A Framework for Large-Scale Train Trip Record Analysis and Its Application to Passengers' Flow Prediction after Train Accidents

Daisaku Yokoyama[1], Masahiko Itoh[1], Masashi Toyoda[1], Yoshimitsu Tomita[2], Satoshi Kawamura[2,1], and Masaru Kitsuregawa[3,1]

[1] Institute of Industrial Science, The University of Tokyo
{yokoyama,imash,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp
[2] Tokyo Metro Co. Ltd.
{y.tomita,s.kawamura}@tokyometro.jp
[3] National Institute of Informatics

Abstract. We have constructed a framework for analyzing passenger behaviors in public transportation systems as understanding these variables is a key to improving the efficiency of public transportation. It uses a large-scale dataset of trip records created from smart card data to estimate passenger flows in a complex metro network. Its interactive flow visualization function enables various unusual phenomena to be observed. We propose a predictive model of passenger behavior after a train accident. Evaluation showed that it can accurately predict passenger flows after a major train accident. The proposed framework is the first step towards real-time observation and prediction for public transportation systems.

Keywords: Smart card data, Spatio-Temporal analysis, Train transportation, Passenger behavior.

1 Introduction

Public transportation systems play an important role in urban areas, and efficient and comfortable transportation is highly demanded, especially in megacities such as Tokyo. Tokyo has one of the most complex train systems in the world, so a major disruptive event can cause congestion and disruption over a wide area. The effect of such events is hard to predict, even for the operating companies.

Our goal is to implement the ability to analyze and predict passenger behaviors in a complex transportation system. In particular, we want to implement a system for understanding daily passenger flows and event-driven passenger behaviors, suggesting itineraries, and preparing for events.

Understanding Daily Passenger Flows. Operating companies want to know how many passengers are using their stations, lines, and trains. An understanding of the spatio-temporal demands of their passengers would help them

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 533–544, 2014.

adapt their train operations to the demands. An understanding of the demands would also enable the passengers to avoid crowded trains by enabling them to change to a less-crowded route or a more favorable departure time.

Understanding Event-Driven Passenger Behaviors. Events, such as natural disasters, public gatherings, and accidents, can create unusual passenger behaviors. To help them recover from disruptive events, operating companies need to know where congestion has occurred and how many passengers are present when an event occurs. Since various types of events occur repeatedly, understanding the changes in traffic that occurred with previous events would help them prepare for the next occurrence. Passengers could also use such information to avoid congestion after a disruptive event.

Suggesting Itineraries. If passenger flows could be observed in real time and if future flows could be predicted, operating companies could recommend itineraries to their passengers that would make their trips more convenient and comfortable. Since the current IT infrastructure for train systems in Japan does not support such real-time monitoring, we used a large-scale dataset of train trip records created from smart card data and examined the feasibility of flow prediction. Such recommendations could make the transportation system more efficient.

Preparing for Disruptive Events. Although operating companies already have some knowledge about the effects of major disruptive events, they still do not know what would happen if two or more of them occurred simultaneously. Such knowledge would help them allocate sufficient staff, trains, and other resources. It would also help them educate their staff by emulating arbitrary events.

Our contribution of this paper is as follows:

1. We propose a framework for analyzing large-scale train trip records as the first step to our goal. While there are various information sources, such as the number of passengers at stations, and train operation logs, we used smart card data for passengers using the Tokyo Metro subway system as such data reflects actual passenger demand. Within this framework, we developed a method for deriving passenger flows from the origin-destination records created from the smart card data and a function for visualizing unusual phenomena.
2. We propose a method to predict how passengers behave after an accident. Our method is based on two models: passenger demand and passenger behavior. The demand model is constructed using the passenger flows derived from the origin-destination records. The behavior model is used to predict passenger behavior after an accident. It uses the Abandonment Rate as a parameter. Our prediction method is based not simply on the results of simulation but on actual trip records.
3. We evaluate the accuracy and effectiveness of the proposed method by comparing the trip records after two major accidents. We find that the prediction accuracy could be improved by using an appropriate Abandonment Rate determined from historical post-accident data.

We describe related work in Section 2 and explain our goal of passenger flow analysis and our developed framework for large-scale trip record analysis in Section 3. Our method for predicting passenger behavior after an accident is explained in Section 4. In Section 5, we evaluate the accuracy and usefulness of the proposed method on the basis of our analysis of smart card data related to two major accidents. Section 6 summarizes the key points and mentions future work.

2 Related Work

Transportation log data, including data from operation logs, smart card logs, and equipment monitoring logs has been analyzed in various studies. Ushida et al. used train operation data to visualize and identify delay events and created chromatic diagrams for one line in the Tokyo Metro subway system with the goal of generating a more robust (delay-resistant) timetable [7]. Smart card data has been used as a data source to analyze the operation of a public transportation system [4]. Trépanier et al. used an alighting point estimation model to analyze passenger behavior from incomplete trip records created from smart card data[6]. Ceapa et al. used oyster (smart card) data to clarify passenger flow congestion patterns at several stations in the London Tube system for use in reducing congestion [1]. Their spatio-temporal analysis revealed highly regular congestion patterns during weekdays with large spikes occurring in short time intervals. Sun et al. provided a model for predicting the spatio-temporal density of passengers and applied it to one line in the Singapore railway system [5].

Previous work using smart card data focused only on a single line or a few stations, mainly because the trip records created from the data did not include transfer station information. We overcame this limitation by determining the most probable transfer station(s) for each trip record created from the origin and destination information and were thus better able to analyze how the effects of a disruptive event in a metro network propagate.

3 Framework for Analyzing Trip Records

The system we are constructing for analyzing train trip records currently uses data collected each night rather than in real time. Our immediate aim is to evaluate the effectiveness of our approach by analyzing historical data.

3.1 System Overview

Entrance and exit information is obtained each time a passenger uses a smart card to enter or exit a station gate (wicket). This information is aggregated on a central server each night. Our system can be used to analyze this information.

The system creates a trip route (start point, transfer points, and end point) for each record and estimates the passenger flows – how many passengers traveled a certain section at a certain time (as explained in detail in Section 3.3).

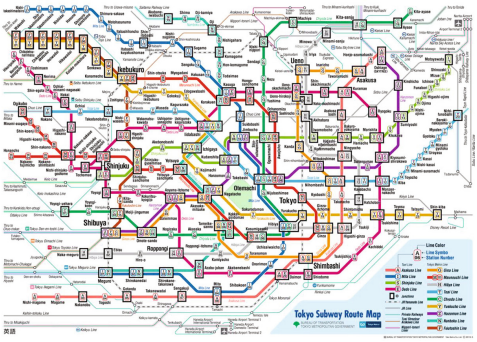

Fig. 1. Tokyo subway map[1]

By analyzing these flows, we can identify disruptive events resulting from various causes. Our interactive visualization framework can be used to identify and understand such events (as explained in Section 3.4).

Our method for predicting passenger behavior after an accident uses a passenger demand model and a passenger behavior model, which describes the effects of an accident (as described in Sections 4.1 and 4.2). Using these models and current traffic condition information, we can predict short-term passenger behavior.

We plan to make a passenger flow simulator that uses the behavior model. We also plan to combine real-time observation data with simulation data to make the predictions more accurate.

3.2 Smart Card Data

We used two years' worth of trip records for the Tokyo Metro subway system created from smart card data. As shown in Figure 1, the train system in Tokyo has a complicated route structure, consisting of lines of various railway companies including Tokyo Metro, Toei Subway, Japan Railway (JR), and many private railroads. We analyzed the Tokyo Metro trip records for almost all of the Tokyo business area, covering 28 lines, 540 stations, and about 300 million trips. The records included lines and stations besides Tokyo Metro ones if passengers used lines of other railway companies for transfers.

In our experiments, we use passengers log data from anonymous smartcards without personal identity information, such as, name, address, age, and gender. From each record, card ID is eliminated. Each record consisted of the origin, destination, and exit time. Since transfer information was not included, we estimated the probable route for each trip (as explained in Section 3.3).

During the week, the trains are mainly used by people going to or returning home from work while they are used more generally on weekends. We thus

[1] `http://www.tokyometro.jp/en/subwaymap/index.html`

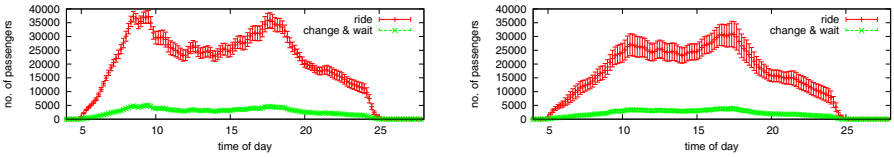

Fig. 2. Average and standard deviation of number of passengers over one year (Apr. 2012 to Mar. 2013) for weekdays (left) and weekends and holidays (right); "ride" represents number of passengers riding on a train, "change & wait" represents number of passengers waiting to change trains, error bars indicate standard deviation.

separated the data between weekdays and weekends and analyzed the two sets independently. National holidays and several days during vacation seasons were treated as weekend days.

Passenger behavior was assumed to follow periodic patterns, especially daily ones, so we statistically analyzed these data to identify the patterns. Figure 2 shows the average and standard deviation of the number of passengers over one year. The plot points were obtained by estimating the total trip time for each trip record (as described in Section 3.3) and then determining the number of passengers who were travelling during each 10-minute time period.

The weekday and weekend demand patterns are clearly different. In the weekday one, there are two distinct peaks corresponding to the morning and evening rush hours. In the weekend and holiday one, there is only one distinct peak, around 5:20 p.m. The deviations in the weekday pattern are smaller than those in the weekend one, indicating that most passengers in the weekday behave in a periodic manner. Therefore, we may be able to detect disruptive events by comparing the differences in the average number of passengers for each section of a line.

3.3 Extraction of Passenger Flows

From the trip records, we can determine how many passengers used a certain station. However, since the data did not include the entrance time, we could not determine how many passengers there were within a certain time period. Moreover, the origin-destination pair information was not enough for estimating the crowdedness of each train or the effects of a disruptive incident at a certain location. We need the trip route for each passenger.

There are usually several possible routes for traveling from an origin station to a destination station. A smart card log contains information about where a passenger touched in and where and when he/she touched out at a station gate. It does not include the entrance time and transfer station information. We therefore assumed that the most probable route for each trip (origin and destination pair) was the one with the shortest total trip time.

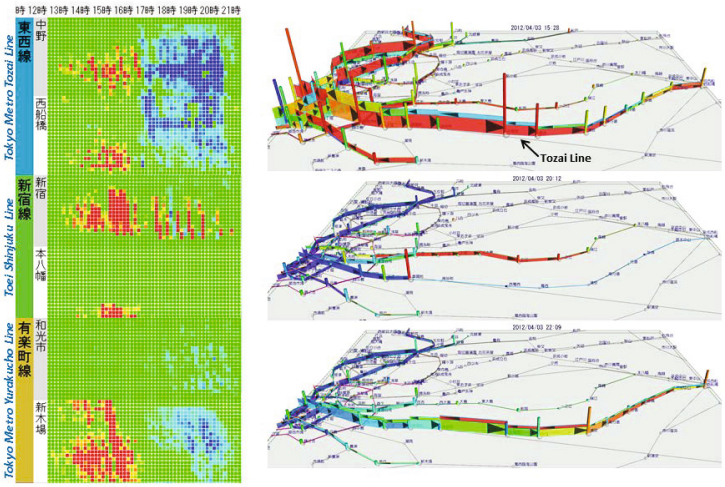

Fig. 3. HeatMap view (left) shows passenger crowdedness during rainstorm in April 2012; RouteMap view (right) shows animated changes in passenger flows and propagation of crowdedness on route map

We defined total trip time $t = T + C + W$, where

- T is the time spent riding, as defined by the timetable,
- C is the walking time when transferring, as determined by the layout of each station and roughly defined using information provided by the train company, and
- W is the time waiting for a train to arrive, as defined by the timetable (average train interval / 2).

Using this definition, we calculated the estimated time for every possible trip route for each origin-destination pair. We then used the Dijkstra algorithm [2] to find the fastest route.

To identify phenomena that differed from the usual cyclical patterns, we first estimated in which section of a line each passenger was during a particular time on the basis of the fastest route and exit time. We then calculated the number of passengers who were in a certain section during a certain (10-minute or one-hour) time period and the average number and standard deviation. The weekday and weekend data sets were independently analyzed as explained in Section 3.2. The average and standard deviation were used to detect unusual patterns, especially in the weekday cyclical patterns.

3.4 Visualization of Passenger Flows

We used a visualization technique [3] to investigate passenger behavior. Our framework provides two visualization views, the HeatMap view and the RouteMap view,

for exploring passenger flows and spatio-temporal propagation of crowdedness extracted in Section 3.3.

The HeatMap view (Figure 3, left) provides an overview of the spatio-temporal crowdedness of sections of the lines in the route map. The RouteMap view (Figure 3, right) shows the animated temporal changes in the number of passengers and the crowdedness of each section. The two views are coordinated – selection of lines and time stamps in the HeatMap view causes the RouteMap view to start showing animated changes in the values for the selected lines and time stamps.

The combination of these views enables users to detect unusual events. Two HeatMap thresholds (high and low) are defined for determining unusually crowded areas and unusually empty areas (for a certain train section and for a certain time period). These unusual geo-temporal areas are concatenated into several chunks separated by normal areas. The size and density of each chunk reflect the largeness of affected area, and the severity of the effect. Many of the large extracted chunks corresponded to unusual events, such as natural disasters, public gathering, and accidents. The example shown in Figure 3 illustrates the effects of a spring storm. The large red and blue chunks in HeatMap view indicate that passenger flows changed drastically during that event. The RouteMap reveals corresponding events by showing the spatio-temporal propagation of the unusual passenger behaviors.

4 Prediction Method

Passenger flows after train accidents are predicted using the demand and behavior models. To predict the number of passengers at the time of an accident without using real-time data, we estimate how many passengers will want to travel during or after the time of the accident by using the demand model. We then estimate how these passengers will change their route by using the behavior model.

Our prediction method requires two kinds of information as input: average passenger flow in the past and the expected time to recover from the accident. The proposed method can predict passenger flows after an accident from this information, without real-time data.

4.1 Passenger Demand Model

We estimated each passenger's entrance time by using the trip time estimation method described in Section 3.3. We then calculated the number of passengers starting a trip, for every origin-destination pair, for every 10 minutes. The results reflected the passenger demand during a certain time period. We calculated the average demand for each time slot over one year (Apr. 1, 2012, to Mar. 31, 2013). As mentioned, data for weekdays and weekends were treated independently.

We used this average demand as the passenger demand model.

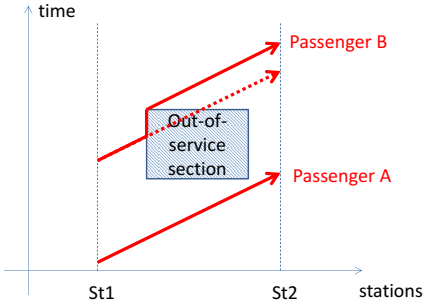

Fig. 4. Model of passenger behavior following an unusual event; red arrows indicate passenger trajectories

4.2 Passenger Behavior Model

In the passenger behavior model, each passenger has an origin station and a destination, and the route taken is calculated as explained in Section 3.3. Example spatio-temporal behaviors of two passengers when service on a section of a line is suspended is illustrated in Figure 4. Two passengers (A and B) want to travel from Station 1 (St1) to St2. We assume that we know about the service suspension and the begin and end times of the suspension (represented in the figure as "out-of-service section"). Passenger A completes passage through that section before service is suspended and is thus not affected. Passenger B is affected and thus must change his route or time of travel:

- If there is another route from St1 to St2 that does not use the out-of-service section, he can change his route, which will likely change the arrival time.
- If there is no such route, he must wait until service is restored (as illustrated in Figure 4).

Since passengers often change their travel method in such situations, such as by switching to travel by bus or taxi, we use the "Abandonment Rate" to capture this behavior. When N passengers are affected by the out-of-service section, we assume only $N \times (1 - \mathit{AbandonmentRate})$ passengers continue to use the train system.

The Abandonment Rate should be based on historical data. The estimation of this parameter is described in Section 5.

5 Evaluation

5.1 Target Data

We found disruptive events for the Tokyo Metro subway system by using the transport information webpage of a third-party company [2]. We checked this page every hour for one year and extracted the train operating condition information.

[2] `http://transit.goo.ne.jp/unkou/kantou.html` (in Japanese).

Here we focus on one line (the "Tozai Line") in order to set the parameter of our model under the simplest conditions. We found that six disruptive events occurred on the Tozai Line during the year. They caused service suspensions on several sections, and it took several hours to restore service.

5.2 Evaluation Metric

Since each trip record contained each passenger's arrival time and not the departure time, we could only determine how many passengers had left from a certain station at a certain time (*Exitnum*). We evaluated the passenger behavior model by comparing these values. We estimated how many passengers left from each station by using our passenger demand and behavior models.

5.3 Test Case 1 — Accident on Nov. 6, 2012

An accident occurred at 17:11 on Nov. 6, 2012. A failure of railway power supply system was found and several sections of the line between Toyocho Station and Kasai Station were closed until 18:34. The accident happened during rush hour, so it affected many passengers.

Figure 5 shows the difference in *Exitnum* between the estimation obtained using the demand model and the actual trip record data. The estimation was made without considering the accident information. Colors are used in the figure to represent the difference in *Exitnum* normalized by the value in the trip record data. Blue represents the situation in which the actual number of passengers was lower than the estimated one, and red represents the opposite case. The large blue area indicates that many passengers were unable to travel as they normally did.

We also estimated *Exitnum* for all the other stations in the metro system. Since the difference was at the highest for stations on the Tozai Line, we focused on that line in our evaluation.

The estimation results when the out-of-service information was considered, with Abandonment Rate = 0, are shown in Figure 6. Most of the blue area (labeled i) disappeared. However, a new large blue area (labeled ii) appeared around 18:50, after the accident.

We adjusted the Abandonment Rate to find the most appropriate setting. We computed the sum of the absolute number of difference in this sections and timespan to estimate the correctness of our model. Figure 8 shows this value normalized by the average number of passengers who used the Tozai Line at this time of the day. The difference for Abandonment Rate = 0 was larger than the case of no information about the service suspension (green line). We can see that the best setting of the Abandonment Rate was 0.9. With the out-of-service information, the difference improved by about 9%. Figure 7 shows the best case, with Abandonment Rate = 0.9. The large problematic blue area (ii) disappeared with this setting.

Even in the best case (Figure 7), we can see several chunks of blue area (iii and iv). We obtained train operation information for that date from the railway

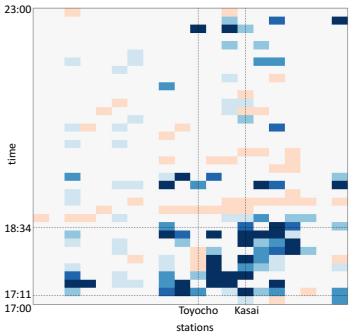

Fig. 5. Difference in number of passengers after accident on Nov. 6, with no out-of-service information

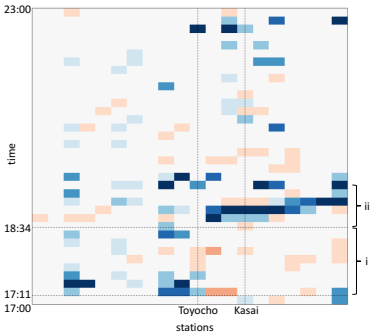

Fig. 6. Difference in number of passengers after accident on Nov. 6, with out-of-service information and Abandonment Rate (AR) = 0.0

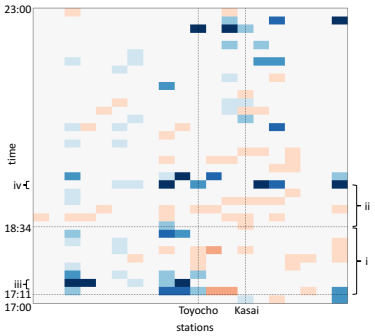

Fig. 7. Difference in number of passengers after accident on Nov. 6, with out-of-service information and $AR = 0.9$

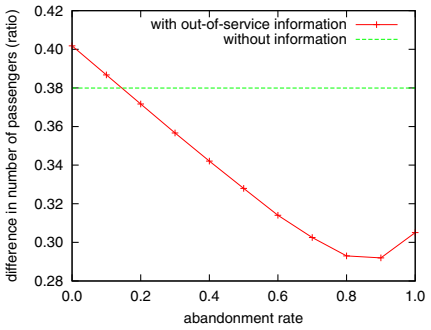

Fig. 8. Difference in number of passengers after accident on Nov. 6

company and determined that most of the decrease in the number of passengers is understandable: the entire Tozai Line was stopped twice for a short time, from 17:11 to 17:28 (iii), and from 19:19 to 19:29 (iv). The blue chunk in Figure 7 is reasonably explained by this information.

5.4 Test Case 2 — Accident on Oct. 10, 2012

Another accident occurred at 11:32 on Oct. 10, 2012. A passenger was injured and sections between Kasai Station and Myouden Station were closed until 13:09. Note that the closed sections and time of day differ from those in the November case (Section 5.3).

Figure 9 shows the difference in *Exitnum* without the out-of-service information. The difference is smaller than in the November case.

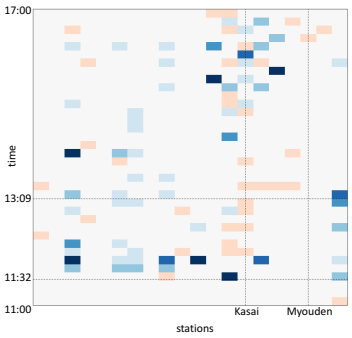

Fig. 9. Difference in number of passengers after accident on Oct. 10, with no out-of-service information

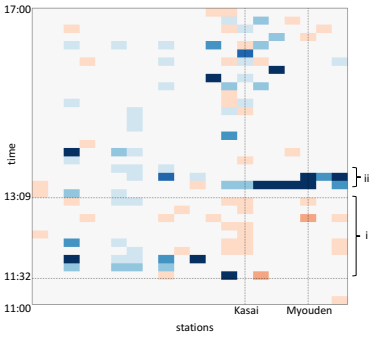

Fig. 10. Difference in number of passengers after accident on Oct. 10, with out-of-service information and Abandonment Rate (AR) = 0.0

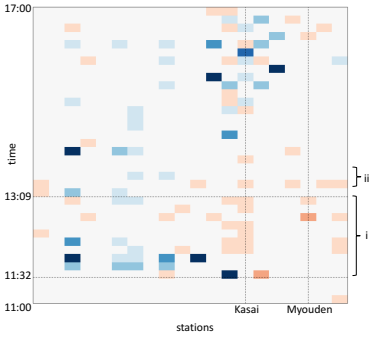

Fig. 11. Difference in number of passengers after accident on Oct. 10, with out-of-service information and $AR = 0.9$

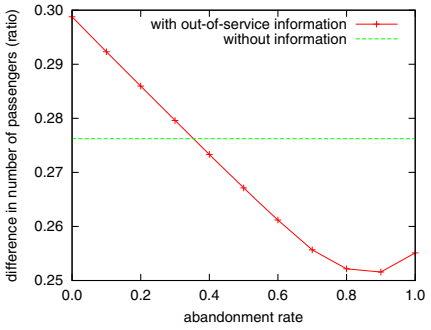

Fig. 12. Difference in number of passengers after accident on Oct. 10

Figure 10 shows the results with the out-of-service information and Abandonment Rate = 0. The difference during the time period of the accident (i) is smaller, but a new blue area (ii) appeared after the accident.

We again adjusted the Abandonment Rate to find the most appropriate setting. Figure 12 shows the total difference in our estimation normalized by the average number of passengers who used the Tozai Line at this time of the day. The best result was achieved when we set the Abandonment Rate to 0.9, the same value as in the November case. This shows that our behavior model can predict user behavior more precisely when the out-of-service information is used.

The best setting (Abandonment Rate = 0.9) results are shown in Figure 11. The large blue chunk in Figure 10-(ii) has disappeared.

5.5 Discussion

As we can see from the results of our two evaluation cases, our behavior model can predict the effects of train accidents. Although the closed sections and time of day differed between the two test cases, the same parameter setting could be used for both cases. In this prediction, all we needed to know was which sections were closed and for how long. Using the average demand and Abandonment Rate parameter, we predicted the number of passengers without using real-time demand information. The use of real-time demand information would of course make the prediction even more accurate.

When an accident occurs, the operating company can usually estimate how long it will take to restore service by using knowledge about previous accidents. Therefore they could use our framework and predict the effects of an accident in real time.

6 Conclusion

The framework we have developed for analyzing passenger behavior in public transportation systems is aimed at gaining an understanding of passenger flows in real time and predicting short-term passenger behavior. With this framework, we can analyze a large-scale dataset of trip records created from smart card data. Various unusual phenomena can be observed by using interactive visualization. Our model for predicting passenger behavior after train accidents was demonstrated to predict passenger flows even without the use of real-time data.

In the evaluation test cases only a single line has stopped. More complex cases in which several lines are affected will be evaluated in future work.

References

1. Ceapa, I., Smith, C., Capra, L.: Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. In: Proc. UrbComp 2012, pp. 134–141 (2012)
2. Dijkstra, E.: A note on two problems in connexion with graphs. Numerische Mathematik 1(1), 269–271 (1959), http://dx.doi.org/10.1007/BF01386390
3. Itoh, M., Yokoyama, D., Toyoda, M., Tomita, Y., Kawamura, S., Kitsuregawa, M.: Visualization of Passenger Flows on Metro. In: IEEE Conference on Visual Analytics Science and Technology, VAST 2013 (2013) (poster)
4. Pelletier, M.P., Trépanier, M., Morency, C.: Smart Card Data Use in Public Transit: A Literature Review. Transportation Research Part C: Emerging Technologies 19(4), 557–568 (2011)
5. Sun, L., Lee, D.H., Erath, A., Huang, X.: Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System. In: Proc. UrbComp 2012, pp. 142–148 (2012)
6. Trépanier, M., Tranchant, N., Chapleau, R.: Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. Journal of Intelligent Transportation Systems 11(1), 1–14 (2007)
7. Ushida, K., Makino, S., Tomii, N.: Increasing Robustness of Dense Timetables by Visualization of Train Traffic Record Data and Monte Carlo Simulation. In: Proc. WCRR 2011 (2011)