# Possibility of Using Tweets to Detect Crowd Congestion: A Case Study Using Tweets just before/after the Great East Japan Earthquake

**Takuya Oki**
Tokyo Institute of Technology
oki.t.ab@m.titech.ac.jp

## ABSTRACT

During large earthquakes, it is critical to safely guide evacuation efforts and to prevent accidents caused by congestion. In this paper, we focus on detecting the degree of crowd congestion following an earthquake based on information posted to Social Networking Services (SNSs). This research uses text data posted to Twitter just before/after the occurrence of the Great East Japan Earthquake (11 March 2011 at 02:46 PM JST). First, we extract co-occurring place names, proper nouns, and time-series information from tweets about congestion in the Tokyo metropolitan area (TMA). Next, using these extracted data, we analyze the frequency and spatiotemporal characteristics of these tweets. Finally, we identify expressions that describe the degree of crowd congestion and discuss methods to quantify these expressions based on a questionnaire survey and tweets that contain a photograph.

### Keywords

Twitter, crowd congestion, time-series analysis, linguistic expression, disaster mitigation.

## INTRODUCTION

If we can quickly and precisely grasp where and to what extent it is crowded with people, it will be useful for prevention of accidents caused by congestion, safe evacuation guidance at large earthquakes, and so on.

There are some data or services (such as "Mobile Spatial Statistics[1)]" by NTT DoCoMo, "Density Map[2)]" by ZENRIN DataCom, "Crowd Radar[3)]" by Yahoo! JAPAN), which enable us to grasp the degree of congestion. In these data or services, the crowd density is estimated based on the location information of users who use some kind of mobile terminal like a smartphone. However, it is difficult to estimate the degree of congestion in each facility or on each street and to grasp congestion in real time because the space-time unit of these data is coarse (250 m x 250 m grid for every hour).

It is also important to analyze the mechanism of congestion that occurs in emergencies and to consider measures to control congestion in preparation for future large earthquakes (such as Tokyo Metropolitan Earthquake or Nankai Trough Earthquake). However, at the time of the Great East Japan Earthquake (11 March 2011 at 02:46 PM JST), which is known for large congestion caused by people who tried to go home on foot or to wait for resuming the service of public transportation in the TMA, there are few previous studies and data quantitatively indicating when, where, and how much local and sudden congestion by the crowd was occurring.

Based on the above background, we focus on the possibility of quantifying the degree of crowd congestion based on information posted to SNSs, which are excellent in immediacy and openness. In this paper, as a first step to discuss the usefulness of posted information for grasping the degree of congestion, we attempt to examine Twitter data about congestion at the time of the Great East Japan Earthquake. More specifically, we extract co-occurring place names, proper nouns, and time-series information from tweets about congestion in the TMA. Using these extracted data, we analyze the frequency and spatiotemporal characteristics of these tweets. Furthermore, expressions describing the degree of congestion are listed from the same tweet data and discuss the method to quantify the expressions based on the questionnaire survey and tweets with a photograph.

## RELATED WORK

There are many prior studies on detecting events, emergency situations, or hotspots of crowds by utilizing social media. For instance, Toepke (2017) developed a system to collect tweets using Twitter API under AWS environment automatically. Using the collected tweet data with geotag (information on the location where users posted tweets) in five cities in the United States, the author analyzed the influence of the length of data collection period on the spatial distribution (heat map) of the number of tweets. Likewise, Toepke and Starsman (2015) compared the heat maps of the number of posts to between Twitter and Instagram.

In some studies, not only text information included in posts but also images are used together for grasping situations. For instance, Panteras et al. (2015) extracted geotagged tweets including hashtags consisting of place names and word ‘fire,’ converted these data to point data on the GIS map, and estimated the angle of view (AOV) by using EXIF information in Flickr images by triangulation. As a result, the authors demonstrated that the estimation accuracy (reproducibility) of the range affected by wildfires could be improved by combining Twitter and Flickr, rather than using only Flickr.

Furthermore, other studies considered social media users as one of the sensors (what is called “social sensor”) for real-time event detection. Gottumukkala et al. (2015) constructed a framework of the real-time evacuation decision-making support system. As information obtained in real time, they proposed to use not only traffic sensors and video images but also hashtags included in social media such as Twitter. More specifically, they suggested it effective to utilize the relationship between hashtags and evacuation behaviors before/during/after disaster for real-time decision-making. For another example, Sakaki et al. (2010) constructed a particle filter for extracting events from geotagged tweets and validated the model comparing the actual seismic centers (also the observed trajectories of typhoons) with the positions estimated by the filter for tweets related to earthquakes (and typhoons).

Most of these studies used geotag information included in posts to SNSs. However, it is difficult to collect enough number of tweets with geotag in real time (or in a brief time) because the percentage of geotagged posts is considerably small. More specifically, the situation of crowd congestion varies from moment to moment, and sometimes the congestion by the crowd occurs locally and suddenly. Given this background, this paper attempts to complement the location information by using proper nouns and place names included in text data posted to Twitter.

Besides, the number of tweets is not necessarily correlated with the number of people actually there because the percentage of Twitter users depends on the composition of the crowd and Twitter users do not always post a tweet in the crowded situation. Therefore, we consider the usefulness of linguistic expressions describing the degree of congestion extracted from tweets as well as the number of tweets.

## EXTRACTION OF TWEETS ABOUT CONGESTION

First of all, we randomly extract 10% of all tweets written in Japanese posted for 24 hours after 11 March 2011 at 02:00 PM JST (just before the Great East Japan Earthquake occurred). To collect the past tweets for the analyses in this paper, we use the Twitter data providing service[4)] (named as “Nazuki no Oto”) by NTT DATA Corporation. Although this is a charged service, it enables us to collect the past tweets (even those posted seven years ago) satisfied with extraction conditions (e.g., keywords, duration, and location). The service ensures sampling rate, and here it is set to 10% due to cost restriction. Additionally, the service supports real-time streaming of tweets. When we extract tweets about congestion and utilize them for emergency response in real time in the future, the Twitter Streaming API will be more suitable in terms of introduction cost for collecting the tweets as well as the related study (Gottumukkala et al., 2012).

From the 10% tweets, we focus on tweets containing the *kanji* ‘混 [5)]’, which is used in words meaning “to mix”, “to blend”, “to confuse”, “to be crowded”, and so on. In other words, it is expected that people often use the *kanji* ‘混’ when they post a tweet about congestion (Figure 1). Next, using the function [Search Words] of KH Coder[6)], we examine the common use cases of the *kanji* ‘混’. As a result, the words that are highly likely to be related to congestion can be roughly limited to the following three words: ‘混む [7)]’ (pronounced as ‘*komu*’, which means “to be crowded”); ‘混乱’ (pronounced as ‘*konran*’, which means “confusion”); and ‘混雑’ (pronounced as ‘*konzatsu*’, which means “congestion”). The composition ratios of these three words in tweets containing the *kanji* ‘混’ are 31.6%, 30.4%, and 20.3%, respectively. Therefore, hereafter, we aim at tweets including any of these three words in the text in the 10% tweets.

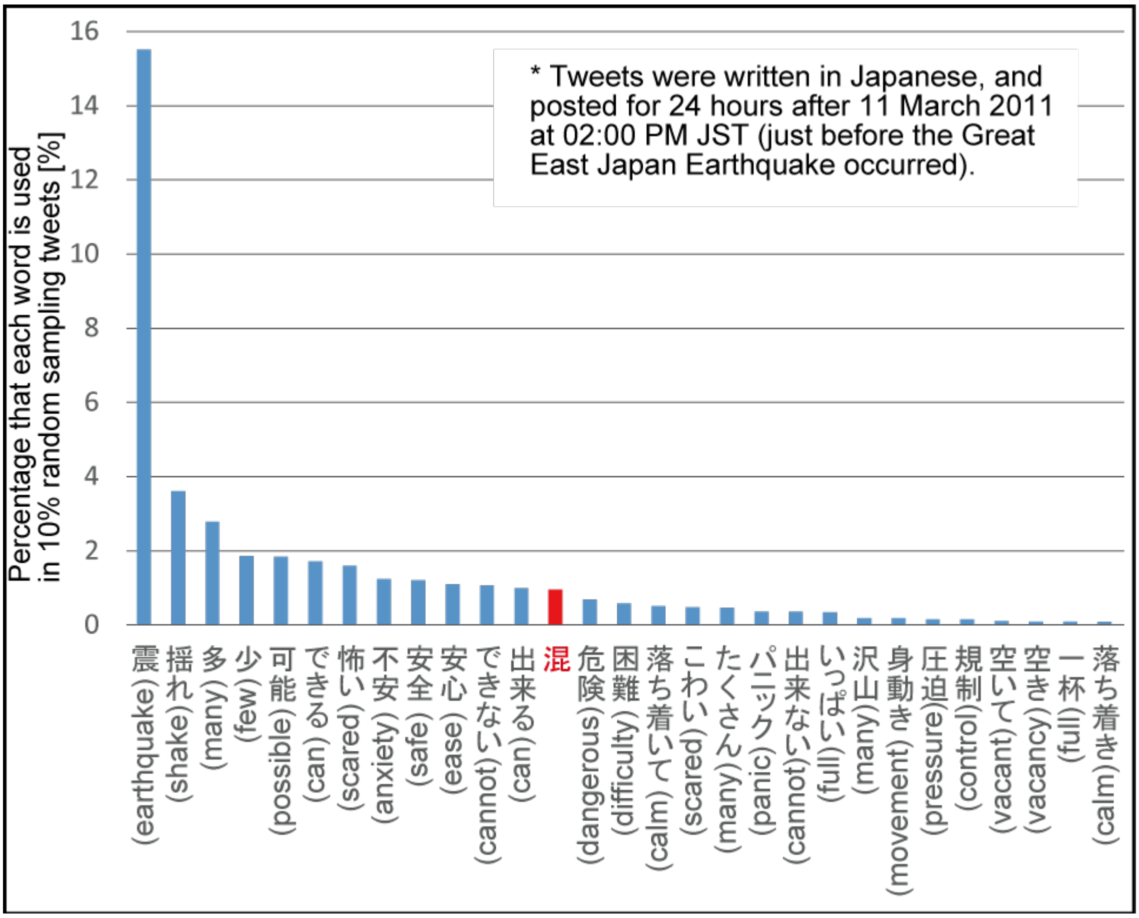


**Figure 1. Frequency of Words Related to Crowd Congestion in 10% Tweets**

## CHARACTERISTICS OF TWEETS ABOUT CONGESTION

### Time Transition of Tweets about Congestion

Looking at the time transition of the number of tweets about congestion (Figure 2), it is found that the total number of the 10% tweets rapidly increased more than five times immediately after the occurrence of the main shock (11 March 2011 at 02:46 PM JST), and then gradually decreased. On the other hand, we can see that the number of tweets about congestion was particularly large after 6:00 pm on 11th (Friday) and after 8:00 am on 12th (Saturday). These time periods corresponded to those when a large number of people tried to go home from places where they were going out. This result suggests that these tweets may well represent the time transition of the congestion.

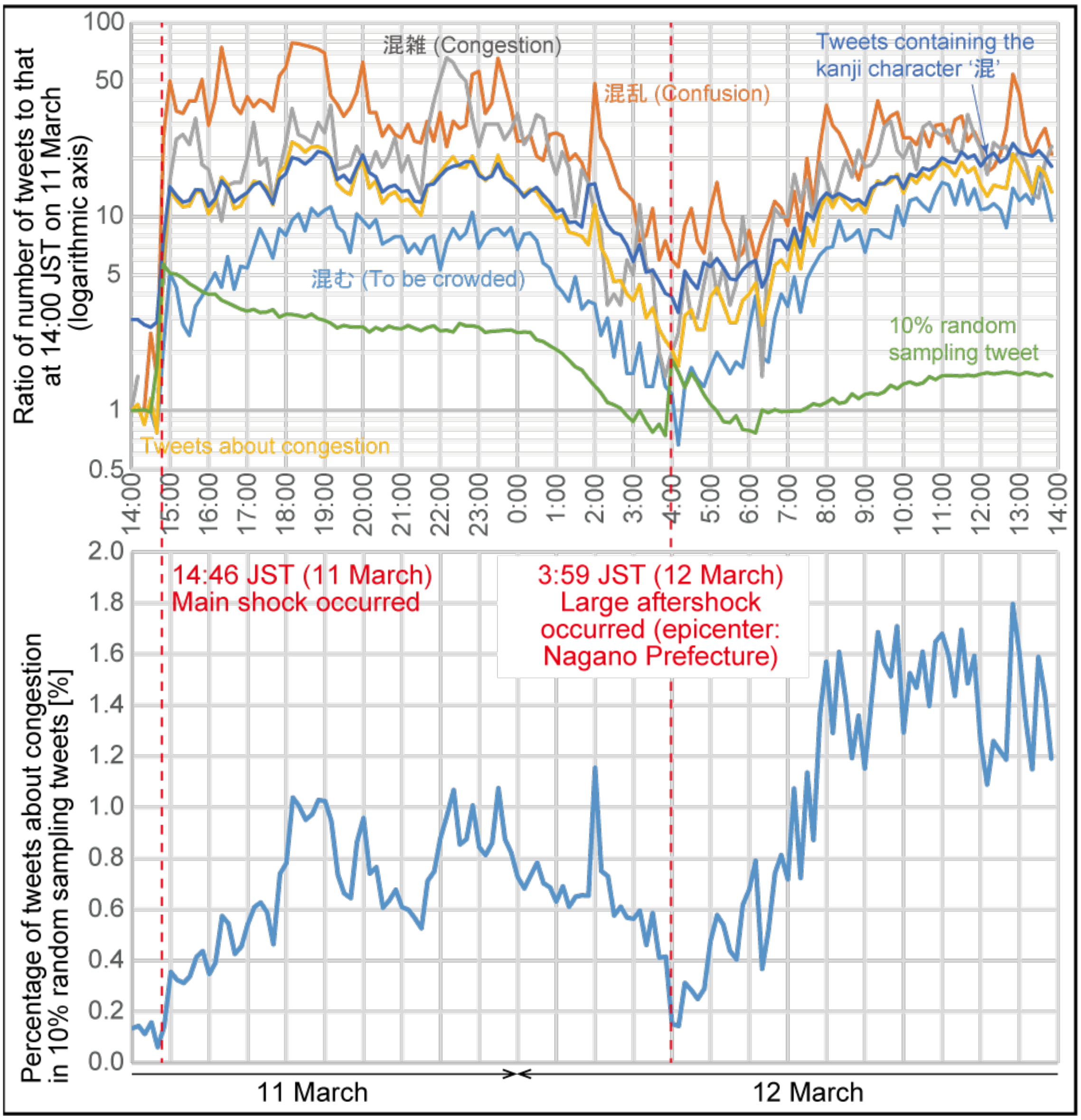


**Figure 2. Time Transition of Number of Tweets (every 10 minutes)**

### Spatial Distribution of Tweets about Congestion

As the percentage of geotagged tweets (with information on the location where users posted tweets) is only about 0.2% of the whole tweets, it is not suitable to use only geotagged tweets for grasping the spatial characteristics of tweets about congestion. Therefore, we extract proper nouns and place names in tweets about congestion by using KH Coder[6)], and attempt to complement location information based on the extracted words (totally 929 words) (Table 1). The word '混雑' tends to co-occur with words related to railway stations and railway lines in the TMA, whereas the word '混乱' tends to co-occur with names of a wide range of area.

These results suggest that: (1) the method to complement location information by proper nouns and place names in the 10% tweets is mainly useful for congestion on public transportation in the TMA; (2) the word '混乱' is not useful for complementing location information.

**Table 1. List of Proper Nouns and Place Names in Tweets about Congestion**

| | **‘混む (*Komu*)’** | | **‘混乱 (*Konran*)’** | | **‘混雑 (*Konzatsu*)’** | |
|---|---|---|---|---|---|---|
| | **Number of unique extracted words = 615** | | **Number of unique extracted words = 353** | | **Number of unique extracted words = 595** | |
| | **Extracted Word** | **Ratio[%]** | **Extracted Word** | **Ratio[%]** | **Extracted Word** | **Ratio[%]** |
| 1 | 日本 (Japan) | 2.6 (3.1) | 淡路 (Awaji) | 2.7 (6.5) | 銀座線 (Ginza line) | 3.2 (11.3) |
| 2 | 東北 (Tohoku) | 1.8 (2.4) | 東京 (Tokyo) | 5.7 (6.1) | 東京 (Tokyo) | 5.7 (10.4) |
| 3 | 愛知 (Aichi) | 0.9 (2.2) | 日本 (Japan) | 2.6 (3.5) | 渋谷 (Shibuya) | 3.0 (6.9) |
| 4 | 横浜 (Yokohama) | 0.9 (1.1) | 東北 (Tohoku) | 1.8 (1.8) | 新宿 (Shinjuku) | 2.5 (4.6) |
| 5 | 京王線 (Keio line) | 0.6 (0.8) | 千葉 (Chiba) | 1.1 (1.7) | 半蔵門線 (Hanzomon line) | 0.7 (2.2) |
| 6 | 東西線 (Tozai line) | 0.7 (0.7) | 関東 (Kanto) | 0.8 (1.6) | 池袋 (Ikebukuro) | 0.8 (2.1) |
| 7 | 山手線 (Yamanote line) | 0.6 (0.7) | 宮城 (Miyagi) | 0.6 (1.2) | 東西線 (Tozai line) | 0.7 (1.8) |
| 8 | 東横線 (Toyoko line) | 0.5 (0.7) | 仙台 (Sendai) | 0.5 (0.8) | 南北線 (Namboku line) | 0.6 (1.6) |
| 9 | バレ (?) | 0.3 (0.7) | 茨城 (Ibaraki) | 0.4 (0.8) | 品川 (Shinagawa) | 0.8 (1.6) |
| 10 | 京浜東北線 (Keihin-tohoku line) | 0.4 (0.5) | 福島 (Fukushima) | 0.4 (0.7) | 東海道線 (Tokaido line) | 0.5 (1.4) |
| 11 | 総武線 (Sobu line) | 0.5 (0.5) | 太平洋 (Pacific ocean) | 0.3 (0.7) | 横浜 (Yokohama) | 0.9 (1.4) |
| 12 | 新宿線 (Shinjuku line) | 0.3 (0.4) | 田端 (Tabata) | 0.2 (0.6) | 千代田線 (Chiyoda line) | 0.4 (1.3) |
| 13 | マック (McDonald’s) | 0.3 (0.4) | 田端新 (?) | 0.2 (0.6) | 北千住 (Kita-senju) | 0.4 (1.3) |
| 14 | 田園都市線 (Den’en-toshi line) | 0.3 (0.4) | 北 (North) | 0.2 (0.6) | 山手線 (Yamanote line) | 0.6 (1.2) |
| 15 | 寄 (?) | 0.2 (0.4) | 気仙沼 (Kesen’numa) | 0.3 (0.4) | 上野 (Ueno) | 0.5 (1.2) |
| 16 | 甲州 (Koshu) | 0.3 (0.4) | 大阪 (Osaka) | 0.2 (0.4) | 銀座 (Ginza) | 0.4 (1.1) |
| 17 | 埼京線 (Saikyo line) | 0.2 (0.4) | 東日本 (East Japan) | 0.3 (0.4) | 有楽町線 (Yurakucho line) | 0.3 (1.1) |
| 18 | 井の頭線 (Inogashira line) | 0.3 (0.3) | 新潟 (Niigata) | 0.3 (0.4) | 荒川線 (Arakawa line) | 0.3 (1.0) |
| 19 | 六本木 (Roppongi) | 0.2 (0.3) | 枝野 (Edano) | 0.2 (0.3) | 京王線 (Keio line) | 0.6 (1.0) |
| 20 | 中央線 (Chuo Line) | 0.2 (0.3) | 日 (?) | 0.2 (0.3) | 吉祥寺 (Kichijoji) | 0.3 (0.9) |
| 21 | 秋葉原 (Akihabara) | 0.2 (0.3) | 中 (?) | 0.2 (0.3) | 三鷹 (Mitaka) | 0.3 (0.9) |
| 22 | 新潟 (Niigata) | 0.3 (0.3) | 名古屋 (Nagoya) | 0.2 (0.3) | 日比谷線 (Hibiya line) | 0.3 (0.9) |
| 23 | 目黒 (Meguro) | 0.2 (0.2) | 近畿 (Kinki) | 0.1 (0.2) | 東横線 (Toyoko line) | 0.5 (0.9) |
| 24 | 朝 (morning) | 0.2 (0.2) | 台東 (Taito) | 0.1 (0.2) | 築地 (Tsukiji) | 0.2 (0.9) |
| 25 | 山手 (Yamanote) | 0.1 (0.2) | 関西 (Kansai) | 0.1 (0.2) | 浅草線 (Asakusa line) | 0.3 (0.8) |

**Notes for Table 1**

* Extracted words are arranged in the decreasing order of proportion that each word is co-occurring with ‘混む’, ‘混乱’ or ‘混雑’.

* Left-side value in each column “Ratio” is the composition ratio (%) of each extracted word in the whole tweets about congestion. Right-side value (in parentheses) in each column “Ratio” is the proportion (%) that each extracted word is used in tweets including the words ‘混む’, ‘混乱’ or ‘混雑’.

* The extracted words which can be considered as the name of railway stations [lines] existing in the TMA are indicated in red [blue] color.

Figure 3 shows the spatial distribution of 588 words existing in the database of names of railway stations/lines[8] out of 929 words extracted above. Among railway lines, there are overwhelmingly many tweets about congestion related to the subway lines (such as the Ginza line, the Tozai line, and the Hanzomon line of Tokyo Metro) and the Yamanote line. Among railway stations, the number of tweets about congestion is large at the major railway terminal stations like Shibuya station, Shinjuku station, and so on.

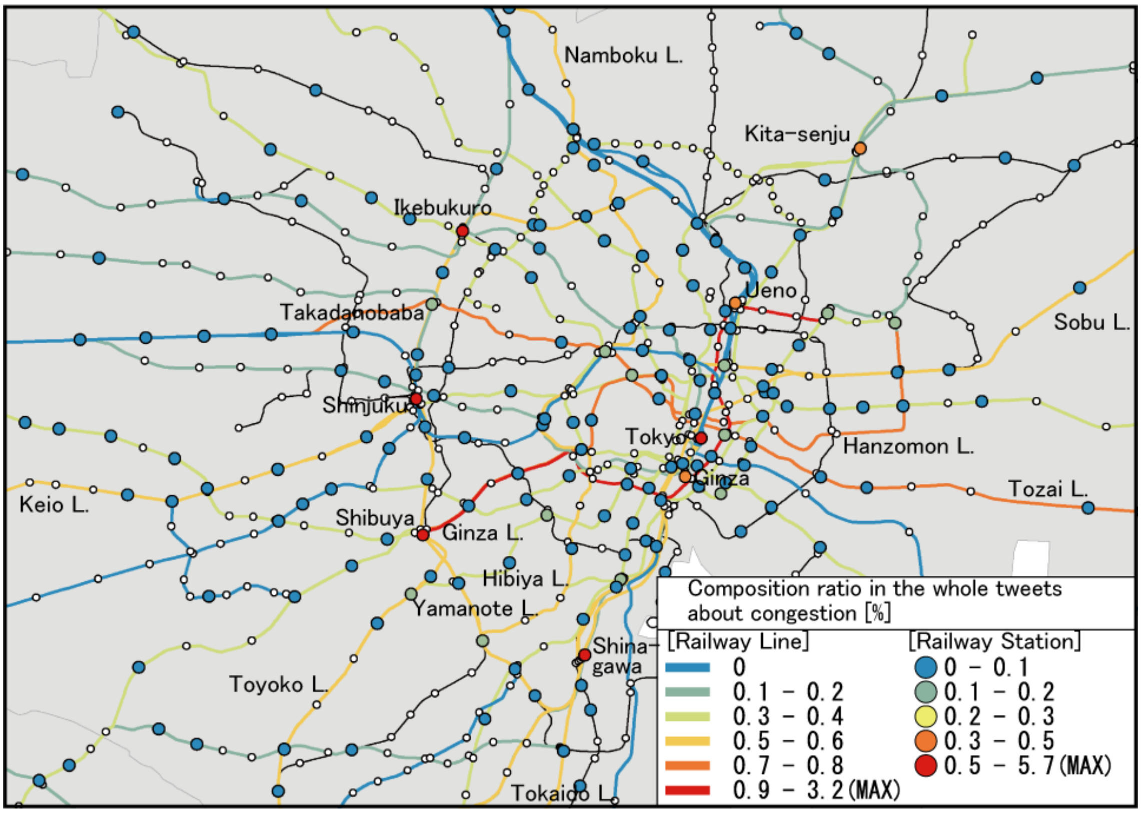
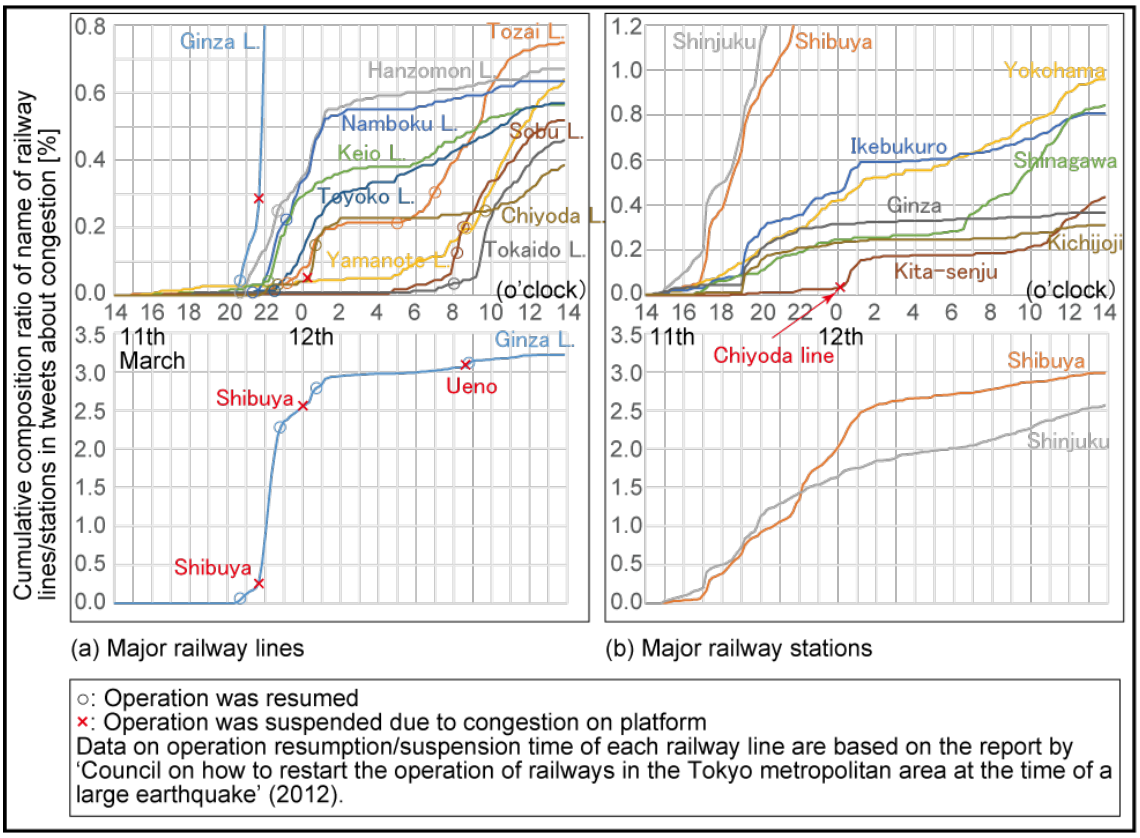


**Figure 3. Spatial Distribution of Number of Tweets about Congestion (around the Main Part of Tokyo)**

The timing when the number of tweets about congestion increases corresponds well with the operation resumption time of railway lines, which suggests the usefulness of such tweets for quick grasp of congestion (Figure 4(a)). Furthermore, paying attention to the trend of increase in the number of these tweets, the number of tweets about the Ginza line drastically increases after 8:00 pm on the 11th, but since then it gradually decreases. It suggests that the congestion on the Ginza line was temporary. By contrast, in Shinjuku, Shibuya, and Yokohama, where the number of tweets about congestion gradually increases from the aftermath of the earthquake occurrence to the next day, there is a possibility that the congestion state lasted for a long time (Figure 4(b)).



(a) Major railway lines (b) Major railway stations

○: Operation was resumed
×: Operation was suspended due to congestion on platform
Data on operation resumption/suspension time of each railway line are based on the report by 'Council on how to restart the operation of railways in the Tokyo metropolitan area at the time of a large earthquake' (2012).

**Figure 4. Time Transition of Number of Name of Main Railway Lines/Stations in Tweets about Congestion**

### Problems in Complementing Location Information with Proper Nouns and Place Names

The number of tweets about congestion seems to correspond to the amount of concern in congestion at a certain time/spot. However, to grasp the degree of congestion, it is necessary to analyze the whole text of tweets in more detail. Moreover, in order to grasp the congestion more densely, it is necessary to: (1) increase the sampling rate of tweets about congestion; (2) utilize useful information such as the name of main roads and landmarks (McDonald's, Disneyland, etc.) out of 341 words (36.7% of all 929 words) other than the names of railway stations/lines extracted by KH Coder[6]. In addition, we will take into account how to deal with the name of places used for various purposes (such as '東京 (Tokyo)') and the name of multiple candidates existing all over Japan (such as '東西線 (Tozai line)').

## EXTRACTION OF CONGESTION LEVEL EXPRESSION

### Problems in Extraction of Congestion Level Expression

To extract linguistic expressions related to congestion, we focus on words co-occurring with the words '混む [7)]', '混乱', or '混雑' same as the 'Extraction of Tweets about Congestion' section. We use KH Coder[6)] to extract co-occurrence words. In using the extracted co-occurrence words, it is necessary to solve the following problems:

*(1) Discrimination of positive/negative and congestion level*

It is an important issue directly related to estimation accuracy of congestion.

*(2) Confirmation of dependency relations*

An extracted co-occurrence word is not necessarily directly related to one of three words '混む', '混乱', and '混雑'. Therefore, dependency analysis (syntax analysis) is required. In general, it is considered that the verb '混む' is modified by adverbs and the nouns '混乱' and '混雑' are modified by adjectives. However, as a result of trying dependency analysis using CaboCha[9)], many modification examples other than mentioned above were seen. This is a problem specific to using tweets in Japanese, where the flexible usage of dependency relations is often seen.

### Training Data on Congestion Level Expression

It is difficult to solve the aforementioned problems by only performing dependency analysis. Therefore, in this paper, we randomly extract 1000 tweets from the tweets about congestion defined in the previous section and attempt to analyze the characteristics of congestion level expressions visually detected from these 1000 tweets. The number of tweets judged as without congestion level expressions is 625 in 1000 tweets, and it is found that a majority of tweets about congestion is not co-occurring with congestion level expression. We will verify whether the number of samples of congestion level expressions is sufficient or not in the future.

Next, we classify the extracted congestion level expressions into two groups: (i) expressions in which he/she wanted to tweet that he/she was facing the congestion (Table 2); and conversely (ii) expressions in which he/she wanted to tweet that it was NOT crowded there (Table 3). Based on the classification result, expressions in Group (i) are more frequently used both in terms of variety and number than expressions in Group (ii). It suggests that crowd congestion can easily urge Twitter users to post the situation of congestion in front of them to Twitter.

In addition to the words that directly modify '混む', '混乱', and '混雑' as mentioned above, there are some expressions by which the degree of congestion is expressed using metaphor or comparison. For instance, expressions related to traffic (such as "difficult to take on/off", "a last train", "rush hour", "crowded train", "commuter hours", etc.), expressions likened to a specific date and time (such as "past 11:00 PM on Friday", "not look like Saturday morning", etc.), expressions compared with the situations in normal times (e.g., "as usual", "not ever", "a usual evening", etc.), expressions related to action/behavior (e.g., "hard to move", "entry restriction", "go down like ninepins", "a long line", etc.). Besides, there are unique expressions such as "there is almost no shopping cart", "like a movie", "over 200 m", and so on.

If we use the knowledge obtained in this section as training data, there is a possibility that congestion level expressions can be automatically and accurately extracted. In future, we are planning to verify the accuracy in extracting congestion level expressions by applying to other tweet data.

**Table 2. Expressions in Which He/She Wanted to Tweet That He/She Was Facing Congestion**

| Expression in Japanese | Meaning | Frequency (Composition ratio) | Expression in Japanese | Meaning | Frequency (Composition ratio) |
|---|---|---|---|---|---|
| 大 (*Dai*) | Heavy | 68 (22.6%) | ごった返す (*Gottagaesu*) | Crowded | 2 (0.7%) |
| 大変 (*Taihen*) | Heavy/Very | 33 (22.0%) | 少し (*Sukoshi*) | A bit | 2 (0.7%) |
| かなり (*Kanari*) | Rather/ Considerably/ Comparatively | 31 (10.3%) | あふれる (*Afureru*) | Full | 2 (0.7%) |
| すごい (*Sugoi*) | Awesome | 29 (9.6%) | 結構 (*Kekkō*) | Pretty/ Quite | 2 (0.7%) |
| 激 (*Geki*) | Extremely/ Terrifically | 21 (7.0%) | 尋常ではない (*Jinjōde wanai*) | Extraordinary | 2 (0.7%) |
| めちゃ (*Mecha*) | Very/Extremely /Excessively | 20 (6.6%) | こんな (*Kon'na*) | Such | 1 (0.3%) |
| すぎる (*Sugiru*) | Too | 14 (4.7%) | とんでもない (*Tondemonai*) | Tremendous | 1 (0.3%) |
| 半端ない (*Hanpa nai*) | Immense/ Enormous/ Tremendous/ Awful/Terrible | 10 (3.3%) | ハイパー (*Haipā*) | Hyper- | 1 (0.3%) |
| ひどい (*Hidoi*) | Heavy/Awful | 10 (3.3%) | 恐ろしく (*Osoroshiku*) | Terribly | 1 (0.3%) |
| ものすごい (*Monosugoi*) | Tremendous | 6 (2.0%) | あんなに (*An'nani*) | So | 1 (0.3%) |
| 超 (*Chō*) | Super | 6 (2.0%) | ビッシリ (*Bisshiri*) | Tightly | 1 (0.3%) |
| やばい (*Yabai*) | Dangerous | 5 (1.7%) | 非常に (*Hijōni*) | Extremely | 1 (0.3%) |
| 激しい (*Hageshī*) | Heavy | 4 (1.3%) | さらに (*Sarani*) | Further | 1 (0.3%) |
| 相当 (*Sōtō*) | Considerable | 4 (1.3%) | 何気に (*Nanigeni*) | Surprisingly | 1 (0.3%) |
| 劇 (*Geki*) | Extremely | 3 (1.0%) | まくる (*Makuru*) | Too much | 1 (0.3%) |
| ちょっと (*Chotto*) | A bit | 3 (1.0%) | ほどほどの (*Hodohodono*) | Moderate | 1 (0.3%) |
| マジ (*Maji*) | Seriously | 3 (1.0%) | やや (*Yaya*) | A little/Slight/ Rather | 1 (0.3%) |
| あまりの (*Amari no*) | Too much | 2 (0.7%) | ありえない (*Arienai*) | Unbelievable | 1 (0.3%) |
| 鬼 (*Oni*) | Very/Extremely /Super | 2 (0.7%) | 若干 (*Jakkan*) | A bit/ A little | 1 (0.3%) |
| 異常 (*Ijō*) | Unusual/ Abnormal | 2 (0.7%) | えらい (*Erai*) | Serious/Awful / Terrible | 1 (0.3%) |

* Total: 301 (100.0%)

**Notes for Table 2**

* In summarizing Table 2, we visually and manually unified the words with notation fluctuation. Some examples are shown below:

(1) Fluctuation of *kanji* notation

(2) Fluctuation of *hiragana* / *katakana* notation

(3) Fluctuation in colloquialism

**Table 3. Expressions in Which He/She Wanted to Tweet That It Was NOT Crowded There**

| Expression in Japanese | Meaning | Frequency (Composition ratio) | Expression in Japanese | Meaning | Frequency (Composition ratio) |
|---|---|---|---|---|---|
| そんなに...ない (*Son'nani … nai*) | Not so … | 4 (13.8%) | 全然...ない (*Zenzen … nai*) | Not … at all | 2 (6.9%) |
| それほど...ない (*Sorehodo … nai*) | Not too … | 4 (13.8%) | そこまで...ない (*Sokomade … nai*) | Not … so far | 1 (3.4%) |
| 大きな...ない (*Ōkina … nai*) | Without any … | 4 (13.8%) | 大して...ない (*Taishite … nai*) | Not too … | 1 (3.4%) |
| あまり...ない (*Amari … nai*) | Not too … | 4 (13.8%) | さほど...ない (*Sahodo … nai*) | Not too … | 1 (3.4%) |
| 特に...ない (*Tokuni … nai*) | Not particularly … | 3 (10.3%) | 驚くほどではない (*Odorokuhodode wanai*) | Not surprising | 1 (3.4%) |
| ...ない (*… Nai*) | Not … | 3 (10.3%) | ゼロ (*Zero*) | Zero | 1 (0.3%) |

* Total: 29 (100.0%)

## DISCUSSION ON QUANTIFICATION METHOD OF CONGESTION LEVEL EXPRESSION

It is an important subject in various research fields how to quantify linguistic expressions about the degree, and there are many previous studies on quantifying level expression. For instance, Shibata et al. (2011) analyzed how much the test subjects adjust the speed of the machine in response to speed adjustment instructions using degree adverbs (such as "just a little", "a little", "quite", and "very"). Based on the experiment results, the authors constructed a model which determines the amount of adjusting speed corresponding to degree adverbs for intelligent machines. Kumamoto (2004) conducted a subject experiment to determine the relation of strength and weakness of 119 words used for strengthening/weakening the impression expressions of music. Furthermore, the author scored each degree word based on the experiment results and evaluated the influence on impression words by using the impression scale. Besides, there are some studies focused on the relationship between words expressing the degree of noise and the actual types of noise (e.g., Yano et al., 2002; Miyagawa and Aono, 2002).

As a method of associating the actual degree of congestion with linguistic expressions, it also can be considered to use the image information posted together with the text to Twitter. It is highly likely that images posted with tweets related to congestion were taken of crowded situations. However, it is found that the percentage of tweets including the URL[10)] of the image posted to "TwitPic", which was one of the main external image posting services as of March 2011[11)], was only about 0.5% of the whole tweets. If it is limited to tweets including any congestion level expression, the percentage will further decrease.

Considering the above background, we are planning to conduct a Web questionnaire survey aiming at ranking congestion level expressions and grasping the expressions selected under crowd congestion. The congestion will be virtually set by showing the test subjects photographs of crowded situations collected in advance. Here, it can be expected to quantify the selection tendency of congestion level expressions according to crowded situation, personal attribute, and individual by requesting the cooperation of test subjects with various personal attributes (such as gender and age) as much as possible (Table 4).

**Table 4. Image of Correspondence Table between Congestion Level Expressions and Actual Degree of Congestion**

| **Congestion Level Expression in Japanese** | **Meaning** | **Actual Degree of Congestion [person/m$^2$]** |
|---|---|---|
| 非常に (*Hijō'ni*) | Very / Extremely / Exceedingly | 2.0 – |
| すごく (*Sugoku*) | Very / Mighty | 1.5 – 2.0 |
| とても (*Totemo*) | Very / Extremely / Exceedingly | 1.3 – 2.5 |
| だいぶ (*Daibu*) | Rather | 1.0 – 1.2 |
| 多少 (*Tashō*) | More or less / Somewhat / A little / A bit / Slightly | 0.5 – 1.0 |
| … | … | … |

## SUMMARY AND CONCLUSIONS

Focusing on the tweets about congestion just before and after the occurrence of the Great East Japan Earthquake, we demonstrated the effectiveness for grasping the time transition of congestion. During the time periods when a large number of people tried to go home, the number of tweets about congestion was also significantly large. This result suggests that Twitter users can play an important role as the social sensors to detect local/sudden congestion by crowds in real time.

Although most of the related studies in the past used the geotag information to identify the location of user's post, the percentage of posts with geotag is considerably small. To overcome this problem, we tried to complement the location information by using proper nouns and place names co-occurring with the tweets. As a result, we showed the possibilities of grasping the spatiotemporal distribution of tweets about congestion by incorporating the proposed method without geotag information into methods in prior work. In the example of the aftermath of the Great East Japan Earthquake, the proper nouns and place names co-occurring with the tweets were mostly related to railway stations/lines in the TMA because many people rushed there after the operation of trains was resumed. Our future work will need to evaluate the applicability of the proposed method to other situations and countries.

Furthermore, we focused on linguistic expressions describing the degree of congestion extracted from tweets because the number of tweets is not necessarily correlated with the actual number of people there. Based on the analysis results of the 'Extraction of Congestion Level Expression' section, there are a certain number of tweets that contain congestion level expressions in the tweets about congestion, and it can be said that the ways of expressing congestion level are diverse. In this paper, the linguistic expressions in the actual tweets written in Japanese were used for analyses, and there were some difficulties specific to Japanese. Although the procedure of analyses in this paper can be easily generalized to tweets written in English or other languages than Japanese, it is necessary to investigate the frequency distribution of expressions similar to Table 2 and Table 3.

If it is possible to quantify congestion level expressions included in tweets, it is expected that it will be possible to estimate the degree of congestion in situations: (i) where/when it is difficult to secure a certain number of mobile terminal users for estimating the degree of congestion at each point; or (ii) where the radio waves are unstable such as in buildings or underground spaces. For that purpose, it is necessary to grasp beforehand to what extent congestion is expressed by each congestion level expression. Therefore, in the 'Discussion on Quantification Method of Congestion Level Expression' section, we discussed various methods for quantifying the extracted congestion level expressions. Of course, it is also important to examine the credibility of information in posted tweets separately. In the future, we will conduct a Web questionnaire survey on the relationship between the actual congestion levels and linguistic expressions, and report the survey results.

## ACKNOWLEDGMENTS

This article was written by adding new analyses and making modifications to the previous papers (Oki, 2016a; Oki, 2016b). This work was supported by JSPS (Japan Society for the Promotion of Science) KAKENHI (Grant-in-Aid for Young Scientists (B)) Grant Number JP15K18176. We also appreciate valuable comments from anonymous reviewers.

## NOTES

1) https://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/ [accessed Mar. 28, 2016]

2) http://lab.its-mo.com/densitymap/ [accessed Mar. 28, 2016]

3) http://map.yahoo.co.jp/maps?layer=crowd&v=3&lat=35.681277 & lon = 139.766266 & z = 15 [accessed Mar. 28, 2016]

4) https://nazuki-oto.com/twitter/ [accessed Mar. 24, 2018]

5) http://tangorin.com/kanji/%E6%B7%B7 [accessed Dec. 10, 2017]

6) http://khc.sourceforge.net/en/ [accessed Dec. 10, 2017]

7) Include conjugations.

8) The database of names of railway stations/lines and Figure 3 were processed by the author based on the National Land Numerical Information (Railways data) as of 2011, which were provided by Ministry of Land, Infrastructure, Transport and Tourism (MLIT).

http://nlftp.mlit.go.jp/ksj-e/gml/datalist/KsjTmplt-N02.html [accessed Dec. 10, 2017]

9) CaboCha: https://taku910.github.io/cabocha/ [accessed Dec. 10, 2017]

10) The URL which starts with “http://twitpic.com/”.

11) Before August 2011, the official image posting service was not supported by Twitter, Inc.

## REFERENCES

Gottumukkala, R., Zachary, J., Kearfott, B. and Kolluru, R. (2012) Real-time information driven decision support system for evacuation planning, *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support* (*CogSIMA*), 206-209.

Kumamoto, T. (2004) Relative ranking of degree modifiers and its application to impression-based content-retrieval by natural language input (in Japanese), *IPSJ SIG Technical Reports*, 2004-NL-164(13), 77-82.

Miyakawa, M. and Aono, S. (2002) Reexamination of rating scales on the impressions of environmental sounds (in Japanese), *Journal of the Acoustical Society of Japan*, 58, 3, 151-164.

Oki, T. (2016a) Characteristics of tweets about congestion status in Tokyo Metropolitan Area immediately after Great East Japan Earthquake (in Japanese), *Proceedings of the 78th National Convention of IPSJ*, Information Processing Society of Japan, 2, 5-6.

Oki, T. (2016b) Extraction and quantification of the expressions of congestion status in information posted to SNS – A case study for tweets just after the Great East Japan Earthquake – (in Japanese), *Proceedings of 30th Annual Conference of the JSAI*, Japanese Society for Artificial Intelligence, 1I2-NFC-01-4in2, 1-3.

Panteras, G., Wise, S., Lu, X., Croitoru, A., Crooks, A. and Stefanidis, A. (2015) Triangulating social multimedia content for event localization using Flickr and Twitter, *Transactions in GIS*, 19, 5, 694-715.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors, *Proceedings of the 19th international conference on World Wide Web* (*WWW '10*), 851-860.

Shibata, S., Yamamoto, T. and Jindai, M. (2011) Fundamental approach on velocity adjustment of intelligent machines using degree adverb (in Japanese), *Japanese Journal of Ergonomics*, 47, 4, 155-159.

Toepke, S. L. and Starsman, R. S. (2015) Population distribution estimation of an urban area using crowd sourced data for disaster response, *Proceedings of the 12th ISCRAM Conference*.

Toepke, S. L. (2017) Temporal sampling implications for crowd sourced population estimations from social media, *Proceedings of the 14th ISCRAM Conference*, 564-571.

Yano, T., Igarashi, J., Kaku, J., et al. (2002) International joint study on the measurement of community response to noise: Construction of noise annoyance scale in Japanese (in Japanese), *Journal of the Acoustical Society of Japan*, 58, 2, 101-110.