A Visual Analytics System for Metropolitan Transportation

Siyuan Liu, iLab, Heinz College, Carnegie Mellon University
syliu@andrew.cmu.edu
Ce Liu, University of Pittsburgh
Cel38@pitt.edu
Qiong Luo, Hong Kong University of Science and Technology
luo@cse.ust.hk
Lionel M. Ni, Hong Kong University of Science and Technology
ni@cse.ust.hk
Huamin Qu, Hong Kong University of Science and Technology
huamin@cse.ust.hk

November, 2011

LARC-TR-01-11

LARC Technical Report Series: http://smu.edu.sg/centres/larc/larc-technical-reports-series/

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Carnegie Mellon University

ABSTRACT

With the increasing availability of metropolitan transportation data, such as those from vehicle GPSs (Global Positioning systems) and road-side sensors, it becomes viable for authorities, operators, as well as individuals to analyze the data for a better understanding of the transportation system and possibly improved utilization and planning of the system. We report our experience in building the VAST (Visual Analytics for Smart Transportation) system. Our key observation is that metropolitan transportation data are inherently visual as they are spatio-temporal around road networks. Therefore, we visualize traffic data together with digital maps and support analytical queries through this interactive visual interface. As a case study, we demonstrate VAST on real-world taxi GPS and meter data sets from 15,000 taxis running two months in a Chinese city of over 10 million population. We discuss the technical challenges in data cleaning, storage, visualization, and query processing, and offer our first-hand lessons learned from developing the system.

A Visual Analytics System for Metropolitan Transportation

Siyuan Liu
iLab, Heinz College, CMU

Ce Liu
University of Pittsburgh

Qiong Luo
HKUST

Lionel M. Ni
HKUST

Huamin Qu
HKUST

ABSTRACT

With the increasing availability of metropolitan transportation data, such as those from vehicle GPSs (Global Positioning Systems) and road-side sensors, it becomes viable for authorities, operators, as well as individuals to analyze the data for a better understanding of the transportation system and possibly improved utilization and planning of the system. We report our experience in building the VAST (Visual Analytics for Smart Transportation) system. Our key observation is that metropolitan transportation data are inherently visual as they are spatio-temporal around road networks. Therefore, we visualize traffic data together with digital maps and support analytical queries through this interactive visual interface. As a case study, we demonstrate VAST on real-world taxi GPS and meter data sets from 15, 000 taxis running two months in a Chinese city of over 10 million population. We discuss the technical challenges in data cleaning, storage, visualization, and query processing, and offer our first-hand lessons learned from developing the system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.8 [**Database Management**]: Database Applications

General Terms

Design, Experimentation, Performance

Keywords

Vehicle trajectory, spatiotemporal data, visual analytics

1. INTRODUCTION

Transportation management has been a world wide challenge in modern urbanization. Particularly, in medium to large-sized Chinese cities, taxi cabs are the most dynamic and challenging transportation means to manage for city authorities. Taxis have several unique characteristics in comparison with other types of transportation means, including public transportation such as buses, subways, and private cars. First, the routes and operating times of taxi cabs are highly flexible and vary greatly. Second, their operations involve the interests of multiple parties - passengers, taxi drivers, taxi operators (companies), and city authorities. To assist in the understanding and future improvement of a metropolitan transportation system, we developed VAST, a visual analytics system for transportation data, specifically GPS and meter data from taxis.

ACM SIGSPATIAL GIS '11 November 1-4, 2011. Chicago, IL, USA.

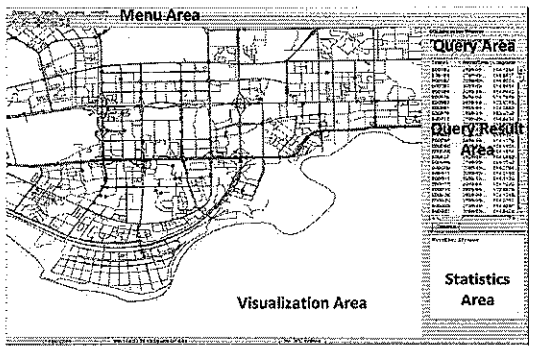


Figure 1: VAST Graphical User Interface

To start the design of the data analytics, we consider a set of queries that are of interest of one or more of the involved parties. For example, city authorities are probably most concerned about the geographic distribution of the taxis, the throughput of taxis along road segments, as well as the evolvement of these information over time. In comparison, taxi companies might want to identify top earning taxis as well as reckless speeding drivers. More modestly, individual taxi drivers wonder what routes to take to have a higher chance of picking up passengers, and passengers need to know where nearby they are most likely to find a vacant cab. Having collected a set of analytical queries most commonly issued on the taxi data, we observe that these queries are different from traditional multidimensional OLAP (Online Analytical Processing) workloads. The major distinction is that there are no regular cubic dimensions along which to perform aggregation; rather, our queries are centered around trajectories along road segments over time. As such, it is necessary to have a visual interface with road network maps to help users pose queries and to display query results. We have developed the visual interface in VAST as shown in Figure 1. In Figure 1, we illustrate the VAST GUI (Graphical User Interface). The menu area lists menu items, including query, visualization, and other functions. In the query area, we provide an interface for users to issue predefined as well as customized SQL queries. The query result area shows results in text whereas the

statistics area in bar charts, pie graphs, and other visual analytical representations as applicable. Finally, the main window of the visualization area displays a 2D digital map of the city road networks as well as other geographic information. When the results of a query are returned, they will be displayed simultaneously in the query result, visualization, and statistics areas as applicable. In particular, taxi trajectories of interest will be embedded into the map in the visualization area.

We have obtained a two-month GPS and meter data set of 15, 000 taxis running in a Chinese city of 10 million population to study with VAST. The first problem we encounter is missing and erroneous records in the data. In particular, notable errors appear in reported locations as well as in the reporting frequency, due to the high mobility of vehicles and the variance in the environment [12, 23]. To deal with such problems, we propose efficient calibration methods utilizing geographic and historical information. The second problem is about query efficiency given the large number of taxis and their trajectories. Considering the typical query workloads, we index on the road segments and embed the taxis GPS data points into the road segments. The third challenge is in visualizing the trajectory data from a large number of taxis. In our work, we utilize queries to reduce the data to be visualized and store previous query as well as visualization results to speed up visualization.

In summary, we have made the following contributions. (1) We develop a system to study large scale transportation data, integrating visualization and data analytics methods, and propose techniques to improve efficiency and scalability. (2) We report our experience and observations in building VAST, and discuss technical challenges. (3) We demonstrate and test our system using real-life data sets for real-world applications.

2. RELATED WORK

Data storage and management: The storage representation of a general trajectory model is crucial for the system performance, because all operations are performed on this structure. Le et al. [11] presented a dynamic spatio-temporal data structure called the Graph Strip Tree (GStree) that can efficiently answer range queries about the current or past positions of moving objects. Ni et al. [17] described a parametric space indexing method for historical trajectory data, yielding a much finer approximation quality than MBRs (Minimum Bounding Rectangles). Recently the topic of extending OLAP with spatial and temporal features has attracted the attention of the database and GIS communities. Several papers investigated how the traditional data cube model be adapted to trajectory warehouses in order to transform raw location data into valuable information. In particular, they focus on the issues that are critical to trajectory data warehousing (TDW). Marketos et al. [16] proposed a framework for the development of TDW taking into consideration the complete flow of the tasks. Vaisman et al. [20] presented a conceptual framework for defining spatio-temporal data warehouses using an extensible data type system. Madden et al. [8, 22] proposed a dynamic storage system optimized for quickly accessing data in a particular spatial-temporal region. This collection of work handles a small set of trajectory data, or the trajectory data of one or a few vehicles. In comparison, our work is done for large scale trajectory data, including both trajectory and road networks data.

Modeling and querying: Compieta et al. [6] introduced a structured inventory of existing exploratory techniques for spatio-temporal data. Alvares et al. [1] proposed a reverse engineering framework for mining and modeling semantic trajectory patterns in a geographic database. Spaccapietra et al. [18] proposed two trajectory modeling approaches, one based on a design pattern, and the other based on dedicated data types, and illustrated their differences in the implementation in an extended-relational system. Gkoulalas et al. [10] presented a privacy-aware trajectory tracking query engine that offers strict guarantees about what can be observed by intrusted third parties with a single continuous polynomial approximating a sequence of movement functions. Chen et al. [5] introduced an Edit distance function on Real sequence (EDR), which is robust against data imperfections. Tiesyte et al. [19] defined distance measures based on similarity and a notion of nearest neighbor, capable of predicting future vehicle movements. In comparison with this line of work, we integrate visualization methods with database queries to help users explore, analyze and understand trajectory data.

Visual analytics: Crnovrsanin et al. [7] presented proximity-based visualization approach for movement traces in an abstract space rather than the given spatial layout. This abstract space is obtained by considering proximity data, which is computed as distances between entities and some important locations. Andrienko et al. [2] defined aggregation methods suitable for movement data and proposed interaction techniques to represent results of aggregations, enabling comprehensive exploration of the data. In our work, we not only visualize large scale spatio-temporal trajectory data sets, but also embed database query results to digital maps.

3. DESIGN AND IMPLEMENTATION

3.1 Data Schema and Predefined Queries

The database schema consists of two tables - TAXITRACK and TAXIMETER. TAXITRACK stores GPS data reported from taxis. Its attributes include taxi ID, reporting time, longitude, latitude, speed, direction, and status (occupied or vacant). TAXIMETER stores meter data with attributes including taxi ID, transaction start time and end time, total distance traveled in this transaction, total waiting time, and fare. This database schema is provided by the available raw data and may not be suitable for the query workload. It would be interesting to study how to redesign the schema to best suit a predefined query workload.

The predefined queries in VAST are as follows.

- Query 1: Find the location or trajectory of a given taxi for a given time or a given time period.
- Query 2: Find the top ten speeding taxis for a given time period.
- Query 3: Find the top 10 taxis by revenue for a given time period.
- Query 4: Retrieve the taxi pick-up/ drop-off locations for a given time period.
- Query 5: Report the top 10 taxis by number of transactions (rides) for a given time period.
- Query 6: Retrieve the taxi distribution in the city for a given time period.
- Query 7: Report the top 10 hot spots (crowded areas) in the city by number of taxis for a given time period.
- Query 8: Compute the throughput in number of taxis for a given area for a time period.
- Query 9: Given a no-entrance zone, find taxis that entered the zone for a time period.

- Query 10: Given a taxi's current location and a time period, find a trajectory that earns the highest fare.
- Query 11: Given the start and destination locations, find the fastest and the shortest routes.
- Query 12: Given a passenger location and a time period, find the nearby vacant taxis.

3.2 Main Techniques

3.2.1 Calibration

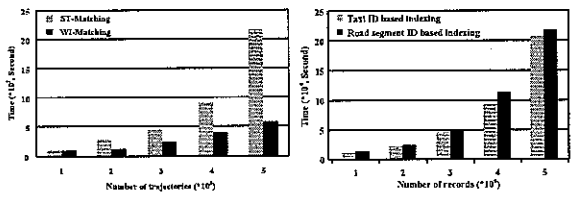
In this work, we propose **WI**-matching, a **W**eighting-based map matching algorithm and an **I**nterpolation algorithm to calibrate the erroneous and low-sampling-rate vehicle GPS trajectory data set. The details are available in our technical report [13]. In our proposed algorithms, we first integrate the vehicle GPS sampling data and digital road networks data, to identify the road where a vehicle traveled and vehicle locations on that road. The weighting-based map matching algorithm considers (1) the geometric and topological information of the road networks and (2) historical trajectory information to efficiently and effectively calibrate the sampling data points. Moreover, we propose an interpolation algorithm to identify the path between consecutive GPS points. Then we estimate the interpolated positions and timestamps of the vehicle along the path, to finally construct a correct and complete trajectory.

3.2.2 Indexing

Observing the dozen of predefined queries, we notice that nearly one half of the queries have a selection condition on the location. Therefore, we first index the road networks map into a 2D grid so that given a location point or bounding box, we can return the road segment IDs using the grid. Then, we create a road segment ID index table (road segment ID, taxi ID, reporting time). Using this road segment ID index, we can then retrieve taxis that are on a given road segment at a specific time. This indexing scheme is efficient and scalable to a large number of trajectories, because the number of road segments is much smaller than and independent from the number of taxi GPS points, and changes little over time.

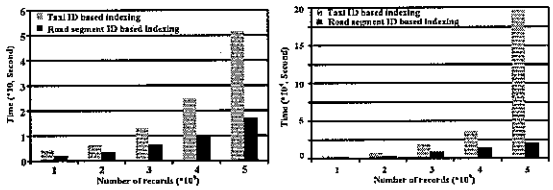
3.2.3 Visualization

The large scale spatial-temporal nature of vehicle GPS data and geographic information increases the visualization complexity. To discover the underlying correlations within the data, we need a versatile visual analytic solution to facilitate the exploration of the data. Some new visual encoding schemes for spatial temporal trajectory data are proposed in our work. To study the vehicle GPS time series in a large scale, we employ a trajectory coloring method to distinguish each vehicle trace. To explore the correlation between different information, for example, the income and the business practice (e.g., driving records, location, time), we first differentiate the taxi drivers into different income groups (i.e., high and low-income) with the help of the time-series plot and parallel coordinates. To visualize the distribution information, for example, the hot spots in the city, we utilize a heat map to represent the statistic data. To speed up visualization efficiency, we utilize the LOD (Level of Details [14]) to investigate the large scale trajectory data in digital map. In VAST, we provide a history scheme for users to track their operations so that users can always go back to any step in their operations. For example, after users conduct a comparison for one trajectory in the detail viewer, they can go back to the selected area with all trajectories. And they can go back from the detail viewer to the overview, holding the query results.



(a) Calibration Performance (b) Indexing Performance

Figure 2: Calibration and Indexing



(a) Trajectory query (b) Nearby taxis query

Figure 3: Query Performance

4. SYSTEM EVALUATION

4.1 Experiment Setup

Hardware and software: We used the Oracle 10g lite as our database server. The server is equipped with an AMD Athlon(tm) II X2 B24 Processor running at 3.01 GHz, 2.00 GB RAM, and a 320 GB disk. We employed JDBC to connect to the database server and Java swing to develop visualization modules [9]. The digital map is a *.shp* file (including bit maps and vectors to describe the objects in the digital map). We divided the map into different layers in VAST with each layer storing one kind of geographic objects, such as buildings, road networks and mountain areas [14].

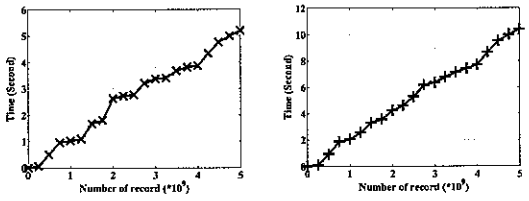
Data sets: We obtained a two-month taxi GPS and meter data set of 15, 000 taxies. The GPS data sampling rate was around 1 minutes and in total there are more than 100 million GPS data points in the raw data. The meter data contains about 2 million transactions. In total, the size of the raw data set in ASCII files loaded into the Oracle database is 270 GB.

Experiment queries: We used three example queries to evaluate the query processing performance. The trajectory query retrieves the trajectory of a given taxi for a given time period. The passenger query finds nearby taxis for a given location and time period. The throughput computes the number of distinct taxis that go through a given location within a given time period.

Evaluation metrics: In our experiments, we pick a few representative queries, e.g., the fastest one (querying single vehicle's trajectory), the slowest one (querying all nearby taxis), and a medium-speed one (querying road throughput), to show the query and visualization time. In addition, we report the raw data loading and calibration performance.

4.2 Results

First, we report the time of loading, calibrating, and indexing data. Then, we evaluate the time performance of query processing and visualization for the three example queries with billions of data records (GPS data points). In our experiment environment, the time of loading the raw data set is nearly 12 hours. The calibration performance is reported in Figure 4.1. We select ST-Matching [15] as



(a) Trajectory query (b) Nearby taxis query

Figure 4: Visualization Performance

a baseline method to compare with our method, WI-Matching [13]. WI-matching is up to four times faster than ST-matching, especially when the method works on a larger trajectory data set. More results (e.g., accuracy) on calibration are available in our technical report [13]. In Figure 4.1, we report the indexing performance. We compare our road segment ID based method with a naïve indexing, taxi ID based indexing. Our indexing is slightly slower than the taxi ID based indexing due to the time cost of integrating the GPS data points into road segments. In Figure 3 (a), we report the time cost in querying a taxi's trajectory for a given 1-hour time period. In Figure 3 (b), the query is on nearby taxis (a circle with a diameter of 3 kilometers, the base fare distance) for a given location in a given 15-minutes time period. Our method is significantly more efficient and scalable than the baseline. Figure 4 shows the visualization performance. Note that the visualization is on the query results from the given size of the original data set. The visualization time is much shorter than the query processing time cost. If we visualize all the GPS data points and then find out the results, the visualization time is about 27 hours in our experiment.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduce a visual analytics system for metropolitan transportation, VAST. The system design, implementation and evaluation are described and the lessons we learned from the system running on real world data sets are also discussed. We are currently working on improving the efficiency of the system as well as providing more queries and visualization functionalities.

6. ACKNOWLEDGEMENT

This research was supported by Shenzhen Transportation Bureau, Guangdong, China. This paper was supported in part by the National High Technology Research and Development Program of China under Grant No. 2011AA010500.

7. REFERENCES

[1] L. O. Alvares, V. Bogorny, J. A. F. de Macedo, B. Moelans, and S. Spaccapietra. Dynamic modeling of trajectory patterns using data mining and reverse engineering. In *Proc. of ER*, 2007.

[2] G. Andrienko and N. Andrienko. Spatiotemporal aggregation for visual analysis of movements. In *Proc. of IEEE VAST*, 2008.

[3] G. Andrienko, N. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Information Visualization*, 7(3), 2008.

[4] G. Andrienko, N. Andrienko, and S. Wrobel. Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl.*, 9(2), 2007.

[5] L. Chen, M. T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *Proc. of ACM SIGMOD*, 2005.

[6] P. Compieta, S. Di Martino, M. Bertolotto, F. Ferrucci, and T. Kechadi. Exploratory spatio-temporal data mining and visualization. *J. Vis. Lang. Comput.*, 18(3), 2007.

[7] T. Crnovrsanin, C. Muelder, C. Correa, and K.-L. Ma. Proximity-based visualization of movement trace data. In *Proc. of IEEE VAST*, 2009.

[8] P. Cudre-Mauroux, E. Wu, and S. Madden. Trajstore: An adaptive storage system for very large trajectory data sets. In *Proc. of IEEE ICDE*, 2010.

[9] R. Eckstein, M. Loy, and D. Wood. *Java Swing*. O'Reilly & Associates, Inc., 1998.

[10] A. Gkoulalas-Divanis and V. S. Verykios. A privacy-aware trajectory tracking query engine. *SIGKDD Explor. Newsl.*, 10(1), 2008.

[11] T. T. T. Le and B. G. Nickerson. Efficient search of moving objects on a planar graph. In *Proc. of ACM GIS*, 2008.

[12] S. Liu, Y. Liu, L. Ni, J. Fan, and M. Li. Towards mobility-based clustering. In *Proc. of ACM SIGKDD*, 2010.

[13] S. Liu, Y. Luo, and L. Ni. *Calibration of Vehicle Trajectory*. Technical report, 2010.

[14] S. Liu, G. Wen, and J. Fan. A 3d geosciences modeling system for large-scale water-diversion projects. *IEEE Des. Test*, 12(1), 2010.

[15] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *Proc. of ACM SIGSPATIAL*, 2009.

[16] G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà, and Y. Theodoridis. Building real-world trajectory warehouses. In *Proc. of ACM MobiDE*, 2008.

[17] J. Ni and C. V. Ravishankar. Indexing spatio-temporal trajectories with efficient polynomial approximations. *IEEE Trans. on Knowl. and Data Eng.*, 19(5), 2007.

[18] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data Knowl. Eng.*, 65(1), 2008.

[19] D. Tiesyte and C. S. Jensen. Similarity-based prediction of travel times for vehicles traveling on known routes. In *Proc. of ACM GIS*, 2008.

[20] A. Vaisman and E. Zimányi. What is spatio-temporal data warehousing? In *Proc. of DaWaK*, 2009.

[21] P. C. Wong and J. Thomas. Visual analytics. *IEEE Comput. Graph. Appl.*, 24(5), 2004.

[22] E. Wu, P. Cudre-Mauroux, and S. Madden. Demonstration of the trajstore system. *Proc. VLDB Endow.*, 2(2), 2009.

[23] H. Zhu, M. Li, Y. Zhu, and L. M. Ni. Hero: Online real-time vehicle tracking. *IEEE Trans. Parallel Distrib. Syst.*, 20(5), 2009.

[24] H. Zhu, Y. Zhu, M. Li, and L. Ni. Seer: Metropolitan-scale traffic perception based on lossy sensory data. In *Proc. of IEEE INFOCOM*, 2009.