this strategy is not systematic, and subject to misidentification. In this paper, we propose a visual analysis method to systematically identify different kinds of data quality problems. Analysts visually explore the dataset to discover quality problems. The system then automatically search for all occurrence of such problems.

A major complexity that we consider is, many quality problems are related to the spatiotemporal relationship between sampling points. This is different from the problems in tabular data, which mostly concerns individual records. For example, when a GPS sampling point is 20 km away from its adjacent point recorded 5 minutes earlier from the same vehicle, there should be something wrong, even though the position and attributes of this individual point seem all right. To capture such spatiotemporal relationship, we partition each taxi trajectory into 5 minutes long sub-trajectories. Then we define a feature vector on each sub-trajectory to reflect the relationship.

2 Related Work

For common tabular data, quality problem identification can be based on detecting rule violations [3]. The rules are usually generated from taxonomies [3], domain expert experience [2], or automatic discovery [1]. For taxi GPS data, such violation detection method cannot be easily adapted to consider the spatiotemporal relationship between sampling points. The most relevant work is from Liao et al. [4]. It combines visualization and machine learning, which allows analysts to detect GPS data anomalies in a semi-automatic manner. However, it only outputs the anomalous sampling points, without identifying the quality problems.

3 Method

We treat each sub-trajectory as a basic analysis unit, trying to label it as normal or having certain quality problem. Our method consists of five steps, illustrated in Fig 1.

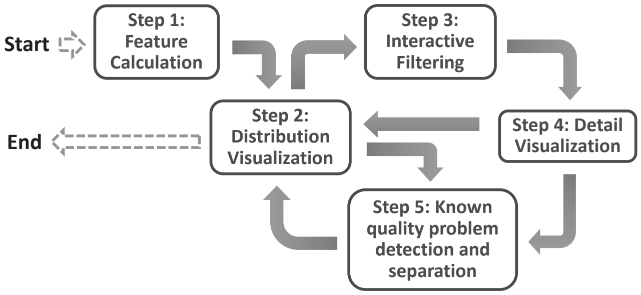

Start
Step 1: Feature Calculation
Step 3: Interactive Filtering
Step 2: Distribution Visualization
Step 4: Detail Visualization
End
Step 5: Known quality problem detection and separation

Figure 1: The pipeline of our method.

Step 1: Feature Calculation. We have defined 19 features for each sub-trajectory to characterize the spatiotemporal relationship between sampling points. That includes: number of sampling points, total travel distance, sum of turning angles, the min/max/recsum/logsum of turning angle, of segment distance, of segment time interval, and of segment average speed. Here each *segment* is the straight-line connection between two consecutive sampling points in a trajectory. *Recsum* relates to the number of zeros, which is calculated as $\sum_{i=1}^{n} \frac{1}{x_i+1}$. *Logsum* relates to the number of large values, which is calculated as $\sum_{i=1}^{n} \log(x_i+1)$. Anomalies on these features may indicate quality problems. For example, small number of sampling points suggests data missing; large max segment speed suggests jumping points; large sum of turning angle suggests signal oscillation.

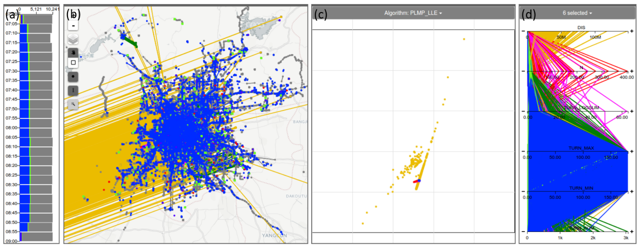


Figure 2: Visualization of data distribution in spatiotemporal domain and in feature space.

Step 2: Distribution Visualization. We assume quality problem corresponds to some sub-trajectory outliers or clusters. We try to visualize them in several views, including: a temporal histogram (Fig 2(a)) showing the temporal distribution, a map (Fig 2(b)) showing the spatial distribuion, a high dimensional projection view (Fig 2(c)) and a parallel coordinates view (Fig 2(d)) showing the

[*] e-mail: zuchao.wang@pku.edu.cn, now at Qihoo 360 Technology Co. Ltd.

[†] e-mail: xiaoru.yuan@pku.edu.cn

[‡] e-mail: yetangzhi66@gmail.com

[§] e-mail: ajihyf@gmail.com

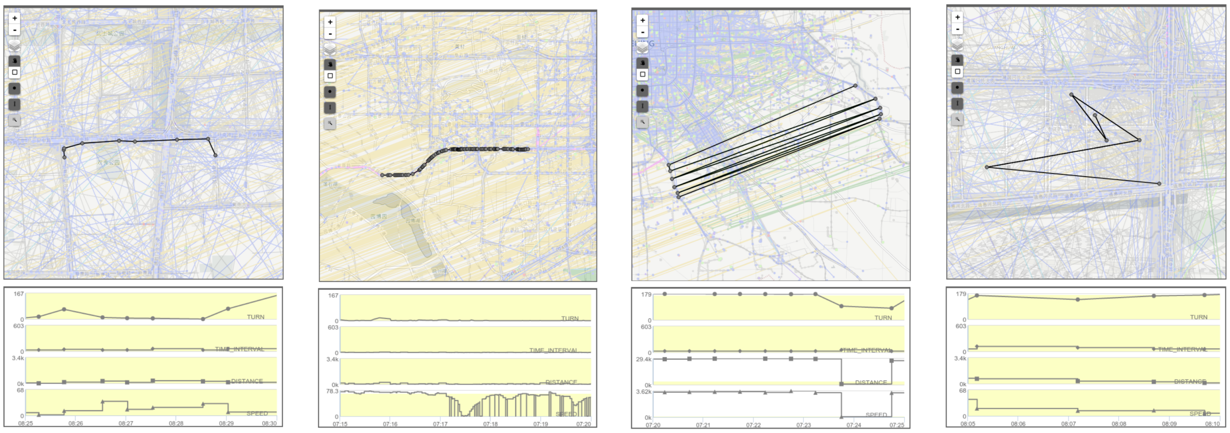
[¶] e-mail: csm@pku.edu.cn

[‖] e-mail: christy.jie@gmail.com

[**] e-mail: qiushengli245@gmail.com

[††] e-mail: haiyangcode@gmail.com

[‡‡] e-mail: wyd028@163.com



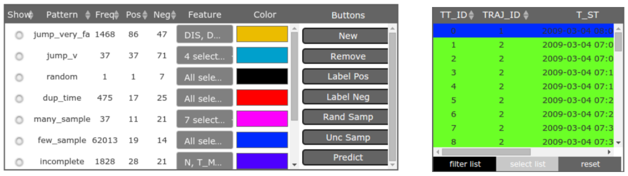
(a) (b) (c) (d)

Figure 3: Experiment result: detail visualization of a normal sub-trajectory and three sub-trajectories with different quality problems.

feature space distribution. The colors of sub-trajectories in all views indicates the corresponding types of quality problems (Fig 4(a)).

Step 3: Interactive Filtering. We allow analyst to extract the outlier or clustered sub-trajectories with a set of filters. They can select a time range in in the temporal histogram view, a rectangular spatial range in the map, a rectangular range in the projected feature space, and value ranges on parallel coordinates axes. Sub-trajectories satisfying all filters will be selected. We do not automatically generate clusters, because the result directly generated by machine may not correspond to quality problems, and can be hard to interpret. We rely on human analysts to find meaningful clusters.

Step 4: Detail Visualization. All filtered sub-trajectories are maintained in a list (Fig 4(b)). Analyst can highlight one and get its detailed information with three visualizations. This helps them decide whether the sub-trajectories are normal or having some quality problems. The map view (Fig 3 top row) now shows their path and shape. The new timeline view (Fig 3 bottom row) shows the temporal change of four attribute: turning angle, segment time interval, segment distance and segment average speed.For each attribute we estimated a valid value range, and mark the range yellow. Values outside the yellow range can be considered as potential outliers.



Show	Pattern	Freq	Pos	Neg	Feature	Color	Buttons
○	jump_very_fa	1468	86	47	DIS, D...	🟨	New
○	jump_v	37	37	71	4 select...	🟦	Remove
○	random	1	1	7	All sele...	⬛	Label Pos
○	dup_time	475	17	25	All sele...	🟥	Label Neg
○	many_sample	37	11	21	7 select...	🟪	Rand Samp
○	few_sample	62013	19	14	All sele...	🟦	Unc Samp
○	incomplete	1828	28	21	N, T_M...	🟪	Predict

TT_ID	TRAJ_ID	T_ST
0	1	2009-03-04 08:0
1	2	2009-03-04 07:0
2	2	2009-03-04 07:0
3	2	2009-03-04 07:1
4	2	2009-03-04 07:1
5	2	2009-03-04 07:2
6	2	2009-03-04 07:2
7	2	2009-03-04 07:3
8	2	2009-03-04 07:3

(a) (b)

Figure 4: The classification interface.

Step 5: Known quality problem detection and separation. For each quality problem, a binary SVM classifier is automatically built, but analysts need to provide the training data. As manual labelling is very expensive, we use active learning strategy to mitigate the labeling effort. That is, the system will ask analysts to label the sub-trajectories that have greater influence on the model accuracy. Besides, analysts can also search for sub-trajectories similar to labelled ones, and label them. These together help to extract sub-trajectories of identified quality problems in a semi-automatic way. Analysts can separate out these sub-trajectories and focus on the remaining ones. When the remaining ones are all normal, our discovery process ends.

4 Experiment Results

We have tested our method on a sample dataset from Beijing taxi GPS data. It spans from 7 am to 9 am at Mar. 4th, 2009. It has recorded the trajectories of 13,080 taxis, covering 21% of licensed taxis in Beijing. There are 747,431 sampling points, and the overall size is 74.2 MB. The sampling interval is 30 seconds, but many points are missing.

We have discovered 8 quality problems. As shown in Fig 4(a), this gives us 10 classes of sub-trajectories: one for normal sub-trajectory, one for partition fragments, and 8 for quality problems. Their distribution is shown in Fig 2.

We first show a normal sub-trajectory in Fig 3(a). Its spatial path in the map seems reasonable, and the timelines are all within the yellow ranges. Then we show three sub-trajectories with different quality problems. The one in Fig 3(b) has many sampling points. The sampling points are very close to each other in the map, and the two timelines in the middle show that the time interval and distance of segments are all near zero. The sub-trajectory in Fig 3(c) has multiple jumps. The map shows that it corresponds to a taxi jumping back and forth from two distant locations. The bottom two timelines show that the distance and average speed on segments are quite high, above the yellow ranges. Very likely this is due to two taxis mistakenly having identical identifiers, therefore their GPS records are mixed together and considered as from one taxi. Finally, the sub-trajectory in Fig 3(d) shows a taxi in random movement. Although it seems normal in timelines, but the map shows it jumping randomly in a small region.

5 Conclusion

In this paper, we have proposed a visual analysis method to systematically detect data quality problems in taxi GPS data. In the future, we would try more features, evaluate the detection rate and false positive rate, and test it on more and larger dataset.

Acknowledgements

This work is supported by NSFC No. 61170204, and partially funded by NSFC Key Project No. 61232012 and the National Program on Key Basic Research Project (973 Program) No. 2015CB352500. This work is also supported by PKU-Qihu Joint Data Visual Analytics Research Center.

References

[1] F. Chiang and R. J. Miller. Discovering data quality rules. *Proc. VLDB Endow.*, 1(1):1166–1177, 2008.

[2] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang. NADEEF: A commodity data cleaning system. In *Proc. ACM SIGMOD*, pages 541–552, 2013.

[3] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee. A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99, 2003.

[4] Z. Liao, Y. Yu, and B. Chen. Anomaly detection in gps data based on visual analytics. In *Proc. IEEE VAST*, pages 51–58, 2010.