Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 96 (2013) 2084 – 2095

Procedia
Social and Behavioral Sciences

13th COTA International Conference of Transportation Professionals (CICTP 2013)

Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data

Lin Xu*, Yang Yue, Qingquan Li

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing
Wuahn University,129 Luoyu Road, Wuhan 430079, China

Abstract

With the increasing amount of traffic information collected through floating car data, it is highly desirable to find meaningful traffic patterns such as congestion patterns from the accumulated massive historical dataset. It is however challenging due to the huge size of the dataset and the complexity and dynamics of traffic phenomena. A novel floating car data analysis method based on data cube for congestion pattern exploration is proposed in this paper. This method is different from traditional methods that depend only on numerical statistics of traffic data. The view of the event or spatial-temporal progress is adapted to model and measure traffic congestions. According to a multi-dimensional analysis framework, the traffic congestion event is first identified based on spatial-temporal related relationship of slow-speed road segment. Then, it is aggregated by a cluster style to get the traffic pattern on a different level of detail of spatial-temporal dimension. Aggregated location, time period and duration time for recurrent and important congestions are used to represent the congestion pattern. We evaluate our methods using a historical traffic dataset collected from about 12000 taxi-based floating cars for one week in a large urban area. Results show that the method can effectively identify and summarize the congestion pattern with efficient computation and reduced storage cost.

Keywords: floating car data; traffic congestion; spatial-temporal progress; clustering; data cube;

1. Introduction

The ability to accurately identify traffic pattern from massive historical traffic data, especially congestion pattern, is very important for traffic management. It can be used to reduce congestion, increase safety, and improve traffic forecasting accuracy. Among of all kinds of existing traffic data collection methods, floating car

* Lin Xu. Tel.: +86-27-6877; fax: +0-000-000-0000 .
E-mail address: xuling717@gmail.com

doi:10.1016/j.sbspro.2013.08.235

is a relatively reliable and cost-effective way to gather traffic data over a wide-area road network (Quiroga and Bullock, 1998). Speed and position data of floating cars, i.e. FCD (floating car data), is collected periodically (e.g., 30s), then processed to generate traffic information for each road segment at regular time intervals. With the growing popularity of FCD, the amount of FCD has begun to increase rapidly, even in a short period. For example, it can generate about 1GigaByte data for 10,000 cars in just one day. Existing research related with FCD are mainly about traffic status estimation, such as, travel time and travel speed, (Kerner and Demir et al., 2005; de Fabritiis and Ragona et al., 2008). Thus, how to archive and summarize massive historical FCD effectively, and extract meaningful traffic patterns from accumulated data to support decision making has become a significant challenge, considering the huge size of the dataset, complexity and dynamics of traffic phenomena.

In practice, most historical FCD or traffic data are organized and stored using RDBMS in which one row of relation table presents a GPS or traffic data record. And the summary description about traffic congestion pattern is usually given by aggregation of numeric measures, e.g. average travel speed and total congestion duration time (Smith and Lewis et al., 2003). Thus, traffic congestion is treated as an independent spatial-temporal fragment, which usually describes congested traffic status for one road segment in one time interval. In fact, urban traffic congestion is usually dynamic spatial-temporal progress or event in which several congested road segments are spatially close and temporally approximate. Moreover, it is associated with several contextual conditions (e.g., weather, time of day, and geographic location) and usually recurrent in the same or similar condition. So, identification and exploration of recurrent congestion progress or event can uncover periodicity of traffic congestion pattern, its evolution over time or space, and correlation with contextual features such as weather.

In this paper, we propose an approach that can identify these urban congestion patterns based on data cube. This method is different from traditional methods that depend only on numerical statistics of traffic data. The view of event or spatial-temporal progress is adapted to model and measure traffic congestions. The traffic congestion progress is firstly identified based on spatial-temporal related relationship of slow-speed road segment. Then, it is aggregated by a cluster style to get recurrent traffic congestion pattern. In order to exploration traffic congestion pattern on different spatial-temporal scale in an scale-able and efficient way, the multi-dimensional data organization and analysis methods based on data cube are introduced. The result shows it can effectively identify and summarize the congestion pattern with efficient computation and reduced storage cost. In the rest of the paper, Section 2 describes related work related FCD and introduces the multi-dimension analysis method based data cube. Section 3 and 4 describes the overall framework and the detailed algorithms separately. Section 5 presents comprehensive experimental results. Section 6 summarizes our conclusions and future work.

2. Related work

2.1. Traffic information collection, traffic congestion identification and traffic pattern analysis

Road traffic status estimation or traffic parameter data acquisition is usually the base of the all kinds of application of FCD. Much work has been done in this aspect, which includes map matching, route recovery and traffic parameter computation, etc (Turner and Eisele et al., 1998; Kerner and Demir et al., 2005; de Fabritiis and Ragona et al., 2008). Moreover, a significant effort has been made on traffic congestion or incident detection. Early works on incident detection (Persaud, 1990), e.g. California algorithm and McMaster algorithm, focused on developing incident detection rules based on the change of traffic volume, speed or occupancy. More recent work has used neural networks, fuzzy logic and support vector machines to attack the problem (Lu and Cao, 2003; Porikli and Li, 2004; Thianniwet and Phosaard et al., 2010). On the other hand, historical traffic data archiving and pattern analysis attracts increasing attention because of the enormous quantities of data collected by all kinds of intelligent transportation systems. Pan and Shahabi (2010) proposed a traffic data stream summary method based on spatial-temporal redundancy/correlation of sensor data.

Different from these existing methods, which focus on the traffic status of single road segment in one-time intervals, the interest of our methods is the progress of traffic congestion, which usually involves several near road segments and lasts for some time. Because the identification and exploration of congestion progress can provide more concise and meaningful information or pattern, e.g. the start/end time, spatial related road segment and detail description of the congestion progress.

2.2. Multidimensional data Analysis based on data cube

Facing detailed and continuously increased data, the ability of having large data set summarized in concise and succinct terms is of great necessary for providing the overall picture of the massive historical data at hand. Multidimensional data cube model is used for this kind of data analysis. It supports data summarization by aggregations at different levels of granularity from different views or dimensions. Data cube can be seen as a conceptual extension of a two-dimensional spreadsheet or cross-table (Pedersen and Jensen, 2001). Instead of displaying data with multiple cells into rows and columns, the data cube organizes data into all possible cross-tabulations and aggregations with respect to the data dimensions and user-specified aggregation hierarchies (Codd and Codd et al., 1993; Gray and Chaudhuri et al., 1997; Inmon, 2005; Song and Miller, 2012). For example, considering a multidimensional historical traffic data database, given a particular measure (e.g., ‘‘volume’’) and some dimensions of interest (e.g., ‘‘location’’, ‘‘hour,’’ ‘‘day’’), a data cube returns the power set of all possible aggregations of the measure with respect to the dimensions of interest. These include aggregations over 0 dimension (e.g., ‘‘total volumes’’), one dimension (e.g., ‘‘total volumes by location,’’ ‘‘total volumes by hour’’, ‘‘total volumes by day’’), two dimensions (e.g., ‘‘total volumes by location and hour’’) and up to n dimensions.

Traffic data is typical multi-dimensional data, which has spatial, temporal, and some related dimensions, e.g. weather. So, many researchers have used data cube/warehouse to archive traffic data and get aggregated summary information of travel speed or traffic volume (Shekhar and Lu et al., 2002; Smith and Lewis et al., 2003; Pfoser and Tryfona et al., 2006; Song and Miller, 2012). Different from these methods based on aggregation of numeric traffic measure, our work focuses on identification and aggregation of traffic congestion progress and The work of Gonzalez and Han (2011), Tang and Yu (2012) is the most similar to us, but we improve it by using level of granularity of road network.

3. Framework

Traffic congestion events occur when traffic demand is greater than the available road capacity. It is a dynamic spatial-temporal process. Congestion usually starts from a single road segment, then expands along the road and influences the nearby roads. As time passes by, those congested fragments shrink slowly, eventually reduce their coverage and finally disappear. In this process, congested road segment are spatially close and temporally approximate. At the same, traffic congestion is related with many factors, and usually shows difference at different location or in different time period and similarity or recurrence at similar conditions.

In order to explorer spatial-temporal characteristics of urban traffic congestion, we propose a multi-dimension analysis framework of traffic congestion based on historical floating car data set. In this framework, a traffic congestion progress is identified as a ‘congestion event’, which consists of spatially and temporally, connected congested road segments. The major components in this framework are illustrated in Figure 1.The **traffic data acquisition and pre-processing module** is in charge of collecting floating car data from GPS-equipped taxi, map matching these data to road network and then getting the travel speed and traffic state on each road segment by aggregation of sample data. The **data-cubing** module firstly takes abnormal or congested traffic data record from the data pre-processing layer and then identify the congestion event based on spatial-temporal connected relationship. Finally, the congestion event is aggregated by a cluster style to get summarized description about recurrent congestion pattern. There has already many related research in the field of traffic data acquisition based on floating car data, so we will focus on the identification and aggregation of congestion event in following sections.

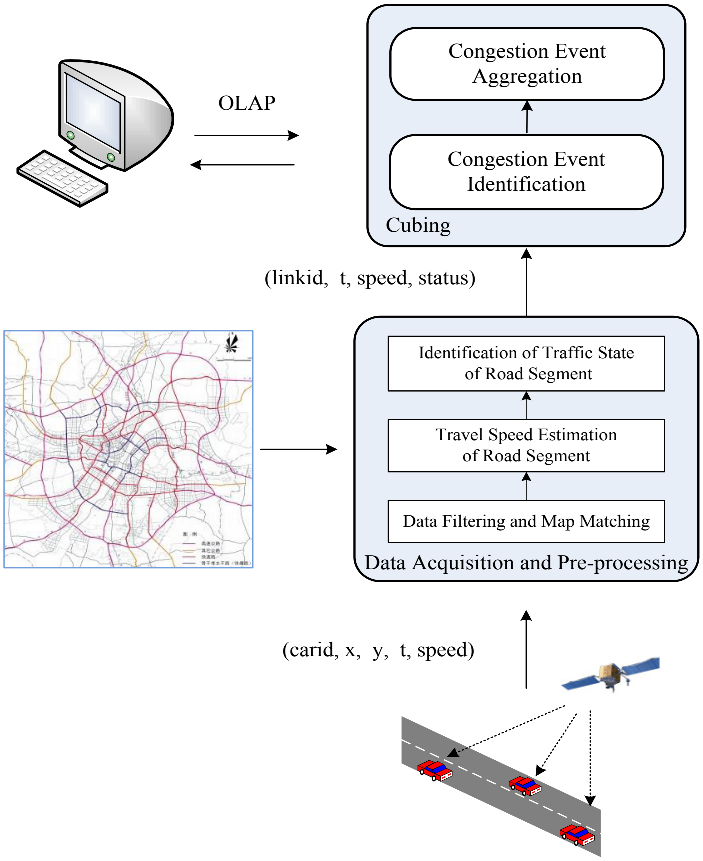

OLAP
Congestion Event Aggregation
Congestion Event Identification
Cubing
(linkid, t, speed, status)
Identification of Traffic State of Road Segment
Travel Speed Estimation of Road Segment
Data Filtering and Map Matching
Data Acquisition and Pre-processing
(carid, x, y, t, speed)

Figure 1. Traffic congestion analysis framework based FCD

4. Congestion event identification and aggregation

In this section, we will formally define the ‘congestion event’, and then identify and aggregate it in multi-level of road network.

4.1. Congestion Event

In the formation of a traffic congestion, the slow traffic records are spatially close and timely relevant to each other if they are in the same congestion, as illustrated in figure 2. Based on this spatial-temporal related relationship, we introduce the concepts of congestion event as following:

Definition 1 (Direct Spatial-Temporal Related): Let $r_i(s_i, t_i)$ and $r_j(s_j, t_j)$ be two slow traffic records, where s_i and s_j are the corresponding road segments, t_i and t_j are the timestamps, respectively; Let T_d be the distance threshold and T_t be the time threshold; r_i and r_j are direct spatial-temporal connected if dist (s_i, s_j) < T_d and $|t_i – t_j|$ < T_t

Definition 2 (Spatial-Temporal Related): Let r_1 and r_n be slow traffic records, if there is a chain of records r_1, r_2, . . . ,r_n, such that ri and r_{i+1} are direct spatial-temporal related; then r_1 and r_n are congestion reachable.

Definition 3 (Congestion Event): Let R be the set of traffic records, T_d be the distance threshold and T_t be the time threshold. Congestion Event E is a subset of R satisfying the following conditions:

1) ∀ r_i, r_j: if $r_i \in$ E and r_j is spatial-temporal connected from ri w.r.t. T_d and T_t, then $r_j \in$ E.

2) ∀ r_i, $r_j \in$ E: r_i is spatial-temporal connected to r_j w.r.t. T_d and T_t.

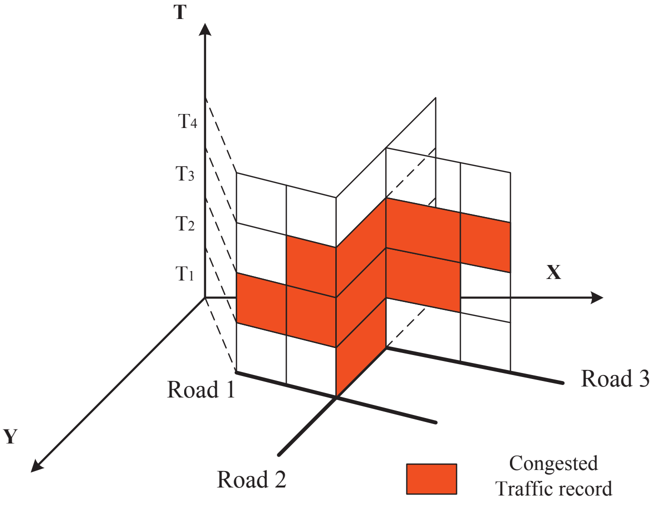


Figure 2. Spatial-temporal related record in congestion

Definition 4 (Congestion Event Feature): In order to summarize and measure the congestion event, a feature vector is defined as

$$CE = \{\mathbf{SF},\ \mathbf{TF},\ \mathbf{STF}\}$$

Using three related features: (1) SF, the spatial feature which describes location of the congestion, such as street or region name, spatial coordinate range; (2) TF, the time feature which describes the start and end time for the congestion; (3) STF, the spatial-temporal aggregated feature which describe aggregated attributes of the progress of formation and propagation of the congestion

4.2. Congestion Event Identification in Road Network

4.2.1. Road Network Spatial Granularity and Linear Reference System

Aggregation according to different level of spatial granularity can reveal different views on different level of detail. The higher the level of spatial granularity the more generalized the result of aggregation is. As for urban road network, the level usually consists of road segment, road and network. As shown in figure 3, a network consists of roads, while a road consists of road segment. We use a ‘node-link’ model to represent urban road network. A ‘Node’ represents the road junctions or starting/ending point of a road segment. A ‘Link’, the part between two connected nodes, represents a road segment. A ‘Multi-Link’, which consists of a set of continuous link, represents one road. The set of all multi-links or all links represents the ‘Network’.

Given that a traffic study is mainly conducted on roads, a linear reference system (LRS) rather than a general coordinate reference system is more useful. A LRS is used to associate attributes or events to locations or portions of linear features, such as highways. There are three major highway linear reference methods – road name and mile mark, control section, and link and node systems. The reference method of road name and mile mark is employed in this study. Then the location of road segment can be described as following:

Link Location = {MultiLinkID, offset1, offset2}

The ‘offset1’ / ‘offset2’ is offset of start/end node for the road segment started from the start point of the road. Figure 3 shows an example road network, which consists of two arterial roads and two minor roads. Road A is 5km long and have 15 road segments, each of which can be referenced by road name and offset from the start point.

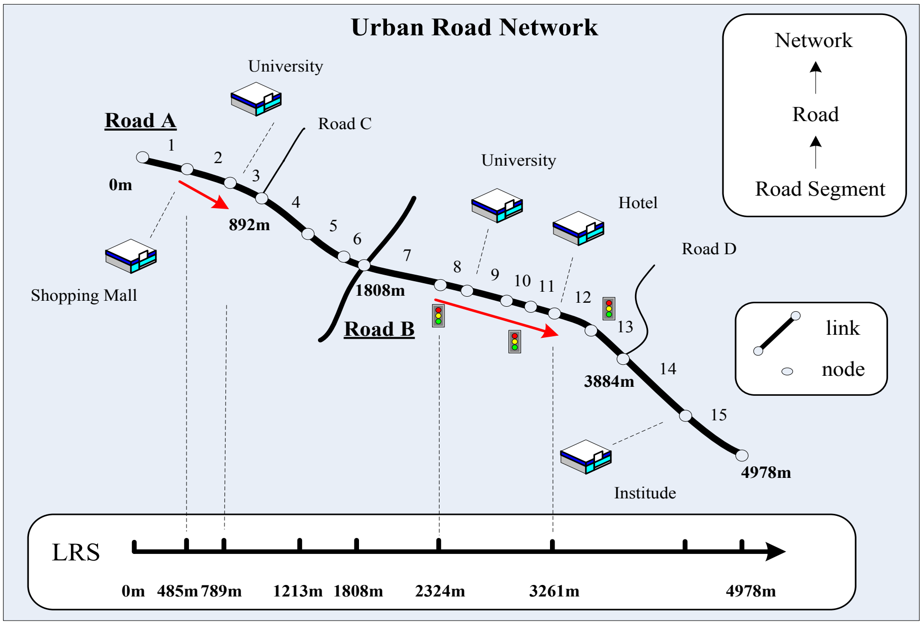


Figure 3. Urban road network space and Linear reference system

4.2.2. Traffic Congestion Event Identification in Multi-Level Road Network

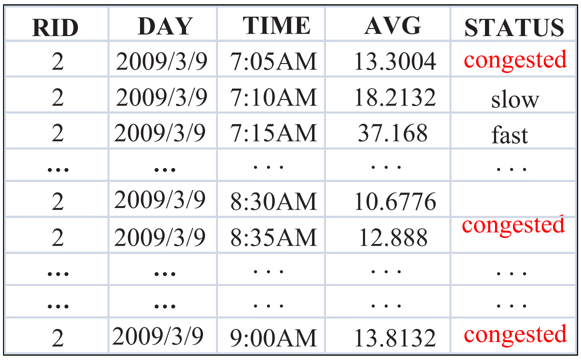
According to the definition of congestion event and road network spatial granularity, we propose our multi-level congestion event identification algorithms. The main idea of this algorithm is to explore the congestion event at different level of spatial granularity and compute the event at higher level using events at low level with regard to the given spatial and temporal distance threshold. The detail of algorithms will be elaborated according to different level in following.

◆ Road Segment Level

A congestion event on road segment shows the time intervals during which the congestion state is continuously observed. The feature vector of the congestion event can be specified as:

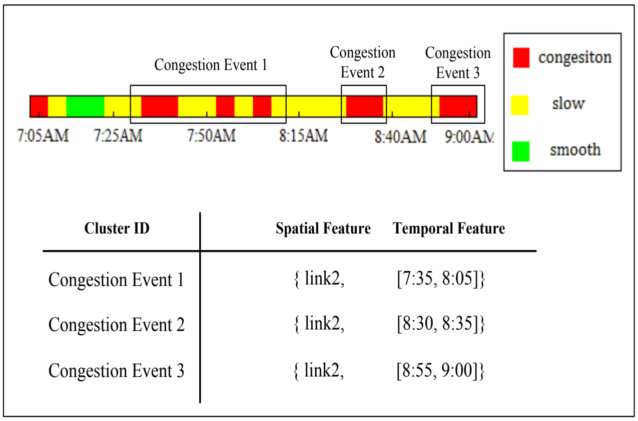
LinkCE = {LinkID, (StartTime, EndTime) }

according to the specification of the Definition 4 *Congestion Event Feature* in Section 4.1. The LinkID describes the spatial location, and the ‘StartTime/EndTime’ is temporal intervals of the traffic congestion. The congestion event identification algorithm randomly picks a congested traffic record from the dataset, then searches and marks all spatial-temporal related records. The computation of spatial distance is the main cost. However, it should be noted that it is much simplified by use of level of granularity, because the spatial distance is zero at the link level. That is one of advantages provided by our methods. Figure 4 give an example of congestion event identification for the segment 2 on road A of figure 3. Three congestion events are identified from 24 detail traffic record based on temporal approximation. Summarized information about traffic congestion is given for each event. Taking the No.1 congestion event for example, it tells us the congestion happened between 7:35am and 8:05 on segment 2 and the duration time for serious congestion is 20mins.



RID	DAY	TIME	AVG	STATUS
2	2009/3/9	7:05AM	13.3004	congested
2	2009/3/9	7:10AM	18.2132	slow
2	2009/3/9	7:15AM	37.168	fast
…	…	. . .	. . .	. . .
2	2009/3/9	8:30AM	10.6776	congested
2	2009/3/9	8:35AM	12.888	
…	…	. . .	. . .	. . .
…	…	. . .	. . .	. . .
2	2009/3/9	9:00AM	13.8132	congested

a. Traffic data record extracted floating car data



Cluster ID	Spatial Feature	Temporal Feature
Congestion Event 1	{ link2,	[7:35, 8:05]}
Congestion Event 2	{ link2,	[8:30, 8:35]}
Congestion Event 3	{ link2,	[8:55, 9:00]}

b. Traffic congestion event identified from traffic data record

Figure 4. Traffic data record V.S. Traffic congestion event

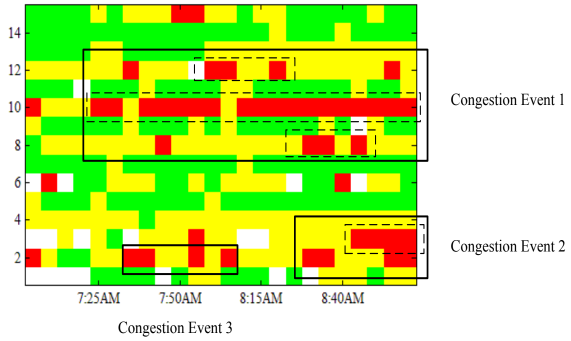
◆ Road Level

A congestion event on one road shows a spatial and time interval, in which the spatial-temporal related congestion state is identified. The feature vector of the congestion event can be described as:

$$\mathbf{MLinkCE = \{SF, TF, \ STF\}}$$

The spatial feature SF = (Multi-Link, StartOffset, EndOffset) describes the linear location of the congestion event on the road. It is an interval from the 'StartOffset' to the 'EndOffset' started from the start point of the 'Multi-Link'. The temporal feature TF= (StartTime, EndTime) is the temporal interval of the whole of congestion progress. The spatial-temporal feature STF = $\{(\text{Link}_1,\text{Dur}_1),\ldots,(\text{Link}_n, \text{Dur}_n)\}$ describes the aggregated duration time of congestion on each link within the SF and the TF.

The congestion event identification on roads is based on the congestion events of all road segments on the same road. Two congestion events on different road segments but within the same road can merge a new congestion event when they are spatial adjacent and intersects in time range. The low cost of distance computation due to the linear location reference improve the algorithm efficiency.



Cluster ID	Congestion Event 1
Spatial Feature	(Road A, 2324m~3561m)
Temporal Feature	[7:25, 9:00]
Aggregated Feature	{ (link8, 15min), (link8, 95min), (link8, 25min) }

Figure 5. Traffic Congestion Event on Multi-Link Level

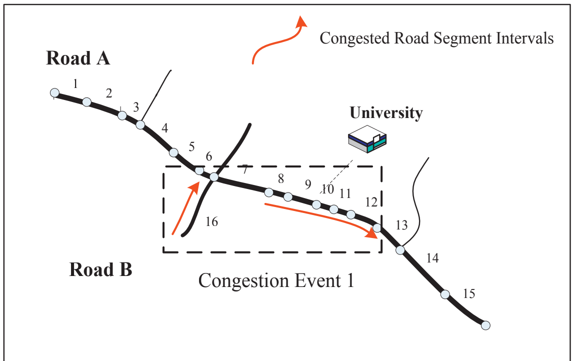
Figure 5 give an example of congestion event identification for the road A in figure 3. Merging of events on all road segments of road A identifies three congestion events. More summarized and meaningful information about traffic congestion is given for each event.

◆ Road Network Level

A congestion event on whole road network shows the spatial and time region where the congestion progress happened. The feature vector of congestion event can be described as:

$$\mathbf{NetworkCE = \{SF, TF, STF\}}$$

The spatial feature SF = MBR describes the Minimum Bounding Box (MBR) of the congestion event on the 2-dimension road network space. The temporal feature TF = (StartTime, EndTime) is the temporal interval of the whole of congestion progress. The spatial-temporal feature STF = $\{(\mathrm{Link}_1,\mathrm{Dur}_1),\ldots,(\mathrm{Link}_n, \mathrm{Dur}_n)\}$ describes the aggregated duration time of congestion on each link within the SF and the TF. The congestion events on the whole network space is obtained by merging of congestion events on all roads, which are spatially close and temporal intersected. Figure 6 give an example of congestion event identification on the road network.



Cluster ID	Congestion Event 1
Spatial Feature	MBR
Temporal Feature	[7:00, 9:00]
Aggregated Feature	{ (link8, 15min), (link10, 95min), (link8, 12min) , (link16, 60min) }

Figure 6. Traffic Congestion Event on the Road Network

4.3. Traffic Congestion Event Aggregation

The objective of traffic congestion event aggregation is to identify recurrent congestions which appear on different days, but at around the same location and time of day. So, the aggregation of traffic congestion event is different from traditional numeric aggregation function, e.g. average, sum or min/max. In this paper, we use a clustering-style aggregation method to get aggregated congestion events. The first task of congestion event aggregation is to measure the similarities between two congestion events. In this paper, we use the Jaccard similarity to measure spatial and temporal similarity separately as shown in Equation 1 and Equation 2. Jaccard similarity measures the percentage of the overlapping areas or intersected time range of the two events in their union coverage. With the spatial-temporal similarity definition, we use the 'ST-DBSCAN' clustering algorithm to aggregate congestion events.

$$\mathbf{Sim_{sf}}\,(\mathbf{CF1},\mathbf{CF2}) = \frac{\mathbf{SF1} \cap \mathbf{SF2}}{\mathbf{SF1} \cup \mathbf{SF2}} \qquad (1)$$

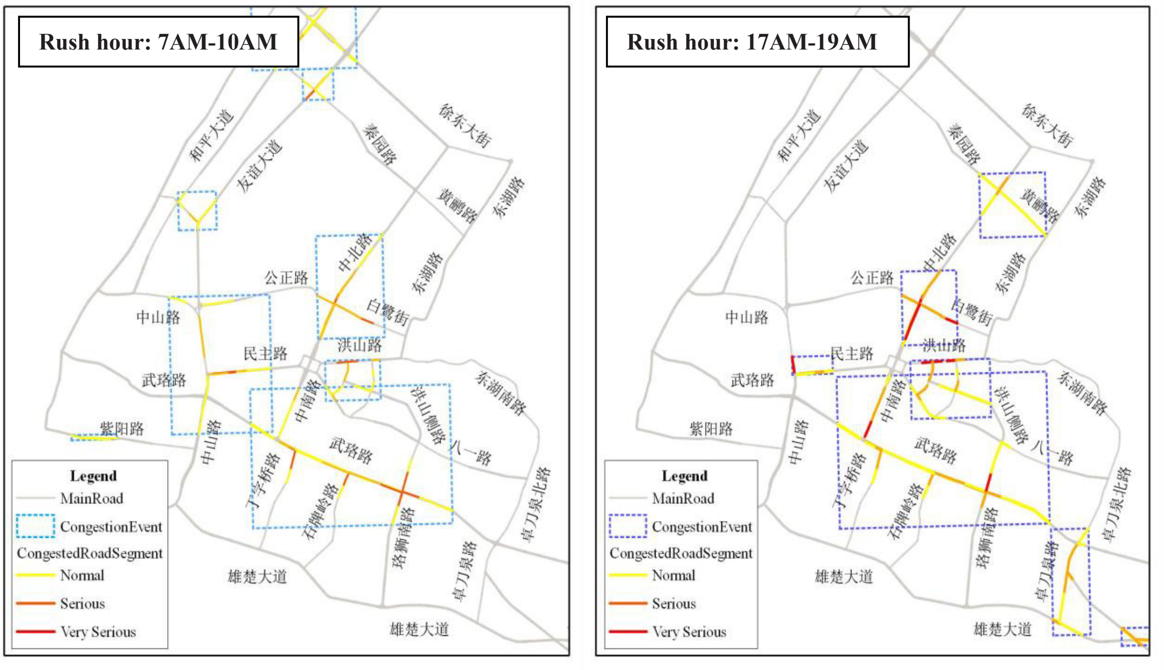
$$\mathbf{Sim_{tf}}\,(\mathbf{CF1},\mathbf{CF2}) = \frac{\mathbf{TF1} \cap \mathbf{TF2}}{\mathbf{TF1} \cup \mathbf{TF2}} \qquad (2)$$

5. Experiment

In this section we perform a brief analysis of our algorithms applied to historical floating car data sets for the city of Wuhan, China. All the experiments were conducted on an Intel Core Dual processor running at 2.53 GHz, with 4 GB of RAM. All the algorithms are implemented in Java on Eclipse 3.7.1 platform with JDK 1.6.0.

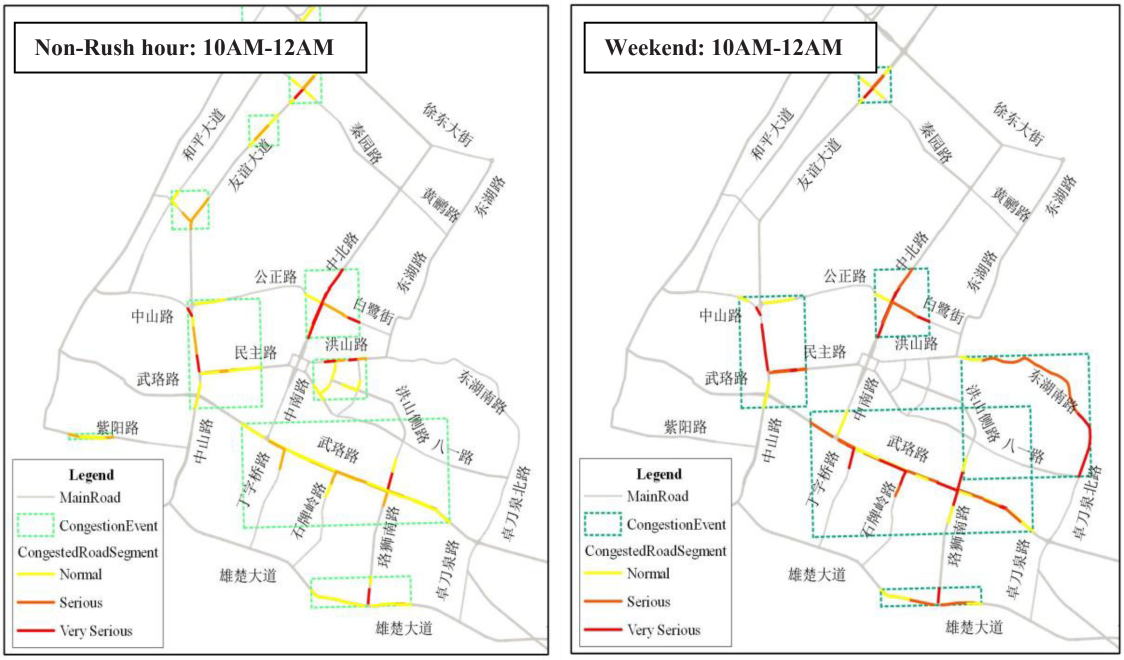
We use the floating car data set of the city of Wuhan in China for one week started from Mar.9th, 2009. This dataset is collected from a fleet of 12,000 taxis, which operates, on a large urban road network with 13,413 nodes and 17,620 links. With considering the characteristics of floating car samples, we choose a sub-network in downtown area. It consists of 37 main roads and includes 924 links, 798 node and 58 multi-links. Each car generates a record every 40s, and the data set takes 1.24GB storage space for 14,400,000 records in one day. The format of raw data is (*carid*, *x*, *y*, *t*, *speed*, *status*), where the *carid* is the identifier of each car, *x* and *y* is longitude and latitude of the location at *t*, *speed* and status are the travel speed and occupation status at the location. By data pre-processing and filter, e.g. map matching, traffic status estimation, we get the traffic data for each link in 5 minutes interval in the format (*linkid*, *t*, *speed*, *status*) which describes the identifier of link, travel speed and traffic status of the link at the time interval.

Figure 7 shows spatial-temporal region of recurrent traffic congestion event on different time periods. Figure 7.a and Figure 7.b shows recurrent congestion at rush hour of week day. Figure 7.c shows recurrent congestion at non-rush hour. Figure 7.d shows recurrent congestion at weekend.



a. Rush hour at morning

b. Rush hour at evening



c. Non-rush hour at morning

d. weekend

Figure 7. Recurrent Traffic Congestion Eevent on the Road Network

6. Conclusion

In this paper, we propose a historical floating car data analysis method based on data cube to identify and explore urban traffic congestion pattern. Different from traditional methods based on numerical statistics of traffic parameter data, traffic congestion is seen as dynamic spatial-temporal progress. It is modeled and measured as a ‘congestion event’, and then aggregated at different level of granularity of space and time. In order to implement this kind of cube, we propose a level-by-level computation method, which make full use of level of granularity of road network space, and linear referencing method. The evaluation on real FCD data set show it is an effective and efficient method.

References

Turner, S. M. and W. L. Eisele, et al. (1998). *Travel Time Collection Handbook*, Texas Transportation Institute.

Quiroga, C. A. and D. Bullock (1998). "Travel time studies with global positioning and geographic information systems: an integrated methodology." *Transportation Research Part C: Emerging Technologies* **6** (1-2): 101-127.

Kerner, B. S. and C. Demir, et al. (2005). Traffic state detection with floating car data in road networks. *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria.

de Fabritiis, C. and R. Ragona, et al. (2008). Traffic Estimation And Prediction Based On Real Time Floating Car Data. *Proceedings of the 11th International IEEE Conference On Intelligent Transportation Systems.*

Persaud, B. N. and F. L. Hall, et al. (1990). "Congestion identification aspects of the McMaster incident detection algorithm." *Transportation Research Record*(1287).

Parkany, E. and C. Xie (2005). A Complete Review of Incident Detection Algorithms & Their Deployment: What Works and What Doesn, *Transportation Center, University of Massachusetts.*

Lu, J. and L. Cao (2003). Congestion evaluation from traffic flow information based on fuzzy logic. *Intelligent Transportation Systems*.

Porikli, F. and X. Li (2004). Traffic congestion estimation using HMM models without vehicle tracking. *IEEE Intelligent Vehicles Symposium*.

Thianniwet, T. and S. Phosaard, et al. (2010). Classification of Road Traffic Congestion Levels from Vehicle’s Moving Patterns. *Electronic Engineering and Computing Technology*. S. Ao and L. Gelman, Springer Netherlands. **60**: 261-271.

Codd, E. F. and S. B. Codd, et al. (1993). Providing OLAP to User-Analysts: An IT Mandate, Codd & Date, Inc.

Gray, J. and S. Chaudhuri, et al. (1997). "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals." *Data Mining and Knowledge Discovery* **1** (1): 29--53.

Inmon, W. H. (2005). *Building the Data Warehouse*, Wiley Publishing, Inc.

Pedersen, T. B. and C. S. Jensen (2001). "Multidimensional database technology." Computer **34** (12): 40--46.

Shekhar, S. and C. T. Lu, et al. (2002). CubeView: a system for traffic data visualization. ITSC 2002, Singapore.

Smith, B. L. and D. C. Lewis, et al. (2003). "Design of archival traffic databases: Quantitative investigation into application of advanced data modeling concepts." *Transportation Research Record: Journal of the Transportation Research Board* **1836**: 126-131.

Pfoser, D. and N. Tryfona, et al. (2006). Dynamic Travel Time Maps - Enabling Efficient Navigation. SSDM 2006, Vienna, Austria, *IEEE Computer Society*.

Song, Y. and H. J. Miller (2012). "Exploring traffic flow databases using space-time plots and data cubes." *Transportation* **39**: 215-234.

Gonzalez, H. and J. Han, et al. (2011). "Multi-Dimensional Data Mining of Traffic Anomalies on Large-Scale Road Networks." Transportation Research Record: *Journal of the Transportation Research Board*, Volume 2215: 75-84

Tang, L. A. and X. Yu, et al. (2012). Multidimensional Analysis of Atypical Events in Cyber-Physical Data. ICDE 2012, Washington, DC, USA, *IEEE Computer Society*.