Location-based Microblog Viewing System Synchronized with Web Pages

Yuanyuan Wang*,Gouki Yasui†,Yuji Hosokawa*, Yukiko Kawai*, Toyokazu Akiyama† and Kazutoshi Sumiya‡
* *Kyoto Sangyo University, Japan*
{yuanw, i1358103, kawai}@cc.kyoto-su.ac.jp
† *Kyoto Sangyo University, Japan*
{i1458085, akiyama}@cse.kyoto-su.ac.jp
‡ *University of Hyogo, Japan*
sumiya@shse.u-hyogo.ac.jp

***Abstract*—Microblogging is a new form of blogging in which microbloggers can share their status in short posts. Microbloggers can be kept up to date with current information from microblog senders; however, Web users cannot be simultaneously updated by microblogs, to obtain the most recent information, whilst they browse Web pages since these are not updated in real time. To circumvent this time lag and provide Web users with up to date information, it is thought that an appropriate method is to associate microblogs and Web pages. Therefore, we have developed a location-based microblog viewing system to support Web users. Our system can present microblogs that are synchronized with Web pages in order to help Web users grasp the current situation in real time, both easily and effectively. In order to map heterogeneous media, we extract the relationship between microblogs and Web pages by generating queries based on common location names. In this paper, we discuss our prototype system and evaluate of our proposed mapping method's effectiveness.**

***Keywords*-location-based microblogs; Twitter; Web pages;**

I. Introduction

In recent years, microblogging platforms such as Twitter[1], Tumblr[2], and Facebook[3], have been rapidly gaining attention. For instance, currently, Twitter has more than 230 million active users, with an approximate 500 million "tweets" on a daily basis[4]. Twitter users can broadcast and share information about their activities, opinions, and statuses with brief texts (tweets), up to 140 characters long, using smartphones at anytime and anywhere. Despite the useful information on Twitter, there still exists a lack of Web users' requirements. That is, Twitter users can gain information about current scenarios from tweet senders; however, Web users cannot simultaneously obtain this information through tweets while they browse Web pages. Furthermore, Twitter supports only limited communication between its users (i.e., followees and followers). Web users cannot access the most recent information from Web pages of popular places (i.e., tourist spots, restaurants, shopping centers etc.) because the Web pages are not updated in real time. Therefore, it is important to focus on associating Twitter feed and Web pages to facilitate real-time information updates for Web users.

In this research, we aim to develop a novel support system for Web users by using location-based microblogs synchronized with Web pages. To achieve this goal, we first acquired geo-tagged tweets based on content analysis and region selection. Therefore, our method can detect tweets if they are related to, or nearby, a target location, even though they do not include the location name, or it can also detect tweets from Twitter users who are not in the target location. Furthermore, the proposed method can filter out tweets from the target location for which the content is not related to the target location. The system then maps the heterogeneous media of the acquired tweets and Web pages. To achieve this, relevant tweets are acquired by a common location name as a query that is generated from the Web page and the relationship between the tweets and the Web page is extracted while the Web user browses a Web page. As depicted in Figure 1, our system efficiently acquires tweets related to an official Website of a location, "Tokyo Skytree," and displays tweets in a synchronized manner on this Web page to help Web users stay up to date with information pertaining to "Tokyo Skytree" in real time, easily and effectively. In brief, our novel contributions are summarized as follows:

- We acquire location-based microblogs (tweets) based on content analysis and region selection.
- We extract the relationship between heterogeneous media on microblogs (tweets) and Web pages by generating queries focused on common location names.

This paper is organized as follows. Section 2 describes our approach and provides a brief summary of related work. Section 3 explains the procedure for mapping location-based tweets and Web pages. Section 4 illustrates the experimental results obtained by using a real dataset of streamed geo-tagged tweets and Web pages with our developed prototype system. Section 5 evaluates our findings and discusses the effectiveness of obtaining tweets based on locations.

[1]https://twitter.com
[2]https://www.tumblr.com/
[3]http://www.facebook.com/
[4]https://business.twitter.com/audiences-twitter

978-1-4799-7361-3/14 $31.00 © 2014 IEEE
DOI 10.1109/SRDSW.2014.18

Figure 1. Conceptual diagram of tweets synchronized with Web pages

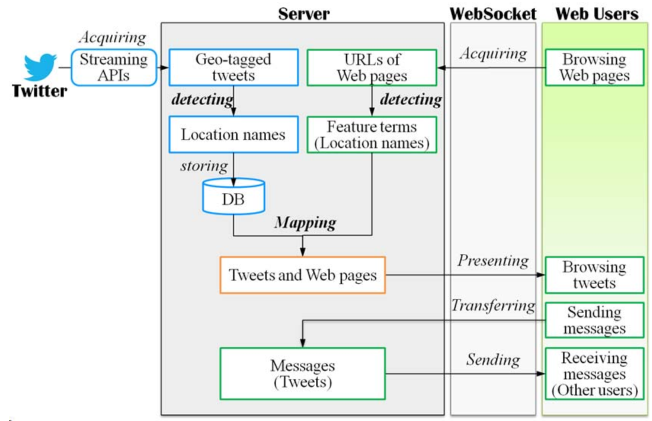

Figure 2. System configuration diagram

II. Our Approach and Related Work

A. Overview of Our System

In this research, we propose a tweet viewing system for Web users by employing microblogs synchronized with Web pages in real time. In this manner, our system enables Web users to stay up to date with information from tweets as they browse associated Web pages by mapping the tweets and Web pages based on common location names. Moreover, when a Web user submits a message through our system, the message can be sent to other Web users who browse the same Web page. In our system, anonymity of all message senders is maintained through a WebSocket server[5].

An example is shown in Figure 1, which depicts a Web user browsing an official Website of Tokyo Skytree in the Web browser of our system. Streaming tweets, e.g., "Very nice viewing!" located in the Tokyo Skytree, are associated with the Web page by generating a query based on a common location name, "Tokyo Skytree," even though the tweets do not include the location name. This allows the Web user to gain insight into the congestion level or gain impressions of the Tokyo Skytree from the presented tweets. In addition, the Web user can send messages through a chat box.

B. Related Work

Recently, event detection on Twitter has been a very popular research area with a large number of applications. Our approach builds on the strength of understanding the current situation from microblogs. Several research efforts have focused on the detection of events from Twitter using tweet content, such as a trend detection system by identifying bursty words [1], [2], tracked events by applying locality sensitive hashing (LSH) [3], a map generation of small sub-events around big, popular events by using conditional random fields (CRFs) [4], the discovery of breaking events by using popular hashtags in Twitter [5], and, topic detection by using tag correlations [6]. Our work is unique from these studies because our analysis aims to explore microblogs that may help in providing current information for users who do not use microblogging platforms.

Many applications of event detection have been developed, for example, forest fires have been detected by mining location-based social networks [7], sporting events have been detected by summarizing an event using only status updates from Twitter [8], unusual geo-social events (local festivals) have been detected from the usual behavior patterns of crowds with geo-tagged microblogs [9], traffic events have been detected by looking for relevant text patterns that indicate the traffic condition in specific locations [10], a web interface for tracking the prevalence of Influenza-like Illness (ILI) in several regions of the United Kingdom has been developed using the contents of Twitter's microblogging service [11], also, earthquake detection by using Twitter as a social-sensor, notifying users of incoming earthquakes much faster than even the Japan Meteorological Agency [12]. These studies focused on location-based microblogs, which provide a real-time stream of updates, opinions, and first-hand reports of where and what is happening; however, they cannot support Web users who do not use microblogging platforms. Thus, our research aims to study the issues in utilizing location-based microblogs (geo-tagged tweets) to provide support for Web users.

III. Presenting Tweets synchronized with Web Pages

The processing flow of our system can be constructed as shown in Figure 2. In this research, in order to associate tweets with Web pages in real time, we first acquire real-time tweets (streaming tweets) and the most recent URLs of Web

[5]http://gihyo.jp/dev/feature/01/websocket/0001

pages that users are accessing. The server side then manages the acquired streaming tweets (geo-tagged tweets) and the acquired Web pages and maps the tweets and the Web pages by detecting common location names. Web users access the acquired Web pages, the server extracts the associated tweets and our system presents them to the Web users in the Web browser. In this case, the Web users need to use an add-on of our developed system in the Web browser. In addition, when the Web users send the messages through the add-on, the server receives and sends them to other Web users who are browsing the same Web pages. Furthermore, anonymity can be maintained by all transmissions through the server in our system.

A. Acquisition of Streaming Tweets

In this paper, in order to associate tweets and Web pages based on locations. At first, it is necessary to acquire tweets, for instance, a method based on a content analysis of the tweets and hashtag search on Twitter [13], it can detect the tweets of Twitter users who are not in the locations in the real world. Also, tweets can be acquired by identifying the latitude and longitude of locations from geo-tagged tweets [14]. However, many detected tweets are still not related to the locations of them, it is difficult to report the current situation from the detect tweets. As described above, in this research, we propose a tweet acquisition method, which combines a content analysis and a region selection. We first obtain streaming tweets, i.e., geo-tagged tweets, including location information (i.e., latitude and longitude) except duplicates from a specified region by using Streaming APIs, and we analyze the content of the tweets focusing on nouns by a morphological analysis. The specified region is determined by a northeast point and a southwest point, then, we can obtain streaming tweets in a rectangular region surrounding these two points. Next, we detect location names within a radius d of a region by using Google Places API v3[6], from the latitude and longitude information of the obtained streaming tweets. Since the detected location names can be generated search queries for obtaining Web pages, and we empirically set $d = 5$m by considering the movement of Twitter users. Then, our server database manages {Twitter user ID, icon URL, latitude, longitude, location name, tweet, word set, acquisition time} in a certain time.

B. Mapping of Tweets and Web Pages based on Locations

To determine if the streamed tweets are related to the detected location names, it is necessary to analyze the content of the tweets and to remove tweets that have a low relation to their locations. Therefore, high-frequency words are extracted from the tweets by the locations at a certain time and within a certain region, and tweets that contain many feature terms related to the locations are then selected.

[6]https://developers.google.com/place

Figure 3. Screenshot of our developed prototype system

In particular, we first obtain a total amount n of tweets in a certain time within a radius d of a region based on a given location of the acquired tweet t. The average of the frequency of each word i that appears in t is then calculated as follows:

$$\sum_{i=1}^{q} \frac{\text{The number of tweets with } i}{\text{The total number } n} \times \frac{1}{q} \tag{1}$$

Here, q denotes the total number of words that appear in tweet t. If a value of Eq. (1) is more than the empirical threshold value, the tweet t related to the location of it is selected.

To extract location names from Web pages, we acquire URLs of the Web pages that Web users are browsing. Then we extract high frequency words from snippets of the acquired URLs as feature terms of the Web pages and detect location names of the Web pages by using a morphological analyzer. The location name can be determined from the place names that have been extracted from the URLs.

Based on the above, when a user browses a Web page, the system can obtain and present tweets that are relevant based on a common location name from the Web page and the tweets. In addition, the server database of the system stores the obtained tweets, the obtained Web pages, and the detected common location names. Furthermore, our system allows Web users to send messages through our server.

IV. Prototype System and Experiments

A. Prototype System

We developed a prototype system to support Web users in which tweets are synchronized with Web pages in real time. The server side is built using Apache httpd 2.4, Java, PHP 5.5. The client side is implemented using JavaScript in Firefox 23.0.1. The prototype system has three stages: analysis, association, and interface. In the analysis stage, we acquired geo-tagged tweets, without duplicates, from a

Table I
EXPERIMENTAL DATASET (9:00-14:00 (JST) AND 16:00-20:00 (JST), DURING 2013/9/16-20)

	Location Name	(Latitude, Longitude)	#tweet (Total: 4,809) within different radius d						
			50m	100m	300m	600m	800m	1km	Total
1	Tokyo Skytree (medium scale)	(35.710023, 39.810702)	58	68	86	148	216	269	269
2	Tokyo Disneyland (large scale)	(35.359796, 38.727598)	10	20	72	342	587	679	679
3	Tokyo Station (large scale)	(35.681178, 39.766085)	64	247	350	682	1945	3093	3093
4	Tokyo International Airport (large scale)	(35.632518, 39.881359)	1	4	115	350	710	768	768

specified region by using The Streaming APIs version 1.1[7]. The specified region is defined as a rectangular region by a northeast point and a southwest point. In the association stage, we mapped the acquired tweets and the acquired Web page based on their common location names. During the interface stage, the Web page browsing interface for Web users is developed into a Web browser by using an add-on for Firefox, and the Web users are connected to our system through WebSocket[8].

For example, a Web user browsing an official Website of Tokyo Skytree through our developed prototype system (see Figure 3), and streaming tweets are shown on the left side of the official Website of Tokyo Skytree. Therefore, the Web user can get up to date information about the Skytree by browsing the tweets that are posted near the Skytree in real time. Additionally, the Web user can browse the tweets in the past, since the tweets related to the each Web page are stored as logs in the server database of our system. Furthermore, the Web user can enter a message in the chat box and our system can send this message to all other users who browse the same Web page through our system.

B. Experimental Dataset

The aim of our evaluation was to verify whether our proposed system of tweets synchronized with Web pages based on locations is useful for supporting Web users. As a test dataset, we prepared a Twitter data collection, whose tweets were intentionally collected and filtered only for geo-tagged messages. The total number of geo-tagged tweets was approximately 1.35 million, almost 30MB, collected for approximately two months between 2013/9/3-9/28, and 2013/12/10-2014/1/13, in an area centered on the Kanto region, including capital of Japan, with longitude range: [132.2, 146.1] and latitude range: [29.9, 46.2]. Indeed, several tweets of them were collected within a radius d in a range of the collected tweets by considering the movement of Twitter users, using Google Place API version 3[9].

In order to evaluate the performance of tweet acquisition and the usability of our system, we narrowed down the test dataset in an area centered on Tokyo with different ranges of locations (i.e., detected location names from Web pages) within different radii d as shown in Table 1. This was 4,809 tweets in total related to Tokyo Skytree (medium scale), Tokyo Disneyland (large scale), Tokyo station (large scale), and Tokyo International Airport (large scale), during 2013/9/16-20 (working days: Monday-Friday) within two periods of time, 9:00-14:00 (JST) and 16:00-20:00 (JST).

There were twelve college students in Faculty of Computer Science and Engineering, Kyoto Sangyo University, who participated in the experiments. We show and discuss the experimental results in the follow sections.

C. Validity of Tweet Acquisition based on Locations

This experiment was designed to assess the mapping of tweets and Web pages based on their common locations for supporting Web users to gain information about that location. Tweet acquisition is an important issue in our system. n order to evaluate the performance of the tweet acquisition based on location information (i.e., latitude and longitude) and the content of the tweets, we used our proposed method to obtain tweets in a range centered on Tokyo with different ranges of locations from the dataset (see Table 1). Five participants identified if the content of the obtained tweets was related to their location names or not. If the tweet related to its location name, the participants gave a score of 1; when the tweet was not related to its location name, the participants gave a score of -1; when unsure, the participants gave a score of 0. Correct answers were defined as those for which that the average score was over 0.6. In addition, the correct answers are not only the messages of the tweets containing their location names, but also the tweets related to the locations containing impressions of the locations. Precision and recall are calculated by using the following formulae.

$$Precision = \frac{\text{The number of correct answers within a radius } d \text{ by our system}}{\text{The total number of tweets within a radius } d}$$

$$Recall = \frac{\text{The number of correct answers within a radius } d \text{ by our system}}{\text{The total number of correct answers within a 1km radius}}$$

The results of obtaining tweets in different ranges of locations within different radii d (=50m, 100m, 300m, 600m,

[7]https://dev.Twitter.com/docs/streaming-apis
[8]http://www.atmarkit.co.jp/ait/articales/1111/11/news135.html
[9]https://developers.google.com/places/

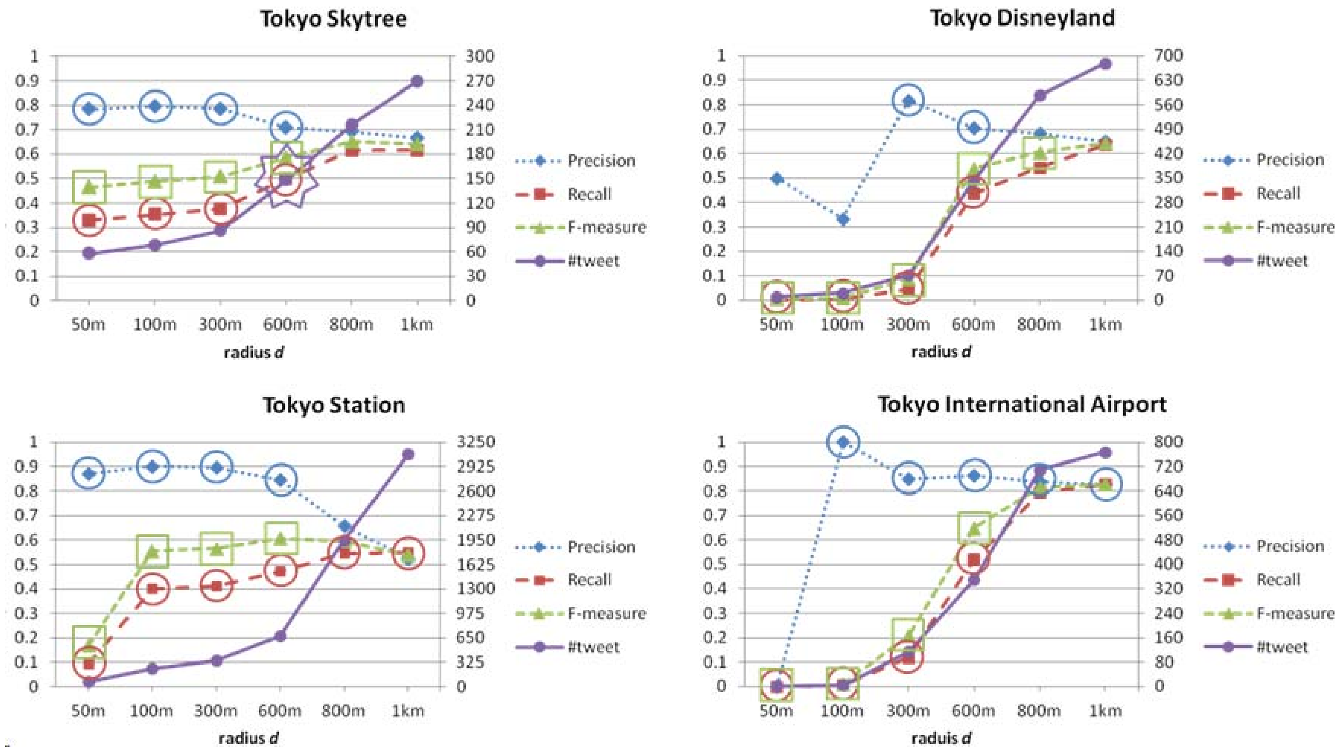

Figure 4. Results of obtained tweets within radius d based on locations

800m, 1km) are shown in Figure 4; the experimental results can be explained as follows:

- The precision of Tokyo Skytree within a range of a radius d (50-600m), Tokyo Disneyland within a range of a radius d (300-600m), and Tokyo Station within a range of a radius d (50-600m), reached 0.70; the precision of Tokyo International Airport within all ranges except a 50m radius, reached 0.80. Thus, the average precision of all locations within all ranges in our test dataset was 0.72.
- The recall of Tokyo Station within all ranges and other locations within a range of radius d (50-600m), were below 0.60. Additionally, the average recall of all locations within a 1km radius, reached 0.66.
- The F-measure of Tokyo Disneyland within a range of radius d (50-800m) and other locations within a range of radius d (50-600m), were below 0.60. Thus, the average F-measure of all locations within a 1km radius in our test data was 0.66.

This experiment indicated that several locations in our test dataset did not obtain enough tweets due to the recall of each location being too low. Since presenting tweets in real time for Web users is a very important feature of our system, over 2, 3 tweets one hour and almost 300 tweets a week must be acquired. However, as Tokyo Skytree is a medium area in Japan, the range of a 600m radius could not acquire enough tweets. Although feature terms can be detected from a tweet, when the tweet contains a lot of words, the average frequency of each feature term is reduced, and it becomes lower than the empirical threshold value by Eq. (1) in Section Ⅲ.B. Therefore, in order to increase the recall of our tweet acquisition method, we improved the method by taking the weight of feature terms related to location names from tweets into consideration.

Although the tweets were acquired by generating search queries based on the location names, from the Web pages, surrounding areas of the target locations must be considered since users often browse the Web pages of the target locations together with the Web pages of the surrounding areas (e.g., stores, restaurants, traffic intersection, etc.). The experimental results also suggested that we can utilize high frequency words that are extracted from tweets to extract the names of the surrounding areas. Future work will see the locations classified into various types of facilities by using a concentration ratio and this will be implemented in our system to verify this as improvement.

D. Situational Awareness of Locations by Tweets

In order to verify the effectiveness of our system for Web users, we first verified if tweets were presented in real time by our prototype system. Our prototype system could present streaming tweets synchronized with Web pages to Web users in 2 s. Finally, we confirmed that the Web users could gain information about the location whilst browsing the Web pages with the related tweets. As an illustration, a tweet “Oh no, don’t climb Skytree today during a typhoon! lol” was posted in Tokyo Skytree on September 16th, 2013, 17:46 (JST), and a tweet “I’ve been waiting 90 minutes now on Sunday···orz” was posted nearby Tokyo Skytree on September 29th, 2013, 19:09 (JST). Since the Web users browsed an official website of Tokyo Skytree with these tweets, they could gain insight into the current situation of Tokyo Skytree in real time, such as that Tokyo Skytree is closed owing to a typhoon, or that many people are attempting to see the night view of Tokyo Skytree, and thus, it is difficult to climb

V. Concluding Remarks

In this paper, we developed a novel tweet viewing system for supporting Web users to gain insight into a location in real time. In this system, the streaming tweets and the most recent Web pages are acquired by generating queries based on common location names from geo-tagged tweets and snippets of the Web pages. Our system supports Web users by mapping the tweets and the Web pages based on the generated queries. As a result, the system can present the tweets to help Web users gain more information about those locations in real time, both easily and effectively. Experimental results show that the system has the potential to efficiently support Web users through tweets that are synchronized with Web pages in real time.

For future work, we plan to enhance the evaluated system based on the experimental results and verification experiments will be carried out for many more locations, their surrounding stores, or restaurants and with many more participants. Furthermore, we will evaluate the usability of Web users viewing the tweets synchronized with the Web pages through the developed system.

Acknowledgment

This work was partially supported by SCOPE (the Ministry of Internal Affairs and Communications of Japan), and a Grant-in-Aid for Scientific Research (B)(2) 26280042 (the Ministry of Education, Culture, Sports, Science, and Technology of Japan).

References

[1] M. Mathioudakis and N. Koudas, “Twittermonitor: Trend detection over the twitter stream,” in *Proc. of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD 2010)*, pp. 1155–1158.

[2] C. Li, A. Sun, and A. Datta, “Twevent: Segment-based event detection from tweets,” in *Proc. of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pp. 155–164.

[3] S. Petrović, M. Osborne, and V. Lavrenko, “Streaming first story detection with application to twitter,” in *Proc. of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT 2010)*, pp. 181–189.

[4] A. Khurdiya, L. Dey, D. Mahajan, and I. Verma, “Extraction and compilation of events and sub-events from twitter,” in *Proc. of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT 2012)*, pp. 504–508.

[5] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, “Discover breaking events with popular hashtags in twitter,” in *Proc. of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pp. 1794–1798.

[6] F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum, “Enblogue: Emergent topic detection in web 2.0 streams,” in *Proc. of the 2011 ACM SIGMOD International Conference on Management of Data (SIGMOD 2011)*, pp. 1271–1273.

[7] B. D. Longueville, R. S. Smith, and G. Luraschi, ““omg, from here, i can see the flames!”: A use case of mining location based social networks to acquire spatio-temporal data on forest fires,” in *Proc. of the 2nd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2009)*, pp. 73–80.

[8] J. Nichols, J. Mahmud, and C. Drews, “Summarizing sporting events using twitter,” in *Proc. of the 17th ACM International Conference on Intelligent User Interfaces (IUI 2012)*, pp. 189–198.

[9] R. Lee, S. Wakamiya, and K. Sumiya, “Discovery of unusual regional social activities using geo-tagged microblogs,” *World Wide Web (WWW) Special Issue on Mobile Services on the Web*, vol. 14, no. 4, pp. 321–349, 2011.

[10] S. S. Ribeiro, C. A. Davis, D. R. R. Oliveira, W. Meira, T. S. Gonçalves, and G. L. Pappa, “Traffic observatory: A system to detect and locate traffic events and conditions using twitter,” in *Proc. of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2012)*, pp. 5–11.

[11] V. Lampos, T. D. Bie, and N. Cristianini, “Flu detector: Tracking epidemics on twitter,” in *Proc. of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III (ECML PKDD 2010)*, pp. 599–602.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: Real-time event detection by social sensors,” in *Proc. of the 19th International Conference on World Wide Web (WWW 2010)*, pp. 851–860.

[13] S. Tajima and T. Ushiama, “A method for composing ad-hoc following networks on twitter for sharing information among event participants,” *Journal of ADADA*, vol. 17, no. 4, pp. 119–124, 2013.

[14] G. Ference, M. Ye, and W.-C. Lee, “Location recommendation for out-of-town users in location-based social networks,” in *Proc. of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pp. 721–726.