CityProphet: City-scale irregularity prediction using transit app logs

Tatsuya Konishi
The University of Tokyo
Tokyo, Japan
tkonishi@miubiq.cs.titech.ac.jp

Mikiya Maruyama
Yahoo Japan Corporation
Tokyo, Japan
mmaruyam@yahoo-corp.jp

Kota Tsubouchi
Yahoo Japan Corporation
Tokyo, Japan
ktsubouc@yahoo-corp.jp

Masamichi Shimosaka
Tokyo Institute of Technology
Tokyo, Japan
simosaka@miubiq.cs.titech.ac.jp

ABSTRACT
Thanks to the recent popularity of GPS-enabled mobile phones, modeling people flow or population dynamics is attracting a great deal of attention. Advances in methods where regular population patterns with respect to factors such as holidays or weekdays are extracted have provided successful results in irregularity detection. With large-scale crowded events such as fireworks, it is crucial that there be enough time to take countermeasures against the irregular congestion, i.e., irregularity prediction. It remains a tough challenge to predict population from GPS trace logs with existing methods.

To tackle this problem, we focus here on route search logs, since aggregation of the location-oriented queries of individual plans serves as a mirror of short-term city-scale events, in contrast to GPS mobility logs. This paper presents a brand new framework for city-scale event prediction: CityProphet. By our observation of data where the route search logs related to a future event are in most cases repeatable and accumulated in proportion as the event draws near, we are able to leverage the divergence between the above two properties to predict city-scale irregular events. We demonstrate through experiments using the transit app logs of over 370 million queries that our approach can successfully predict city-scale crowded events one week in advance.

Author Keywords
City-scale irregularity prediction; Urban computing; User schedule information; Transit app logs;

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous

UbiComp '16, September 12 - 16, 2016, Heidelberg, Germany

ISBN 978-1-4503-4461-6/16/09…$15.00

DOI: http://dx.doi.org/10.1145/2971648.2971718

INTRODUCTION
Of especial interest in terms of awareness of user context is location data, which previously had been collected only by special terminals exclusively for that purpose [1, 6, 12, 15]. Thanks to the groundwork laid by prior research, we are able to grasp the regular characteristics of regions by modeling the location data acquired intermittently and massively, which is useful for applications related to the opening plans of stores or property sales [3, 4, 6]. For the dramatically changing modern society, however, responsive specific urban dynamics is strongly needed in addition to the regular population patterns. When a large-scale disaster such as an earthquake or a hurricane occurs, for example, visualizing the people flow enables us to comprehend the situation and thereby save lives and help recover the city functions [4, 7, 10, 13, 14]. In response to these issues, there have been many studies focusing on irregularity "detection".

However, since irregularity detection essentially follows its occurrence, it is disadvantageous in the sense that there is no time to take countermeasures against the detected irregularities caused by artificial events, which of course are the most important times to take actions quickly. Technology for irregularity "prediction" is useful for optimizing the placement of guides and planning the unexpected issuance of public transport when artificial events such as festivals or concerts cause temporary overconcentration of the population [19].

Although existing research on the extraction of general population patterns can predict the regular population patterns in regions, these methods are less successful at predicting the fluctuating population brought about by irregular events because they are specialized to fit the regular population patterns. Moreover, although the latest pioneering work [5] has reported success in predicting irregularity, it can only make predictions about an hour before the target irregularity. This begs the question of why it is so difficult to predict the irregularity. The reason is that the amount of data of unusual days with some events is so much less than that of regular days with no event that regular population patterns for events are difficult to extract. Due to the difficulty in precisely predicting usual and unusual populations simultaneously, the existing methods have mostly given up on irregular prediction and focus instead upon irregularity detection by comparing the obtained real population with predicted regular population.

In the present research, our objective is to develop a method that can predict irregularity. We call the proposed scheme "CityProphet". Since previous methods had difficulty in precisely predicting the future population by using only GPS traces and extracting patterns from them, we have built a prediction model from the schedule data of locations for accurate population prediction. The key point here is that the events causing an unusual population are certainly irregular for the region, while at the same time, the attendance at the events is in exact accordance with the scheduled action of the individual user. Besides, the usual population with no event will also match the regular schedule of the users—in other words, user schedule information is expected to model regional populations that reflect the number of those who investigate their routes in advance following their schedule regardless of whether or not events are held there. It is also a fact that regular patterns affected by factors such as day of the week or the weather are observed from the actual population in regions. Although previous works on irregularity detection have focused on the difference between regular population and actual population in the moment, the difference between predicted regular population and schedule in terms of location that is strongly correlated to the actual population serves as irregularity prediction. In terms of user schedule information, this paper focuses on the user route search results.

The contributions of this research are as follows.

- This is the first report in the world of an irregularity prediction model, CityProphet, that compares a regular population with a short-term population from location schedule information.

- Through experiments using actual big query data, we found that the proposed model can achieve irregularity prediction that is fast enough to take the appropriate countermeasures. This is in contrast to state-of-the-art works (e.g., [5]) can only predict irregularity an hour in advance.

RELATED WORKS

Isaacman, et al. developed a population estimation method at the city-scale by constructing models from each user's favorite regions and routes [8]. Since this assumes, however, that each user repeatedly selects the same routes, it is difficult to estimate population in regions where they do not often go. Song, et al. suggested an evacuation route estimation method using data of the Great East Japan Earthquake [17]. However, because the training data used can be collected only after the disaster actually occurs, estimation prior to a disaster cannot be done, even though this is the time when help is most urgently needed. Additionally, it falls short in practical terms in that the method cannot reuse data that have been collected in previous disasters. Sasaki, et al. constructed a system that can detect the occurrence of an earthquake and estimate the epicenter within one second by using data from Twitter[1] [13].

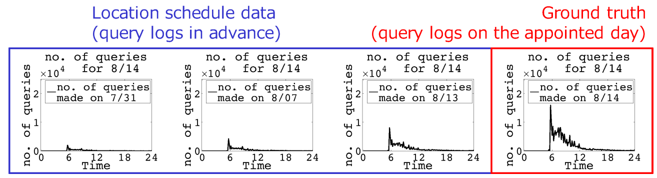

Figure 1. Possible prediction by using aggregated schedule information concerning Kokusai-tenjijō Sta. for 14 August 2015.

This is a good example of how the actions on an SNS are reflected in the real world, which might be beneficial for use with prediction technology to develop evacuation routes.

Even though there have been many methods for irregularity detection using location data, irregularity prediction has remained difficult [18]. A framework on irregularity prediction beyond detection has been proposed in the recent works [9, 11], but irregularity prediction on the city-scale level has yet to be resolved. As a state-of-the-art work, Fan, et al. focused on the fact that people flow is much different from usual at times shortly before large events are held and predicted the next hour's population an hour in advance [5]. However, one hour in advance does not leave enough time to take appropriate countermeasures against events that are likely to cause irregularity in the city-scale population. Although previous works have focused on location data to detect irregularity, using schedules pertaining to location will help us predict irregularity accurately and far enough in advance compared to using only location data. Although there are already a few look-ahead services targeted at individuals (such as Google Now[2]) that utilize calendar information, as yet there has been no city-scale analysis that exploits schedule information.

IRREGULARITY PREDICTION BY EXPLOITING USER SCHEDULES

Transit app log dataset overview

As one example of user schedules in modeling the prediction of a population, in this paper we use the log data of a route search service (specifically, we use the log data of Yahoo! JAPAN[3]). The log data contains the time of passing each train station on the results of route searches, so the number of people at any time and at any station can be obtained quantitatively. The most characteristic feature of this service is that users are able to specify any day they want to query (e.g., they can query routes for 8 April when it is still 1 April). Hence, the results of searching routes at future dates can be exploited as the user schedules. In fact, as shown in Fig. 1, in their scale, all of the queries concerning Kokusai-tenjijō Sta. for 14 August 2015 that were made on 31 July 2015, 7 August 2015, and 13 August 2015 are as relevant as the queries made on the actual day (14 August). In this paper, we focus on user schedules accumulated in this way and use them to construct the population prediction model. On the other hand, as shown in Fig. 2, the population of a region metamorphoses its scale and pattern strongly depending on various factors such as day of the week, and hence, this model can predict the regular population from these factors. Hereafter, we refer

[1] https://twitter.com/

[2] https://www.google.com/intl/en/landing/now/

[3] http://transit.yahoo.co.jp/

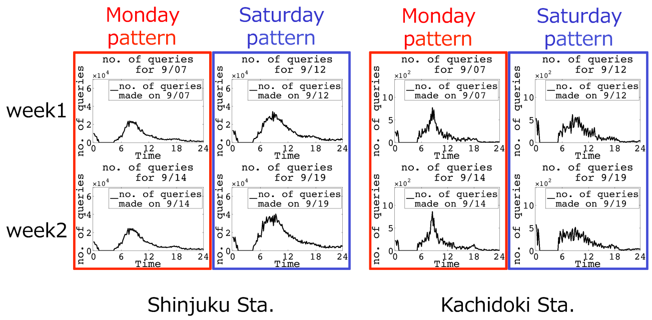

Figure 2. Population changes dramatically depending on region or day of the week. Left: Shinjuku Sta., right: Kachidoki Sta.

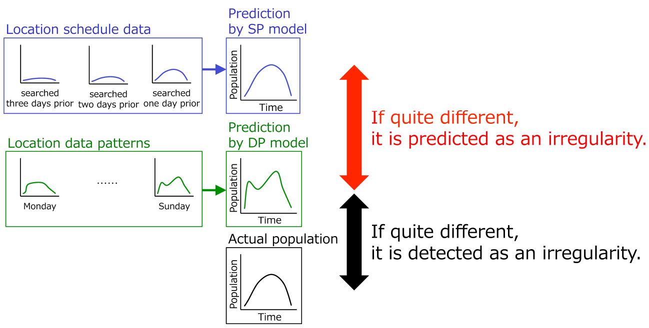

Figure 3. Irregularity prediction by SP and DP models.

to the former short-term population model as the schedule-based population (SP) model and to the latter regular population model as the descriptor-based population (DP) model.

The result of a route search denoting the log of the timetable on data d at time t from the searched date $d-i$ (i.e., at the point in time i days in advance) is defined as $x_{d,t|d-i}$. The ground truth of prediction in our method is the number of search results for a date that was searched on the same date. Although in fact the number of people who searched routes passing through a station is strictly different from the actual population of the region, we assume a correlation here so that there is nothing wrong with irregularity prediction, namely our main purpose. On this assumption, in this paper, we represent the number of those as "population".

Methodology overview

In previous works, irregularity is "detected" as follows. First, the general population of target regions is predicted by using methods of extracting the regular population patterns of the regions. Second, the actual population is obtained in real time or later. Finally, the above two populations, predicted and informed, are compared and if they are quite different, irregularity is "detected". In turn, the population given at the second step should be changed from actual to a predicted value in order not to "detect" but to "predict" irregularity.

To achieve irregularity prediction, in this work, we replace the second part with a highly accurate short-term population prediction model that takes into account user schedule information. As shown in Fig. 3, irregularity is "predicted" with the SP model as follows. The first step is the same as above. Second, a highly accurate short-term population prediction model considering the user schedule information, which is route search logs, is built and we try to predict the population correctly regardless of whether or not any event is being held. Finally, the above two populations predicted by these models are compared and if they are quite different, irregularity is "predicted".

Proposed model

Quantitative irregularity prediction using transit app logs

In order to judge the occurrence of irregularity at a certain date and time, there should be a quantitative index representing the deviation between the DP and SP models. In this paper, we define the deviation measure $\nu_{d,t} = \frac{|\hat{x}^{\mathrm{SP}}_{d,t|d} - \hat{x}^{\mathrm{DP}}_{d,t|d}|}{\hat{x}^{\mathrm{DP}}_{d,t|d}}$ for date d and time t. If this measure exceeds the threshold $\bar{\nu}$, it is judged as an irregularity. The time variable t is discretized by the same interval Δ, which is adjusted empirically. Moreover, $\hat{x}^{\mathrm{SP}}_{d,t|d}$, $\hat{x}^{\mathrm{DP}}_{d,t|d}$ represent the predicted population at day d and time t by the SP and DP models, respectively. If $\hat{x}^{\mathrm{DP}}_{d,t|d} = 0$, ν is not calculated.

SP: Schedule-based population model

As stated above, our objective is to build a model that can predict the population regardless of whether or not people are attending special events by means of user schedule information such as the results of route searches. We focus on the user schedule information because we assume that although an event is irregular for a region, in actual fact, it is within the scope of the plan (i.e., is not irregular) for those who are going to the event. Our approach is to model the correlation between the ground truth and the number of logs for a date searched a few days in advance that can be regarded as the user schedule information.

Our proposed SP model formulates the following equation by referring to the auto regressive model [2],

$$\hat{x}_{d,t|d} = \sum_{i=p_d}^{p_d+p_w} \sum_{j=-q}^{q} \varphi_{i,j} x_{d,t-j|d-i}, \tag{1}$$

where q is the range of time over which the data affects the population target (i.e., only the result of route search for ${}^{\forall}t', t-q \leq t' \leq t+q$, which is searched on day $d-i$, has an effect on the prediction of time t). Similarly, the ground truth, namely the search results for day d, which is searched on day d, is expressed as $x_{d,t|d}$, the same as above. The parameters are optimized by minimizing the negative log likelihood with penalized l_2 norm regularization.

DP: Descriptor-based population model

No matter how accurately our proposed SP model, which considers user schedules, predicts the population of regions, it is still difficult to "predict" irregularity by using only the result predicted by the SP model. Therefore, the usual population at regions should also be predicted and then compared with those predicted by the SP model. As one approach to predicting the usual population, we can exploit previous models connected to research on urban dynamics. We use features such as the weather, day of the week, whether it is a holiday, and so on to predict the patterns of a region.

In this work, we use the bilinear low-rank Poisson regression model [16], which is known as a the state-of-the-art method.

In this model, the population distribution is assumed to follow the Poisson distribution and population is predicted with the model parameter $\lambda_{\boldsymbol{f},\boldsymbol{g}} = \exp(\boldsymbol{f}^{\mathsf{T}}\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{g})$, where this model makes the bilinear features $\boldsymbol{f} \in \mathbb{R}^{M}$ and $\boldsymbol{g} \in \mathbb{R}^{S}$, which are designed in terms of outer factors such as day of the week or the weather and in terms of temporal factors, respectively. In addition, $\boldsymbol{U} \in \mathbb{R}^{M \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{S \times K}$ are parameter matrices. The estimated population $\hat{x}_{d,t|d}$ is generated from a Poisson distribution as follows. Note that K and M satisfy $K \ll M$.

$$\hat{x}_{d,t|d} = \mathbb{E}\left[\mathrm{Pois}(x_{d,t|d}|\lambda_{\boldsymbol{f},\boldsymbol{g}})\right] = \lambda_{\boldsymbol{f},\boldsymbol{g}} \tag{2}$$

EXPERIMENT

Dataset details

The dataset we used for the experiments covers 73 days from 20 July 2015 to 30 September 2015 and comprises roughly 370 million queries, about 96% of which are the results of future-date-oriented searches, namely schedule information. In general, users sometimes search routes to know only routes rather than the timetable of a train. Since such data cannot be part of any schedule, only the query data that are searched with a specific date and time are used.

Short-term population prediction using the SP model

Before evaluating the effectiveness of irregularity prediction, which is the main purpose of this research, we check whether our proposed SP model can predict population regardless of whether city-scale events are held or not.

Descriptor-based population prediction using the DP model

Since it will be compared with the SP model in irregularity prediction, the DP model is only expected to accurately predict regular population, not irregularity population. The evaluation was performed with two changing parameters: the dimensions of the parameter matrix k and the coefficient γ, which are relevant to the regularization. As the outer factor $\boldsymbol{f}$, we use 14 dimensioned features including the tensor product of day of the week (seven dimensions) and whether it is a holiday (two dimensions).

Results

A large-scale event, Comiket88[4], was held in August 2015 and no less than 550,000 people attended over the course of three days. To determine if there was any difference between the accuracy of the SP and DP models, we predicted the population of both the last day of the event and the day one week after the event (i.e., with no event in this region), plotting the results in Fig. 4. Under the use case to take measures against the event in advance, the population is predicted at the day one week before, that is, $p_d = 7$. On one of the regular days (right side of the figure), there is no difference between the SP and DP models. In contrast, on an irregular day with 100 times more population as regular days (left), the SP model could predict accurately but the DP model could not follow the population explosion. The fact that there is a difference between the SP and DP models only on an irregular day supports our hypothesis that comparing the SP and DP models will enable us to predict irregularity successfully.

[4]http://www.comiket.co.jp/index_e.html

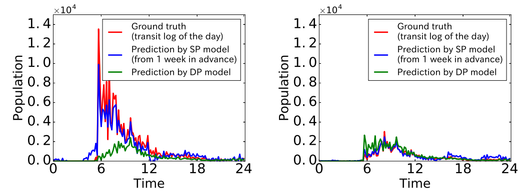

Figure 4. Population prediction for a city-scale event, Comiket88. Left: the day of Comiket88, right: the day one week after Comiket88.

Table 1. Details of each event.

No.	Name (Sta.)	Scale
1	Fireworks (Kachidoki)	∼700 thousand
2	Comiket88 (Kokusai-tenjijō)	∼200 thousand
3	Fireworks (Katase-Enoshima)	∼20 thousand
4	Baseball game (Korakuen)	∼50 thousand
5	Typhoon (Haneda Airport)	—

Prediction of city-scale events by our model

We confirmed that exploiting user schedule information enables us to predict short-term population in specific regions. To determine whether it is possible to predict irregularities, we performed another experiment in which we compared the DP model, which represents the regular population, with the SP model, which accurately represents short-term population regardless of whether an event is held. In addition to large-scale events such as Comiket88, small-scale events and a typhoon were selected as target events in this experiment. An overview of all selected events is given in Table. 1. Concerning the typhoon, the scale is not described. Due to the use case in predicting the typhoon, its prediction is done on one day before the target date (i.e., $p_d = 1$), and for the others, predictions are done a day one week before (i.e., $p_d = 7$). So as to predict irregularity, the deviation rate ν is compared with the threshold $\bar{\nu} = \{5.0, 1.0, 0.5\}$. Additionally, to examine the irregularity predicted result compared with the irregularity detected result, the deviation rate of the detection $\mu_{d,t} = \frac{|\hat{x}^{\mathrm{DP}}_{d,t|d} - x_{d,t|d}|}{x_{d,t|d}}$ is defined and compared with the threshold $\bar{\mu} = \{0.8, 0.5, 0.3\}$, as with the irregularity prediction.

Results

The results concerning five events are plotted in Fig. 5. The top graph represents the ground truth and the population predicted by the SP and DP models. The second graph represents the calculated deviation rate and the threshold of judgment of irregularity. To ensure readability, values higher than 8 are not plotted. The third graph represents the time zone in which ν exceeds the threshold $\bar{\nu}$ and is judged as an irregularity. The fourth graph represents the results of detection of irregularity the same as the prediction. For city-scale events such as Nos. 1 and 2, even with a large threshold ($\bar{\nu} = 5.0$), we could predict irregularity. For the smaller fireworks display (No. 3), although it showed the same trend, the irregularity around noon could not be predicted, probably because of the small scale of the event. However, the irregularity caused by the baseball game (No. 4) could not be predicted. We assume this is because the region has a stadium where baseball games and concerts are held every day, the game was not technically "irregular". As for the typhoon (No. 5), we could not predict it as accurately as we could the artificial events, even though we knew a day in advance that the typhoon was coming.

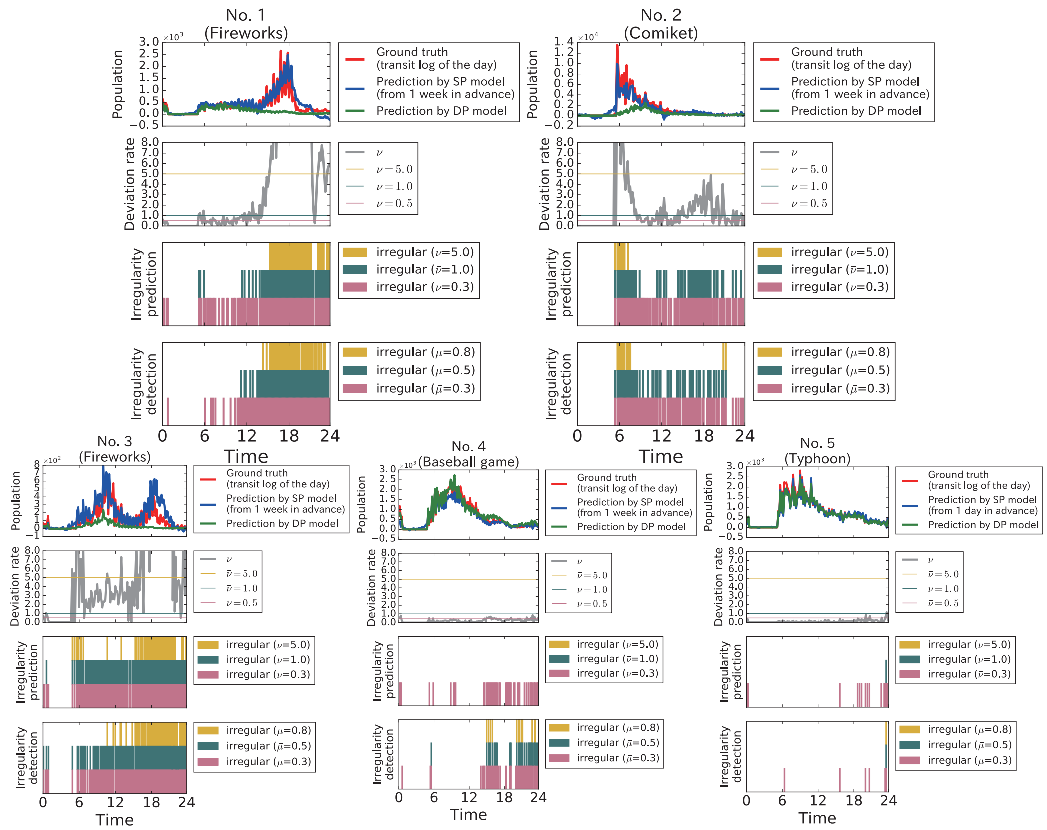

Figure 5. Irregularity prediction of Nos. 1, 2, 3, 4, 5. To "predict" irregularity quantitatively, the deviation rate is calculated by comparing the SP and DP models. The result of irregularity "prediction" (the 3rd graph) is about the same as that of irregularity "detection" (the 4th graph).

Table 2. F-measures of irregularity prediction and detection.

	No. 1	No. 2	No. 3
F-measure	0.84	0.73	0.74

When we assume that, additionally, irregularity "detection" works well, we calculated the F-measures regarding the irregularity detection result as the ground truth in order to verify the performance of our proposed method of irregularity "prediction". The F-measures of irregularity prediction and detection of Nos. 1, 2, and 3, in which artificial events were accurately predicted, are shown in Table. 2. From these high values, it is clear that our proposed method can predict irregularity one week in advance while maintaining the accuracy of irregularity detection. We utilize a simple and naive irregularity prediction based on the fixed threshold ($\bar{\nu} = 0.5$): however, the results imply that our method is able to handle multiple scale city events (Nos. 1 and 2) so as to achieve accurate irregularity prediction. We should be able to improve the accuracy with a more sophisticated approach (e.g., adaptive thresholding), which will be handled in future works.

CONCLUSION

In this paper, we have proposed a brand new scheme that can predict irregularities one week in advance (which earlier methods cannot do) by exploiting user schedules involving future-oriented location data. We presented the observation of this kind of data and then provided a framework with two basic prediction models: 1) the DP model, which predicts regular population patterns, and 2) the SP model, which predicts short-term population by considering user schedule information. Although the SP model is defined similarly to the conventional auto regressive model, it differs in that we utilize user schedules as input, resulting in accurate prediction regardless of whether events are held. Moreover, our proposed framework is advanced in that we can use other schedule information to predict irregularity. Although our proposed method has difficulty in predicting sudden natural events because our approach depends on user schedules and intentions, it can usually predict irregularity caused by low-frequency, large, and deliberate events. In this paper, the deviation rate is formulated with a simple equation, leading us to conclude that a better expression will enable us to predict different particle size irregularities more accurately.

Certainly, irregularities caused by large events have already been known well in general, however, our approach is able to provide a quantitative measure of irregularities even when the first event (such as new railway opening) is held. That is, to the best of our knowledge, our framework is the first framework for real world application that provides the quantitative measure of irregularity caused by events in advance by using user schedules, which enables us to take countermeasures against it. In this work, we predicted irregular population per station, but irregularity per train line might also be predicted from our data, for example, and the exploding popularity of products could be also predicted from query logs pertaining to them. We are considering the utilization of other schedule information and will examine the feasibility of this in future works.

REFERENCES

1. Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. Towards a better understanding of context and context-awareness. In Proc. of HUC (1999), 304–307.
2. Akaike, H. Fitting autoregressive models for prediction. AISM (1969), 243–247.
3. Cheng, C., Yang, H., Lyu, M. R., and King, I. Where You Like to Go Next: Successive Point-of-Interest Recommendation. In Proc. of IJCAI (2013), 2605–2611.
4. Fan, Z., Song, X., and Shibasaki, R. Cityspectrum: A non-negative tensor factorization approach. In Proc. of UbiComp (2014), 213–223.
5. Fan, Z., Song, X., and Shibasaki, R. CityMomentum : An Online Approach for Crowd Behavior Prediction at a Citywide Level. In Proc. of UbiComp (2015), 559–569.
6. Gao, H., Tang, J., Hu, X., and Liu, H. Content-Aware Point of Interest Recommendation on Location-Based Social Networks. In Proc. of AAAI (2015), 1721–1727.
7. Haddawy, P., De Felice, G., Frommberger, L., Charkratpahu, P., Saengpao, S., Kanchanakitsakul, P., and Kauppinen, T. Situation awareness in crowdsensing for disease surveillance in crisis situations. In Proc. of ICTD, No. 38 (2015).
8. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., and Willinger, W. Human mobility modeling at metropolitan scales. In Proc. of MobiSys (2012), 239–252.
9. Lian, D., Xie, X., Zheng, V. W., Yuan, N. J., Zhang, F., and Chen, E. CEPR: A collaborative exploration and periodically returning model for location prediction. TIST (2015), 1–27.
10. Ludwig, T., Reuter, C., Siebigteroth, T., and Pipek, V. Crowdmonitor: mobile crowd sensing for assessing physical and digital activities of citizens during emergencies. In Proc. of CHI (2015), 4083–4092.
11. McInerney, J., Stein, S., Rogers, A., and Jennings, N. R. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. Pervasive and Mobile Computing (2013), 808–822.
12. Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., and Ratti, C. Activity-Aware Map: Identifying human daily activity pattern using mobile phone data. HBU (2010), 14–25.
13. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proc. of WWW (2010), 851–860.
14. Salfinger, A., Girtelschmid, S., Proll, B., Retschitzegger, W., and Schwinger, W. Crowd-Sensing Meets Situation Awareness: A Research Roadmap for Crisis Management. In Proc. of HICSS (2015), 153–162.
15. Schilit, B., Adams, N., and Want, R. Context-aware computing applications. In Proc. of HotMobile (1994), 85–90.
16. Shimosaka, M., Maeda, K., Tsukiji, T., and Tsubouchi, K. Forecasting Urban Dynamics with Mobility Logs by Bilinear Poisson Regression. In Proc. of UbiComp (2015), 535–546.
17. Song, X., Zhang, Q., Sekimoto, Y., Horanont, T., Ueyama, S., and Shibasaki, R. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In Proc. of SIGKDD (2013), 1231–1239.
18. Triantakonstantis, D., and Mountrakis, G. Urban growth prediction: a review of computational models and human perceptions. JGIS (2012), 555–587.
19. Wang, L., Zhang, D., and Xiong, H. effSense: energy-efficient and cost-effective data uploading in mobile crowdsensing. In Proc. of UbiComp (2013), 1075–1086.