Detecting Train Delays using Railway Network Topology in Twitter

Yuanyuan Wang[1], Yusuke Nakaoka[2], Panote Siriaraya[2], Yukiko Kawai[2], and Toyokazu Akiyama[2]

[1] Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi, 755-8611 Japan
y.wang@yamaguchi-u.ac.jp
[2] Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto, 603-8555 Japan
{g1444936,kawai,akiyama}@cc.kyoto-su.ac.jp, spanote@gmail.com

Abstract. This paper presents a novel train delay detection method based on topic propagation analysis of geo-tagged tweets between railway stations. Our goal is to detect traffic accidents and to predict train delays in railway network topology by tracing how relevant tweets propagate in real space and cyberspace. In our method, we utilize railway network as the topology of real space, and extract the topology of social network that is mapped on the railway network. This permits observing the influence of delays on stations with a few tweets, or predicting related tweets of affected stations even if the tweets contain indirect topics about delays.

Keywords: topic propagation, railway topology, delay detection

1 Introduction

Many researches have been recently undertaken to detect topics over space from Twitter [2]. Previous works have a common target for detecting real-world events based on geographical areas or location mentions, such as a spatio-temporal event visualization system [3], an earthquake reporting system [5], and a traffic detection system [4]. However, the location-based analysis has been mainly limited to the analysis of tweets based on their location stamps, as given by GPS coordinates. In particular, traffic event detection only focused on the time and place of accidents from Twitter analysis [1]. However, it is hard to analyze a few distributed tweets on complexity road in real time, so they have not achieved the forecast on the impact of delays by the tweets.

To address the above, we propose a train delay detection method based on spatio-temporal topic propagation analysis by considering the relationship between stations in railway network topology using geo-tagged tweets. By using our proposed method, it can detect delays at stations with a few tweets, and it can also detect delays with indirect tweeting. Therefore, we aim to clarify to what extent the corresponding events can be detected from tweets by labeling the events with only the attribute of event occurrence time durations.

In this paper, we utilize 488 stations of 62 routes in Tokyo area in Japan which might be one of the most complex railway networks in the world, to verify the effectiveness of accident delay detection by the proposed method.

Table 1. Example of detected stations, routes and their expected impacts

Rank	Station	Route	Expected Impact
1	Tokyo Sta.	Yamanote Line	Large (90%)
2	Aoyama 1-Chome Sta.	Tokyu-Denentoshi Line	Medium (81%)
3	Ginza Sta.	Tokyo Metro Ginza Line	Small (63%)

2 System Overview

When traffic accidents or train delays occur, we analyze geo-tagged tweets of each station and analyze topic propagation by considering both the topologies of real space and cyberspace. For this, we first collect and analyze Twitter data from stations. Then, we extract the time and the stations where accidents or delays occurred by applying neural networks for learning the past train status information collected from a route website (Jorudan). Specifically, each station is assigned as a label , and vector words are learned by using a *BoW* (Bag of Words) approach. Vectorization is optimized using *TF-IDF* weights and dimension reduction by *LSI* (Latent Semantic Indexing).

Table 1 shows an example of detected stations, routes, and their expected impacts when the delay occurred at "Ueno Station" on the Yamanote Line in Tokyo. We can detect stations or routes that are less affected by delays, and traffic congestion can be predicted by topics such as accidents or delays.

3 Topic Propagation Analysis

In order to analyze topic propagation between railway stations, we construct the topology of the social network by calculating virtual connections between stations. In our study, we learn collected tweets by applying neural networks to generate a dictionary. In the dictionary, vector words of the learning data are weighted based on *TF-IDF*. Then, important vectors can be extracted for calculating the topology of stations based on dimension reduction by *LSI*.

3.1 TF-IDF Weighting

We first acquire geo-tagged tweets of each station within a radius r (r=50m) by using Twitter Streaming API. The learning data is transformed into feature vectors using *BoW*, and vector words are weighted by *TF-IDF* as follows:

$$\frac{\#i \text{ in each station}}{\text{total \#words in each station}} \cdot \frac{\text{total \#tweets in all stations}}{\#\text{tweets with } i}$$

Therefore, important words can be extracted by calculating the cosine similarities between stations. Here, the number of dimensions is the total number of words and it is enormous, dimension reduction is needed.

3.2 Dimension Reduction by *LSI*

Dimension reduction is aimed at reducing learning costs in neural networks. *LSI* can compress synonyms into one vector by indexing the latent meanings of words. Through the *LSI*, we reduced the number of dimensions of a vector space to 300 dimensions as the number of dimensions of input.

3.3 Topology Construction of Social Network

We extract the influence on each station where accidents or delays occurred, and then simultaneously compute the relevance to other stations. Specifically, tweets of each station where accidents OCCURRED or NOT are converted into a feature vector set, and then assigning them with the label of accident AFFECTED or NO accident affect to learn by the neural network. Therefore, we construct the topology of the social network that is mapped on the railway network of the real space by analyzing the propagation situations of accident topics.

4 Evaluation

In this section, we evaluate the accuracy of delay detection by the proposed method on datasets derived from Twitter data with the actual delay information acquired from a route website, called Jorudan[3].

4.1 Experiment Environment

The Twitter data used in a dictionary has been collected from 488 railway stations of 62 routes in Tokyo area during one year. For the learning data, we used a tweet set of each hour of the date when the topic of accident delays occurred, and we verified the influence of accidents that occurred in the Yamanote Line. The learning data is shown in Table 2, We assigned three labels of accident AFFECTED on the learning data as follows:

a. Tweets when accidents OCCURRED are INCLUDED in the learning data
b. Tweets when accidents OCCURRED are NOT included in the learning data
c. Tweets when accidents NOT occurred are NOT included in the learning data

4.2 Experimental Results

The accuracy of the input test data **a.** for the learning data is 85% and the accuracy of **c.** is 75%. The accuracy of **b.** is 43%, it is lower than those of **a.** and **c.**. Because very few tweets in the test data for the learning data with only three days in this experiment, especially for comparing **b.** with **c.**, the tweet data size of **b.** for the learning data is extremely small. We need to improve the accuracy of the prediction by increasing the tweet data size for the learning data.

[3] http://www.jorudan.co.jp/unk/

Table 2. Learning data

	Date	Time Period	Route
1)	2016/10/08	09:00-11:00	Yamanote Line
2)	2016/10/30	13:00-16:00	Yamanote Line
3)	2016/11/06	06:00-08:00	Yamanote Line

5 Conclusion

In this paper, we have proposed a topic propagation analysis of geo-tagged tweets based on railway network topology for observing the influence on railway stations when a delay occurs at a station, or predicting related tweets of the delayed station even if the tweets contain indirect topics about delays. Experimental results show that our proposed method can effectively detect delays and the influence of stations when delays occurred, compared with the actual delay information.

For future work, We will try to analyze topics of tweets about delays on each hour or each minute to observe the speed of topic propagation. Furthermore, we plan to expand the current analysis method to recommend routes or stations to avoid traffic congestion and delays.

Acknowledgments

This work was partially supported by MIC SCOPE (0159-0089), and JSPS KAKENHI Grant Numbers 16H01722, 17K12686, 17H01822.

References

1. D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F.: Real-time detection of traffic from twitter stream analysis. IEEE Transactions on Intelligent Transportation Systems 16(4), 2269–2283 (2015)
2. Goonetilleke, O., Sellis, T., Zhang, X., Sathe, S.: Twitter analytics: A big data management perspective. SIGKDD Explor. Newsl. 16(1), 11–20 (2014)
3. Itoh, M., Yoshinaga, N., Toyoda, M.: Spatio-temporal event visualization from a geo-parsed microblog stream. In: Companion Publication of the 21st International Conference on Intelligent User Interfaces. pp. 58–61. IUI '16 Companion (2016)
4. Kokate, S., Bhosale, S.: Traffic detection using tweets on twitter social network. International journal of advance research in computer science and management studies 4(7), 84–88 (2016)
5. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering 25(4), 919–931 (2013)