World Wide Web (2011) 14:321–349
DOI 10.1007/s11280-011-0120-x

Discovery of unusual regional social activities using geo-tagged microblogs

Ryong Lee · Shoko Wakamiya · Kazutoshi Sumiya

Received: 15 July 2010 / Revised: 16 January 2011 /
Accepted: 3 February 2011 / Published online: 1 March 2011

Abstract The advent of microblogging services represented by Twitter evidently stirred a popular trend of personal update sharing from all over the world. Furthermore, the recent mobile device and wireless network technologies are greatly expanding the connectivity between people over the social networking sites. Regarding the shared buzzes over the sites as a crowd-sourced database reflecting a various kind of real-world events, we are able to conduct a variety of social analytics using the crowd power in much easier ways. In this paper, we propose a geo-social event detection method by finding out unusually crowded places based on the conception of social networking sites as a social event detector. In order to detect unusual statuses of a region, we previously construct geographical regularities deduced from geo-tagged microblogs. Especially, we utilize a large number of geo-tagged Twitter messages which are collected by means of our own tweets acquisition method in terms of geographic relevancy. By comparing to those regularities, we decide if there are any unusual events happening in monitoring geographical areas. Finally, we describe the experimental results to evaluate the proposed unusuality detection method on the basis of geographical regularities which are computed from a large number of real geo-tagged tweet dataset around Japan.

Keywords geo-social event detection · microblog · socio-geographic analytics

R. Lee (✉) · K. Sumiya
School of Human Science and Environment, University of Hyogo, Himeji, Japan
e-mail: leeryong@shse.u-hyogo.ac.jp, Lee.Ryong@gmail.com

K. Sumiya
e-mail: sumiya@shse.u-hyogo.ac.jp

S. Wakamiya
Graduate School of Human Science and Environment, University of Hyogo, Himeji, Japan
e-mail: nd09a025@stshse.u-hyogo.ac.jp

1 Introduction

Microblogging services represented by Twitter (http://twitter.com/), Jaiku (http://www.jaiku.com/) and Plurk (http://www.plurk.com/) are explosively attracting a great deal of attention all over the world. According to a survey conducted by Sysomos Inc. in June 2009 [19], Twitter has shown an explosive growth over the last few years; its population already surpassed 11.5 million users increasingly spreading out its geographic coverage to the world-wide area beyond borders. Comparing to conventional blogging services, the most important feature of microblogging sites would be probably that they realized instant sharing of massive personal updates by the significant short length of messages. Behind the successful growth of the microblogging services, we can't miss out the critical contribution of smart phone technologies which enable us to write microblogs anytime and anywhere. In particular, on behalf of location positioning functions such as GPS-enabled smart phones and the location tagging ability of Twitter services to their messages, so called tweets, we are able to embed geographic location information as an additional tag or as a part of textual message. This slight evolution of the mobile functionality is, however, greatly encouraging crowds to share their life logs on the microblogs from outdoor environments; beyond simply remaining daily personal logs, people started sharing updates of incidents which they are experiencing every day by the short texts as well as much vivid photos or videos. In fact, some critical social incidents such as Iran's protest [20] and India's terror [16] as illustrated in Figure 1 have exemplified the power of microblogging enough as a noble media of the fastest uncensored public voices. In respect to the world-wide coverage and the diversity of users of microblogging sites, there will be more and more numerous crowds who are affected by various social and natural events and, after that, are willingly and voluntarily writing posts from the places.

On the perspective of socio-geographic analytics, this new type of massive messages emitted by massive publics via the microblogging sites must be a valuable

Figure 1 Geo-social events found in Twitter: the post-election protests from Iran [20] (*left*) and the terror from Mumbai [16] (*right*).

socio-geographic resource which is able to be used for revealing other uncharted social dimensions on a large scale. Based on the geo-social database constructed from microblogging services, it is possible to look into when/where crowds exist and what they think or feel almost in real time

In this paper, we present a geo-social event detection method based on geographical regularities of local crowd behavior through Twitter. For the purpose of the detection of unusual socio-geographic events, we first decide what the statuses of local crowd behavior in a geographical region are usual in terms of Twitter. After mapping the geo-tagged tweets onto relevant locations on a map, we focus on the following points: First, a sudden increase or decrease in the number of tweets happening in a geographical region can be an important clue. Secondly, the increasing number of Twitter users in a geographical region for a short time period can apparently notify the occurrence of a local event. Probably, even for a small town which the number of Twitter users is usually not significant, an unusual growth in the number of users can tell us that there may now be a crowd congregating; furthermore, the reason for the increase can be clear if we perform further complex content analyses. Lastly, from time-stamped geo-tagged messages to the public, we can trace the movement histories of crowds and grasp the overall degree of activities of local crowds. For instance, if the movings of the local users unexpectedly become highly activated, we may speculate that there is a high probability of some unusual local events, happening in the monitoring region.

In order to achieve our goal of the analysis of regional phenomena using the content of micro-blogging sites, we developed a microblogging monitoring system that can effectively obtain a large amount of geo-tagged and time-stamped microblog data. Based on the dataset crowded by the monitoring system, we performed a cluster analysis to examine the spatial distribution of the crowd and to determine the presence of usual and unusual patterns. Finally, we detect unusually crowded locations. Specifically, we define usual movement patterns of mass microbloggers upon a time axis and detect locations where regional crowds show unusual patterns of tweet writing and moving.

The remainder of this paper is organized as follows: section 2 describes our initial motivation and provides the brief description of our research model with a summary of related work. Section 3 explains a geographic microblog monitoring system for collecting mass volumes of microblog data. Section 4 presents the detail of the method used in the discovery of regional events. Section 5 illustrates an experimental evaluation of the proposed method, in which we examine geographic regularities and contrast them with a target test dataset. Section 6 describes our conclusions and suggestions for further work.

2 Social event analysis through micro-blogging sites

2.1 Microblogging sites as a geo-social event detector

From the point of view of conventional geographical information systems (GISs), location-based social networks accessed by a significantly large number of people from all over the world promise an uncharted and profitable realm where we can easily conduct various socio-geographical analytics using the massive crowds' thoughts, experiences, or even feelings. Certainly, the new research field can enable us to explore several sophisticated socio-geographical phenomena.

In this work, we aim to perform a nation-wide geo-social event detection using a considerable number of messages from Twitter. When we build a socio-geographical event detection system, we have to consider methods for monitoring crowd buzzing and for discovering unusual events occurring in a geographical region as depicted in Figure 2. In particular, to exploit crowd power through microblogging sites, we developed a geographical microblog collecting system and successfully accumulated a significant number of geo-tagged tweets. Indeed, we examined the dataset to explore the real-world events and carefully investigated the on-the-spot crowd's

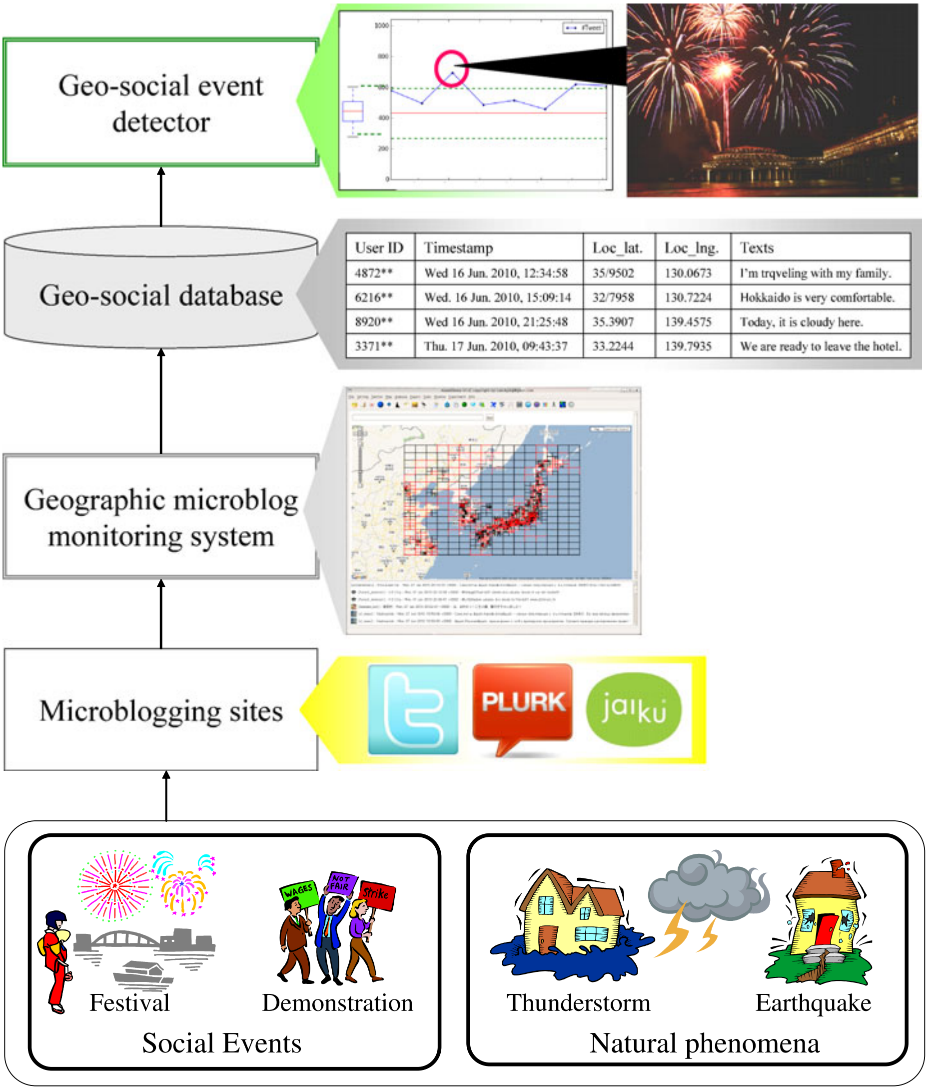


Figure 2 Microblogging sites as a geo-social event detector.

reaction to these events, which they shared via Twitter. Although many of the collected messages were very personal and it was difficult to understand what was happening, these messages obviously reflected the existence of a considerable number of people and their activities from all over the world. Accordingly, we can benefit from exploiting the massive crowd behavior logs to find out and comprehend what is going on in our societies.

2.2 Research model

In order to utilize the crowd's life logs obtainable through Twitter for the detection of geo-social events, we have to consider several important issues; first, we need to monitor tweets for a geographic range. However, in terms of geographic search, Twitter's search functions are limited due to the extendibility to a large area. Next, detecting unusual geo-social events reversely requires the usual status of a region. However, because there is no exploit and definitive source to define normal status of local crowd behavior, we will need to construct such usual status of local crowd behavior through Twitter likewise. Finally, for the practical use of our event detection system, we need a fast detection method to compare a current status to the past usual status. For this, we will provide a simple and logical event detection framework which can be easily manageable and be extendible for a variety of socio-geographic event analytics.

As shown in Figure 3, we proceed with our work in the following three steps: (1) collecting geo-tagged tweets by a unique Twitter monitoring system, (2) identifying socio-graphic boundaries of Twitter users and measuring geographical regularities of crowd behavior, and (3) detecting geo-social events through a comparison to the regularities.

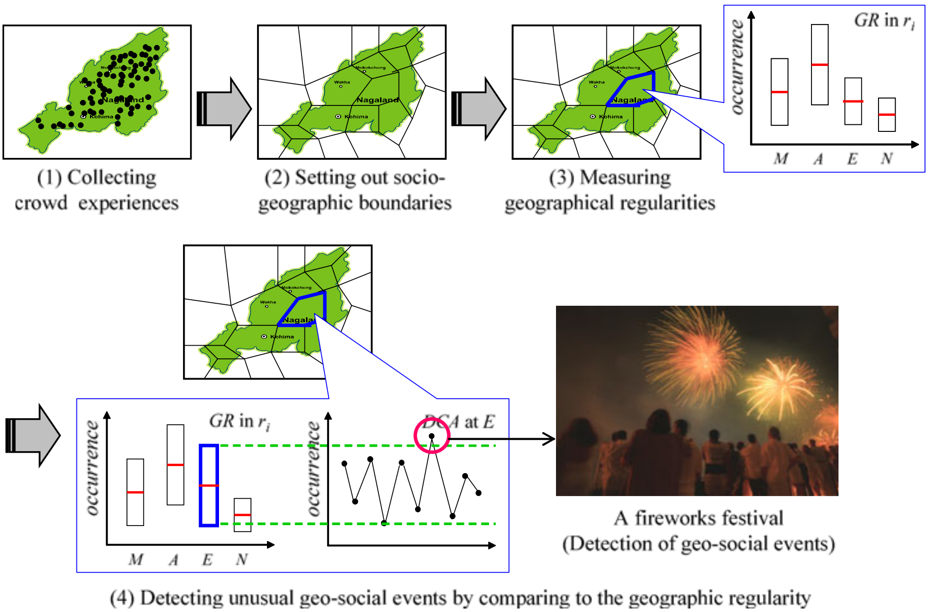


Figure 3 Process of geo-social event detection.

We attempt to utilize the collective experiences of local crowds and materialize them as a geographical regularity that indicates the usual crowd patterns. Specifically, a normal status will be defined for a geographical region; for instance, usually, crowded stations have a steady geographical regularity of many moving populations. On the basis of the unique characteristics of each town, we can easily detect unusual incidents such as town festivals or unexpected natural disasters if they motivate many local crowd to write more microblogs.

(1) Collecting crowd experiences via Twitter
While it is possible to automatically collect tweets from Twitter, it takes a considerable amount of effort to gather a significant number of geo-tagged tweets from Twitter because of some limitations; first, Twitter's open API (Twitter open API, http://apiwiki.twitter.com/Twitter-Search-API-Method%3A-search) solely supports the simplest near-by search by means of the specification of a center location and a radius. Furthermore, each query can only obtain a maximum of 1,500 tweets. To overcome these restrictions and perform periodical monitoring of any user-specified region, we developed a geographical tweet gathering system [6] that can collect significant amounts of geo-tagged microblog data for a region of any size.

(2) Establishing natural socio-geographic boundaries
Next, to detect unusual local events for a given large area, we first need to determine how to partition the target region into sub-areas by establishing socio-geographic boundaries in order to estimate the usual patterns of the crowds in the areas.
In order to configure socio-geographic boundaries conveniently, we adopt a clustering-based space partition method that can reflect a geographical distribution of a dataset and better deal with heterogeneous regions differently. In detail, we adopt the K-means clustering method [12] based on the geographical occurrences of a dataset. Then, the K-partitioned regions over a map are later represented by a Voronoi diagram using the centers of the clusters. As a result, we achieve an appropriate socio-geographic boundary setting for the target region by distributing the actual occurrences and consequently focusing on the major hot-spots much effectively.

(3) Estimating geographical regularity of local crowd behavior
Before making a decision of the unusuality of crowd behavior that we utilize as a clue for the final geo-social event detection, we need to decide what status is ordinary or not in each socio-geographic boundary. For this, we estimate the geographical regularity for a socio-geographic boundary during a certain time period focusing on the following points: (1) How many tweets are posted? (2) How many users are there? (3) How active are the movements of the local crowd? On the basis of these observations, we define the geographical regularity as a normal tendency for each region. Furthermore, we look into these factors every 6-h by splitting a day into four equal time periods of morning, afternoon, evening, and night.

(4) Detecting unusual geo-social events
Finally, on the basis of a geographical regularity constructed from the usual behavior patterns of crowds, we detect unusually crowded regions for a test dataset. In our previous work [5], we first monitored the significant crowd movements such as aggregation and dispersion to detect crowded places.

We also extended the measurement to decide unusual statuses of geographic regions [11]. More specifically, depending on statuses obtainable from Twitter indirectly, we finally decide whether a geographical region has a normal status or not. In particular, the regularities in the form of boxplots [13] can detect unusual statuses.

2.3 Related work

Microblogging services are still on the way of rapid evolution, simultaneously inspiring many academic and practical challenges. As early work, Java et al. [9], Zhao et al. [22] and Krishnamurthy et al. [10] examined the usage patterns of Twitter in relation to its impact on lifestyles and topical discussions. Iwaki et al. [8] also considered the discovery of useful topics from microblogs. These studies largely paid attention to the content analyses of textual messages and the link structure of followers among users. To sum up, they analyzed the trends in remarks and discovered tastes based on the context of the messages and link structures. Our study differently focuses on the location and time of microblogs, in order to detect unusually crowded places from crowd movement patterns.

For geo-social characteristics analysis with blogging sites, Moriya et al. [15] developed a system that estimates images, impressions, or atmosphere felt by bloggers about a region from texts, in relation to geographic information provided by blogs, and displayed the results on a digital map. This work is similar to our approach in terms of geo-social analysis, but in our research we shed light on social events and habitual movement patterns.

In order to extract the social movements of crowds, Wang et al. [21] analyzed a trajectory dataset from GPS-equipped taxies and found "hot places" where passengers were often picked up or dropped off to illuminate the movement of human flows. However, this kind of work requires lots of efforts to collect the dataset with the support of taxi companies. In our work, we developed an autonomous tweet gathering system that is easily able to collect enormous number of data for any geographic region. That is, we can analyze movement histories in various regions with many volunteers over a social network site.

As another interesting analysis of movement of crowds, Otsuka et al. [17] and Mohan et al. [14] presented a relationship between the real world and the virtual (networked) by analyzing how web behavior reflects real world events. Our study also detects geo-social events, but based upon microblogs, where individuals' movements are directly reflected in the tweets they wrote.

As recent work, Sakaki et al. [18] made a significant effort to develop a natural disaster alarming system by using Twitter occurrences from the event locations for some special type of cases; earthquakes in Japan with a concept of Twitter user as a sensor. In this work, they depend on a message analysis approach on Twitter as a cue to detect geo-social events. For example, in the earthquake case, they made a query set $Q =$ 'earthquake' and 'shaking' and in the typhoon case, they made a query set $Q =$ 'typhoon', but for the real-time events that they want to discover changes, they must change an input query manually. Therefore, they couldn't extend their approach to a various kind of socio-geographic analytics involving many unknown and unexpected events.

Without doubt, the web space becomes a main stage where we can conduct various human-centered analytics [3]. Especially, in our work, we regard the social networking sites as a critical geo-social database which we had not had so far, differently from the usual web resources relevant to real world [1]. Based on this novel type of database growing autonomously by numerous volunteers around the world, we need new ways to handle the massive volume of data.

3 A geographic microblog monitoring system

In this section, we introduce a microblog monitoring system which was developed to support various kinds of social geographic analyses using micro-blogging sites. As shown in Figure 4, this system can support various requests from socio-geographic analysts who wish to obtain and analyze data on a large region from a small town to the whole globe. An analyst first needs to specify a region for examination. Then, the system then locates the geographic region, accesses the microblog site and obtains data autonomously. For the simple discussion, we assume that a microblog post mb_i has a set of properties of explicit *user ID*, *timestamp*, *location*, and *texts* which are easily written by location-aware smart phones. The texts part can actually include any kind of media with the help of external supporters such as Twitpic (http://twitpic.com/) for photos and TwitVideo (http://twitvideo.com/) for videos. By analyzing content of the texts and its external contents, we can identify what kinds of thing or thoughts are talked. Then, each microblog message mb_i will be the most fundamental unit to be monitored in our work. Although each person's privacy should be kept, we also simply assume that the tweets which we are dealing with here are only tagged with no confidential policy towards the public. In addition, while each user's microblog's location can be incorrect or imprecise and lack of the written

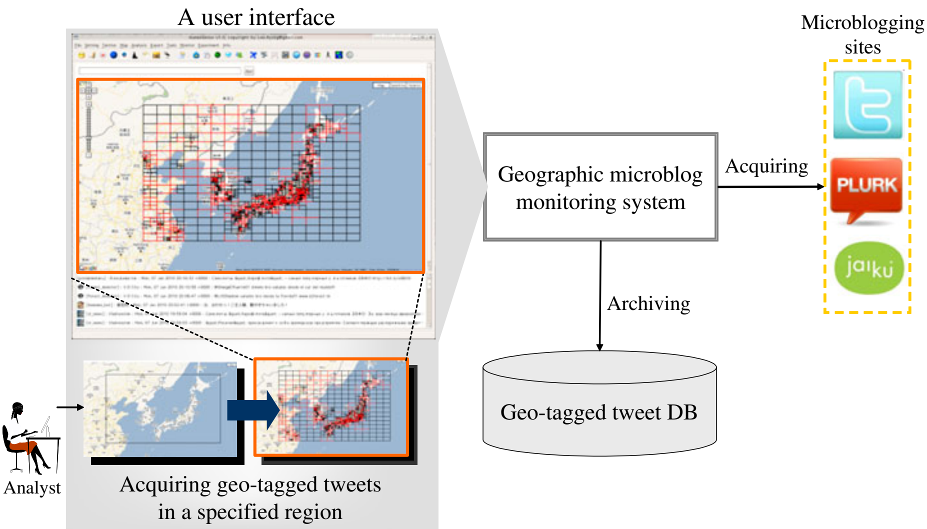


Figure 4 Outline of geographic microblog monitoring system.

Table 1 Example of geo-tagged tweets obtained from Twitter

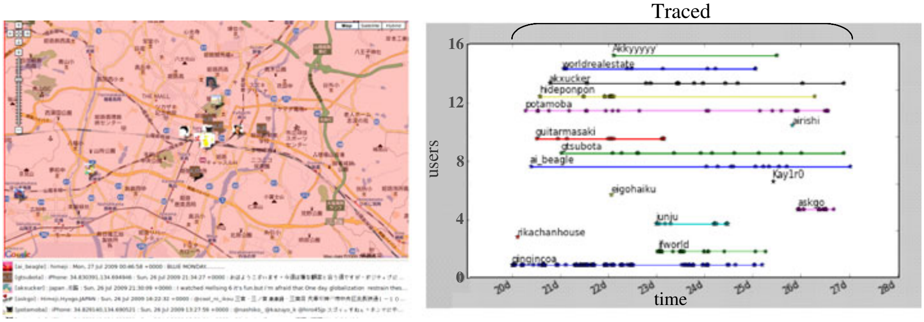
User ID	Timestamp	Loc _ lat.	Loc _ lng.	Texts
4872**	Wed 16 Jun. 2010 12:34:58	35.950251	139.067394	I'm traveling with my family.
6216**	Wed 16 Jun. 2010 15:09:14	32.795897	130.722657	Hokkaido is very comfortable.
8920**	Wed 16 Jun. 2010 21:25:48	35.390788	139.457521	Today, it is cloudy here.
3371**	Thu 17 Jun. 2010 09:43:37	35.905418	139.793515	We are ready to leave the hotel.
2609**	Thu 17 Jun. 2010 13:16:29	33.224421	130.320621	I'm in here. http://twitpic.com/4irwfg

contents, we also abbreviate such issues in this paper to focus on the proposed issue. We show some of tweets geo-tagged tweets which were collected in our experiment in Table 1.

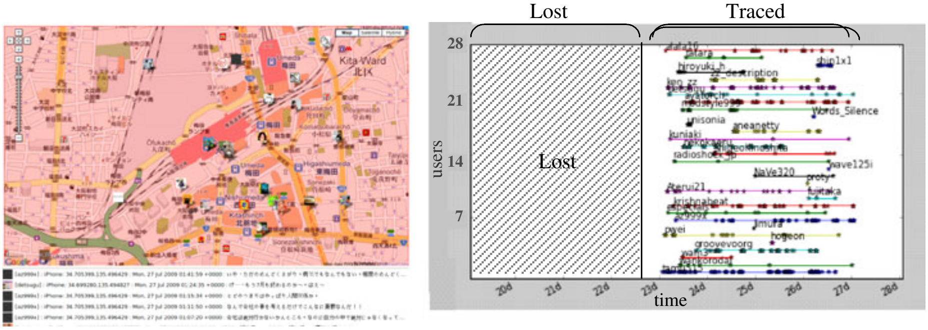
In fact, the location can be received in either the raw text form or in terms of very precise location coordinates. Hence, in the former textual style, we need to perform geo-coding to identify the exact coordinates by translating place names into the corresponding exact locations. We can easily solve the problem by using another mash-up service with Google Map's geo-coding service [7]. We directly transferred the place names to this thankful conversion service and received the precise coordinates. Therefore, we can accurately know when and where each tweet was written.

3.1 Difficulty of geographic microblog monitoring

In fact, with Twitter's Open API, we can search for the most up-to-date local tweets by a query with a geographic range composed of a central location and a radius. Based on the query framework, we are able to make a geographic search query, as like placing a virtual radar station on a center of any region of our interests to monitor what events are actually happening from there. In our first primitive experiment, a query to count tweets in our small suburban city, Himeji in Japan, as shown in Figure 5a (requested with a query centering in Himeji Castle with a 15.0 km radius boundary) got solely about 540 tweets, comparing with Osaka Station with a 1.0 km radius, where is one of the busiest areas in Osaka was over 1,500 shown in Figure 5b. For the two regions illustrated in Figure 5, temporal distribution of each dataset is shown together. The first dataset centering in Himeji Station can be traced up to the whole week, but the later cases had lost some portions. Since the current Twitter's API can return only up to 1,500 answers during a past week, the maximum number of answers with a time span shorter than one week such as Figure 5b means that we may have lost some portion of messages for the past week. Therefore, for better monitoring performance to cover a possibly minimal area and keep up the whole occurred messages, we need an efficient deployment method of virtual radar stations, which are actually realized as on-line queries consisted of a center and a radius to microblogging sites. The specification of circular queries is quite general in most web open APIs for the 'near-by' searches. Thus, we expect our approach applies to



(a) Himeji Station (less populated area) by 15.0 km radius



(b) Osaka Station (highly populated area) by 1.0 km radius

Figure 5 Geographic tweets search and temporal traceability.

monitor real-world events through microblogging sites in most sites, if they support at least a circular range query.

3.2 Efficient monitoring by rapid focus on tweet occurrence areas

We now explain a framework to gather those sensing reports from crowds spread over the world. For microblogs' occurrence patterns in spatial and temporal dimensions, we first assume a random distribution, while there may be predictable biases such as intensive weekdays or populated urban areas. Then, we need to monitor evenly whole the region of our interests without any pre-determined deployment setting. For the generic setting which can be extendable enough to apply any place or time period, we need an autonomous deployment method to reduce the total costs of observing microblogs by the virtual stations described above. Especially, for the rapid focusing on a sudden event in a region on the earth, we should perform fast traces for the changes while identifying where to focus on. However, to effectively deploy our radar stations considering the network politeness issue and coverage area in a limited time, we have to consider what places in the region of our interest are best suitable, what a large area each radar should cover for monitoring, and how often each one needs to monitor; possibly, we had better to generate a small number of queries having an optimal setting by best locations and smaller non-overlap ranges to micro-blogging sites. In our first test described above, our suburban city can be covered with about a 15.0 km radius sufficiently only by one radar station for full-tracking of the past week's messages, but the latter two busiest cities were not enough by one radar even for the 1/255 times smaller geographic area. For such densely populated cities where many tweets are published in a relative short period, we need to deploy much more small radars and conduct much more frequent observations up to several times in a week or in a day, if we want to catch up whole tweets in a specified area without any loss of them. In order to establish a theoretical framework for deploying the virtual radar stations to monitor massive geographic microblogs, we first built a circumcircle-based target space covering and applied a quadtree-based space splitting [4] as shown in Figure 6, where a space is recursively split into four the nearly same-size rectangular area until each space has a size larger than a minimum radius permitted in the query specification and the number of found results is in bound of the identifiable number. (Here, for a practical processing and simple discussion, we will deal with a rectangular region as a basic shape in this work.)

Hence, to examine a rectangular region, a circumcircle region enclosing the rectangle is asked instead. Although the larger area increased by the circumcircle could deteriorate search results due to four unintended coverings on four sides of the rectangle, this method can help us simplify the deployment problem and utilize general planar based space partitioning and indexing algorithms when asking for on-line APIs which usually support a nearby circular query only. Furthermore, for the exactness of the results, we can also filter out unwanted results existing in the extended areas by examining detailed coordinates of the results. Eventually, the required number of radar stations and its coverage can be greatly improved as shown in Figure 7a, if spatial distribution of data is well agglomerative. In Figure 7, 8×8 square cells can be monitored only by 4 radars, compared to expected 64 radars by the smallest size cells. With the quadtree, we split each cell whenever the number of tweets is reaching the maximum number of answers, since such cell can have much more higher-resolution data in their further splitting subordinate cells.

Based on the monitoring system, we actually obtained geo-tagged microblogs for approximately one and a half months (2010/06/04–07/30) around Japan with the

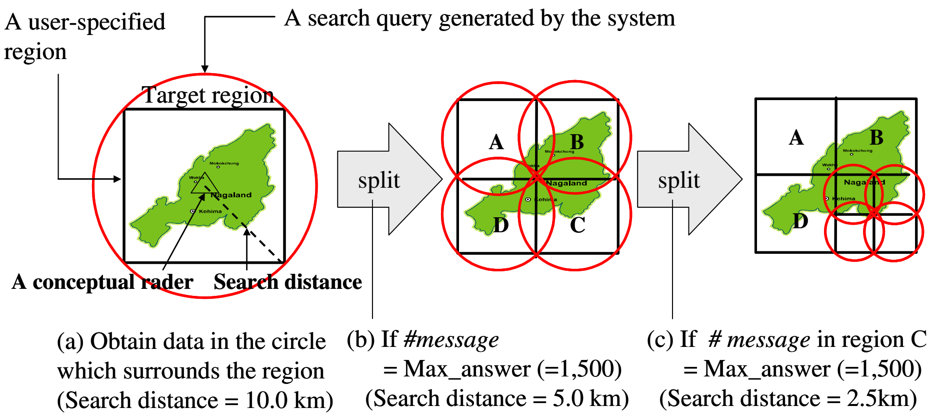
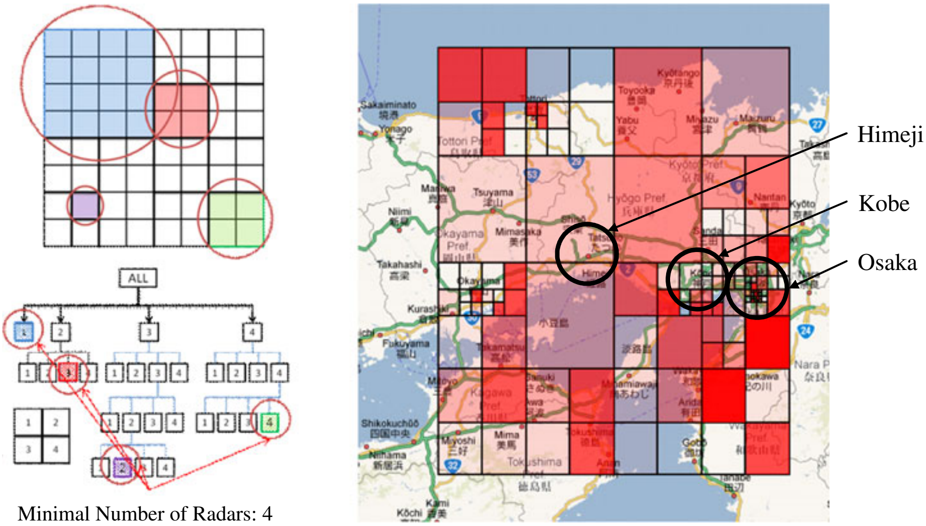


Figure 6 Quadtree adaptive space partitioning by tweet distribution.



(a) Quadtree-based radar deployment

(b) Adaptive radar deployment around an area centering in Himeji City

Figure 7 Optimal deployment of geographic tweets search queries by quadtree.

latitude range [30.004609:45.767523] and the longitude range [116.27921:148.381348] by our geographical tweets collecting system. As a result, we could gather 20,378,372 geo-tagged tweets from 287,145 distinct users. Figure 8 depicts our monitoring results

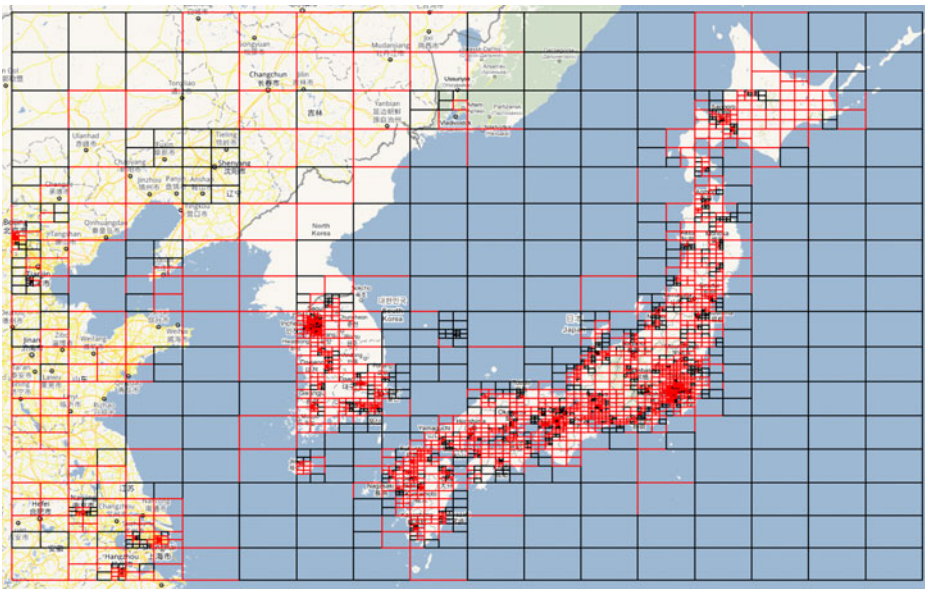


Figure 8 Geographic distribution of tweets represented by a quadtree (2010/06/04–07/20).

targeting a nationwide area around Japan and Korea. However, in this experiment, we utilize the tweets found only in the area of Japan.

4 Process of geo-social event detection

In this section, we address the process of the proposed event detection method in detail. We first describe a platform which is designed to realize the proposed method based on the two critical functions of (1) geographic regularity construction and (2) final event detection. In the following subsections, we will explain the process separately from the configuration of socio-geographic boundaries to the final detection.

4.1 Overview of event detection process

The event detection system we have developed has the two core functions as depicted in Figure 9.

(1) **Geographic regularity estimator**
In the function, the training data for a period of time are clustered with predetermined geographic boundaries. Then, for each region, 6-h usual statuses

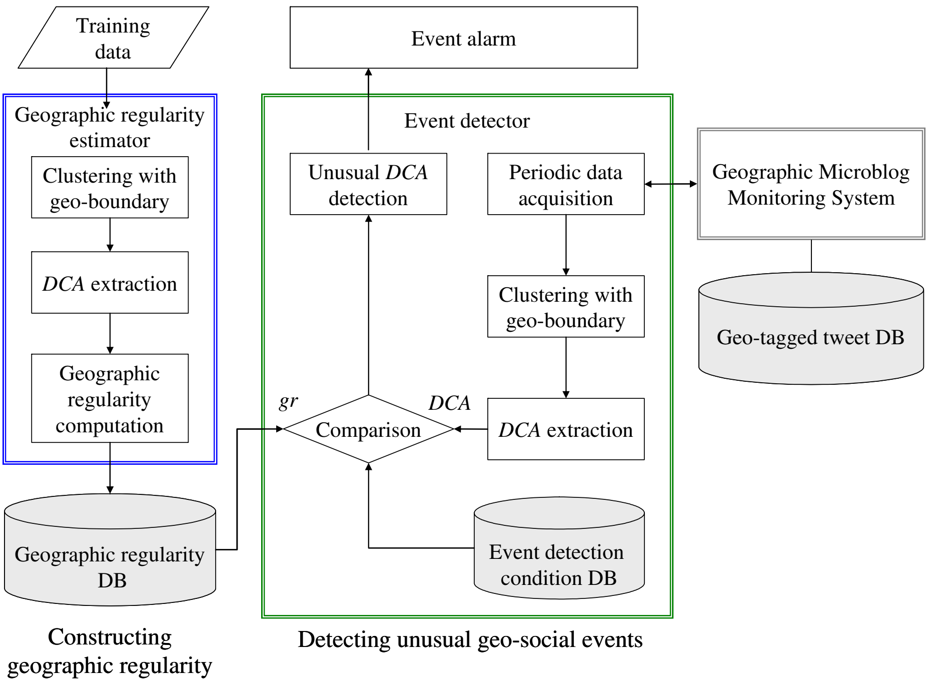


Figure 9 Platform of geo-social event detection system.

called DCA (this will be described later) are generated. Finally, the geographic regularities are computed and saved into the geographic regularity DB.

(2) **Event detector**
This function has an important role for the final event decision. Like the above estimator, it also computes 6-h current statuses through DCA extraction. After that, these statuses are compared to the regularity of each region. If the comparison results satisfy with the event conditions which are defined at Event detection condition DB, it will alert the detected events finally.

4.2 Configuration of socio-graphic boundaries

We initially need to determine a set of test socio-graphic boundaries from the occurrence places of the geo-tagged dataset, which we have accumulated in advance by our geographical tweet monitoring system. The target region can be partitioned using grid-based equally sized cells or administrative districts such as towns or cities. In the case of grid-based cell splitting, the adequate cell size is very difficult to determine. For instance, if we split a region into excessively small cells, most sub-urban areas will consume considerable unnecessary monitoring costs, even though the probability of tweet occurrence is generally very rare. Furthermore, since this approach does not consider the geographical distribution of tweets, the balance over the target region becomes inefficient and consequently results in a poor detection performance. On the other hand, partitioning on the basis of administrative districts also has a weakness since we cannot determine whether crowd activity regions are strongly relevant or almost dependent on the administrative districts. In addition, if two neighboring districts are strongly connected to each other in terms of social crowd activities, simply splitting them into two different groups will not be a good choice.

Therefore, to establish socio-graphic boundaries in a simply way, we classify geo-tagged microblogs on the basis of latitudes and longitudes by K-means [12]. Specifically, we become to partition a target space on the basis of the point distribution of actual microblogs. Then, for performing the K-means clustering method, we need to decide how to adequately set the number of K. We empirically set $K = 300$ in our later experiment. Furthermore, we formed a Voronoi diagram [2] using the center points (lat., lng.) of the K-means results and regard the formed regions as a set of socio-graphic boundaries. Lastly, we constructed a set of socio-graphic boundaries formed by our method using tweets in the rectangular region shown in Figure 10. While we could gather tweets for surrounding areas, we only use the tweets written in Japanese for the focused event detection. Consequently, small socio-graphic boundaries of densely populated areas appeared around major cities (Tokyo, Osaka, and Nagoya); in contrast, large socio-graphic boundaries of sparsely populated areas are spread over the other sub-urban areas or surrounding sea.

4.3 Measuring geographical regularities

In an assumption that we can recognize local events from the crowd behavior, we present a method to summarize the status of these behavior with a succinct representation.

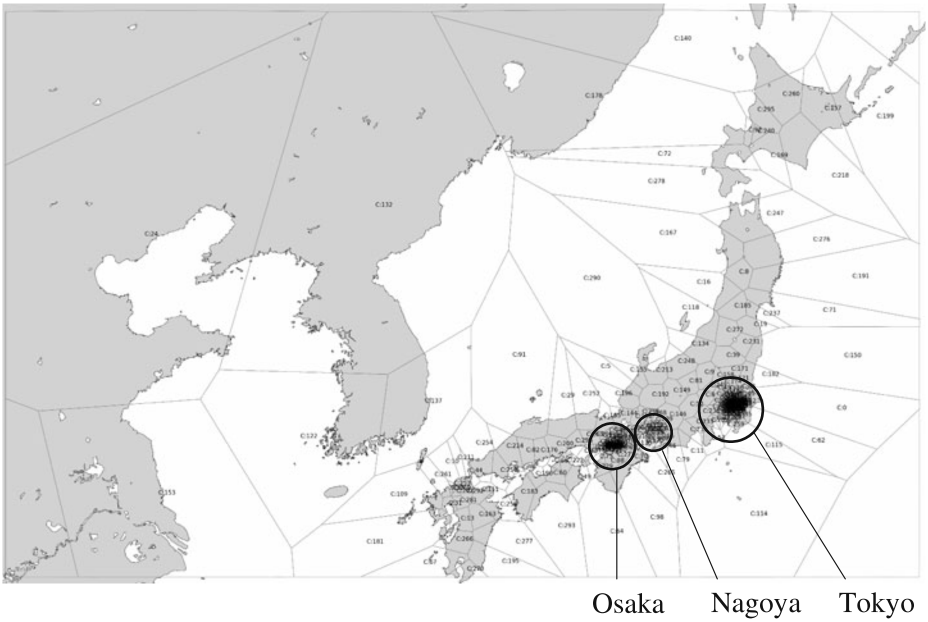


Figure 10 Socio-graphic boundaries represented by a voronoi diagram.

4.3.1 Estimating the degree of crowd activity

Based on the conception of microblogging services as a geo-social event detector, we actually need to determine how we can estimate the degree of crowd activity from the obtainable dataset. Generally, the written texts by crowds would have enough information detailing the real situations of geo-social events. However, detecting events depending only on textual information is quite limited, because the intrinsic property of the short length comparing to other usual web pages or conventional blogs makes it difficult to extract information enough to decide the occurrence of events. Especially, even for messages about an event, it is hard to find event names directly, since people would write about their thoughts or feelings. However, one critical point is that people who are experiencing events have a tendency to write more and act more during the time. Thus, by focusing on such indirect but explicit information which can be easily derived from the dataset such as the increasing occurrences of tweets, users and their moving would be much effective measures, when detecting the occurrence of local events quickly. Therefore, in this work, we primarily focus on the crowd activities reflecting on Twitter as the first geo-social event detector. Specifically, we define three different types of crowd activity indicators as follows (Figure 11).

(1) **Degree of Crowd Activity based on Tweets (DCA_T)**
The number of tweets that were written in a socio-graphic boundary within a specific period of time. For this, we count the total occurrence of tweets whose geo-spatial occurrence places are inside of a socio-graphic boundary during a specific period of time.

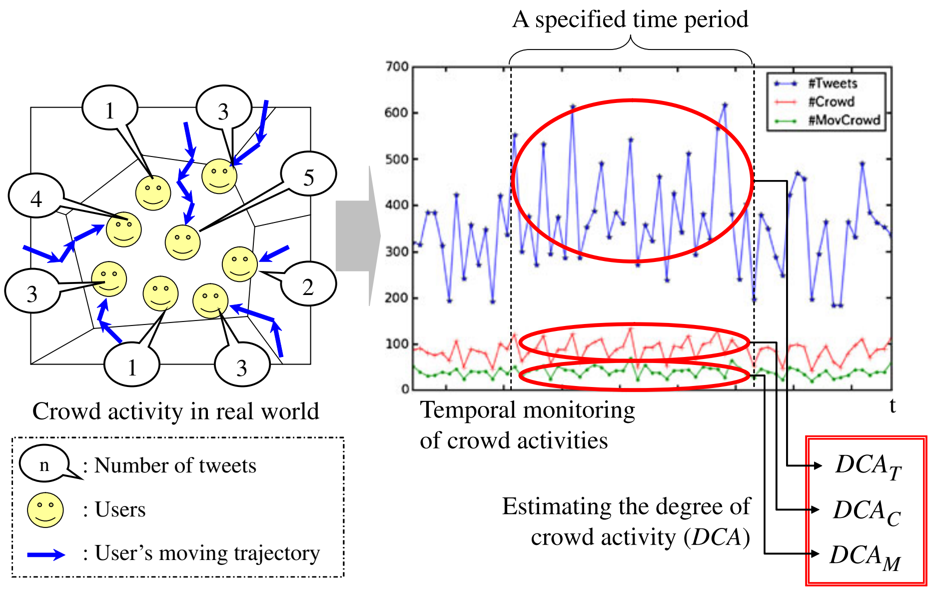


Figure 11 Acquisition of crowd activity data.

(2) **Degree of Crowd Activity based on Crowd (**DCA_C**)**
The number of Twitter users found in a socio-graphic boundary within a specific time period. Likewise, we count the total number of users who wrote the tweets counted for obtaining the value of the abovementioned indicator. In general, the in-equality $DCA_C \leq DCA_T$ is valid since any individual can write one or more tweets during a specific period of time.

(3) **Degree of Crowd Activity based on Moving crowd (**DCA_M**)**
The number of moving users related to a socio-graphic boundary within a specified period of time. In terms of partitioned socio-graphic boundaries, there are three types of moving user group: (1) Inner: A crowd in a socio-graphic boundary moves only inside the region without going outside; (2) Incoming: There are some people coming from outside; and (3) Outgoing: Conversely, some people move outside the socio-graphic boundary. To simplify the cases as much as possible, the moving crowd that we took into consideration included the inner and incoming user groups. Also, another inequality $DCA_M \leq DCA_C$ stands since a part of crowds actually shows movings.

4.3.2 Construction of geographic regularity

As aforementioned, we define three indicators for representing a current status of a socio-graphic boundary. Next, for the above indicators, we need to derive their usual patterns from a long-term training dataset. For the sake of simplicity, we deal with all the indicators in a statistical manner by using a boxplot [13], which is primarily used for explicitly visualizing a data distribution as shown in Figure 12. In its simplest form, a boxplot presents five sample statistics—the minimum, the lower quartile, the median, the upper quartile, and the maximum—in a visual

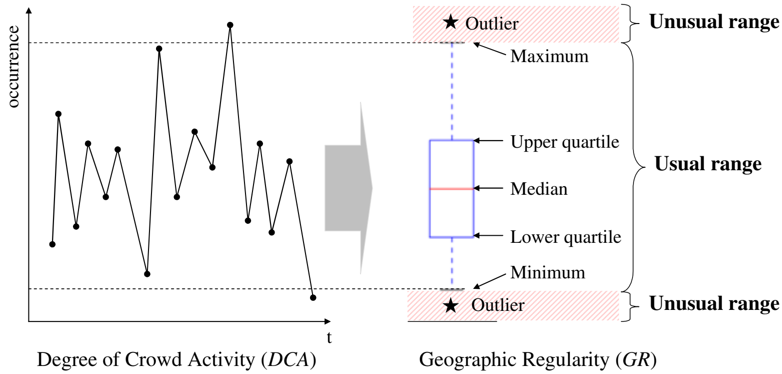


Figure 12 Boxplot-based geographical regularity construction.

display. The box of the plot is a rectangle that encloses the middle half of the sample, with an end at each quartile. A line is drawn across the box at the sample median. Whiskers sprout from the two ends of the box between the sample maximum and minimum.

For temporal logs during a specific time period, the boxplot generates a bounding range which is used as a geographic regularity with the range between the minimum and the maximum as shown in Figure 12. In fact, it is a sufficiently succinct representation to achieve our goal of recognizing an unusuality on the basis of a usuality range constructed from a real dataset. Specifically, in the figure, we regard the range inside of a box as a stability domain. The part that grows like a mustache expresses the minimum and the maximum values excluding the outliers and the range between these values is defined as a permissible zone. In contrast, we regard the upper and lower ranges except the permissible zone as unusual ranges. Therefore, if higher or lower limits of the permissible zone outreach, we extract a socio-geographic boundary r_i as an unusually crowded region.

As depicted in Figure 13, we finally build geographical regularities from the three indicators. That is, for a socio-graphic boundary during a specific time period, gr_T, gr_C, and gr_M represent the geographical regularities for DCA_T, DCA_C, and DCA_M, respectively. We considered that individual socio-graphic boundaries could have a different quantity of patterns and hence established a geographical regularity for each socio-graphic boundary from these three indicators by the notation $gr(r_i) = (gr_T, gr_C, gr_M)$, where $r_i \in$ socio-graphic boundaries.

4.4 Detection of unusual local activities

In order to detect unusually crowded regions that can eventually tell us the occurrence of unusual geo-social events, we compare the established geographical regularity with a newly observed 6-h count for the entire target socio-geographic boundaries. In particular, we determine the cases that we need to detect from among the combinations of the 6-h based three indicators.

(a) Normal cases
First, we explain what cases would be normal. Case (a) shows a typical type of normal tendency in the three indicators. In fact, people would write tweets

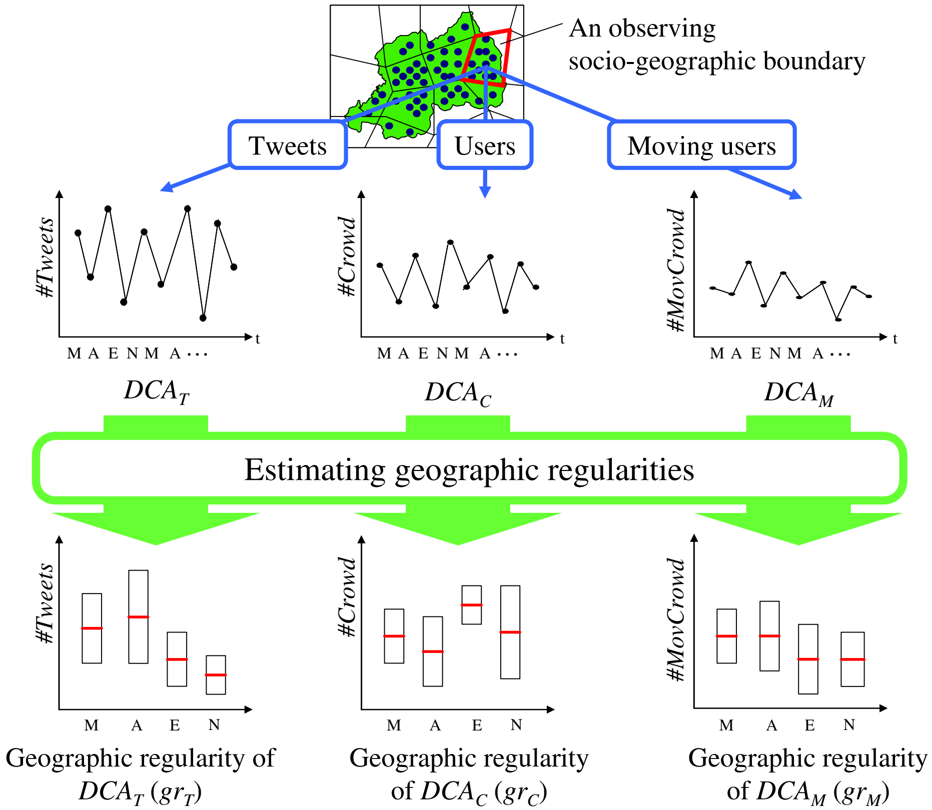


Figure 13 Estimation of geographical regularity.

usually in an expected range when there is no unusual incident in the region. Next, case (b) expresses a possibility that the number of movements is only abnormal, but it does not make the final decision (F) abnormal. We considered that this tendency would be often seen observed in urban stations. In such places, it is usual that many people show a considerable amount of movement. Hence, there are no likely geo-social events in these stations. If an event is held in the places, it is very hard to be detected. Thus, we simply assume that the case (b) is normal. In case (c), it represents a possibility that the number of crowd is only abnormal. For example, some users may write a significant number of messages every day, while others would only write few messages in holidays. In this case, the type says that the timing of the tweets overlaps by chance in a certain time span. Hence, we ignore this condition as a normal case. Then, in case (e), only the number of tweets is abnormal. In other words, people who sent messages on a routine basis write more tweets. If a user experienced a happy occasion or an unfortunate accident, the user would share his/her feelings with many users and more messages than usual are sent. Therefore, we do not consider this case to be abnormal with respect to geo-social events. Lastly, in case (g), both DCA_T and DCA_C are abnormal. This may indicate many topical incidents. For example, when World Cup 2010 began on Jun. 11, 2010, many people in Japan saw this event at home and sent lots of messages related to this topic. In this case, the number of tweets and crowd increased as compared to a usual condition. We also ignore such a case as it is not connected to the geo-social event detection.

(b) Abnormal cases

In case (h) as illustrated in Figure 14, it is apparent that all the indicators show an abnormality as shown in Figure 15. That is, there many moving crowds who usually write many tweets. Next, in case (f), DCA_T and DCA_M are unusually activated as compared to their geographical regularities. As a representative example, we can consider local festivals that usually bring about the movement of local crowds but fail to aggregate many visitors.

Finally, in case (d), DCA_C and DCA_M unusually increased as compared to their geographical regularities. This case can be regarded as a variant of case (h). In this case, only the total number of tweets differs. We can reason this by assuming that some long national holidays. However, they do not write so many tweets in only one place; rather, they write many tweets while moving around. Therefore, each socio-geographic boundary's average message will not be significant, while the other two indicators (gr_C and gr_M) increase significantly.

In our research, we want to detect geo-social events that result in a congregation of people. For this, we define that a socio-geographic boundary is under an unusual condition when its indicators satisfy the following logical function:

$$F = TM + CM + TCM = CM + TM = M(C + T) \tag{1}$$

In formula (1), the unusual conditions are denoted by T, C, and M. The final condition F implies that DCA_M is consistent in an unusual condition, Tweets is under a usual condition, and DCA_C is under an unusual condition, or DCA_C is under an unusual condition. That is, in the case of Table 2, (d), (f), and (h) become unusual conditions.

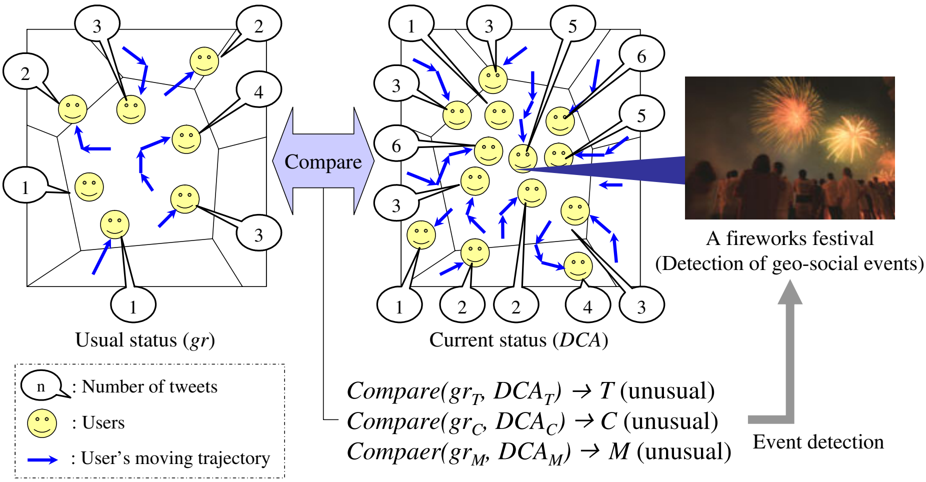


Figure 14 Example of detection of unusually crowded region based on case (*h*) in Table 1.

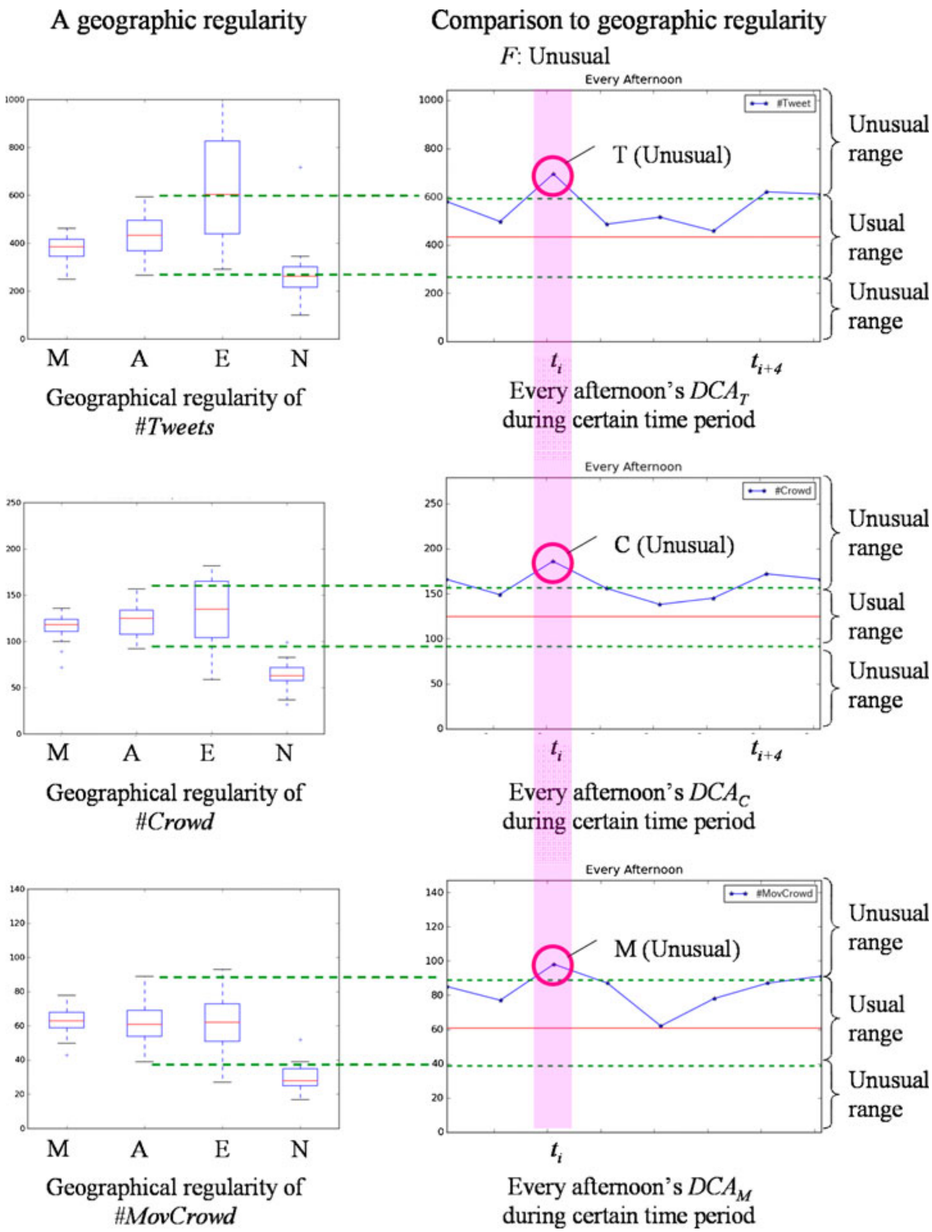


Figure 15 Detecting unusually crowded status with every-afternoon regularity.

5 Experiments and evaluation

In the experiments, we verify how the proposed method is successful in the discovery of geo-social events. For this, we prepared a list of festivals in Japan and tested our method to evaluate the detection performance by our proposed method.

Table 2 Final decision by a combination of unusuality detection indicators

	DCA_T	DCA_C	DCT_M	Final decision (F)
(a)	False	False	False	False
(b)	False	False	True	False
(c)	False	True	False	False
(d)	False	True	True	**True**
(e)	True	False	False	False
(f)	True	False	True	**True**
(g)	True	True	False	False
(h)	True	True	True	**True**

False: usual, *True*: unusual

5.1 Data preparation and socio-geographic boundary construction

Before evaluating the ability of our event detection method, we need to prepare a true set about a local area's geo-social events. Actually, making such list would be a bit contradictory, since at first we do not know what kind of geo-social events would be found best. In general, there are two types of events; (1) expected events and (2) unexpected events. Expected events are usually scheduled in advance such as festival; using this type of events as the true set can benefit well-arranged schedule table. Consequently, it becomes so convenient to get the table from the web and evaluate the detection rate easily. The other case of unexpected events takes time and efforts to list up the events, usually much later after the occurrence of real events. In this paper, for the convenience of experiment and analysis, we prepared a list of expected town festivals in Japan during 2010/07/16–07/21 from a local event guide site (Rurubu.com, http://www.rurubu.com/event/index.aspx) as shown in Table 3. In the table, we listed 50 town festivals that had taken place in Japan during a testing period. We will test how many events in the list would be found by our geo-social event detection method and what unexpected events are actually found by our method. As for the experimental dataset, we utilize the tweets gathered in the geographic microblog monitoring system described in Section 3. Then, we constructed a set of test socio-geographic boundaries using the obtained data. As explained in Section 4.2, we split a specified space by a K-means clustering method using the data with a condition of $K = 300$. We showed the constructed geographic boundaries in Figure 10.

5.2 Estimating geographical regularities

We estimated the geographical regularities for the socio-geographic boundaries in terms of gr_T, gr_C, and gr_M for every fixed time slot; empirically, we split a day into four equal time slots—morning (M), afternoon (A), evening (E), and night (N)—for the period of 2010/06/14–06/30. In Figure 16, the center graph shows temporal changes of DCA for the training time period; we computed the boxplots shown in the right side of the figure. In our experiment, we estimated the geographical regularities of socio-geographic boundaries = 300 for every 6-h time slots basically, we discover unusually crowded regions using the predefined thresholds by the boxplot-based usual status ranges.

Table 3 Experimental result: 50 town festivals (Rurubu.com, http://www.rurubu.com/event/index.aspx) held in Japan during 2010/07/16–07/21 are investigated

Event ID	Event name	Date	Place	Cluster ID	Result
1	Sawara summer festival	7/16–18	Katori, Chiba	0	T
2	Hakusan festival	7/17–18	Hakusan, Ishikawa	5	T
3	Kawasuso festival	7/17–19	Shinonsen, Hyogo	29	F
4	Suigo festival	7/20	Yurihama, Tottori	29	T
5	Sanoyoi fire carnival 2010	7/18	Arao, Kumamoto	31	F
6	Gion festival Yamahoko procession	7/17	Chukyo, Kyoto	34	T
7	Yawata drum fstival	7/18	Yahata, Kyoto	38	F
8	Kogura Gion drum	7/16–18	Kitakyusyu, Fukuoka	44	T
9	Toyohashi Gion festival	7/16–17	Toyohashi, Aichi	46	F
10	Summer festival	7/17–19	Kobe, Hyogo	48	T
11	Hiwasa turtles festival	7/17	Minami, Tokushima	49	T
12	Shishikui Gion festival	7/16–17	Kaiyou, Tokushima	49	T
13	Ajiko shrein annual festival	7/19–20	Atami, Shizuoka	57	F
14	The 45th Kanon shrein Zenigata festival	7/17–19	Kannnonji, Kagawa	60	T
15	The 45th Kanon shrein festival	7/18	Kannnonji, Kagawa	60	F
16	Kashiwano Tenno festival	7/17	Taiki, Mie	70	F
17	The 20th shine pia Shiogama	7/18	Shiogama, Miyagi	71	T
18	Yamana shrein Tennou festival	7/16–18	Mori, Shizuoka	79	T
19	Nobori of Take at Bessho hot spring	7/18	Ueda, Nagano	81	F
20	Tadanoumi gion and	7/18	Takehara, Hiroshima	82	F
21	Ojiro festival	7/17–18	Kihoku, Mie	98	T
22	The 58th Kyushu fireworks festival	7/18	Karatsu, Saga	109	T
23	South Uonuma city Kanetsugu kou festival	7/17–19	Minami Uonuma, Niigata	134	T
24	The 32nd Ocean Expo Park fireworks	7/17	Motobu, Okinawa	141	F
25	The 24th Seaport Chatan carnival	7/17–18	Chatan, Okinawa	141	F
26	Omiya shrein annual celebration	7/16–18	Agematsu, Nagano	146	F
27	Hiyoshi summer festival 2010	7/18	Nantan, Kyoto	147	F
28	Doroinkyo festival at Hirakata	7/18	Ageo, Saitama	160	F
29	Kawasaki Daishi Wind bell fair	7/17–21	Kawasaki, Kanagawa	162	T
30	Kamaishi summer festival	7/18	Kamaishi, Iwate	191	F
31	Nishio festival	7/16–18	Nishio, Aichi	208	F
32	Sadamitsu Gion summer festival	7/18	Tsurugi, Tokushima	222	F
33	The 22nd Tamamura firefestival	7/17	Tamamura, Gunma	245	T
34	Onishi summer festival	7/17–18	Fujioka, Gunma	245	T
35	Honjyo Gion festival	7/17–18	Honjyo, Saitama	245	T
36	Towada lake festival	7/17–18	Towada, Aomori	247	T
37	Mt. Myoko Sekiyama shrein fire festival	7/17–18	Myoko, Niigata	248	T
38	Goddess of children grand festival	7/18	Miyazu, Kyoto	252	F
39	The 32nd Akiyoshidai sightseeing festival fireworks	7/17	Mine, Yamaguchi	254	T
40	2010 Kurihama Perry festival	7/17	Yokosuka, Kanagawa	259	T
41	The 58th National fireworks festival	7/17	Ise, Mie	265	T
42	Tenno festival at Wataka	7/20–21	Shima, Mie	265	T
43	The 55th Renryu festival	7/18	Minamata, Kumamoto	266	T
44	Kitakata summer festival	7/17–18	Kitakata, Fukuoka	272	T

Table 3 (continued)

Event ID	Event name	Date	Place	Cluster ID	Result
45	The 129th Echigawa Gion fireworks	7/17	Aisho, Shiga	280	T
46	Hikiyama festival	7/17–18	Oomihachiman, Shiga	280	T
47	Sennichi festival	7/17–18	Higashioomi, Shiga	280	T
48	Nanao port festival	7/17–18	Nanao, Ishikawa	290	T
49	Iida Toroyama festival	7/20–21	Suzu, Ishikawa	290	T
50	Muroto festival	7/18	Muroto, Kochi	293	T

T: found, *F*: not found

5.3 Detection of unusually crowded regions

We evaluate whether the proposed method can detect the geo-social events listed in Table 3. For this, we detected the socio-geographic boundaries that generated unusual activities as compared to the geographical regularity of each time slot for a test time period of six days from 7/16 to 7/21. As the procedure, we first compared gr_M, and if it was evaluated to be unusual, we compared gr_T or gr_C according to the logic function we presented in formula (1). Whenever the final decision (F) becomes abnormal, we extract r_i as an unusual crowded region.

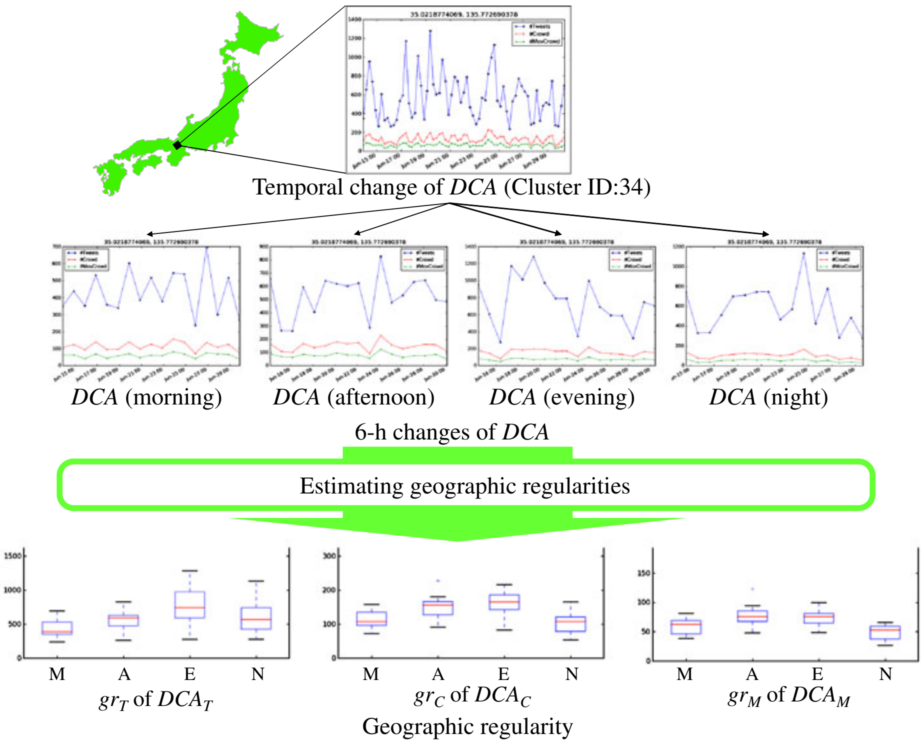


Figure 16 Example of estimated geographical regularities (2010/06/14–06/30) for a region.

5.3.1 Analysis of expected events

Furthermore, we examined the messages that were sent from this region and detected four festivals in the event list. We show some of the microblogs related to them in Figure 17. In these case, we found not only messages but also some photos.

On the whole, we confirmed unusually crowded regions from 7,200 (300 clusters $\times$ 6 days $\times$ 4 periods (6-h)) spatio-temporal dataset for 4 time slots during six test days (7/16–7/21) in 300 socio-geographic boundaries. The total number of socio-geographic boundaries that were evaluated as unusual was 903 (morning: 296, afternoon: 296, evening: 187, and night: 124). Therefore, 12.5% (= 903/7,200) of all the test time slots were answered as unusually crowded regions. In the detected socio-geographic boundaries, we confirmed how many events in the prepared event list were found. As a result, we could find out 32 events among the expected 50 events in Table 3 showing a satisfactory performance of recall = 32/50 (62.0%) in spite of the small number of the testing events lits.

Interestingly, we were able to find out many crowds' tweets who would like to share their excitements and experiences through Twitter as shown in Figure 17. For instance, we could observe some significant festivals involving local crowds' active updates as shown in Figures 18 and 19.

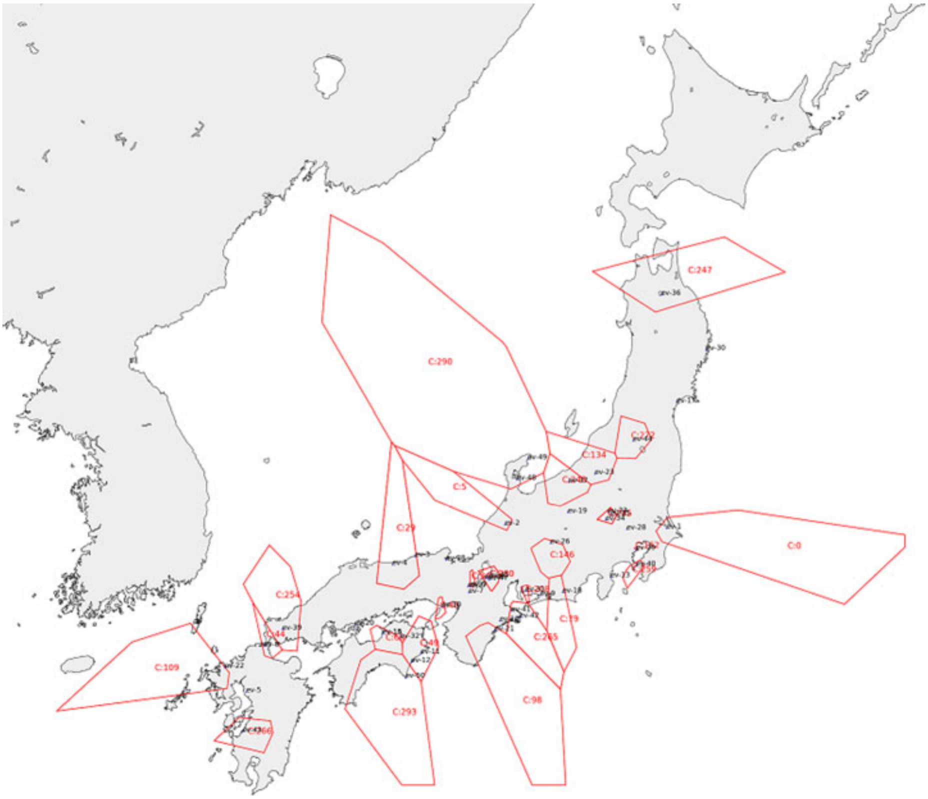


Figure 17 Experimental result: geographic boundaries whose events were found are displayed.

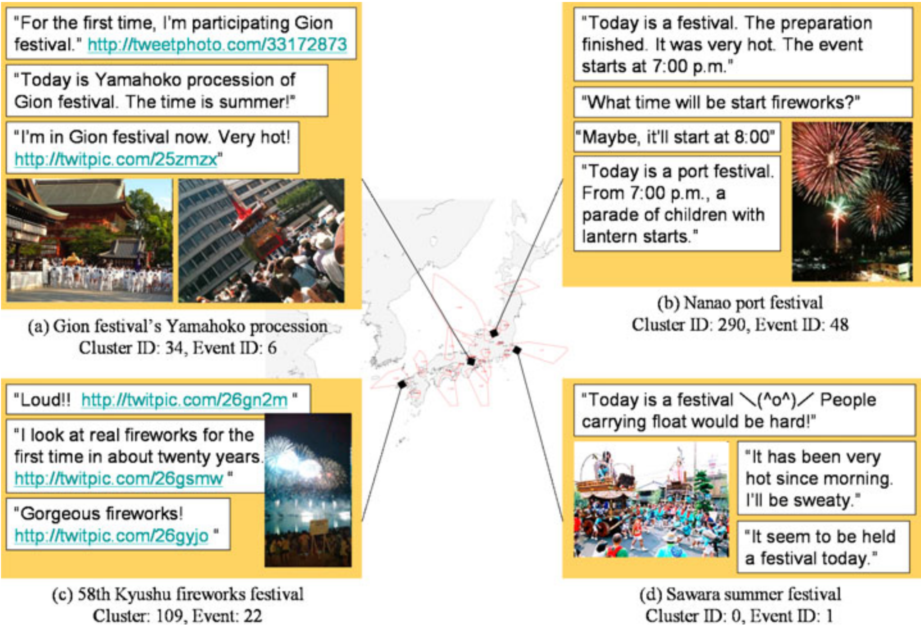


Figure 18 Example of real tweet messages about expected events.

(a) Gion festival's Yamahoko procession (Event ID: 6), Chukyo, Kyoto (Cluster ID: 34)
Gion festival which is a famous traditional festival in Japan was held on July 17th. Especially, the Gion festival's Yamahoko procession we found in this experiment is one of the most famous events and atttacted lots of people' attentions in Japan. In this event, a Japanese float "Yamahoko," a parade with Japan float "Yamahoko" was held at Gion in Kyoto. As for the traditional event, in terms of crowd activity based on Twitter messages, it obviously aggregated lots of people. In case of this event, it is difficult to observe unusual status of tweet occurrences. However, the degree of DCA_T was almost beyond the unusual status. When "Yamahoko" passed near the participants and there were participants who were following with it, hence the degree of crowd activity in terms of moving users are higher than the usual.

(b) The 58th Kyushu fireworks festival (Event ID: 22), Karatsu, Saga (Cluster ID: 109)
Our method can also detect the 58th Kyushu fireworks festival in the evening on July 18th. We found the unusual degree of crowd activity in terms of the number of users and moving users. Actually, the fireworks started at 20:00. People viewing fireworks did not write Twitter messages so frequently during the event time because of being attracted the beauty of fireworks. However, in order to participate in the event, lots of people visited from other regions and write Twitter message, so the distinct number of users definitely increased significantly on the event day.

(c) Nanao port festival (Event ID: 48), Nanao, Ishikawa (Cluster ID: 290)

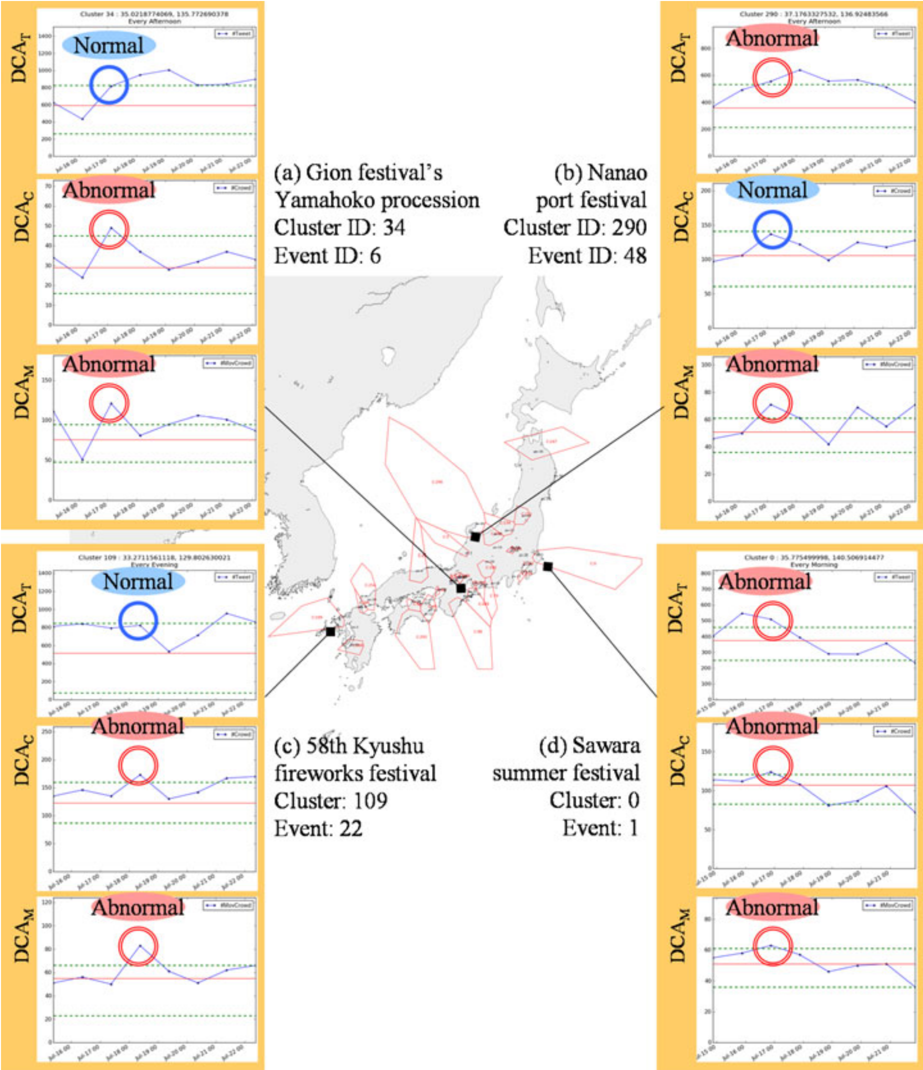


Figure 19 Expected events and their crowd activity data.

The event was held during two consecutive days, on July 17–18th. This event primarily was held in the evening time. In the experiment, an unusual status is detected in the morning period (06:00–12:00), because people who would like to announce the event information published the relevant tweets before the event. As for this local event, in terms of crowd activity, the number of tweets and moving users increased more than usual status. We can speculate that it was a kind of local events attracting people mostly from near-by area.

(d) Sawara summer festival (Event ID: 1), Katori, Chiba (Cluster ID: 0)
Sawara summer festival which is also a traditional festival began three hundred years ago. In this event, as like the Gion festival's Yamahoko procession, a

gorgeous float parade was held in the city. In the experiment, an unusual status is detected in the morning period (06:00–12:00) involving many pre-event activities and anouncements. In other words, even though before the event time, people were writing Twitter messages concerning the event held later.

Furthermore, we examined the messages that were sent from this region and detected four festivals in the event list. We show some of the microblogs related to them in Figure 17. In these case, we found not only messages but also some photos. Among the detected 903 time slots as unusual, 42 slots (4.65% = 42/903) are relevant to the prepared event list. This can be regarded as a precision rate. Even though the rate is quite low, it was not so poor result, since we could not expect all the events actually occurred in the experiment. We assumed that only 50 festivals can happen in Japan. In fact, we could find out many other unexpected events from the result.

5.3.2 Analysis of unexpected events

Among the detected spatio-temporal list as unusual, there were many other unexpected regions which seemed like unusually crowded regions. Hence, we examined the others to know why they were detected as unusual. Then, we newly discovered geo-social and natural events that did not appear in the event list of Table 3. We detected some other unexpected events as drawn in Figures 20 and 21. For instance, we could detect two kinds of events; (1) social event: a baseball game was held in a stadium and (2) natural event: a sudden thunderstorm let crowds write tweets more.

(a) A baseball game in Yahoo Dome in Fukuoka (Cluster ID: 93)
For instance, we detected an aggregation with supporters of a professional baseball team to watch a baseball game in a stadium. Specifically, many fans congregated there to cheer their favorite team writing a lot of tweets as shown in Figure 20a.

(b) A sudden thunderstorm in Niigata (Cluster ID: 134)
In general, natural phenomena would be difficult to expect of their happenings. Therefore, if the natural incidents are occurred, lots of people try to notify their situations to others. We detected the occurrences of a sudden heavy thunder

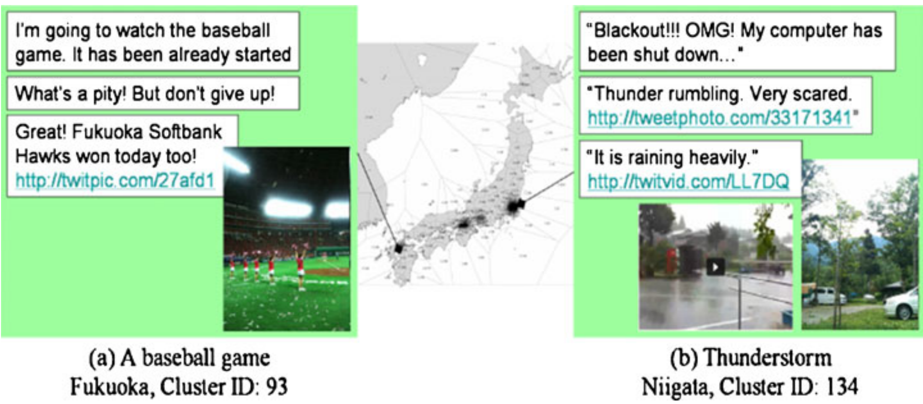


Figure 20 Example of real tweet messages about unexpected events.

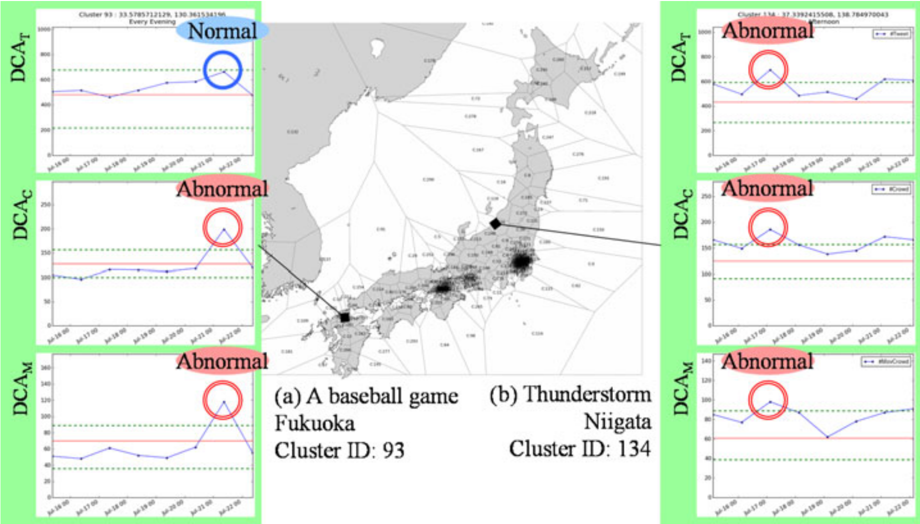


Figure 21 Unexpected events and their crowd activity data.

and lightning as a natural local incident. In Figure 20b, we illustrated the actual messages and photos related to the thunderstorm that we found in the analyzed tweet messages.

6 Conclusion and future work

In this paper, we proposed a geo-social event detection method based on the geographic regularity which reflects a geographic region's usual status through crowd behavior observable on Twitter. We described the detail of the geographic microblog monitoring system which is a unique platform enabling us to realize this exploring geo-social analytics. In particular, focusing on local crowd behavior found in Twitter, we were successfully able to find out many expected events and discover unexpected events in the experiment.

In the future work, we will further examine various measurable patterns as well as their causes and effects using better clustering methods with different levels of temporal and spatial granularity. We also plan to perform a content analysis of microblogging messages to understand the reasons of social events, and aim to identify additional regional phenomena based upon their association with geo-tagged and time-stamped data.

Acknowledgement This research was supported in part by a Grant-in-Aid for Scientific Research (B) (2) 20300039 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Asadi, S., Zhou, X., Yang, G.: Using local popularity of Web resources for geo-ranking of search engine results. World Wide Web **12**(2), 149–170 (2009)

2. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv. (CSUR) **23**(3), 345–405 (1991)
3. Dong, H., Khadeer Hussain, F., Chang, E.: A human-centered semantic service platform for the digital ecosystems environment. World Wide Web **13**(1–2), 75–103 (2010)
4. Finkel, R., Bentley, J.L.: Quad trees a data structure for retrieval on composite keys. Acta Inform. **4**(1), 1–9 (1974)
5. Fujisaka, T., Lee, R., Sumiya, K.: Detection of unusually crowded places through micro-blogging sites. In: Proc. of the 6th International Symposium on Web and Mobile Information Services (WAMIS 2010), pp. 467–472 (2010)
6. Fujisaka, T., Lee, R., Sumiya, K.: Discovery of user behavior patterns from geo-tagged micro-blogs. In: Proc. of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010), pp. 246–255 (2010)
7. Google Geocoding API: http://code.google.com/intl/en/apis/maps/documentation/geocoding/
8. Iwaki, Y., Jatowt, A., Tanaka, K.: Supporting finding read-valuable articles in microblogs. In: DEIM Forum 2009 A6–6 (in Japanese) (2009)
9. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding micro-blogging usage and communities. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65 (2007)
10. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proc. of the First Workshop on Online Social Networks, pp. 19–24 (2008)
11. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proc. of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN 2010), pp. 1–10 (2010)
12. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
13. Mcgill, R., Tukey, J.W., Larsen, W.A.: Variations of box plots. Am. Stat. **32**(1), 12–16 (1978)
14. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In: Proc. of the 6th ACM Conference on Embedded Network Sensor Systems, pp. 323–336 (2008)
15. Moriya, K., Sasaki, S., Kiyoki, Y.: A dynamic creation method of environmental situation maps using text data of regional information. In: DEIM Forum 2009 B1–6 (in Japanese) (2009)
16. Mumbai attacks. http://www.telegraph.co.uk/news/worldnews/asia/india/3530640/Mumbai-attacks-Twitter-and-Flickr-used-to-break-news-Bombay-India.html
17. Otsuka, S., Takaku, M., Kitsuregawa, M., Miyazaki, N.: A study for analysis of user behavior in the free magazine site for women. In: DEIM Forum 2009 B8–4 (in Japanese) (2009)
18. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proc. of 18th International World Wide Web Conference (WWW 2010), pp. 851–860 (2010)
19. Sysomos. Inside Twitter: An In-depth Look Inside the Twitter World, 10 June 2009. Retrieved on 23 June 2009
20. TIME: Iran protests: Twitter, the medium of the movement. http://www.time.com/time/world/article/0,8599,1905125,00.html. Accessed on 17 June 2009
21. Wang, H., Zou, H., Yue, Y., Li, Q.: Visualizing hot spot analysis result based on mashup. In: Proc. of the 2009 International Workshop on Location Based Social Networks (LBSN 2009), pp. 45–48 (2009)
22. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proc. of the ACM 2009 International Conference on Supporting Group Work (GROUP 2009), pp. 243–252 (2009)