A Web-based Platform for Visualizing Spatiotemporal Dynamics of Big Taxi Data

Hui Xiong[a], Lei Chen[a], Zhipeng Gui[a,b *]

[a] School of Remote Sensing Information and Engineering, Wuhan University, 129 Luoyu Rd., 430079 Wuhan, China – (huixiong, lei_chen, zhipeng.gui)@whu.edu.cn
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China – zhipeng.gui@whu.edu.cn

Commission V, WG V/4

ABSTRACT: With more and more vehicles equipped with Global Positioning System (GPS), access to large-scale taxi trajectory data has become increasingly easy. Taxis are valuable sensors and information associated with taxi trajectory can provide unprecedented insight into many aspects of city life. But analysing these data presents many challenges. Visualization of taxi data is an efficient way to represent its distributions and structures and reveal hidden patterns in the data. However, Most of the existing visualization systems have some shortcomings. On the one hand, the passenger loading status and speed information cannot be expressed. On the other hand, mono-visualization form limits the information presentation. In view of these problems, this paper designs and implements a visualization system in which we use colour and shape to indicate passenger loading status and speed information and integrate various forms of taxi visualization. The main work as follows: 1. Pre-processing and storing the taxi data into MongoDB database. 2. Visualization of hotspots for taxi pickup points. Through DBSCAN clustering algorithm, we cluster the extracted taxi passenger's pickup locations to produce passenger hotspots. 3. Visualizing the dynamic of taxi moving trajectory using interactive animation. We use a thinning algorithm to reduce the amount of data and design a preloading strategyto load the data smoothly. Colour and shape are used to visualize the taxi trajectory data.

KEY WORDS: Taxi trajectory, GPS, Visualization, Hotspots, Colour, Shape

1. INTRODUCTION

With the fast development of traffic, various forms of transportation provides the public with a lot of convenience, but it also brings about problems, such as traffic accidents, air pollution, congestion and other negative effects. To address these problems, a number of measures have been taken, like intelligent transportation systems (ITSs), public transportation systems, safety seat belts, etc. Among these solutions, ITSs are universally accepted and considered to be promising in the future because they enhance the efficiency and functionalities of transportation systems with advanced information technology (Figueiredo et al., 2001). Especially, the role of data in ITS becomes more and more important because the of coming of big data era. The data contains valuable information and can be used to generate new functions and services in ITS (Zhang et al., 2011).

Data visualization employs visual channels to represent dataset (Hansen and Johnson, 2004). It transforms various types of data into appropriate visual representations, so the data can be understood and analysed effectively. The advantage of data visualization is that it incorporates human capabilities into an intuitive visual interface, thus combining machine intelligence and human intelligence together (Chen et al., 2015). Scientific visualization, information visualization and visual analytics are three major fields in data visualization. Scientific visualization illustrates structures and evolutions of physical or chemical properties in the spatial domain. Information visualization focuses on the representation of abstract, unstructured, and high-dimensional data, including business data, social network data, textual data, etc. Visual analytics is a new analysis strategy that integrates human intelligence and data analysis (Thomas and Cook, 2005). Traffic datasets are usually high-dimensional or spatial-temporal, so visualization of them mostly employs information visualization and visual analytics.

In many cities, a large amount of taxis move around streets transporting people among urban cores, business centres, tourist attractions, transportation hubs, and residential territories. By installing a vehicle GPS device on a taxi, a real-time moving path of the taxis can be recorded as a series of positions sampled with small periodic interval. At each location, the information recorded includes geographical coordinates, operation status, speed, direction, occupied or vacant status regarding whether the taxi has a customer, etc (Ding Chu et al., 2014). There is a lot of valuable information behind these data. And it remains a far reaching goal to read the information completely hidden in the complex, dynamic behaviour of such large population. During the recent years, trajectory and movement data have been studied with various approaches, including visual analysis (Andrienko et al., 2008), machine vision (Vijverberg et al., 2007), clustering (Andrienko et al., 2009), feature extraction (Andrienko, Dykes, et al., 2008) and movement taxonomy (Dodge et al., 2008). To better analyse and trajectory data, analyst in each domain require effective visualization to help them explore the data intuitively and derive pertinent insights.

A lot of work on visualization of trajectory data has been conducted. Liu et al. (2009) developed a real-time visualization system for the bus, metro and taxi trajectory data in Shenzhen, south China. The system provided more accurate and dynamic method to understand daily urban mobility patterns and explore the relationship for mobility with land use and social-economic changes. Yuan et al. (2010) built a system based on a real-world trajectory dataset generated by over 33,000 taxis in a period of 3 months to recommend the fastest route for the user. Guo et al. (2011) presented interactive visual system for exploring and analysing complex traffic trajectory data to investigate and analyse microscopic traffic patterns and abnormal behaviours. The users are equipped with a carefully designed interface to inspect data interactively from spatial, temporal and multi-dimensional views in the system. Liu et al. (2013) proposed the VAIT system to monitor and analyse complex traffic conditions in large cities. Ferreira et al. (2013) proposed a new model that allows users to visually query taxi trips and supports origin-destination queries that enable the study of mobility across the New York City. Landesberger et al. (2012) presented a new approach for visual analysis for spatio-temporal categorical data by algorithms for selection of globally and focally representative time steps based on categorical changes. Tominski et al. (2012) designed a hybrid 2D/3D solution around the principle of stacking trajectory bands to visualize trajectory attribute data.

The visualization of big data becomes a research hot spot that appeared in recent years, however, the visualization for taxi trajectory data is rare. Based on the results of various

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W7-1407-2017 | © Authors 2017. CC BY 4.0 License.

visualization methods and the characteristics of taxi data, there are some problems that attract our attention. First, each system is expressed in a single form, thus an integrated framework is needed. Second, most of the current work focuses on the analysis of taxi data and visualizes the results after analysing, while visualization of the raw taxi data is worth trying. Third, the speed and passenger information cannot be reflected in the current visualization, which limits the amount of information that can be expressed.

The remainder of this paper is organized as follows. Section 2 describes the data source and processing. Section 3 introduces the design and implementation of the visualization of taxi data. In section 4, we present the results of visualization of taxi data analysis. Finally, we conclude this paper and outline the possible future work.

2. DATA SOURCE AND PROCESSING

2.1 Data Source

The dataset we used in this paper is the taxi trajectory data in Wuhan, China, covering May in 2014. The dataset consists of 31 text files, each of which contains all the taxi trajectory data in a day. There are 424,036,598 records in total. The structure of each record of the data is shown as follows:

<Taxi ID, time, longitude, latitude, direction, speed, ACC status, operation status, passenger loading status>. Each record has latitude and longitude as its location. And it has a string as its time stamp. Taxi ID is used to identify which taxi it is. Attributes like directions and speed indicates the movement information about the taxi. ACC is short for Adaptive Cruise Control, which is a system universally employed on vehicles for safety considerations. The ACC is usually on when the car is fired. The operation status tells whether the taxi is in operation or not. Attribute like passenger loading status can tell whether the taxi has a customer.

2.2 Data Pre-processing

2.2.1 Data Cleaning and Processing

Since errors and missing records are common in the raw data (Liu et al., 2010), we first remove the drift records. Fig. 1 is an illustration on how to remove the drift data. Second, when the driver is at rest, the GPS device aboard the taxi will continue to record data, which is redundant literally. Thus, this kind of data also needs to be eliminated. Fig. 2 shows the detail to eliminate the redundant data.

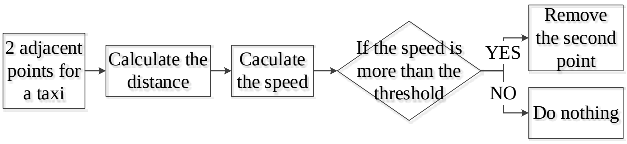

Figure 1. The procedure to remove drift data.

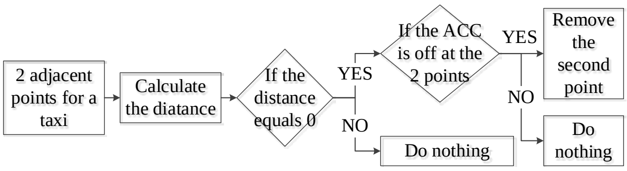

Figure 2. The procedure to remove redundant data.

In order to visualize the hotspots where most people take taxis, we need to find all the pickup points where passengers get on and cluster them with the help of clustering algorithm. In our work, we cluster the pickup points extracted every hour from trajectory data. For each taxi trajectory in an hour, the pickup points are determined based on the passenger information of the adjacent two points. Fig. 3 shows in detail how we select the pickup points. Clustering analysis is a widely used data processing method, which can aggregate some data based on its similarity and separate it from the others. In this research, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm (Ester et al., 1996) is employed to cluster pickup points based on the distances among the pickup points and produce hotspots for the pickup points.

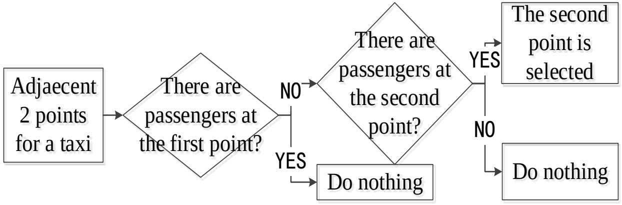
Adjaecent 2 points for a taxi
There are passengers at the first point?
NO
There are passengers at the second point?
YES
The second point is selected
NO
Do nothing
YES
Do nothing

Figure 3. The procedure to extract pickup points.

The time interval between 2 adjacent sampling points for a taxi is not a constant. Sometimes the interval is about 10 seconds, sometimes the interval is about 1 minute. When we visualize the taxi trajectory data dynamically, we choose the records whose time interval is 10 seconds due to their high accuracy. Thus, we extract the records whose interval is 10 seconds for the dynamic visualization of taxi trajectory.

2.2.2 Data Storage and Organization

After the pre-processing of the taxi data, the amount of data is still huge, and the data loading speed through the web side is not very fast. At the same time, different visual effects need to load different data, so it is necessary to store the data in many indexed tables and so that they can be quickly accessed, queried, and displayed very smoothly in the browser.

The database we adopted in our system to store the taxi data is MongoDB, which is a document database with scalability and flexibility and is a typical representation of No-SQL database. Fig. 4 shows the structure of database stored in MongoDB. Each collection in MongoDB corresponds to a table in traditional relational database, and a document corresponds to a record. In traditional relational database, each record in a table must share the same fields, while each document in a collection is allowed to have different fields in MongoDB.

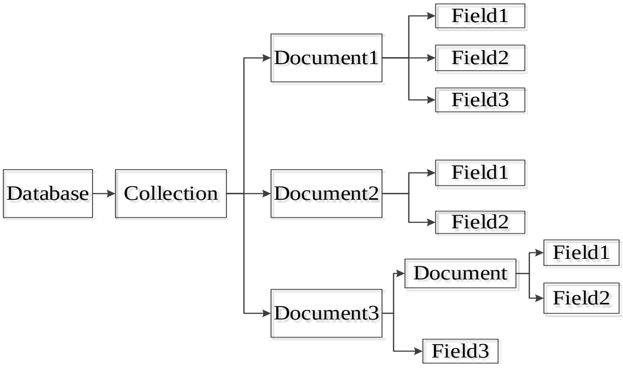

Figure 4. The structure of database stored in MongoDB.

The amount of taxi trajectory data per day is so big that we decide to store the records for one hour into a collection in MongoDB to improve the query efficiency. Thus, all the taxi trajectory data that have been pre-processed are stored in many different collections. To improve the efficiency of querying and accessing further, we index the longitude field, latitude field and taxi ID field in each collection. On the contrary, the hotspots data for a month are all stored in one collection, in which every document contains different hotspots data in different time and map scale. Also, the time field and map scale field are indexed to improve the performance of the system.

3. DESIGN AND IMPLEMENTATION

3.1 The architecture of the system

The visual framework design of this paper is based on Web development techniques, and adopts typical Browser/Server architecture. The main framework is composed of storage level, interface level, presentation level and client. The database is used to store all the data needed to be visualized. To improve the efficiency of the system, some fields are indexed in the database.

Servlet, whose full name is Java Servlets, is a widely accepted server-side technology for building dynamic content for web-based applications. Typically, the servlet takes an HTTP request from a browser, generates dynamic content (such as by querying a database) and provides an HTTP response back to the browser. To deploy and run a servlet, a web container must be used. A web container (also known as a servlet container) is an essential component of a web server that interacts with the servlets. The web container is responsible for managing the lifecycle of servlets, mapping a URL (Uniform Resource Locator) to a particular servlet and ensuring that the URL requester has the access rights. In our system, the web container we used is Apache Tomcat, which works as both a web server and a servlet container.

Leaflet is the leading open-source JavaScript library for mobile-friendly interactive maps and provides the functionality to add markers, popups, overlay lines, etc. It works efficiently across all major browsers in desktop and mobile platforms. Ajax (Asynchronous JavaScript and XML) is a universally adopted development technology for creating interactive web applications. The kernel of Ajax is the XmlHttpRequest JavaScript object, which can make a request to the server and process the response without blocking the user. We design an asynchronous data loading strategy based on Ajax technology. Both Leaflet and Ajax are used to display the data fetched from the database on the browser.

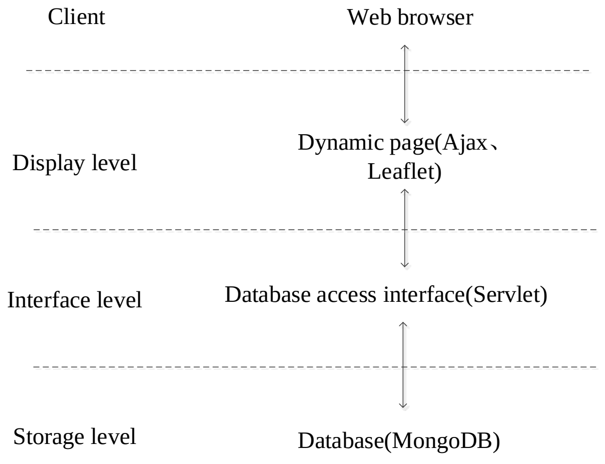

Figure 5. The architecture of the visualization system

3.2 Visualization for taxi trajectory data

3.2.1 Usage of colour and shape

The general visualization for taxi trajectory only expresses three-dimensional information, which is, time, longitude, and latitude. In addition to these information, there are passenger and speed information in the taxi trajectory data.

To present the passenger loading status for a taxi, we use a hollow circle to represent that there is no passenger in the taxi, and a solid circle means otherwise. As for the speed information of a taxi, we divide the speed values into six classes and use six different colours to represent different speed levels. Fig. 3 shows the way we display speed and passenger information for taxis. Generally speaking, red colour can make people feel nervous, thus it is appropriate to use it to represent high speed levels. On the contrary, green colour can make people feel peaceful, so we use it to represent low speed levels.

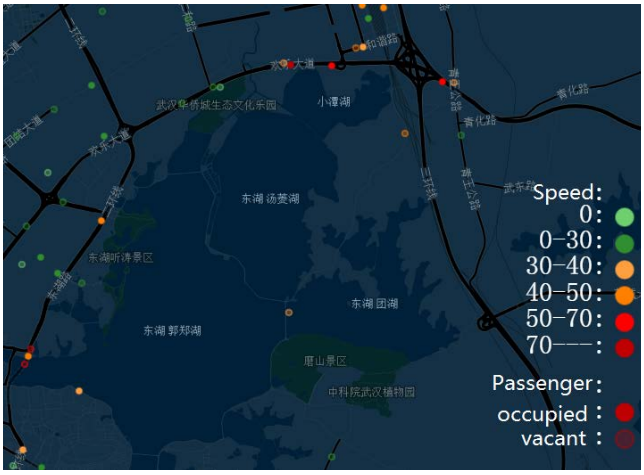

Figure 6. Visualization of speed and passenger information.

3.2.2 Two ways to visualize taxi trajectory data

When the map scale is very small, the data volume is very big and it can be difficult to fetch the trajectory data needed from the database for next period and display it immediately. To fix the problem, first we come up with the solution that the trajectory data for next period can be fetched when the current data is being displayed on the map. Thus, we adopt the asynchronous data loading strategy based on Ajax technique. Second, Lee (2007) proposed a trajectory data extraction method based on feature points. According to this research, we design a thinning algorithm to reduce the amount of data while retaining the main features of the taxi trajectory data. Fig. 7 is an illustration on how to pick up feature points.

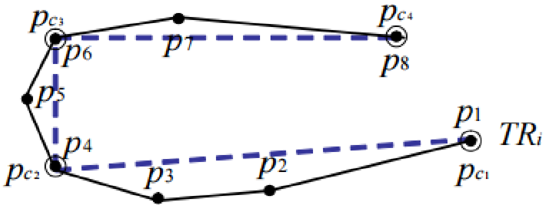

Figure 7. A trajectory and its segmentation.

In our work, we visualize the taxi trajectory data statically and dynamically. When the trajectory data is displayed statically, we first obtain the current map range of the boundaries of latitude and longitude and set the time range for 10 minutes, then these information will be used to access the database and we can get the trajectory data we need. After that, we select the first points of each taxi trajectory and display them on the map for 3 seconds. During 3 seconds, the trajectory data for next 10 minutes will be fetched from the database for next display on the map, which is actually the application of asynchronous data loading. When the trajectory data is displayed dynamically, the mechanism to load and display the trajectory data is almost the same. However, some differences do exist between these two ways. Firstly, the time interval between two adjacent points for a taxi trajectory on the map is 10 seconds rather than 10 minutes. Secondly, the interval between two adjacent displays is very small instead of 3 seconds to keep the taxi moving on the map. Thirdly, when the map scale is too small, the thinning algorithm we discussed above will be used to keep the smooth flow of taxi on the map.

3.3 Visualization for hotspots of pickup points

To display the hotspots for taxi pickup points properly, we first find the corresponding hotspots data stored in the database according to the current time and map scale. Then, each pickup hotspot is positioned by its latitude and longitude attributes. And the shape of each hotspot is a circle, whose radius is determined by the number of pickup points at the hotspot. The more pickup points are grouped, the larger the radius is.

In the visualization framework implemented in this paper, we visualize the hotspots data in one day and one week respectively. For the first case, each hotspot for pickup points is displayed in the same colour. For the second case, we display all the hotspots in 7 different colours, each of which corresponds to a day in a week.

We display the hotspots data in one hour every 3 seconds. For the convenience of user operation, a pause button is designed to stop the loop for detailed observation. When the user clicks one hotspot on the map, all the pickup points that make up the hotspot will be displayed. In this way, the user can clearly see the distribution of pickup points at any time and may find some potential information.

4. RESULTS AND ANALYSES

4.1 Taxi trajectory data visualization

4.1.1 Static visualization for trajectory data

Fig. 8 is a sample result of static visualization for trajectory data on the two sides of the Yangtze River. We can take this kind of visualization as one way to show the distribution of taxis. We divide the speed values into 6 levels. As we have discussed before, different colours represent different speed levels. Even though the time stamp at the bottom shows that it is 0 o'clock, there are many high-speed taxis on the streets. We infer that it has something to do with the Labour Day (May 1 in China, a national holiday). Another conclusion we can draw based on the figure is that the faster the taxi, the more likely there is a passenger on the taxi.

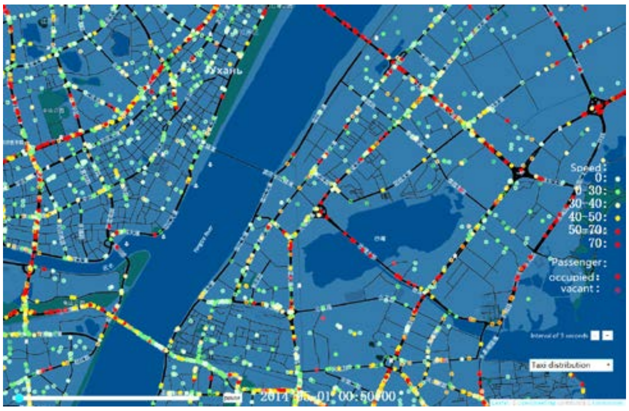

Figure 8. Taxi distribution at 00:50:00 on May 1.

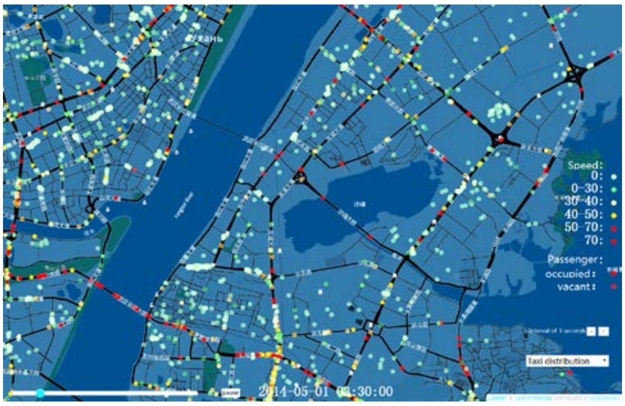

Figure 9 Taxi distribution at 03:30:00 on May 1.

Compared with the taxi distribution at 00:50:00 on May 1, the number of taxis becomes less and most taxis are at low speed at 03:30:00 on May 1. Besides, many taxis are empty at that time. Because most people will be asleep at 03:30 even on Labour Day. However, a small number of taxis are still at high speed, which may suggests that some people are active at late night. Finally, users can see the density of taxis on the streets easily through the static visualization of taxi trajectory.

4.1.2 Dynamic visualization for trajectory data

Taxi is an important means of transport for people, and its trajectory contains a wealth of crowd moving information. The activity of a single person reflects his/her own behavioural characteristics, while the activity of a large number of people in the same city reflects the social activity characteristics of the whole city. The dynamic visualization for trajectory data can be a tool for analysts to better understand the crowd mobility and find the hidden patterns behind it.

When we visualize the trajectory data dynamically, the time interval between two data displays is very short, which makes people have the impression that the taxis are moving on the map and an animation is being played. Since we don't use all the data when we visualize the trajectory data dynamically, the number of taxis displayed on the map is much less than that of static visualization. Fig. 10 shows a snapshot of one moment during the animation. When users want to stop the animation and make further inspection, a pause button can be clicked at the bottom of the map and then users can click any taxi to get more information about the taxi. Fig. 11 shows the query interface in our system. If the users are interested in the trajectory of a specific taxi, they can input the taxi ID and time in the interface, then a single trajectory for one specific taxi will be visualized dynamically on the map.

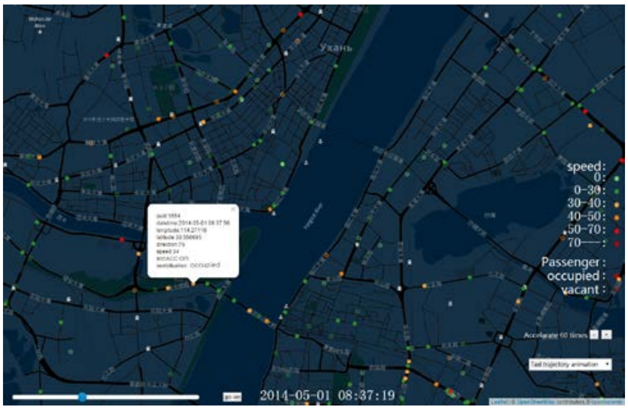

Figure 10. Detailed information of a taxi.

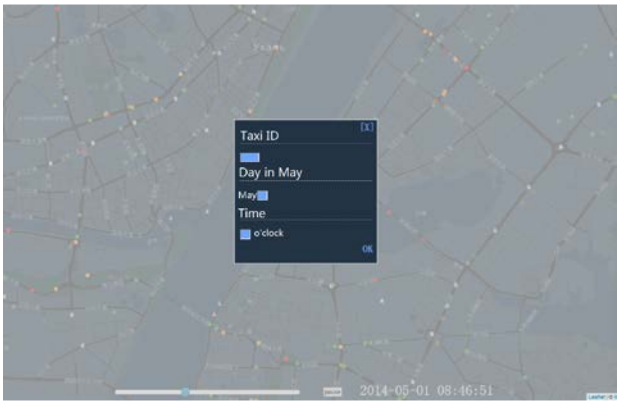

Figure 11. Display of the query interface.

4.2 Hotspots data visualization

4.2.1 Hotspots visualization for one day

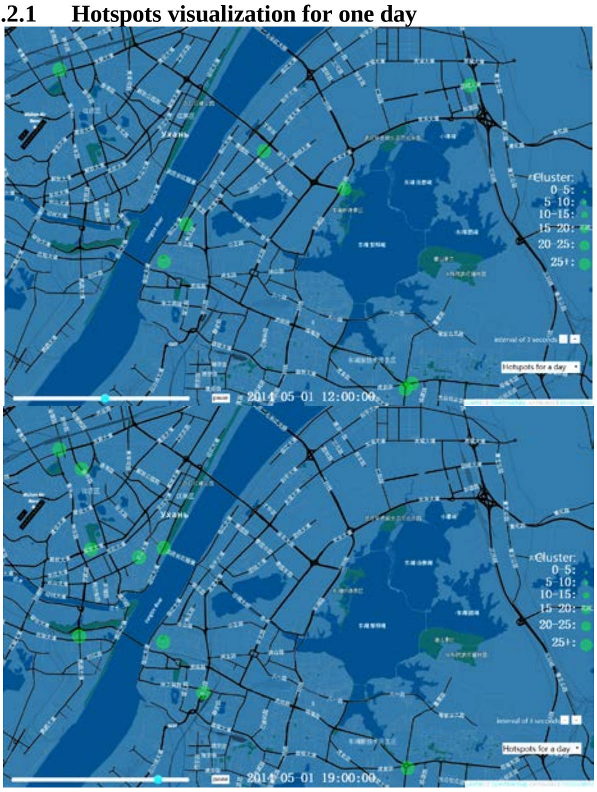

Figure 12. Hotspots for pickup sites at different time on May 1.

Fig. 12 is the visualization of hotspots data at different time on May 1. Each circle on the map is a hotspot for pickup points, and its radius is determined by the number of pickup points that make up the hotspot. All the hotspots on the map are in the same colour, indicating that the hotspots are on the same day. Wuchang and Hankou Railway Station are hotspots at 12:00 noon and 19:00 on May 1, which probably means that many people came to Wuhan city that day. Besides, Guanggu Square (a commercial zone) is another hotspot both at 12:00 noon and 19:00, indicating that there are many people in Guanggu Square on May 1.

4.2.2 Hotspots visualization for one week

Fig. 13 shows the visualization of hotspots data at different time for a week. In this case, we display all the hotspots in 7 different colours, each of which corresponds to a day in a week. And the radius of each hotspot is determined by the number of pickup points that make up the hotspot. No matter what time it is, there are always hotspots around the Wuchang, Hankou, Wuhan Railway Station and Guanggu Square. These four locations are the main taxi pickup points in Wuhan. Such information may be useful for the taxi drivers.

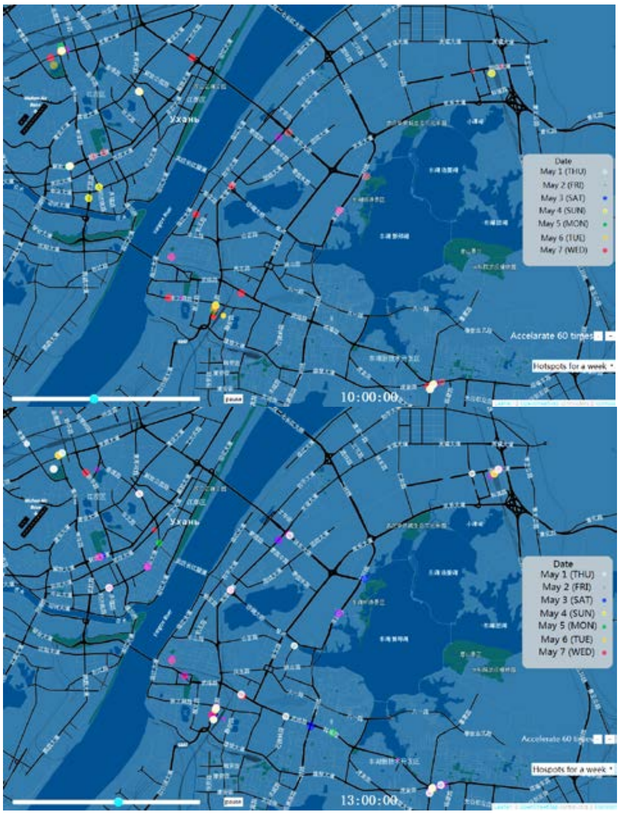

Figure 13. Hotspots for pickup sites at different time for a week

5. CONCLUSION AND FUTURE WORK

In this paper, we present a visualization system for taxi trajectory data. Compared with the existing visualization system, our system can display the speed information and passenger loading status successfully. Besides, our system can visualize the taxi trajectory data in four ways. First, the taxi trajectory are visualized statically and dynamically. Second, the hotspots data produced by clustering pickup points are visualized day by day and week by week respectively. After experimenting, it turns out that our system can help users analyse the taxi data intuitively and make further investigation. In the future, we would like to explore the potential of the system by making it real-time and meet the needs of intelligent transportation. Thus, the system will be more powerful and can help the government departments deal with traffic accidents.

REFERENCES

Andrienko, G., & Andrienko, N. 2008. A Visual Analytics Approach to Exploration of Large Amounts of Movement Data. International Conference on Visual Information Systems: Web-Based Visual Information Search and Management (Vol.5188, pp.1-4). Springer-Verlag.

Andrienko, G. L., Andrienko, N. V., Dykes, J., Fabrikant, S. I., and Wachowicz, M. 2008. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. Information Visualization, 7(3), 173-180.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., and Giannotti, F., 2009. Interactive visual clustering of large collections of trajectories. In Proceedings of IEEE Symposium on Visual Analytics Science and Technology, pp. 3–10.

Chen, W., Guo, F.Z., Wang, F.Y., 2015. A survey of traffic data visualization. IEEE Transactions on Intelligent Transportation Systems. 16 (6), 2970-2984.

Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., et al. 2014. Visualizing Hidden Themes of Taxi Movement with Semantic Transformation. Visualization Symposium. IEEE, 137-144.

Dodge, S., Weibel, R., and Lautenschütz, A.-K., 2008. Towards a taxonomy of movement patterns. Information Visualization, 7(7), 240–252.

Ester, M., peter Kriegel, H., S, J. and Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

Ferreira, N., Poco, J., et al. 2013. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. IEEE Transactions on Visualization and Computer Graphics, 19(12): 2149-2158.

Figueiredo, L., Jesus, I., Machado, J.T., Ferreira, J., and de Carvalho, J.M. 2001. Towards the development of intelligent transportation systems. IEEE Intelligent Transportation Systems. Proceedings, pp. 1206–1211.

Guo, H.Q., Wang, Z.C., Yu B.W., Zhao, H.J., Yuan, X.R., 2011. TripVista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. IEEE Pacific Visualization Symposium, 1-4 March, Hong Kong, China, pp. 163-170.

Hansen, C. D. and Johnson, C. R., 2004. The Visualization Handbook. San Diego, CA, USA: Academic.

Lee, J.G., Han, J.W. and Whang, K.Y., 2007. Trajectory clustering: a partition and group framework. ACM-SIGMOD International conference on Management of Data. ACM, 593-604.

Liu, L., Biderman, A., Ratti, C., 2009b. Urban mobility landscape: real time monitoring of urban mobility patterns. Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management. 1-16.

Liu S.Y., Pu J.S., Luo Q., et al. 2013. VAIT: A visual analytics system for metropolitan transportation. IEEE Transactions on Intelligent Transportation Systems, 14(4): 1586-1596.

Liu, S., Liu, Y., Ni, L., Fan, J. and M. Li. 2010. Towards mobility-based clustering. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 919–928.

Thomas, J. J. and Cook, K. A., 2005. Illuminating the Path: The Researchand Development Agenda for Visual Analytics. Richland, WA, USA: National Visualization and Analytics Center.

Tominski, C., Schumann, H., Andrienko, G. and Andrienko, N., 2012. Stacking-based visualization of trajectory attribute data. IEEE Transactions on Visualization and Computer Graphics. 18(12):2565–2574.

Vijverberg, J., de Koning, N., Han, J., de With, P., and Cornelissen, D., 2007. High-level traffic-violation detection for embedded traffic analysis. IEEE International Conference on Acoustics, Speech and Signal Processing, volume 2, pp. 793–796.

Von Landesberger, T., Andrienko, G., Andrienko, N., Tekusova, M. and Bremm, S., 2012. Visual analytics methods for categoric spatio-temporal data. Visual Analytics Science and Technology(Vol.48, pp.183-192). IEEE.

Yuan J., Zheng Y., Zhang C., et al., 2010. T-drive: Driving directions based on taxi trajectories. Proceedings of the 18th GIS Spatial International conference on advances in geographic information systems. ACM, 99-108.

Zhang, J., et al., 2011. Data-driven intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems. 12(4), pp. 1624–1639