Understanding requirements and issues in disaster area using geotemporal visualization of Twitter analysis

A. Murakami
T. Nasukawa
K. Watanabe
M. Hatayama

During disasters, requirements and situations on the ground change very rapidly. Moreover, they depend on timing and location; thus, it is very hard to understand them in a timely manner. Social media may contain such information with the posted time and the location information. However, it is difficult to extract situational requirements from numbers of conflicting sources. In this article, we propose a system that enables us to find out such useful information from social media and visualize it to understand the data easily. The system is divided into two steps. The first step is to extract requirements and issues from textual data, such as "We cannot buy gas here" or "We are short of batteries," using natural language processing (NLP) technologies. The system also uses NLP to extract geolocation information, such as city names and location landmarks. The second step is to visualize the results in a timely and geolocated manner. We show the system results with using real Twitter data from the Kumamoto Earthquake in 2016. By visualizing the information, the personnel in the disaster area, such as the local governments and/or volunteer organizations, can utilize this information very effectively. For instance, they can decide how to distribute food and water in the disaster area and also how to implement and responsed to their logistics.

Introduction

Social media data are one of the important data sources we can get for recognizing actual information in a timely manner. We can find the operation status of trains and airplanes from Twitter searches, and we can also get good insights for elections from social media surveys. In the domain of disaster recovery area, social media data also plays an important role as data resources for understanding the actual status of affected people and areas.

During disasters, requirements and situations on the ground change very rapidly. Moreover, they depend on timing and location; thus, it is very hard to understand them in a timely manner. Social media may contain such requirements information. However, it is not easy to extract situational requirements from noisy unstructured data.

In 2011, there was a huge earthquake in the northern part of Japan, named the "Great East Japan Earthquake." At that time, some IBM employees (including the authors) tried to extract information useful for assisting affected areas from social media. The results indicated what kinds of requirements frequently occurred at the disaster-affected areas, and how the requirements changed over time. However, if the information were associated with location in a timely manner, we could have used this information more effectively. Thus, after the disaster, we proposed "Geolocational" analysis to extract the requirements and location information at the same time and visualize them to understand the situation appropriately and quickly.

One of the challenges for the goal is how to extract such information from textual documents. Requirements are represented as expressions such as "I cannot buy any batteries" or "We are short of bottled waters." The first step

Digital Object Identifier: 10.1147/JRD.2019.2962491

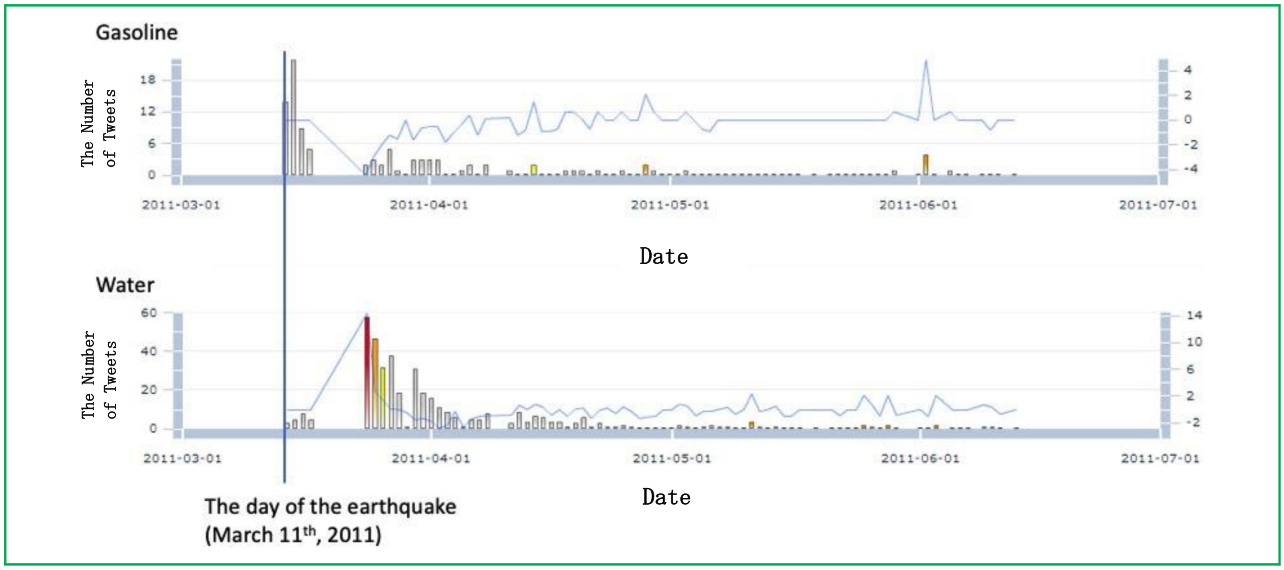

Figure 1

Time series of Tweet frequency containing shortages of "gasoline" and "water."

of our system is to extract such information by using natural language processing (NLP) technologies. Some social media has the capability to add location information. In the case of Twitter, users can add their location to their Tweet, and some users may contain their profile location (Profile Geo) information in their profile, so we can estimate the Tweet location from this geolocation information. However, Twitter decided to remove accurate Tweet location (longitudes/attitude) from their system since most of the users did not use location tag [1]. Thus, we extracted location information such as a city name and a location landmark in textual data as well.

The second step of our system is to visualize the results in timely and geolocated manner. Our system has a capability to show a time-series graph, so we can see the increase or decrease of the number of Tweets with specific contents based on the change of time. Also, we enabled our system to display the map by importing the GeoJSON, which is known for the general format for encoding a variety of geographic data structures. As a result, we can analyze the relevance between the Tweet and location.

In order to take advantage of our system for supporting people in disaster-affected areas, we collaborated with a volunteer nonprofit organization (NPO), named Information Technology Disaster Assistance and Response Team (IT DART), at the Kumamoto Earthquake in 2016. With the real Twitter data at the disaster time, we could understand what the requirements in each affected area were. The NPO investigated the effectiveness of the analysis results and applied that in deciding how to distribute food and water in the disaster area and also how to implement logistics for disaster response.

Lesson learned from our effort for Great East Japan Earthquake

As we described in the Introduction, we tried to extract the requirements of people affected by the Great East Japan Earthquake from social media. For detecting the requirements, it is not enough to count just expressions that indicate goods possibly required at the disaster area since people sometimes talk about goods even if they are sufficient in the disaster area. The following sentences are examples of Tweets during the disaster:

> *"We cannot buy **bottles of water** at the store, but they are selling plenty number of **Onigiri** (rice ball)."*

> *"We found **gas** at the gas station, but **dry cells** are sold out."*

In these examples, the sufferers in the disaster area could buy **Onigiri** and **gas**, but they could not buy **bottles of water** and **dry cells**.

To extract such information, we used NLP technology; the details will be covered in the next section. We used IBM Content Analytics [2] to extract these expressions and analyzed their trends. We found that various types of goods were required at the disaster time, for instance, "water" and "gasoline." **Figure 1** shows the trends of Twitter data that contain the shortage expressions for "water" and "gasoline."

The figure illustrates that many people wanted to buy gasoline just after the disaster time, but the shortage of gasoline did not last for a long time. On the contrary, people did not express a strong desire to buy water just after the earthquake. However, toward the end of March 2011, the requirements for water were increasing very rapidly since the contamination of tap water caused by the Fukushima Nuclear Power Plant accident was reported. Please note that we could not keep crawling the Twitter data because of the scheduled power outage due to troubles at power plants, and much of the data for approximately one week from March 18 is missing.

Thus, we captured the actual trend of missing items and requirements at the disaster time as in Figure 1. However, we noticed that most of the Tweets were tweeted around the Tokyo area, instead of the severely affected area (northern part of Japan), for several reasons. First, this Great East Japan Earthquake induced a big tsunami, so the affected-area people could not use the Internet at the time nor tweet for a long period. Second, as the disaster was in 2011 when Twitter was not so popular all over Japan, most of the users were in urban areas, so the user distribution was biased toward the urban area. Therefore, we formulated the idea that the system needs to know from which area they tweet about and remove this users bias of user distribution to pay more attention to the rural area.

System for understanding requirements with location information

In this section, we describe our proposed method for understanding requirements in the disaster-affected area in timely manner. Based on the previous section, we extracted requirements from social media data that enabled us to understand the trend based on time-series analysis. However, in the previous system, not all Tweets with requirement expressions were tweeted about the affected area. Thus, we enhanced the system for the requirements with location information. Our new system can extract requirements and the Tweet location information simultaneously, and it visualizes the situation in the affected area with this information.

There are two steps for the new proposed system: information extraction from Tweets part and visualization of Tweet information with location information part.

Approach for understanding requirements in the affected area

Information extraction from Tweets

In this step, the system extracts textual expression for requirements and location information using NLP. For the requirements, as described in the previous section, we used IBM Content Analytics in the Great East Japan Earthquake, and we also used the same technologies with IBM Watson Explorer [3] for the new system. Watson Explorer is a successor product of IBM Content Analytics and has the same capability of extracting information from textual documents. It does not only have the capability to capture the word expressions, but also has the capability to extract expressions matched by syntactic rule patterns, such as the following:

X is/are <missing>
X is/are <not enough>
X is/are <unavailable>
X is <sold out>
X is <ran out>
...

Moreover, we can extract just nouns that appeared in X as unavailable items by specifying grammatical features of X. This way, we can extract "*gasoline*" and "*water*" for shortage analysis as in Figure 1.

The Tweet we wanted to analyze at the time of the Kumamoto Earthquake when we tried the new system was written in Japanese, so the actual syntactic rule patterns were all Japanese. The Japanese language is very flexible for word ordering, so we should care about the syntactic dependence with case information, such as subject and object, rather than the word ordering. Watson Explorer uses Custom Rule Files for these syntactic analyses and applies them to their dependence parser outputs [4]. We customized this capability not only for extracting the matched expressions but also for identifying the goods of requirements with the sample patterns above.

For the location information, we tried to capture the Tweet location using "Tweet Location" provided by Twitter. Tweet Location is attached to each Tweet when users add it by themselves. There are several levels of location information. The most detailed one contains longitude and latitude, city level, area level, and nation level. However, very few Tweets include "Tweet Location," so location information just based on "Tweet Location" is not reliable because of their sparseness.

On the other hand, we can find the location information from textual messages in Tweets. For example, some Tweets contain expressions indicating their location as in the following examples, where ***bold italic expressions*** are proper nouns for location.

*"I'm in **Mashiki-machi** Elementary School now. In this shelter we are lack of water. . ."*

*"In **Kikuchi city** all convenience stores are running out of Onigiri and bread. I saw some of diapers and dry cells."*

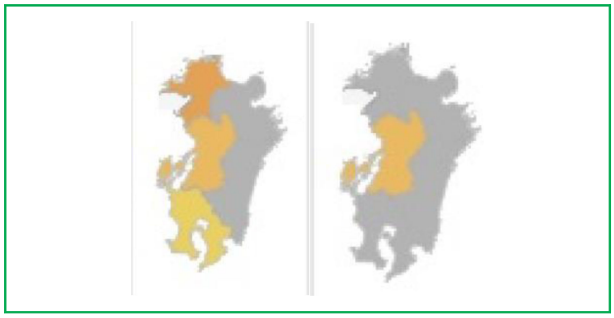

Figure 2

Map visualization with document frequency and relevancy index.

Watson Explorer has the out-of-the-box named entity recognition that identifies the location name automatically. However, in order to improve accuracy both in precision and recall, we created dictionaries to identify the specific location in the disaster area. These dictionaries contain information for handling synonym expressions for the location, such as "Elementary School" and "E.S."

Visualization of extracted requirements with maps

When the extracted location information indicates areas such as city names, state names, etc., the system can visualize them within maps. The system returns the frequency of documents that contain the location information and visualizes the frequency based on colors and shades within a map.

Although it is important to understand how many Tweets were mentioned for each area, it is often more important to understand the proportion of the count to the population or how popular the area is for removing the distribution bias. When we focus only on the count, sometimes we miss some important insights. Thus, Watson Explorer allows us to analyze the document with "*relevancy index.*" The relevancy index is one of the most important ideas of the Watson Explorer, whose value shows the strength of relevance toward the current search criteria. The relevancy index R is calculated by the following formula:

$$R = \frac{F_P}{F_A} \tag{1}$$

where F_A means the frequency of the word in the whole documents and F_p is frequency of the part of documents.

Figure 2 shows an example of the frequency analysis and the relevancy index analysis. The left-side map shows which areas have a large count of documents including the matched expressions, and the right-side map shows which areas have high relevancy index in the area. These examples are analysis results of security reports issued by stock companies. In these maps, frequencies are high in Fukuoka, Kumamoto, and Kagoshima areas. On the other hand, the right-side map shows that only Kumamoto has high relevancy index. This means that Kumamoto is more relevant to the specific search criteria than other areas because the basic distribution of the counts in other areas is higher than in Kumamoto.

Table 1 Top requirements in the disaster area.

Requirements	Number of Tweet
stock	3,206
relief supplies	2,556
water	2,162
food	1,866
personnel	1,690

Effort for Kumamoto Earthquake

As shown in the previous section, we enabled our system to provide the capability for extracting requirements from social media data and visualizing them in timely manner. However, even if the system can provide such information useful for supporting people in a disaster area, we cannot utilize the information without collaborating with local governments and/or volunteers who can take appropriate actions in the disaster recovery area. They can decide how to distribute food and water in the disaster area based on such requirements with location information.

As some of the authors had a strong desire to support disaster recovery with IT technologies, we participated in the founding of a volunteer organization, an NPO named IT DART, in 2015. In the same year, a big earthquake hit Kumamoto, the western part of Japan. Right after the earthquake, IBM decided to provide IT environments with these Watson Explorer capabilities to this NPO as a part of their social contribution activities. Twitter, Inc., also provided Tweets related to the Kumamoto Earthquake around the days its occurrence. IT DART and IBM jointly analyzed the Twitter data for future disaster response.

Twitter provided us a massive amount of Tweet data (around 800 million Tweets) related to the Kumamoto Earthquake for three months. Among them, we analyzed the Tweets with keywords related to earthquake and Kumamoto area for two weeks since the day the first big earthquake hit (from April 9 to 19, 2016). The number of Tweets was 10,941,108 without the ones retweeted officially. **Table 1** shows the top five requirements we identified in the Tweets.

Then, we created two types of dictionaries for identifying the location information in the Tweets: city names and names of shelters name in Kumamoto prefecture. Using

Table 2 Top five city names in Twitter data.

City Name	Number of Tweet
阿蘇市 (Aso Shi)	142,062
益城町 (Mashiki Shi)	92,815
南阿蘇村 (Minami-Aso Mura)	41,177
宇土市 (Uto Shi)	20,164
八代市 (Yatsushiro City)	18,700

these dictionaries, we could classify the Tweets with mention of the requirements to various locations very easily. **Table 2** shows the top five city names in the Tweets. Aso Shi and Minami (South) Aso Mura are the most affected areas in this earthquake, and the epicenter of this earthquake was located in Mashiki Machi, so Tweets containing these location names were very frequent.

To understand what kind of supplies are highly requested in each area, we can narrow down the requirement and see the count of their Tweets. **Table 3** shows the number of Tweets with city names after narrowing down to "shortage of water." The relevancy index with this condition is also presented in the table.

The result shows that water was highly requested in Mashiki Machi, but we cannot understand the regional trend from this table without being familiar with location of each city. Thus, we decided to visualize them onto a map. At that time, we enhanced Watson Explorer's capabilities to show these Tweet counts and their relevancy index with a map. The map in Watson Explorer is generated from GeoJSON, which is known for the general format for encoding a variety of geographic data structures [5]. As a result, it can visualize any kind of regions. **Figure 3** is a map of the Kumamoto prefecture with the number of Tweets containing city names after narrowing down to "shortage of water."

Table 3 Top five city names in Twitter data after narrowed down to "shortage of water."

City Name	# of Tweet	Relevancy Index
益城町 (Mashiki Machi)	44	2. 24
阿蘇市 (Aso Shi)	23	0.66
宇土市 (Uto Shi)	11	1.76
嘉島町 (Kashima Machi)	10	8.32
南阿蘇村 (Minami-Aso Mura)	8	0.48

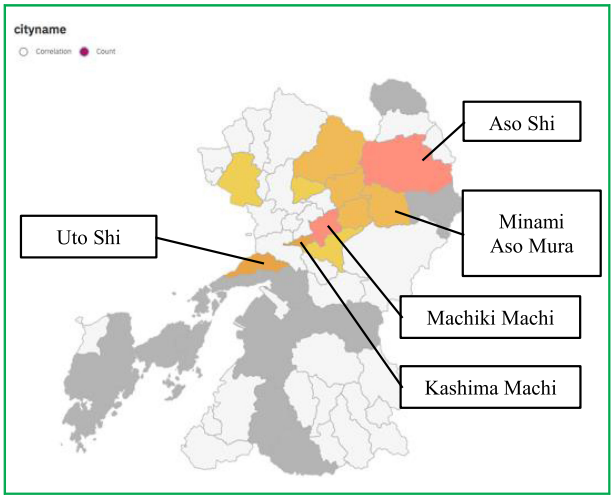

Figure 3

Map visualization with Twitter count after narrowing down to "shortage of water."

From this map visualization, we can easily understand that the requirements for water are frequent in the northern part of Kumamoto prefecture, around Aso Shi. However, the number of Tweets in general is also high around this area, so actual demand might be better understood with relevancy-index-based visualization. **Figure 4** shows a map with relevancy-index-based visualization for the data with the same conditions as in Figure 3. It tells us that Mashiki Machi and Kashima Machi might be the areas with the highest requirements for water during the disaster.

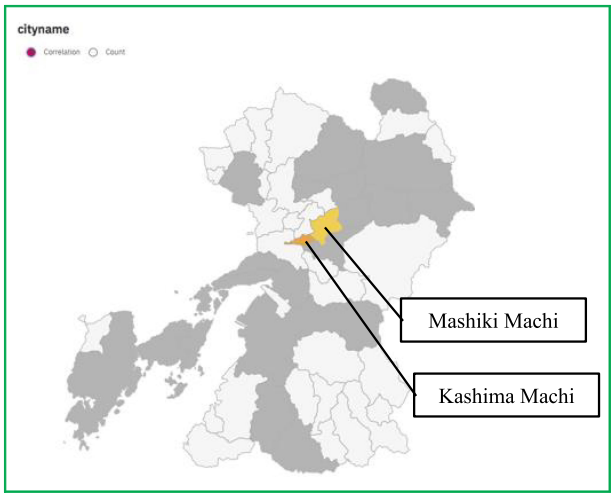

Figure 4

Map visualization with relevancy indexes after narrowing down to "shortage of water."

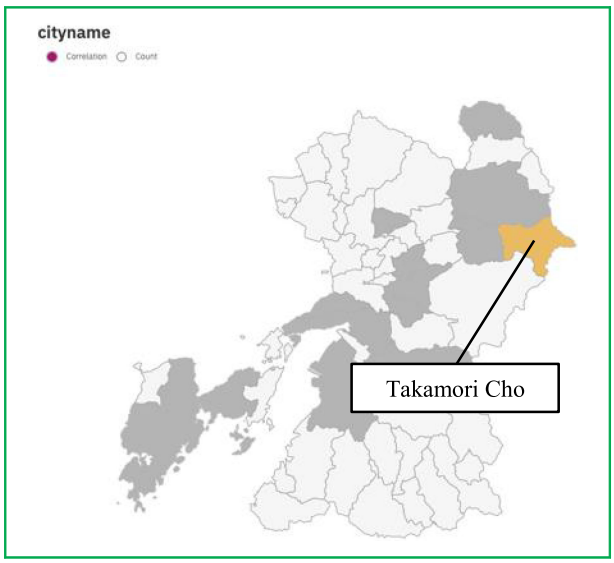

Figure 5

Map visualization with relevancy indexes after narrowing down to "cannot buy gasoline."

Water outage is the one of the main causes of water shortage. In fact, several days after the earthquake, the people in the affected area could use tap water. However, in some parts of the Kumamoto area, people were utilizing underground water as tap water. One report [6] said the underground water became muddy due to the earthquake. Moreover, there was no water quality inspection organization in Mashiki Machi, so it might cause the severe water shortage in Mashiki Machi.

Another visualization also shows an interesting result. **Figure 5** shows a relevancy-index-based visualization of the data after narrowing down to "cannot buy gasoline."

According to this figure, Takamori Cho is the city with the highest value in terms of relevancy index to "cannot buy gasoline" among other cities even though the number of Tweets for this city name is not so large.

Based on a news article on April 19, the earthquake cut off the road providing gasoline to gasoline stations in the city of Takamori Cho, so the people in the city actually had much trouble for getting gasoline.

Discussion

Just after the disasters, supplies, such as water, food and clothing, were very limited. Moreover, the road might be cut off due to the disaster, so it was difficult to transport such supplies to the disaster-affected areas. Under such conditions, local governments and/or volunteer organizations should decide how to distribute supplies in disaster areas with appropriate prioritization and how to distribute such supplies.

As shown in the previous section, the relevancy-index-based map visualization is effective to understand which areas/cities are affected more severely than others in terms of specific requirements.

We also tried to extract information on requirements co-occurred with shelter names. However, the number of Tweets containing shelter names was very small. Thus, it was hard to identify the trend of supply shortage based on shelters names. Only four Tweets were tweeted with "shortage of food" expression with certain shelter names. In this article, we have not tried to map shelter names to the city areas. However, if we migrated these shelter-name-based requirements into appropriate city areas, we could have taken advantage of their information for better understanding of trend in supply shortage.

One important aspect of social media analysis is to consider reliability of the information for avoiding false and misleading information. Our current approach is to check user profiles and other Tweets of the same user for suspicious Tweets. In addition, as we focus on analysis of regional and timely trends from multiple Tweets with the similar messages rather than a single exceptional message, we feel that the risk of fraud is relatively low. Moreover, if false and misleading information is actually being spread in the disaster area, it is also important to take some kind of action to deal with such information for avoiding troubles based on that.

If local governments and/or volunteer organizations got supply shortage information in a timely manner, they could have provided better support for each affected area based on appropriate prioritization. IT DART is a volunteer organization in Japan that aims to support local governments/volunteer organizations in affected areas from an IT perspective. It has a strong connection with Japan Voluntary Organizations Active in Disaster (JVOAD), which enhances the collaboration among national bodies including the government sector, the NGO/NPO sector, and the corporate sector in order to respond more effectively and collaboratively against future emergencies in Japan. Thus, IT DART can collaborate with JVOAD to support volunteer organizations by providing such information from social media based on the technology described in this article.

For disaster recovery, timing is one of the most important aspects. Affected people in a certain shelter may need water and food, but they may not need any clothes. If a shelter has plenty of food, delivering food to the shelter may cause trouble for disposal of surplus food, or it may cause shortage of food in other shelters. At the disaster time, this information on requirements should be provided to local governments/volunteer organizations in a timely manner, and the system should be available right after the disaster. For enabling such an environment, the system should be used not only at the disaster time but also during normal

times. Furthermore, collaboration among supporting information-sharing mechanisms is important for quick actions at the time of a disaster time.

Related works

Text mining is a technology that enables us to discover patterns and trend automatically from huge amounts of textual documents [7, 8]. The target of text mining is widely spread as a life science area [9].

Identifying location information from social media data is a challenging task. Han et al. [10] predicted user geolocation information based on the textual data of Twitter data.

In the domain of disaster management, several systems, such as Ushahidi [11] and CrisisTracker [12], were developed and deployed for helping NPO and disaster relief professionals. Combining location data in social media and event is one of the approaches [12, 13] for alerting people about crisis incidents with maps.

Our approach is not just for identifying the events and/or items and pointing them out on a map. We also tried to visualize data based on statistical relevancy for better understanding of trends in events and/or items on a map. People can make better decisions for the next steps of disaster recovery based on the trend. Right after the beginnings of a disaster, there is little information about the affected area, so it should be useful for decision-making.

Conclusion and future work

In this article, we propose a system to visualize requirements at the disaster area based on their statistical relevancy index. This system can extract requirements and location information from textual data automatically and visualize them in a map. The results of visualization are useful for local governments and volunteer organizations to make better decisions for distribution of supplies in the disaster area in a timely manner.

For identifying the location, we used dictionaries of city names and area names for matching them with textual data. However, many of the Tweets do not contain the registered names of cities or areas directly. Instead of mentioning the names of cities or areas, Tweets often contain names of shelters or landmarks inside the areas. Thus, if we can combine the information by mapping the names of shelters or landmarks into appropriate cities or areas, we can gather more exhaustive requirements for each area, and the information will become more reliable.

Acknowledgment

We would like to thank Twitter, Inc., for providing Twitter data and IBM and IBM Corporate Citizenship for providing their analytics products and environment for IT DART. We also would like to thank members of IT DART for providing their professional insights for disaster recovery.

References

1. C. Shu, "Twitter will remove precise location tagging in tweets, citing lack of use," Tech Crunch, Jun. 2019. [Online]. Available: https://techcrunch.com/2019/06/18/twitter-will-remove-location-tagging-in-tweets-citing-lack-of-use/
2. IBM Content Analytics. Accessed: 2020. [Online]. Available: https://www.ibm.com/support/knowledgecenter/SS5RWK_3.5.0/com.ibm.discovery.es.nav.doc/iiypofnv_prodover_cont.htm
3. IBM Watson Explorer. Accessed: 2020. [Online]. Available: https://www.ibm.com/support/knowledgecenter/SS8NLW/SS8NLW_welcome.html
4. IBM Knowledge Center, "Custom rule files for content analytics collections." Accessed: 2020. [Online]. Available: https://www.ibm.com/support/knowledgecenter/SS5RWK_3.5.0/com.ibm.discovery.es.ta.doc/iiysatextanalrules.html
5. GeoJSON. Accessed: 2020. [Online]. Available: http://geojson.org/
6. "熊本地震を受けた被災地の地下水に関わる対応 (Support activities related to ground water in the Kumamoto Earthquake affected area)," (in Japanese), Kumamoto Ground Water Foundation. Nov. 2016. [Online]. Available: http://www.chikasuinet.sakura.ne.jp/2016_13_4_koga.pdf
7. M. Hearst, "Untangling text data mining," in *Proc. 37th Annu. Meeting Assoc. Comput. Linguistics*, 1999, pp. 3–10.
8. T. Nasukawa and T. Nagano, "Text analysis and knowledge mining system," *IBM Syst. J.*, vol. 40, no. 4, 2001, pp. 967–984.
9. N. Uramoto, H. Matsuzawa, T. Nagano, et al., "A text-mining system for knowledge discovery from biomedical documents," *IBM Syst. J.*, vol. 43, no. 3, 2004, pp. 516–533.
10. B. Han, P. Cook, and T. Baldwin et al., "Text-based twitter user geolocation prediction," *J. Artif. Intell. Res.*, vol. 49, 2014, pp. 451–500.
11. Ushahidi. Accessed: 2020. [Online]. Available: http://ushahidi.com
12. J. Rogstadius, M. Vukovic, C. Teixeira, et al., "CrisisTracker: Crowdsourced social media curation for disaster awareness," *IBM J. Res. Dev.*, vol. 57, no. 5, pp. 4:1–4:13, Sep. 2013.
13. F. Morstatter, H. Gao, and H. Liu, "Discovering location information in social media," *IEEE Data Eng. Bull.*, vol. 38, no. 2, pp. 4–13, 2015.

Received January 16, 2019; accepted for publication November 14, 2019

Akiko Murakami *IBM Japan Ltd, Tokyo 103-8510, Japan (akikom@jp.ibm.com)*. Ms. Murakami is a Software Engineer with IBM Japan and an Architect and Lab Advocate with IBM Watson NLP Products (Watson Explorer, Watson Discovery, and Watson Knowledge Studio). She joined IBM as a Researcher in 1999 and was a part of members of the various natural language processing (NLP) research projects. She made contributions to NLP core technologies of Watson Explorer, and moved to Software Development and became a Software Engineer to enhance her core research. Her research interests include not only NLP/AI but also disaster recovery. She was a volunteer when the disaster happened in Japan, and founded an NPO, named Information Technology (IT) Disaster Assistance and Response Team, which enhanced IT technologies in disaster affected areas.

Tetsuya Nasukawa *IBM Research—Tokyo, Tokyo 103-8510, Japan (nasukawa@jp.ibm.com)*. Dr. Nasukawa received a master's degree from Waseda University, Tokyo, Japan, and a Ph.D. degree from Waseda University in 1998 for his work on natural language processing. He joined IBM Research—Tokyo in 1989. He started a text mining project to take advantage of natural language processing technology developed through machine translation projects, and developed a text analysis and knowledge mining system, named IBM TAKMI. For his work on the development of text mining technology, he received the 2012 Commendation for Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology of Japan. His research interests include cross-lingual and multimedia mining.

Kenta Watanabe *IBM Japan Ltd., Tokyo 103-8510, Japan (wataken@jp.ibm.com)*. Mr. Watanabe is a Software Engineer with IBM Japan. He joined IBM in 2017 and is involved in the development of Watson Explorer, especially the user interface. He implemented the map visualization feature of Watson Explorer.

Michinori Hatayama *Kyoto University, Kyoto 611-0011, Japan (hatayama@dimsis.dpri.kyoto-u.ac.jp)*. Dr. Hatayama received a Ph.D. (Dr. Eng.) degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2000. Since 2016, he has been a Professor with the Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan. He is also a board member with the Information Technology Disaster Assistance and Response Team, Tokyo. His current research interests include disaster risk management and crisis management with information technology.