# MEGA CITY METRO NETWORK DATA ANALYSIS USING HADOOP

[1]Janarthanan.R
[2]Raghu.G,[3]Nandhinisaraswathi.C, [4]Yamini.T, , [5]Supraja.R

[1]Hod[2]Asst.Professor[3][4][5]UG Student, Department of Computer Science and Engineering

T.J.S Engineering College

[1] hodcse@tjsenggcollege.com, [3]nandhucns96@ gmail.com, [4]yamininaidu1996@gmail.com,, [5]subharam2323@gmail.com

Abstract—Transportation systems in mega-cities are often affected by various kinds of events such as natural disasters, accidents, and public gatherings. Highly dense and complicated networks in the transportation systems propagate confusion in the network because they offer various possible transfer routes to passengers. Visualization is one of the most important techniques for examining such cascades of unusual situations in the huge networks. This paper proposes visual integration of traffic analysis and social media analysis using two forms of big data: smart card data on the Tokyo Metro and social media data on Twitter. Our system provides multiple coordinated views to visually, intuitively, and simultaneously explore changes in passengers' behavior and abnormal situations extracted from smart card data and situational explanations from real voices of passengers such as complaints about services extracted from social media data. We demonstrate the possibilities and usefulness of our novel visualization environment using a series of real data case studies and domain experts' feedbacks about various kinds of events.

Index Terms—Information visualization, visual analysis, smart card, big data

## 1 INTRODUCTION

PUBLIC transportation systems, such as railways and metros, in mega-cities are always required to increase their resilience to extreme situations caused by various events. For instance, Tokyo, which is the biggest mega-city in Japan, will host the 2020 Summer Olympics and Paralympics, which will cause large scale movements of people over the wide area around Tokyo. Powerful inland earthquakes are also estimated to possibly occur in the Tokyo metropolitan area. Public transportation systems are now preparing responses for these events. To increase the resilience of the systems, lessons must be learned from past events to understand how the systems are affected by changes in passengers' behaviors. Integration of smart card data and social media data enables us to replay past events and to discover abnormal situations of transportation systems, propagations of abnormalities over transportation networks, and passengers' complaints or dissatisfaction about which even train system operators and station staff do not know. While some analysis systems have utilized both mobility data and social media data [1], [2] to understand human behavior or traffic anomalies, they cannot support both finding abnormal situations from wide spatio-temporal

space and exploring spatio-temporal propagation of them,and interactively exploring their reasons in detail. Developing a visual environment for exploring passenger behaviors in a complex transportation system using transportation logs and social media stream is still a challenging task. For supporting effective exploration, the environment needs to satisfy the following requirements:

1) Discovering unusual phenomena from the wide range of temporal overviews that are derived from differences between daily and event-driven passenger behaviors. The techniques for intuitively verifying effects of known events and discovering trouble unknown to even train system operators are desired.

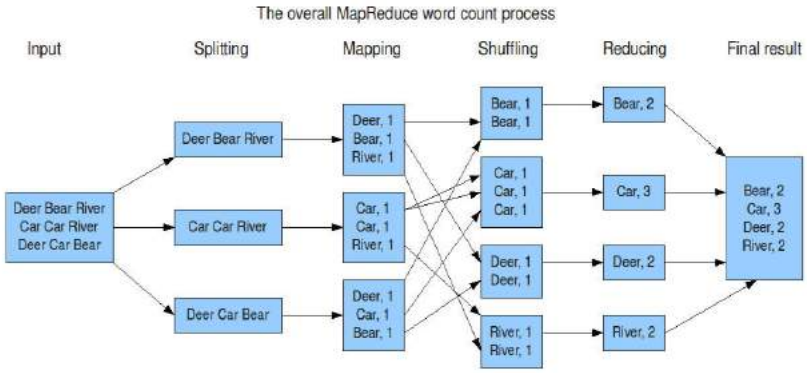
2) Understanding changes in passenger flows and spatial propagation of unusual phenomena in each time period on a wide area metro network. A visual exploration environment is necessary to intuitively understand the route, speed, and range of propagation of the unusual phenomena such as abnormal crowdedness. These are difficult for the train system operators to understand because the transportation system network in Tokyo is extremely dense and complicated.

3) Exploring reasons for unusual phenomena or their effects from real users' voices. A system is required for exploring information about passengers' complaints, activities such as use of taxis or buses, and confusing situations instations,which often cannot be obtained from customer support or operation trouble databases.

This paper proposes a novel visual fusion analysis system that can support ex post evaluations of trouble in a metro system by using two forms of big data: archived transportation logs from the smart card system of the Tokyo Metro and social media data from Twitter. Knowledge acquired through the visualized results mostly reflects real situations such as disasters, accidents, and public gatherings. To address the above requirements, we built the analysis system by integrating the following visualization components: 1) HeatMap view provides a temporal overview of unusual phenomena in passenger flows, 2) AnimatedRibbon view visualizes temporal changes in passenger flows with spatial contexts and propagation of unusual phenomena over the whole metro network using animation, and 3) TweetBubble view provides an overview of trends of keywords explaining the situation

_ M. Itoh, D. Yokoyama, and M. Toyoda are with the University of Tokyo,
Tokyo, Japan. E-mail: {imash, yokoyama, toyoda}@tkl.iis.u-tokyo.ac.jp.
_ Y. Tomita is with the Tokyo Metro Co., Ltd, Tokyo, Japan.
E-mail: y.tomita@tokyometro.jp.
_ S. Kawamura iswithPASMOCo., Ltd, Japan.
E-mail: s.kawamura@pasmo.co.jp.
_ M. Kitsuregawa is with the National Institute of Informatics, the University
of Tokyo, Tokyo, Japan. E-mail: kitsure@tkl.iis.u-tokyo.ac.jp.
Manuscript received 30 Mar. 2015; revised 14 Dec. 2015; accepted 9 Mar.
2016. Date of publication 31 Mar. 2016; date of current version 27 May 2016.
Recommended for acceptance by Y. Zheng.
For information on obtaining reprints of this article, please send e-mail to:
reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TBDATA.2016.2546301
IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 1, JANUARY-MARCH 2016 85
2332-7790 _ 2016 IEEE. 

during the unusual phenomena.
We demonstrate the usefulness of our novel visualization system through a series of case studies extracted from real data related to natural disasters, accidents, and public gatherings. These case studies show how our visualization system enables users such as domain experts in the metro operating company to explore hidden knowledge based on data-driven analysis and visualization that were previously unattainable.
In summary, we have made the following contributions. (1) We introduce one of the first visual analysis systems that integrate smart card data including origin-destination data and textual social media data. (2) We provide three coordinated views, HeatMap, AnimatedRibbon and TweetBubble view, to help analysts to understand changes in passengers' behavior in the complex transportation systems. (3) AnimatedRibbon view provides a novel visualization technique to dynamically represent changes in multiple attributes values of both nodes and edges in a network while embedding them in a spatial context. (4) The case studies using real data and domain experts' feedbacks strongly highlight the effectiveness of our system and three visualization components.
In what follows, we give an outline of related work in Section 2. We offer information about our data set in Section 3. We introduce the overview of our system in Section 4. We then describe a method for extracting passenger flows in Section 5 and situational explanations in Section 6. We introduce our novel exploration environments in Section 7. We present some case studies in Section 8. Section 9 presents reviews from train operating system experts. This article ends in Section 10 with a conclusion.




**EXISTING SYSTEM:-**

Existing concept deals with RDBMS which contains lot of drawbacks
data limitation is that processing time is high when the data is huge and once data is lost we cannot recover.
DRAWBACKS
Existing concept deals with RDBMS which contains lot of drawbacks data limitation is that processing time is high when the data is huge and once data is lost we cannot recover.

PROPOSED SYSTEM
It deals with providing database by using hadoop tool we can analyze with no limitation of data and simple add number of machines to the cluster
we get results with less time, high throughput and maintainance cost is very less and we are using joins , partitions and bucketing techniques in hadoop .

## 2 RELATED WORK

### 2.1 Smart Card Data Analysis

Smart card data is one of the data sources to analyze operation of public transportation systems [3], [4]. Ceapa et al.

focused on congestion patterns of some underground stations in London to reveal station crowding patterns to avoid traffic crowdedness [5]. They utilized data of oyster cards, the smartcards used on the London Underground. Their spatio-temporal analysis showed a highly regular crowding pattern during the weekdays with large spikes occurring in short time intervals. Sun et al. provided a model to predict spatio-temporal density of passengers and analyzed it for one MRT line in Singapore [6]. However, previous work only focused on a single selected line or some stations. One reason is that most smart card data does not include transfer station information. Our work speculates the most probable path of each trip from origin and destination in smart card data and succeeds in visualizing propagation of effects of trouble on the metro network.

Zeng et al. [7] provided a visualization system to explore passenger mobility in a Singapore public transportation systems including Metro system and public bus network. They estimated various mobility-related factors such as waiting time, riding time, transfer time, and travel efficiency using data including passenger journey data via RFID card, transit line schedule data, and transportation network data. They then visualized them to explore geographical accessibility, time-efficient routes and their temporal variations along the origin-destination journeys. Although they focused on visualizing mobility-related information along routes from a specific origin in a tree structure, our work focus on visualizing spatio-temporal propagation of crowdedness or emptiness in a complicated network. As far as we know, there has been no research on the visualization of propagation of influences spreading over a wide range of public transportation systems such as metro networks

.

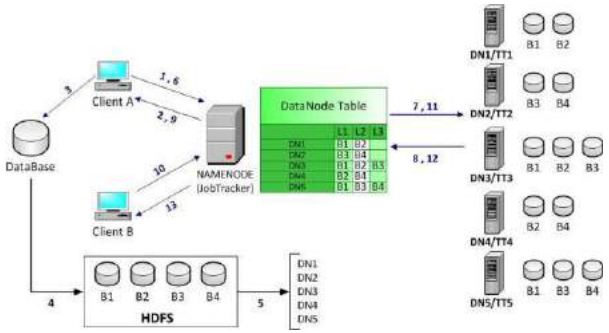
### 2.2 Spatio-Temporal Information Visualization

There have been some research on and systems developed for the visualization of geo-spatial and temporal values on a map. Andrienkos and Slingsby et al. utilized multiple heatmaps in a map divided into regular rectangles [8], [9]. Each heatmap in the grid represented temporal overview of traffic volume in a specified area. Wang et al. also used heatmaps to visualize temporal changes in traffic speeds in selected road sections [10]. Introducing 3D icons into a map in place of heatmaps is one of the alternative approaches for representing temporal overview of attribute values in specified spots [11], [12]. Their approach focused on describing changes in values at independent points or areas and did not provide a method for representing temporal changes in values between two points or flows.

There has been some research on analyzing mobility data and extracting and visualizing important events or mobility patterns. Doraiswamy et al. introduced techniques to extract urban events from large spatio-temporal data such as taxi trips in New York City1 [13]. Wang et al. extracted and visualized traffic jams and their propagation from data of taxi trips in Beijing [14]. Andrienko et al. extracted and characterized important places from mobility data such as GPS tracks of cars and flight trajectories and visualized them in 3D spatio-temporal space [15], [16]. Unlike these cases, smart card data in our system includes origin-destination (OD) data without trajectory information. We therefore need to speculate the most probable route for each trip from OD data and visualize aggregated passenger behaviors.

Ferreira et al. provided a visual environment for comparing temporal changes in values such as trip duration or number of trips between selected OD regions using taxi trip data in New York City [17]. Flowstrates visualized a temporal overview of flow magnitudes among multiple OD pairs by usingheatmap and two separate maps [18]. Jiang et al. introduced Circular pixel graphs to represent spatio-temporal patterns of OD distributions from or to a selected region using circularheatmaps [19]. Although their approach can visualize temporal changes in flows among two separated regions, it

1. Although they also utilized MTA subway data in New York City, the subway data does not includes trip information of passengers but delay information of each train.






cannot represent changes in flows in the specific trajectories or routes on the map such as lines of trains or roads.

Tominski et al. showed the usefulness of 3D trajectory bands to visualize trajectory attribute data [16], [20]. In their visualization, attribute data of individual trajectories was visualized as color-coded bands and sets of trajectories were visualized by stacking the bands. Cheng et al. also utilized 3D staked bands to represent overview of spatio-temporal changes in attribute data on a road network [21].

Stacked and color-coded 3D bands are useful for representing spatio-temporal changes in an attribute value on the map, but they cannot represent two or more kinds of attribute values or their scale such as the number of people. Our approach utilizes 2D heatmaps for overviewing temporal changes in flows and 3D animated ribbons for simultaneously visualizing changes in absolute counts such as the number of passengers and relative counts such as the deviation from the average and how these propagate in a complicated network.

Visualization of time-varying changes in the number of passengers in a Metro network is a kind of dynamic graph visualization. Andrienkos proposed methods for representing generalized movements as network of flows [22], but they did not provide a method for visualizing temporal changes in flows. Although there have been some research on dynamic graph visualization [23], [24], [25], as far as we know, there has been no research to be able to simultaneously provide insights into multiple attributes of both nodes and edges in a graph while embedding them in their spatial and temporal context as our AnimatedRibbonviewdoes.

### 2.3 Spatial Tweet Visualization

LeadLine [26] detected events from social media data, extracted information about 4 Ws (who, what, when, and where) related to the events, and then visualized the information in coordinated views. SensePlace2 [27] provided an integrated environment for filtering and visualizing spacetime-theme information from twitter streams. Thom et al. provided visual analysis system for detecting spatio-temporal anomalies from geo-located tweets and visualizing them as word clouds representation on a map [28]. Their approaches focus on exploring events from social media data without using other data resources.

Pan et al. provided a system for traffic anomaly detection from human mobility data and anomaly analysis using social media data [1]. They used term clouds to visualize terms related to the detected anomalies. Although in their approach visualization is only used for showing detected results, our work focus on providing interactive environments for finding anomalies and exploring them in detail by using two forms of data from smart card system and social media.

## 3 DATASETS

### 3.1 Smart Card Data

We use a large scale data set of travel records from March 2011 on the Tokyo Metro extracted from the smart card system.

Tokyo has the complicated train route map.2 It consists of lines of various kinds of railway companies including Tokyo Metro, Toei Subway, Japan Railway (JR), and many private railroads. We analyze large scale log data covering almost all of the business area of Tokyo. It consists of 28 lines, 540 stations, and over 410 million trips. This includes 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. Card ID is eliminated from each record. Each record consisted of the origin, destination, and exit time.3 Since transfer information was not included, we estimated the probable route for each trip (as explained in Section 5).

Trains in Tokyo are mainly used by working people, so the usage patterns of trains on weekdays and weekends may be different. We separate the data into weekdays and weekends and analyze them independently. National holidays and some other days in vacation seasons are treated as weekends.

Passengers are expected to behave with some periodic patterns, especially daily ones, thus we try to do a statistical analysis of this data. Fig. 1 shows the average and standard deviation of the number of passengers at every time period of the day through one year, from April 2012 to March 2013. The error bar of each point indicates standard deviation. To extract Fig. 1, we first estimate the trip time length of each trip log (mentioned in Section 5) and then accumulate the number of passengers who were travelling at a certain time period. The time periods are divided every 10 minutes. Weekdays and weekends have clearly different demand patterns. The deviations of weekdays are considerably smaller than those of weekends. This means that most passengers actually behave in a periodic manner, so we may be able to detect some irregular accidents or events by comparing the differences with the average number of passengers at each section. We try to confirm this hypothesis in the following sections.

### 3.2 Social Media Data

Social media immediately reflects real world events such as accidents. In this paper, we utilize Twitter as a social media

Fig. 1. Average and standard deviation of number of passengers over one year (April 2012 to March 2013): (a) weekdays, and (b) weekends and national holidays.

2. http://www.tokyometro.jp/en/subwaymap/index.html 3. No records contain trip start times.



data resource. We have been crawling through more than four years' worth of Twitter data from Twitter API from March 11, 2011. Our crawling started from 30 famous Japanese users by obtaining their past timelines. Then we repeatedly expanded the set of users by following retweets and mentions appeared in their timelines. We have continuously performed the user expansion and tracking of their timelines. We then obtained data of more than 2 million active users and 25 billion tweets on 2015.

## 4 OVERVIEW OF SYSTEM WORKFLOW

We construct an analysis framework which can integrate both smart card and twitter data to explore passengers' behavior. Fig. 2 shows the design of our system workflow. Smart card data are transferred to our analysis system once a month. We estimate passenger flows from the one month worth of data, compute average flows for every sections from updated last one year worth of flows, and compute z-scores (difference normalized by standard deviation) of

the last one month (described in detail in Section 5.1). This computation requires less than one hour in total under the current implementation. When we try to simulate passengers' behavior under accident case, flows are recomputed by the smart card data with constraints; taking into account of suspended lines, for example (described in detail in Section 5.2). The recomputation process requires several minutes. Passengers' reactions are obtained from tweet archive by extracting situational explanations (described in detail in Section 6). All tweets which are related to traffic conditions are archived. Visualization components described in Section 7 access extracted information in on demand manner to keep high interactivity.

## 5 EXTRACTION OF PASSENGER FLOWS

With the recorded smart card data, we can understand how many passengers used a certain station. However, that data does not include the entrance time, so we could not see how many passengers are there within a certain time period. Moreover, if we try to estimate the crowdedness of each train, or effects of an accident at a certain location, the origin-destination pair is insufficient. We must figure out the travel path of each passenger for such requests.

### 5.1 Estimating Daily Passenger Flows

There are several possible paths to take 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. It does not include the entrance time or transfer stations' information. We therefore speculate the most probable path for each trip (origin and destination pair) by assuming that they take the shortest time path.
We assume that total travel time (t) of each trip is defined as $t = T + C + W$, where:
_ T is the time while passengers are riding trains. It defined by using the timetable.
_ C is the walking time while passengers are transferring trains. It relates to the structure of the station, so it differs at every station. We roughly define these times by using the information fromthe train company.
_ W is the time waiting for a transfer. We define this as (average train interval / 2) extracted from the timetable. It differs on every line.
With this model, we can calculate the estimated travel time of any travel path. We then search for the shortest time path of every origin-destination stations pair by using the Dijkstra algorithm.
We want to find unusual phenomena that differ from the usual cyclical patterns of the passengers. For this purpose, we first estimate in which section of a line a passenger passes in a particular time period from the speculated shortest time path and exit time. We then accumulate the number of passengers who travelled a certain section in a certain time period (every 10 minutes or one hour). Data for weekdays and data for weekends are separately analyzed because weekdays and weekends show clearly different patterns as shown in Section 3.1. After that, we calculate the simple moving average (SMA) of the previous one year for each month and calculate standard deviation using the same time window. SMA reflects daily cyclical patterns, and unusual patterns can be detected by comparing it with log data.
All passenger flows from one-day smart card records take several minutes to estimate using one CPU core. This process can be easily parallelized because path estimation of every trips are totally independent. We can efficiently use 20 CPU cores on one server, therefore we achieve to extract one month worth of passenger flows within several minutes. The amount of computation time is acceptable for current usage. Many parts of our current system are experimentally implemented and have large room for improvement in terms of the execution performance.

### 5.2 Estimating Passenger Flows after Accidents

Accidents sometimes cause service suspensions at several sections, making passengers take detours. In such situations, the route estimation method proposed in Section 5.1 cannot calculate an appropriate route because the shortest path would be changed by service suspensions.

We can recompute the shortest paths considering the suspension information such as suspended sections and time. This refines passenger flow estimation to make it more appropriate to describe what happened at that time. We provide interfaces to input constraints of suspended lines and/or sections and start and end times of suspension

Fig. 2.Overview of our system workflow.



on our visual exploration environment shown in Section 7. The visual exploration environment visualizes the recomputed result. We can then visually check how passengers take detours and concentrate on particular lines.

Fig. 3 compares passenger flows with and without suspension information on 27 November 2013, the day on which an accident resulting in injuries happened at Machiya station on the Chiyoda Line. Fig. 3a shows passenger flows without suspension information. Fig. 3b shows recomputed passenger flows using suspension information based on factual information. In Fig. 3a, the flow of passengers continues to exist even after the Chiyoda Line service is suspended. In Fig. 3b, sections from Kita-Senju to Yushima are suspended from 9:59 to 10:37. We can find out that passenger flows concentrate on the Hibiya and Ginza Lines to avoid selected sections on the Chiyoda Line.4

We can use suspension information obtained from an external information resource such as the metro operating company or the transport information webpage as inputting constraints for our system. The metro operating company holds information about events that disrupt their subway system. We also collect the train operating condition information from the transport information webpage of a thirdparty company.5

Our passengers flow model is constructed on the assumption that every passenger will take the fixed shortest path. Their real behavior is more diverse; they will also consider travel fees, crowdedness, ease of train transfer, etc. Using a probabilistic behavior model may be more appropriate to reflect such diversity. When considering the cases of accidents, our method makes another train scheduling assumption: trains that were unaffected by the accidents keep travelling on time. Trains were stopped at certain sections in the case of serious accidents, but more modest actions such as delaying trains or partially eliminating services would have happened in many cases. We have already tried to construct a preliminary behavior model of passengers after accidents have happened [29] and plan to improve the model by introducing such details.

Since the current smart card system does not have the information of entrance time or transfer points of each trip, we cannot evaluate the preciseness of the estimated passenger flows directly. We interviewed some of the staff of the train operator and found that the extracted flow seems to correspond to their knowledge of the daily operation. We plan to evaluate the preciseness through comparison with other statistical survey results such as traffic censuses.

## 6 EXTRACTION OF SITUATIONAL EXPLANATION

**Soc**ial media enables people to post information about what they saw, thought, and did during and after events such as accidents. We can extract more precise or fine-grained information about the events that sometimes cannot be obtained by operating companies.

We extract a set of words (weighted by word frequencies based on the measure similar with tf-idf) for overviewing

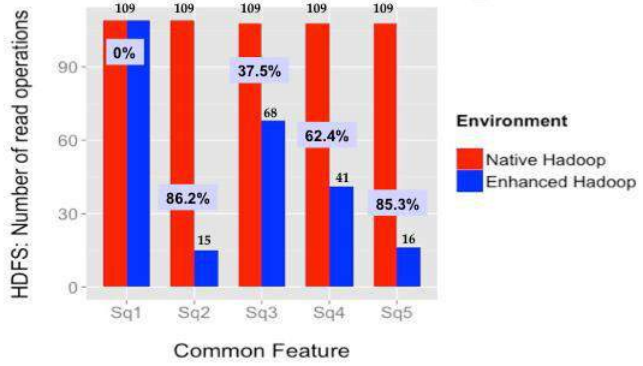
and explaining situations. For this purpose, we first calculate word frequencies for every co-occurring word for each station name or line name on each date and time from the data set described in Section 3.2 as tfðword; station=line; date and timeÞ(if we specify a start time and an end time, we use the sum of the word frequency between them (tfðword; station=line; timewindowÞ)).
We then count the number of days when each word appears for each station or line and treat it as dfðword; station=lineÞ. In this case, we treat a set of tweets on one day including the name of a station or line as one document for each station or line. It is used for decreasing importance of words commonly used all the time for each station or line. Small accidents or short delays happen almost every day around Tokyo. Therefore, if we use dffor all documents that are related to all stations and lines, words related to trouble may not be treated as important. Moreover, the characteristics of co-occurring words that appear routinely are different among stations or lines. Therefore, we calculate dffor individual stations or lines.
We finally calculate weightðword; station=line; date and time=timewindowÞas tf_ idfðword; station=lineÞ(s.t.idf¼ logð jdatej
dfðword;station=lineÞÞ þ 1).

## 7 EXPLORATION ENVIRONMENT FOR PASSENGER FLOWS

In this section, we describe how we design visualization views for supporting exploration requirements described in Section 1. We provide three types of visualization views: HeatMap view and AnimatedRibbon view to explore unusual phenomena of passenger flows and spatio-temporal propagation of them extracted by the methods mentioned in Section 5, and TweetBubble view to explore situational explanations extracted by the method described in Section 6. These views are coordinated with each other. Such coordinated multiple views not only combines the

Fig. 3. Passenger Flows after accident on 27 November 2013, the day
on which an accident resulting in injuries happened at Machiya station
on the Chiyoda Line. The height of 3D ribbons represents the number of
passengers in these examples. Colors represent relative crowdedness
or emptiness compared with the average situation; red indicates crowdedness,
blue indicates emptiness, and green indicates a mostly normal.

4. The meaning of each visual element is described in Section 7.
5. http://transit.goo.ne.jp/unkou/kantou.html (in Japanese)






advantages of existing visualizations but also extend their use for the simultaneous analysis of multiple aspects.

### 7.1 HeatMap View

For easily discovering unusual phenomena in passenger flows on a particular line, over multiple lines over one day and exploring their temporal characteristics from the

wide range of temporal overviews, HeatMap view provides functions for overviewing deviation from average passenger flows in each time bin on each section for every line over one day. It is used for spotting interesting phenomena by using patterns of colors. Although it does not provide spatial context, after finding out interesting temporal spots showing crowdedness or emptiness, we can explore spatial changes in them by combining HeatMap views with AnimatedRibbon views. Moreover, we can observe their causes and effects by combining HeatMap views with TweetBubble views.
HeatMap view uses the x-axis for the timeline and the y-axis for lines. The timeline is divided every 10 minutes (Fig. 4). Each line is represented by different colors, and both directions (up and down) are treated separately. Up and down lines are separated by color, and labeled by the starting stations. Each up/down line consists of sections that are pairs of origin and destination stations as shown in Fig. 4.6 The order of lines can be manually changed. HeatMap view also can be zoomed and panned interactively. Users can interactively select lines or times to visualize in other views such as AnimatedRibbon view and TweetBubble view.

### 7.1.1 Design Alternatives for HeatMap View

Arranging multiple heatmaps [8], [9], [10] on the map is one of the design alternatives to visualize and compare multiple temporal overviews. However, it is difficult to arrange multiple heatmaps for all sections along all metro lines to compare changes in passenger flows over the whole metro lines. It would suffer from occlusion and clutter by the substantial number of heatmaps because of highly dense and complicated networks. Although arranging multiple 3D icons or walls on the map is another possible design alternatives [11], [12], [16], [20], [21] to show multiple temporal overviews, it is clearly impossible to compare temporal overviews for multiple lines or routes on the map (the limitations of this approach are discussed in Section 7.2.1 in more detail). Flowstrates [18] shows one solution for arranging the huge number of heatmaps vertically for representing temporal changes in flows on multiple OD pairs. HeatMap view, the solution which we propose, also takes the similar approach that vertically arranges heatmaps for every section.

### 7.1.2 Color Encoding on HeatMap View

The color code for each cell in the HeatMap represents relative crowdedness or emptiness of each section compared with the average situation. For this purpose, we calculate z-scores (difference normalized by standard deviation) of each section for each time bin by SMA and standard deviations shown in Section 5. Red represents a higher z-score indicating crowdedness, blue represents a lower z-score indicating emptiness, and green represent a middle z-score that indicating a mostly normal situation. Two types of thresholds (one for smaller value (S-th) and the other for larger value (L-th)) can be manually defined to emphasizessmall differences or change the range for viewing z-scores. For instances, Fig. 5a uses S-th¼ 2:5 and L-th¼ 9:0, and Fig. 5b uses S-th¼ 2:0 and L-th¼ 5:0. Z-scores for each block are normalized by using S-thand L-th, and then the color code is defined. If the absolute value of a z-score is smaller than S-th, then green is used. If the absolute value of a z-score is larger than L-th, the cell becomes red/blue. Color is adjusted between green and red or blue. S-thand L-thvalues for specifying color code can be used for specifying colors in AnimatedRibbon view.

### 7.2 AnimatedRibbon View

For understanding changes in passenger flows and spatial propagation of unusual phenomena in a complex metro network, AnimatedRibbon view provides the functions for

Fig. 4.HeatMap view 11 March 2011, the day on which the Great East

Japan Earthquake occurred. All lines were suspended just after the earthquake
at 14:46 (shown as blue time bins). Some lines resumed around
20:40, causing concentration of passengers (shown as red time bins).
Fig. 5.HeatMap views related to the spring storm in April 2012 with
different threshold values (S-thand L-th): (a) for emphasizing only time
bins with large z-score, and (b) for emphasizing small differences.
6. We omit labels for origin and destination stations in other figures.

displaying animated temporal changes in the number of passengers and crowdedness or emptiness of each section in the Metro network (Fig. 6). It dynamically shows changes in two attribute values (absolute number of passengers for both directions on each section every 10 minutes by using height of 3D stacked ribbons and deviation from average by using color-coding) simultaneously while maintaining geographical context in the Metro network. A 3D bar on each station presents the number of passengers who exited the station every 10 minutes.
If we have information about service status such as in operation and suspended operation, AnimatedRibbon view changes the width of each section according to the statuses (for instance, Fig. 9c uses thin lines and bold lines for representing suspended operation and resumed operation respectively).

### 7.2.1 Design Alternatives for AnimatedRibbon View

Three functions are required for exploring changes in passenger flows and spatial propagation of them on the metro network: (1) representing propagation of flows, (2) representing the number of passengers and crowdedness or emptiness simultaneously, and (3) showing temporal changes in them.
Graph visualization techniques are one of the best solutions to represent spatial propagation of flows on the metro network. Three types of techniques can be considered to represent temporal changes in passenger flows on the network: using small multiples [30], 3D wall map [20], [21], and animation.
Small multiples, which displays visualizations belonging to different time steps as small thumbnails in parallel, is standard approach to compare multiple situations, however, it is difficult to represent long-term changes because of the limitation of screen space.
3D wall map representation using 3D stacked bands over the map [20], [21] is one solution for representing spatio-temporal changes in an attribute value such as crowdedness, however 3D wall map would have difficulty in showing multiple attributes values simultaneously. It is necessary for exploring the scale of passenger flows, particularly those with huge spikes, and exploring propagation of abnormal situations on the Metro network at the same time. As far as we know, there has been no research to simultaneously represent changes in both absolute and relative values in flows in a network.
3D wall map would quickly suffer from occlusion problem. We have developed a prototype of the 3D wall map view as shown in Fig. 7. This figure shows only two lines such as the Marunouchi Line and the Ginza Line, however, we can find that it is difficult to read a wall for the Marunouchi Line because of the occlusion. It is important for us to show multiple lines while keeping readability because our work focuses on visualizing spatio-temporal propagation of changes in passenger flows on a wide area metro network. We therefore conclude that 3D wall map representation does not fit for our purpose.
Utilizing 2D bands such as those used by Andrienkos [22]

is one solution for representing the number of passengers. However, 2D bands would quickly suffer from severe overplotting [20] and the occlusion problem, especially around highly connected stations or caused by extremely big values. Dang et al. showed that utilizing heights in the 3D space to represent the magnitude of values in dense data area is useful to avoid over-plotting instances [31]. Our AnimatedRibbon view simultaneously utilizes heights in the 3D space for representing the number of passengers and colors for the level of crowdedness on the metro network, and utilizes animation for representing dynamical changes in them. Height is more suitable than color for representing values that have huge spikes such as the number of passengers shown in Fig. 13b because color does not have the dynamic range to permit extreme magnitude [31].

### 7.2.2 3D Design Issues, Solutions, and Limitations

We utilize the metro network map on the basis of real geographical positions. The metro network illustrated in the AnimatedRibbon view is very complicated. Utilizing 3D ribbons and bars in such a complicated network sometime

Fig. 6. Animated changes in passenger flows and propagation of crowdedness

onAnimatedRibbon view related to the spring stormin April 2012.

Fig. 7. A prototype of the 3D wall map view that shows two lines such as

theMarunouchi Line and the Ginza Line. The displayed example shows

the 3D wall map has low readability for displaying multiple lines.



causes cluttering of visualization results and the occlusion problem. To reduce such occlusion, users can zoom, rotate, and pan the 3D space to interactively change the region being focused on. Moreover, users can hide ribbons or bars and select lines to show bars and ribbons on the selected lines. Users can also change the ratio of height for ribbons or bars related to the number of passengers to reduce occlusion. Users can also pan and zoom the route map in the 2D plane. There are some regions in the real route map where the station is extremely dense such as around Tokyo, Shinjuku, and Ueno stations. The occlusion problem of 3D elements easily occurs in such dense areas. Zooming the region in a 2D map is one solution to avoid occlusion in dense areas. However, we sometimes lose the overview of a wide area in zoomed route maps. We therefore implement a map distortion technique using fisheye view [32]. Users can see details in the dense area, which can be specified by interactively selecting a station or a point in the 2D map, and the overview of the surrounding area while maintaining geographical context to some extent (Fig. 11a).

Although such techniques can enhance readability, some occlusion has still occurred on the dense network. Perspective foreshortening makes it difficult to compare the heights of ribbons and/or bars in the 3D space in different places from the camera. To avoid the problem, our system supports an orthogonal projection in which the ribbons and/or bars in different places that are the same height look completely the same.

### 7.2.3 Color Encoding for 3D Ribbons and Bars

The color of each 3D ribbon for each direction is defined using z-scores, S-th, and L-thspecified in Section 7.1.2. The color of each 3D bar also shows deviation, which is defined by z-score normalized by thresholds, from the average number of passengerswho exited each station in the sameway as passenger flows. In both cases, red represents higher than average, blue

represents lower than average, and green represents the normal situation in the samewayasHeatMap view.
We can also use transparency to represent z-score normalized byS-thand L-thdefined in HeatMap view. In this case, ribbons that have a z-score lower than S-thcan be hidden. This emphasizes important sections on which users should focus.

### 7.3 TweetBubble View

For exploring reasons for unusual phenomena or their effects from real users' voices such as what passengers saw, heard, and felt in the situation, TweetBubble view provides an overview of trends of keywords from people's tweets related to specified times and stations or lines, which can be selected byHeatMap view or AnimatedRibbon view. It uses a bubble chart to represent the popularity of important words and Sparklines [33] to show time-series of the appearance of each keyword in each hour for finding bursting timing.
In this view, the center node represents a selected station or line, and other nodes around the center node represent words co-occurring with the station or line name (Fig. 87). Each node holds tfvalue for each hour and dfvalue described in Section 6. We can interactively filter nodes (other than the center node) by total tfvalue of all hours in the day and time window using range sliders shown in Fig. 8.
A TweetBubble view changes the size of nodes in accordance with the weight defined in Section 6 for the selected time window as
weight
p
(r: constant). We
adopt an automatic and dynamic graph layout algorithm based on a force-directed model [34] to visualize bubble charts. Nodes are colored differently in accordance with parts of speech (noun: green, verb: sky blue, adjective: pink).
TweetBubble view embeds Sparklines [33], which are small line charts, into every node to present variation of tf values for words over time (from 0:00 to 24:00). Parts of lines corresponding to the selected time window are highlighted in red.
We can read original tweets including the selected station or line name and word in the selected time window.
The tweets are displayed in the bottom of the view by clicking an arbitrary node. These are sorted by time. Tweets are colored differently in accordance with their types (normal tweet: black, mention: blue, retweet: red).
Utilizing word-cloud representations on geographical maps is one of the design alternatives to visualize trends of keywords related to specified places extracted from social media [28]. However, we have already used the map for displaying passenger flows as AnimatedRibbon view. Overlaying word-cloud on the AnimatedRibbon view would cause serious overplotting and occlusion problem.
Fig. 8.TweetBubble view related to Toyocho station on 3 April 2012, on
which the spring stromcame. It consists of a bubble chart and sparklines to
represent importance ofwords and changes in appearance frequencies.
7. We omit English labels for proper nouns or words that are too common.
92 IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 1, JANUARY-MARCH 2016
8 CASE STUDIES
In this section, we demonstrate how our system can be used to gain insight into changes in behavior of passengers and influences of various kinds of events such as natural disasters (Sections 8.1, 8.2.1 and 8.2.2), accidents (Sections 8.3.1 and 8.3.2), or public gatherings (Sections 8.4.1 and 8.4.2).

### 8.1 The Great East Japan Earthquake

The earthquake struck off the northeastern coast of Japan at 14:46. It had a seismic intensity8 of 5-upper in Tokyo. A public report9 notes that many public transportation systems suspended operation after the earthquake, so most people could not travel until midnight or the next morning. The report stated that the Tokyo Metro Ginza Line and part of the Tokyo Metro Hanzomon Line resumed at 20:40. The Tokyo Metro Nanboku Line also resumed at 21:20. The Toei Oedo Line, a part of the Toei Asakusa Line, and a part of the Toei Mita Line resumed at 20:40, 21:20, and 21:15, respectively. Train operating companies' stuffs knows which lines in their company were suspended operation and resumed during and/or after huge earthquakes. In this case study, we demonstrate whether such operation was really appropriate, and what kinds of problems are occurred by the operation.

Fig. 9 visualizes passenger flows on 11 March 2011, the day on which the Great East Japan Earthquake occurred. Fig 9a shows the situation just before the earthquake. We can find almost all lines were operating normally because their color is mostly green. We can see that almost all lines suspended operation after the earthquake from Figs. 4 and 9b. There are large blue areas in HeatMap view just after 14:46 in Fig. 4. The color of each section turns blue in AnimatedRibbon view in Fig. 9b.

From the red areas shown in the upper-right part of Fig. 4 and the red ribbons in Fig. 9c, we can find a huge number of people were concentrated on the Ginza Line and moving to Shibuya or Asakusa. We can explore the situation in which many people tweeted information such as "Ginza Line is running again" before and after it resumed as shown in Fig. 9d. The spread of such tweets might have accelerated the concentration of people to the Ginza Line and Shibuya station.

We also found that the number of passengers who went to and exited Shibuya rapidly decreased around 21:50 after the concentration using AnimatedRibbon view in Fig. 9e. Such rapid and short-term decreases cannot be shown in HeatMap in Fig. 4. We searched for the reason by reading original tweets around 21:50 using TweetBubble view shown in Fig. 9d. We then found many tweets such as "Ginza Line resumed once, but it is suspended again because of confusion at Shibuya station" from the tweets related to "resuming".

Our result shows importance of controlling the passenger flows after resuming lines and operating together with other public transportation companies during huge disasters.

### 8.2 Typhoons and Storms

Many typhoons pass through Japan every summer and autumn. Moreover, we have many extreme storms even in spring recently. In this section, we show two case studies to demonstrate whether such typhoons and storms cause the similar confusion, what the difference is, whether the measures to typhoons were properly work, and what kinds of problems still remain.

Visualizations related to typhoons and spring storm show similar results. Extreme confusion in Toyocho station observed from these case studies gave people in the operating company one piece of evidence to help them discuss

Fig. 9.Visualizations of passenger flows and tweets on 11 March 2011, the day on which the Great East Japan Earthquake occurred.

8. http://en.wikipedia.org/wiki/Japan_Meteorological_Agency_seismic_intensity_scale

9. http://www.mlit.go.jp/tetudo/tetudo_fr8_000009.html (in Japanese)



improving the transportation system around the east side of

Tokyo. It led to deeper analysis of passengers' behavior that were affected by suspension of Tozai Line [35]. Information obtained from Twitter is very important for understanding the influence of service suspension on activities of people to improve service operation, including cooperation with other transportation services.

### 8.2.1 Typhoon Roke 2011

Wind and rain began to be strengthened from afternoon and became rainstorm about 15:00 on 21 September 2011 in Tokyo. Fig. 10 shows the route map of east side of Tokyo. The Tokyo Tozai Line runs above-ground between Minamisunamachi station and Nishi-funabashi station, which includes railroad bridges. Therefore it often suspends operation during a strong wind. In the case of Typhoon Roke, operation information provided by Tokyo Metro reports that it suspended operation between Minami-sunamachi and Nishi-funabashi from 15:53 to 18:04 and from 20:40 to 21:41, and then it resumed at 21:41. Toei Shinjuku Line also runs above-ground between Higashi-Ojima station and Funabori station. It then suspended operation around 18:30 because of a strong wind.

AnimatedRibbon views in Fig. 10 visualizes changes in passenger flows on 21 September 2011, the day which Typhoon Roke came. Fig. 10a shows that many passengers therefore exited from Toyocho station, and lots of passengers started to use Toei Shinjuku Line to move to eastern area. Fig. 10b shows the situation that most passengers lost methods to move to eastern area because both the Tozai Line and Toei Shinjuku Line were suspended operation. Fig. 10c shows many passengers who were left in Tokyo central started to move again by Tozai Line after it resumed.

### 8.2.2 Spring Storm April 2012

A spring storm that had the same intensity as a typhoon hit the Japanese mainland on 3 April 2012. Many companies in Tokyo urged employees to go home early that day through the experiences of the Great East Japan Earthquake and the Typhoon Roke last year.

The HeatMap view in Fig. 5 and the AnimatedRibbon view in Fig. 6 visualize changes in passenger flows on 3 April 2012. The Tozai Line suspended operation between Minami-sunamachi and Nishi-funabashi around 17:20. It resumed at 21:05.

Figs. 5b-i) and 6a show the Tozai Line, Toei Shinjuku Line, and Tokyo Metro Yurakucho Line became very crowded before the normal rush hours. We can find many passengers exited Toyocho station in Fig. 6b, and passengers started to use Toei Shinjuku Line to move to eastern areas in Figs. 5b-ii and 6b after suspension of the Tozai Line. Red and blue stripes on the Toei Shinjuku Line in Fig. 5b-ii show it could not maintain normal operation. Many people therefore had no routes to take to eastern areas of Tokyo. Fig. 6c shows passengers who had been left in central Tokyo started to move again on the Tozai Line after it resumed.

TweetBubble view in Fig. 8 shows words related to Toyocho station from 15:00 to 24:00. Words shown as huge nodes mainly represent abnormal situations of service such as suspension and free transfer, or their causes such as strong wind. These also include related words such as taxi, bus, and walk that represent passengers' real behavior, how they traveled from Toyocho to their destinations, during the storm. Original tweets including "taxi" are shown under the bubble chart. Most of these tweets said that there was a long line of people at the taxi stand.

## 8.3 Effects of Accidents

Indirect effects of accidents are hard to understand. In this section, we show two examples of suspension of the JR Yamanote Line10 by accidents to demonstrate the influence of other railway companies' accidents on Tokyo Metro, whether any common phenomena occurred, and what the difference is.

From the results, we can confirm similar changes in passenger flows occurred in two case studies. It shows possibility to predict changes in flows during accident in a complex transportation networks. However, these also show the differences among two studies caused by other factors, e.g., in the case of Section 8.3.1, many passengers moved not to Shibuya but to Meiji-Jingumae because of new year's visit to a Shinto shrine. We also recognized that many passengers decided their routes according to tweets by others.

### 8.3.1 Fire at Yurakucho on 3 Jan. 2014

The fire started at around 6:30 a.m. on 3 January 2014 and sent plumes of black smoke over Yurakucho station, which is an important gateway to famous business, shopping, and nightlife districts such as Yurakucho, Tokyo, Ginza, and Shinbashi (as shown in Fig. 11a, which uses the distortion technique mentioned in Section 7.2.2). It caused suspension of the JR Yamanote Line, JR Tokaido Main Line,11 and Keihin-Tohoku Line.

Fig. 10.AnimatedRibbon view around eastern area of Tokyo on 21 September
2011, the day which Typhoon Roke came.

10. The JR Yamanote Line is a loop line that connects most of the major stations in Tokyo.

11. The JR Tokaido Main Line runs from Tokyo, stops at Shinbashi, Shinagawa, and then eventually terminates at Kobe.



Fig. 11 visualizes passenger flows after the fire around JR Yurakucho station. Figs. 11b and 11c show many people changed their routes to their destinations mainly by using the Fukutoshin, Marunouchi, and Chiyoda Lines. We can observe that many passengers switched to the Tokyo Metro Fukutoshin Line in place of the JR Yamanote Line in Fig. 11b. The number of passengers increased mainly between Ikebukuro and Meiji-Jingumae. Many passengers switched from the JR Yamanote Line to the Chiyoda Line (from Kita-Senju to Meiji-Jingumae). Passengers changed to the Tokyo Metro Marunouchi Line to go to Tokyo station.

Fig. 11c shows changes in passenger flows on the Chiyoda Line from Kita-Senju station. Although many people normally transfer from the JR Joban Line to the JR Yamanote Line at Ueno station, they got off at Kita-Senju station and transferred to the Chiyoda Line to go to central Tokyo in this situation.

### 8.3.2 Accident at Ueno Station on 5 February 2013

Fig. 12 visualizes changes in passenger flows after an accident at JR Ueno station on 5 February 2013.12 After the accident, the JR Yamanote Line suspended. The accident happened during the rush-hour, so it affected many passengers.

Figs. 12b and 12c shows many people changed their routes to their destination. For instance, in the route between Shibuya, Shinjuku, and Ikebukuro, many passengers switched to the Fukutoshin Line in place of the JR Yamanote Line. Passengers changed to the Marunouchi Line to go to Tokyo station.

### 8.4 Large Public Gathering Events

In this section, we show two case studies of large public gathering events. First study demonstrates whether we can observe the movement of people when the huge number of people gathers into one region. Second study demonstrates whether we can trace the behavior of people who moved from places to places when large events sequentially occurred in different places. Similar events may occur in the 2020 Tokyo Olympics and Paralympics, so the knowledge obtained from such kinds of case studies would be useful in formulating plans in preparation for them.

#### 8.4.1 A Parade by London Olympic Medalists in Ginza

Fig. 13 visualizes changes in passenger flows on 20 August

2012, the day on which a parade by London Olympic

Fig. 11.Effect on the fire around Yurakucho station on 3 January 2014.

Fig. 12.Effect of the accident at JR Ueno station on 5 February 2013.

12. The position of Ueno station is out of the range of Fig. 12. It can be identified in Fig. 13.



medalists was held in Ginza. The parade lasted about 20 minutes from 11:00, and about 500,000 people gathered.13 Figures show a massive amount of people gathered in Ginza before the parade started and quickly left from Ginza after the parade ended.

By using AnimatedRibbon view shown in Figs. 13b and 13c, we can recognize extremely huge waves of passenger flows occurred before and after the parade. Figs. 13a-i and 13b show that many people moved toward the Ginza area from various quarters after 9:00. We can also find they started to leave Ginza just after the parade ended in Figs. 13a-ii and 13c. This is a surprising result, because Ginza is one of the most famous shopping districts in Japan, but most people did not stay there for long. Fig. 13c shows that many passengers exited Shibuya, Shinjuku, Ikebukuro, Ueno, and Asakusa stations.

### 8.4.2 Tokyo Marathon

Tokyo Marathon is held every February from 2007. Almost 35,000 runners participated and more than one million people supported or cheered from roadside every year. The course of the Tokyo Marathon14 takes the route moving from the west to east, and making round trips between the north and south sides of Tokyo. Many citizen runners finished the race in three to seven hours.

Fig. 14 visualizes changes in passenger flows and tweeted keywords on 26 February 2012, the day on which Tokyo Marathon 201215 was held in Tokyo. It shows that many spectators moved between main spectator viewing points to cheer runners and tweeted situations. AnimatedRibbon view in Fig. 14a shows that many people gathered inIidabashi which is just after the first water supply point after the start from Shinjuku at 9:10. We can also find many people gathered in Hibiya which is the finish of 10 km Race in Fig. 14a. We select a time window in which a term "marathon" is bursting in each TweetBubble view shown in Figs. 14i, 14 ii, 14 iii, 14 iv, 14v. We can find that the sparklines of tweets related to Iidabashi have short peak times, and such peak times becomes longer as the spectator viewing points approach the goal. Fig. 14b shows many people moved from Shinjuku and Iidabashi to Ginza. Ginza is the halfway point. Runners make the turn at Asakusa (Fig. 14b), then pass through Ginza again. Ginza is therefore the most popular spectator viewing points. Fig. 14c shows that many people move from Ginza to Toyosu that is a transfer point to Tokyo Big Sight (Goal of Tokyo Marathon).

## 9 EXPERTS REVIEW

The main target users for our system are staffs of train operating companies. We therefore interviewed four domain experts (all males and ages 40-60) who specialized and had expert knowledge in train operating systems of Tokyo Metro, and obtained their feedbacks of our exploration system. All experts were familiar with visualization systems, but only one of them was familiar with 3D software.

We first briefly explained our system overview and visual encoding, and then demonstrated to them the four case studies that we presented in Sections 8.1, 8.2.2, 8.3.1, and 8.4.2.

Most of them thought that each view (HeatMap, AnimatedRibbon, andTweetBubble view) is useful tools for them, and the system can be a useful tool for their work.

Our system can support the evidence-based improvements of customer services, the results of visualization may throw light on facts that even station staff did not know or give evidence for the situations that they understood somehow.

Experts commented that HeatMap view enabled them to roughly grasp crowded time periods and sections, however it was difficult to understand what happened in the

Fig. 13. Visualization of passenger flows on 20 August 2012, the day on which a parade by London Olympic medalists was held in Ginza.

13. http://www.joc.or.jp/english/londonolympics/parade.html
14. Tokyo Marathon 2012 course, http://www.tokyo42195.org/2012_en/map/
15. http://www.tokyo42195.org/2012_en/



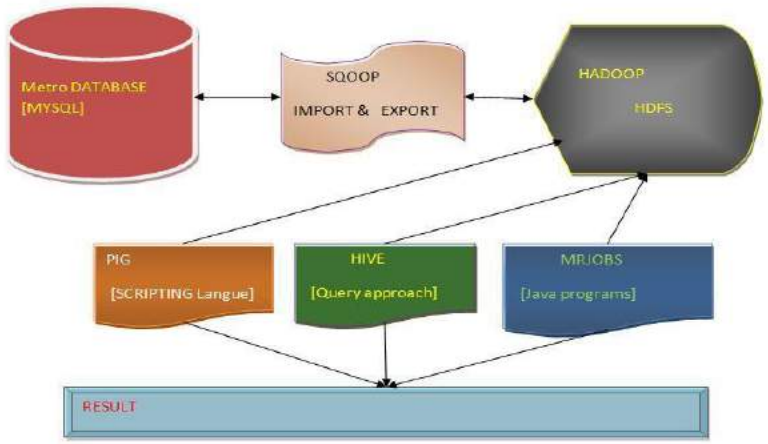
concrete such as some lines resumed. AnimatedRibbon view helped them to understand amount of flows and the overview of movement on the whole network than using HeatMap view. Fig. 11 shows only some selected lines, but, it is better to show all lines on AnimatedRibbon view to understand relationships with other lines. However, 3D view in AnimatedRibbon view was hard to see detailed situations when multiple lines were complicatedly crossing each other.

Some experts commented that TweetBubble view was hard to see, and recommended that it was better to just arrange in order from highest frequency keywords. Tweet-Bubble view enabled them to understand the situation during the unusual events, and some experts suggested that they wanted to easily access to more detailed information, e.g., by categorizing them (such as outside situation, transfer methods, or improvement status), drilling down keywords, or filtering by multiple keywords. Moreover, one expert recommended that it would be easy to recognize what happened in each situation by simultaneously displaying AnimatedRibbon view and TweetBubble view, e.g., by overlaying word-cloud on AnimatedRibbon view.

Two experts commented that they wanted to visually explore which routes passengers used for taking detours after disasters in more detail. One experts commented "many similar situations were occurred by earthquakes or accidents, but they were not the same. So, if we can compare multiple situations and recognize the differences, the system will become more useful."

Some experts also commented "By using this system, we can understand influence of passenger flows among different lines to some extent, and obtained knowledge can be used for optimal operation of transportation systems, and navigation of passengers.", "We recognize extraordinary congestion in Ginza area, Shibuya, Shinjuku, Ikebukuro stations (in the case of Section 8.4.1), such kinds of information can be used for allocation of sufficient staff and other resources in each station.", and "Understanding of passenger flows per time would have possibility to help to plan extra trains."

SYSTEM ARCHITECHTURE




**Data Preprocessing Module:**

In this module we have to create Data set for bank dataset it contain set of table such that customer details, account details, transaction details overall marks details for last year

**Data Migration Module with Sqoop**

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop

**Data Analytic Module with Hive**

Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs

**Data Analytic Module with Pig**

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop. The language for Pig is pig Latin. Pig handles both structure and unstructured language

**Data Analytic Module with MapReduce**

MapReduce is a processing technique and a program model for distributed computing based on java. The

MapReduce algorithm contains two important tasks, namely Map and Reduce.

The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

GENERAL ALGORITHM

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted maprogrammers to use the MapReduce model

## 10 CONCLUSION

We proposed a novel visual fusion environment to explore

changes in flows of passengers on the Tokyo Metro and their causes and effects by using more than four years' worth of data extracted from the smart card system and Twitter.

Our novel approach enables us to extract and visualize (1) passenger flows on a complicated metro network from large scale data from the smart card system and (2) unusual phenomena and their propagation on a spatio-temporal space.

Moreover, (3) we integrated two forms of big-data (data from the smart card system and Twitter) into a visual exploration system to explore causes and/or effects of unusual phenomena. The case studies and reviews by train operating system experts showed the possibilities and usefulness of our system to observe real situations during the events.

We plan to provide mechanisms for automatic detection and prediction of events, and prediction and control of passenger flows on wide and complex transportation networks through fusing various kinds of big data streams including train trips information.

## REFERENCES

[1] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst., 2013, pp. 334–343.

[2] R. Kr€uger, D. Thom, and T. Ertl, "Visual analysis of movement behavior using web data for context enrichment," in Proc. IEEE Pacific Vis. Symp., 2014, pp. 193–200.

[3] M.-P. Pelletier, M. Tr_epanier, and C. Morency, "Smart card data use in public transit: A literature review," Transp. Res. Part C: Emerging Technol., vol. 19, no. 4, pp. 557–568, 2011.

[4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 1–55, 2014, Art.no. 38.

[5] I. Ceapa, C. Smith, and L. Capra, "Avoiding the crowds: Understanding tube station congestion patterns from trip data," in Proc. ACM SIGKDD Int. Workshop Urban Comput., 2012, pp. 134–141.

[6] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system," in Proc. ACM SIGKDD Int. Workshop Urban Comput., 2012, pp. 142–148.

Fig. 14. Visualizations of passenger flows and tweets on 26 February 2012, the day on which Tokyo Marathon 2012 was held in Tokyo.



[7] W. Zeng, C. Fu, S. M. Arisona, A. Erath, and H. Qu, "Visualizing mobility of public transportation system," IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1833–1842, Dec. 2014.

[8] G. L. Andrienko and N. V. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in Proc. IEEE Symp. Vis. Anal. Sci. Technol., 2008, pp. 51–58.

[9] A. Slingsby, J. Wood, and J. Dykes, "Treemap cartography for showing spatial and temporal traffic patterns," J. Maps, vol. 6, no. 1, pp. 135–146, 2010.

[10] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 1813–1822, Dec. 2014.

[11] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D information visualization for time dependent data on maps," in Proc. 9th Int. Conf. Inf. Vis., 2005, pp. 175–181.

[12] S. Thakur and A. J. Hanson, "A 3D visualization of multiple time

series on maps," in Proc. 14th Int. Conf. Inf. Vis., 2010, pp. 336–343.
[13] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. T. Silva, "Using topological analysis to support event-guided exploration
in urban data," IEEE Trans. Vis. Comput. Graph., vol. 20, no. 12, pp. 2634–2643, Dec. 2014.
[14] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. van de Wetering, "Visual traffic jam analysis based on trajectory data," IEEE Trans. Vis. Comput. Graph., vol. 19, no. 12, pp. 2159–2168, Dec. 2013.
[15] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel,
"From movement tracks through events to places: Extracting and characterizing significant places from mobility data," in Proc. IEEE Conf. Visual Anal. Sci. Technol., 2011, pp. 161–170.
[16] G. L. Andrienko, N. V. Andrienko, P. Bak, D. A. Keim, and S. Wrobel, Visual Analytics of Movement. New York, NY, USA: Springer, 2013.
[17] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of New York city taxi trips," IEEE Trans. Vis. Comput.Graph., vol. 19, no. 12, pp. 2149–2158, Dec. 2013.
[18] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, "Flowstrates: An
approach for visual exploration of temporal origin-destination data," Comput. Graph.Forum, vol. 30, no. 3, pp. 971–980, 2011.
[19] X. Jiang, C. Zheng, Y. Tian, and R. Liang, "Large-scale taxi O/D
visual analytics for understanding metropolitan human movement patterns," J. Vis., vol. 18, no. 2, pp. 185–200, 2015.
[20] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "stacking-based visualization of trajectory attribute data," IEEE Trans. Vis. Comput. Graph., vol. 18, no. 12, pp. 2565–2574, Dec. 2012.
[21] T. Cheng, G. Tanaksaranond, C. Brunsdon, and J. Haworth, "Exploratory visualisation of congestion evolutions on urban transport networks," Transp. Res. Part C: Emerging Technol., vol. 36, no. 0, pp. 296–306, 2013.
[22] N. V. Andrienko and G. L. Andrienko, "Spatial generalization and
aggregation of massive movement data," IEEE Trans. Vis. Comput. Graph., vol. 17, no. 2, pp. 205–219, Feb. 2011.
[23] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J.
vanWijk, J. Fekete, and D. W. Fellner, "Visual analysis of large graphs: State-of-the-art and future research challenges," Comput. Graph.Forum, vol. 30, no. 6, pp. 1719–1749, 2011.
[24] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, "The state of the art
in visualizing dynamic graphs," in Proc. EuroVis STAR, 2014, pp. 83–103.
[25] S. Hadlak, H. Schulz, and H. Schumann, "In situ exploration of large dynamic networks," IEEE Trans. Vis. Comput. Graph., vol. 17, no. 12, pp. 2334–2343, Dec. 2011.
[26] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "LeadLine: Interactive visual analysis of text data through event identification and exploration," in Proc. IEEE Conf. Visual Anal. Sci. Technol., 2012, pp. 93–102.
[27] A. M. MacEachren, A. R. Jaiswal, A. C. Robinson, S. Pezanowski,
A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter analytics support for situational awareness," in Proc. IEEE Conf. Visual Anal. Sci. Technol., 2011, pp. 181–190.
[28] D. Thom, H. Bosch, S. Koch, M. W€orner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages," in Proc. PacificVis, 2012, pp. 41–48.
[29] D. Yokoyama, M. Itoh, M. Toyoda, Y. Tomita, S. Kawamura, and
M. Kitsuregawa, "A framework for large-scale train trip record analysis and its application to passengers' flow prediction after train accidents," in Proc. 18th Pacific-Asia Conf. Adv. Knowl. Discovery

Data Mining, 2014, pp. 533–544.
[30] E. R. Tufte, The Visual Display of Quantitative Information. Cheshire,
CT, USA: Graphics Press, 1983.
[31] T. N. Dang, L. Wilkinson, and A. Anand, "Stacking graphic elements
to avoid over-plotting," IEEE Trans. Vis. Comput. Graph.,
vol. 16, no. 6, pp. 1044–1052, Nov./Dec. 2010.
[32] M. Sarkar and M. H. Brown, "Graphical fisheye views," Commun.
ACM, vol. 37, no. 12, pp. 73–83, 1994.
[33] E. R. Tufte, Beautiful Evidence. Cheshire, CT, USA: Graphics Press,
2006.
[34] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by
force-directed placement," Softw. Practice Experience, vol. 21,
no. 11, pp. 1129–1164, 1991.
[35] S. Kawamura, Y. Tomita, M. Itoh, D. Yokoyama, M. Toyoda, and
M. Kitsuregawa, "An effective use of Tokyo metro passengers
flow by visualization of smart card ticket 'PASMO' origin-destination
data for public transport network to be sustainable," in Proc.
WECC, 2015.

Masahiko Itohreceived the PhD degree in information sceience from Hokkaido University, Japan in 2007. He is a specially appointed associate professor of the Institute of Industrial Science, the University of Tokyo, Japan. He was a postdoctoral fellow at Hokkaido University from 2007 to 2009, and a research associate at the University of Tokyo from 2009 to 2014. His research interests include information visualization and 3D graphical user interface. He is a member of the IEEE.

Daisaku Yokoyama received the PhD degree in science from the University of Tokyo, Japan in 2006. He is a specially appointed research associate of the Institute of Industrial Science, the University of Tokyo, Japan. He was a research associate from 2002 to 2007 and from 2009 to 2014, and a specially appointed research associate from 2007 to 2009 at the University of Tokyo. His research interests include parallel and distributed computing and search algorithm. He is a member of the IEEE.

Masashi Toyoda received the PhD degrees in computer science from the Tokyo Institute of Technology, Japan, in 1999. He is an associate professor of the Institute of Industrial Science, the University of Tokyo, Japan. He worked at the Institute of Industrial Science, the University of Tokyo, as a specially appointed associate professor from 2004 to 2006. His research interests include web mining, user interfaces, information visualization, and visual programming. He is a member of the IEEE.

Yoshimitsu Tomita received the BA degree in behavioral science from Chiba University in 1992. He is a deputy manager of ICT Strategy Dept. at Tokyo Metro Co., Ltd., Japan. He works at Tokyo Metro Co., Ltd. (formerly known as the Teito Rapid Transit Authority) from 1992. Now he is in charge of information systems, mainly security management and effective use of smart card data.

98 IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 1, JANUARY-MARCH 2016

Satoshi Kawamura received the BE degree in electrical engineering in 1978 and the ME degree in electronics engineering from the University of Tokyo in 1980, and is currently working toward the PhD degree in the University of Tokyo, Japan. He is a director of PASMO Co., Ltd., Japan. He

worked at Tokyo Metro Co., Ltd., Japan (formerly known as the Teito Rapid Transit Authority) from 1980, and was a director of Information Systems. Now he is in charge of project planning and system development. His research interests include passenger service improvement using trip data extracted from the smart card PASMO system.

Masaru Kitsuregawareceived the PhD degree in information engineering in 1983 from the University of Tokyo. He is the director general at the National Institute of Informatics (NII) in Japan and is also a professor at the University of Tokyo. In 1983, he joined the Institute of Industrial Science, the University of Tokyo as a lecturer. He was the president of Information Processing Society of Japan from 2013 to 2014. His research interests include high-performance database engineering and big data system. He received the ACM SIGMOD E. F. Codd Innovation Award in 2009. He was serving as a science advisor to the Ministry of Education, Culture, Sports, Science, and Technology in Japan. He is a fellow of the ACM and the IEEE.

" For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

ITOH ET AL.: VISUAL EXPLORATION OF CHANGES IN PASSENGER FLOWS AND TWEETS ON MEGA-CITY METRO NETWORK 99