Pers Ubiquit Comput (2013) 17:605–620
DOI 10.1007/s00779-012-0510-9

ORIGINAL ARTICLE

Urban area characterization based on crowd behavioral lifelogs over Twitter

Ryong Lee · Shoko Wakamiya · Kazutoshi Sumiya

Received: 1 October 2011 / Accepted: 1 February 2012 / Published online: 30 March 2012

Abstract Recent location-based social networking sites are attractively providing us with a novel capability of monitoring massive crowd lifelogs in the real-world space. In particular, they make it easier to collect publicly shared crowd lifelogs in a large scale of geographic area reflecting the crowd's daily lives and even more characterizing urban space through what they have in minds and how they behave in the space. In this paper, we challenge to analyze urban characteristics in terms of crowd behavior by utilizing crowd lifelogs in urban area over the social networking sites. In order to collect crowd behavioral data, we exploit the most famous microblogging site, Twitter, where a great deal of geo-tagged micro lifelogs emitted by massive crowds can be easily acquired. We first present a model to deal with crowds' behavioral logs on the social network sites as a representing feature of urban space's characteristics, which will be used to conduct crowd-based urban characterization. Based on this crowd behavioral feature, we will extract significant crowd behavioral patterns in a period of time. In the experiment, we conducted the urban characterization by extracting the crowd behavioral patterns and examined the relation between the regions of common crowd activity patterns and the major categories of local facilities.

Keywords Urban characteristics · Crowd behavior · Location-based social network sites · Microblogs

R. Lee (✉) · S. Wakamiya · K. Sumiya
University of Hyogo, Himeji, Japan
e-mail: lee.ryong@gmail.com

S. Wakamiya
e-mail: ne11n002@stshse.u-hyogo.ac.jp

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

R. Lee
National Institute of Information and Communications
Technology (NICT), Kyoto, Japan

1 Introduction

Lifelogging increasingly becomes one of our common and daily habit undisputedly exemplified by today's social network sites. In fact, from the early work by Steve Mann's laborious lifelogging with wearable computing systems [1] to Gordon Bell's MyLifeBits [2] for digitizing every moment of individuals, storing and recalling our lifetime memory have been intensively studied well. However, the recent advances of social networks encourage us to write our lifelogs much easily and share them instantly with any other people around the world. Accordingly, individual lifelogs are not bound only to personal memory. The shared memories of enormous crowds over the open space are extending the individual lifelogs to community experience logs, which can vividly reflect many important social and physical real-world events or phenomena. Specifically, on behalf of the rapid distribution of smartphones and the location-based microblogging sites such as Twitter [3], Foursquare [4] and Gowalla [5], we can now share our daily activities as well as our minds instantly from any place clarifying where we are located in the world. In particular, this kind of global trend will be delivering lots of novel applications that can benefit from exploiting the shared crowd lifelogs in terms of the huge volume of geo-tagged data and their heterogeneity of contents about almost every kind of crowd activities in the real world.

In this work, motivated by the fact that crowd's lifelogs over the social networks can include real-world location information with the shared messages, we attempt to analyze urban characteristics from the crowd-sourced data over Twitter. Indeed, it is a critical issue to characterize urban space to support various real-life decision makings in the space. For instance, when we have to look for a house to move in an unfamiliar city, it would require a bothering effort to quickly grasp the living environment while drawing the image of the city in mind such as "Where are the most popular places for living with good educational environments?" or "Where is the downtown area attracting together many people on weekends?" Furthermore, in many practical geo-business or geo-politics applications, we frequently have to examine overall features of local areas as quickly as possible. These kinds of questions we would often face will require lots of efforts to obtain, since the possible answers need to investigate many updating sources about ever changing information of urban areas. In other words, surveying such characteristics of urban space usually takes huge costs and time involving off-line on-the-spot observation or gathering information by the questionnaire method. Hence, quick and massive scale investigation could not be successfully done due to the extensibility problem such as monitoring massive tourists congregating in a city [6]. Consequently, we are limited in characterizing urban areas if we only depend on the lazy static survey results. Otherwise, we only had to depend on the general conception of urban areas previously formed by the mass media or self-experience or rough statistics from national census data.

However, compared to conventional urban characterization research, location-based social networks definitely have many significant advantages and benefits; first, there are enormous populations around the world who are voluntarily publishing their daily activities, particularly, acting in urban space. Second, the crowd-sourced data have various information from trivial travel experience to crowd's seasonal movement trajectories. These heterogeneous types of data shared over the social networks can help us conduct various quantitative and qualitative research studies and realize practical systems. Therefore, the unprecedented scale of crowd lifelogs in urban space will promise many challenging and beneficial issues in the near future.

In general, crowd behavior in urban area is a critical factor to understand urban space. We can easily imagine the strong relationship between characteristics of the real urban spaces and the activities of citizen. For instance, in terms of space syntax techniques by Jiang et al. [7], which attempted to describe urban spaces by crowd behavior, people expectedly become more active in morning and evening hours intensively in residential districts in the sense of integration. Particularly, in a city which we have usually developed for our convenience, our activities such as commuting, working, shopping, educating will characterize urban areas clarifying how we behave and live in the real space. Therefore, with crowd's daily lifelogs over social networks, we can conduct urban characterization by observing crowd behavior and finding significant behavioral patterns depicting how crowds are using our urban space. For this purpose, we present a model to explain and study the current situation relevant to social networking sites as illustrated in Fig. 1. Here, we intended to construct

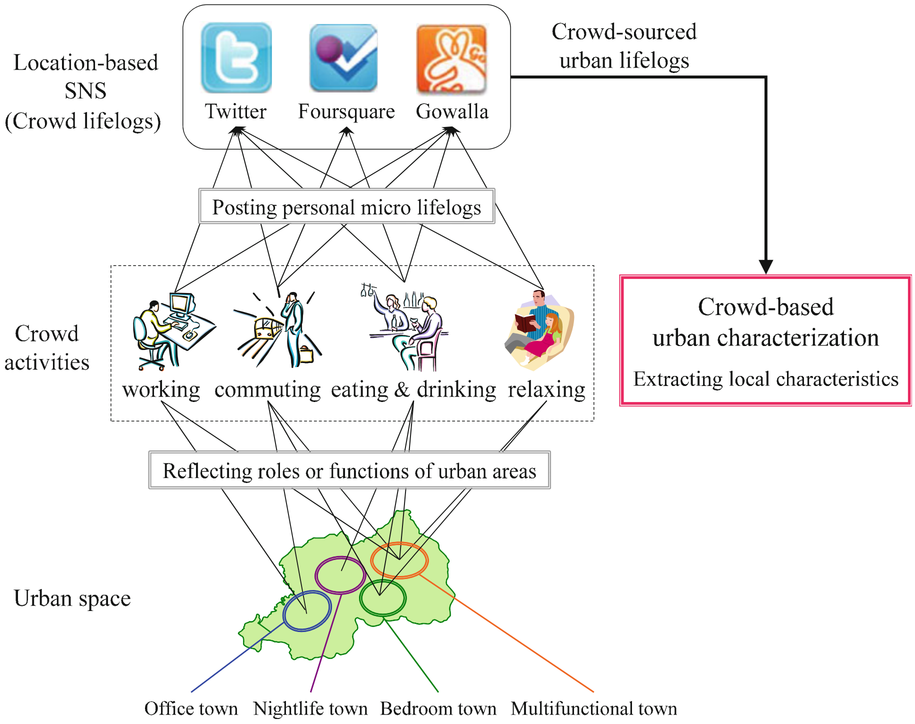

Fig. 1 Research model to characterize urban area using crowd behavior extracted from location-based social networking sites

much more flexible and elastic model, which can comprehensively reflect on the heterogeneity of the participating entities to explain the current situation relevant to social networking sites; that is, the real space (including real-world phenomena and social events as well as geographic physical environment), crowds (generally, people and their various capabilities of sensing, acting, thinking, emotionally feeling, etc.), and the virtual space (while this work limitedly focused on the Twitter as a representative location-based social network).

In the conventional social network studies, graph-based models [8, 9] are often adopted to focus on the relationship of users on the social networks. However, we think that each entity would require its own modeling and operations (e.g., in case of the real space, various spatial and neighboring concepts are often useful, but harder to consider such thing only in a graph). Especially, we would like to focus preliminarily on the influences from the real space to crowds and their reflecting activities on the cyber space by publishing information. In this respect, we presented a simplified model enough to represent the influential relationship among the entities.

Furthermore, in order to examine the dynamic nature of the crowd behavior observable from social network sites, we primarily focus on extracting urban characteristics in terms of crowd behavior in the real world that can be largely available by exploiting geo-tagged tweets from Twitter. Specifically, we compute a crowd behavior feature focusing on temporal changes of periodic occurrence of geo-tagged tweets for a geographic region. We also examine significant crowd behavioral features for reasoning urban characteristics. For this, we experimentally extract geographic crowd behavioral patterns from a large number of actually gathered geo-tagged tweets in Japan and report a comparison with real socio-geographic features of each region as an evaluation.

The remainder of this paper is organized as follows: Sect. 2 presents related work on capturing and utilizing crowd lifelogs for various purposes. Section 3 presents our challenge to make the most of the location-based social networking sites for extracting socio-geographic characteristics, which significantly represent how people use their living urban spaces. Section 4 describes an overall process focusing on exploration of normal crowd behavioral patterns and clustering urban areas where similar patterns of crowd behavior. Section 5 illustrates the experiment that we conducted to extract crowd behavioral patterns as urban characteristics with a huge volume of data gathered from Twitter and to reason the urban characteristics with categories of local facilities. Finally, Sect. 6 concludes this paper with a brief description of the future work.

2 Related work

In recent years, social networking sites can be regarded as a novel source, where we are able to easily monitor a great deal of daily crowd lifelogs. Obviously, this kind of unprecedented popularity by numerous people around the world is reflecting drastic changes toward reciprocal benefits among people through sharing personal lifelogs. While there would still be seemingly endless concerns about privacy related to personal location sharing, it must be a critical issue more and more to exploit the crowd-sourced data in the academic field as well as the business world in terms of social and individual benefits from the new open sharing space.

As for location identification or labeling to user-written messages over Twitter, on behalf of prevalence of location-sensing devices including recent GPS-embedded smartphones, several granularity levels of location can be automatically attached into the user messages from the fine latitude/longitude coordinates to the city name as a coarse representation. In case of a study conducted by Cheng et al. [10] for the purpose of getting more geographically related tweets, they attempted to reveal users' location information only from the written text by referring to other geo-tagged tweets with fine location coordinates. On the other hand, according to the survey by Sysomos [11] in 2010, there are already 73 % users who have submitted geo-tagged tweets compared to 44 % in 2009. While we still have to be aware of the privacy concern in exploring individual data, Barkhuus et al. [12] exemplified the usefulness of open sharing of individual updates including locations through mobile devices for a group of people in a field study, where participants can unexpectedly take advantages of awaring friends' updates less interrupting each other's daily activities. Furthermore, Hightower [13] discussed the importance of semantic location labeling by taking into account the activities of people on the places.

For exploiting the further useful use cases, lots of research work have been studied. Zheng et al. [14] presented a method to utilize the social network as a location information search framework by extracting knowledge over the location data based on GPS and users' comments at various locations to answer two typical socio-geographic questions: If we want to do something, where shall we go? If we visit some place, what can we do there? By modeling a location-activity relation into a matrix, the former question is answered by activity recommendation to given a location query, and the latter one is also resolved by location recommendation given an activity query. In this work, authors focused on determining the correlation between locations and crowd activities, while we are distinctively focusing on regular crowd activities on locations.

Interestingly, some researches focusing on cooperation with Twitter for analyzing some natural incidents in the real world have been introduced. De Longueville et al. [15] analyzed the temporal, spatial and social dynamics of Twitter activity during a major forest fire incident in the South of France in July 2009. Sakaki et al. [16] constructed an earthquake reporting system in Japan using tweets which are posted from each Twitter user regarded as a sensor. In this method, tweets which are reporting the occurrence of earthquakes are extracted by using textual information. Cheng et al. [17] studied human mobility patterns revealed by the check-ins over location sharing services and explored the corresponding factors that influence mobility patterns, in terms of social status, sentiment, and geographic constraints. This study focused on analysis of massive personal trace data based on characterizing geographic facilities. In our previous work [18, 19], we also proposed a method to discover local social and natural events by monitoring unusual statuses of local users' activities utilizing geo-tagged crowd lifelogs over Twitter. In order to attempt to measure TV ratings based on Twitter by looking into Twitter messages including TV-related words made of TV program titles or some verbs such as 'view' or 'watch' and 'TV', we examined the usefulness of Twitter as a means to carry out a poll [20, 21].

3 Characterizing urban areas with crowd lifelogs over social networks

In this section, we explain what we mean by the term 'Urban Characteristics' in this paper, which will be explored from social network sites. Especially, we describe what kinds of contexts we could obtain from the crowd-sourced social network sites. Fundamentally, in this work, we will depend on simplified crowd behavioral logs focusing on location and time of user existence without deeply analyzing textual messages from Twitter.

3.1 How can crowd lifelogs reveal urban characterization?

Generally, in urban space, there are a lot of facilities for living and working such as housings, transportation, offices, schools, parks, and shopping centers. In such complicated space, we can easily observe some daily routines of residents such as commuting, working, eating and drinking, and relaxing at home by exploiting crowd-sourced lifelogs over social networking sites as illustrated in Fig. 1. In addition, these crowd activities let us know roles or functions of the urban space; an urban area which observed crowds who are commuting and working would be conjectured as an office town. On the other hand, if crowds commuting, eating and drinking, and relaxing at home are monitored in the urban area, we might regard there as a bedroom town. Likewise, we are able to capture the image of urban space by means of crowd behavior.

In fact, geographic characteristics have been studied so far from various perspectives by physical geographic shape or diverse objects such as streets or landmarks, or by cultural and structural aspects such as residential, commercial, and industrial districts. These two different views have been well studied in many research fields. Kevin A. Lynch's seminal contribution in his book titled "The Image of the City" [22] defined five fundamental elements of a city: paths, edges, districts, nodes, and landmarks. Based on these elements, Lynch thought that we could characterize our living space within the appearance of a city to imagine ourselves living and working there as shown on the left side in Fig. 2. In another remarkable work describing a way to extract geographic characteristics, Tezuka et al. [23, 24] extracted geographic objects and their roles frequently mentioned on the Web contents, which can be regarded as a mirror of the crowds' minds to the real world as shown in the middle of Fig. 2.

These two different types of urban characteristics may be useful to derive the image of the city. However, in a sense, they are focusing only on static elements of the city and did not take into account the most dynamic and important element of the city; that is, "crowds" living there would be a critical factor for observing urban characteristics. For instance, in the case of the recent terrible disaster, the earthquake and tsunami in Japan on March 11, 2011, although lots of static characteristics in some part of the devastated Fukushima prefecture area remain unchanged, the image of the city based on crowds living there was totally changed because the sequel nuclear accidents

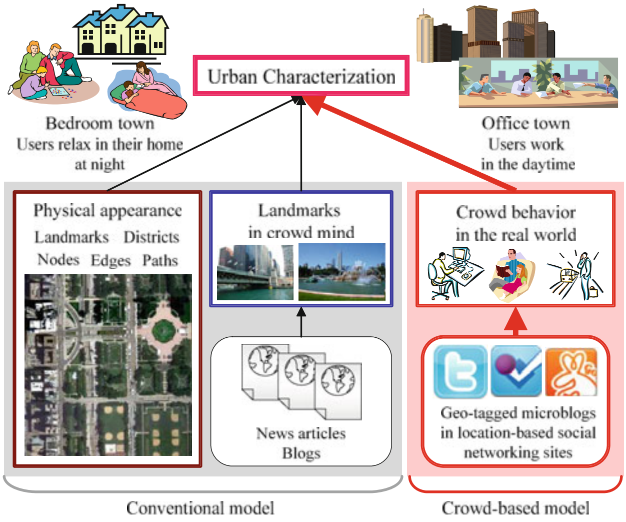

Fig. 2 Approaches of urban characterization

caused by the disasters prevented people from accessing the towns for a long time.

In this paper, we distinguishably attempt to derive a new kind of social geographic characteristics based on crowds, especially in terms of their behavior, by utilizing the location-based social network sites as shown on the right side of Fig. 2. In fact, geo-tagged microblogs over social networking sites can be easily collected since these are shared in the open spaces. We can utilize the free data to extract the image of cities based on crowd activity in the real world.

3.2 Modeling Twitter as a crowd lifelog source

Sharing personal lifelogs over social networks such as Facebook [25] and Twitter [3] is a common phenomenon worldwide spreading deeply into our daily lives. Behind the scene of those buzzing trends, we can find some crucial clues to understand why the trivial logs can be very useful to monitor the overall crowd lifelogs. First of all, we can observe three fundamental elements of crowd activity monitoring, that is, user, location and time. These spatio-temporal logs by enormous crowds are possibly appearing from any place where users can write their lifelogs with the help of automatic location-sensing functionality of recent smartphones. The implication of this kind of crowd behavior means various facts from simply a possibility of some geo-social event occurrences in a region to a trajectory of a tourist's travel by examining a history of time-variant footprints.

From personal to society level, we can even extract much broader and invisible crowd activity patterns. Of course, the surficial existence of crowds and their moving histories would be the foremost important features we can derive from the massive number of crowd lifelogs. In addition, if we can find much detailed about crowd actual behavior, crowd lifelogs over the social networking sites can be used for various investigation of social trend or pseudo census to people in a city or a nation. We may ask crowd opinion on a variety of social topics simply by referring to crowd messages focusing on some specific words such as the name of political parties or topical keywords. As for the extended crowd activities, we can utilize the written texts by analyzing what kinds of contents are actually written inside. In case of Twitter, though the message field is only bound in term of the length up to the 140 characters, it can have various types of contexts: (1) a hyperlink to external media such as some photo links through "http://twitpic.com" would represent that the correspondent user is probably taking a picture at a place, (2) #hashtag is a promise often used to mean the written textual message is related to a certain topic indexed by the tag, and (3) user networks are extractable, if there are patterns such as @user_id or RT (re-tweet) terms; the first pattern is used for sending a message to specified user, while the second pattern is to represent that the current written message is sourced from other user. Therefore, combining these features with the location and time information can give much more clear picture of what kinds of crowd activities are happening in a geographic area of interest. However, for the direct and straightforward approach, we attempt to model crowd behavior based on three basic metadata of geo-tagged lifelogs over most location-based social networks; time stamp, location information, and user ID, without analyzing textual messages.

4 Extracting urban area characteristics based on crowd behavior over Twitter

In order to achieve our goal of crowd-sourced urban characterization, we describe the further details of methodology about how we can collect crowd's geo-tagged micro lifelogs from Twitter in Sect. 4.1 as shown in Fig. 3a, how we can partition a region by setting socio-geographic boundaries in Sect. 4.2 as shown in Fig. 3b, how we can model crowd behavioral features based on crowd's lifelogs in Sect. 4.3 as shown in Fig. 3c, and how we can extract characteristics of urban areas grouped by crowd behavioral patterns in Sect. 4.4 as shown in Fig. 3d.

4.1 Location-based social network as a source for crowd activity monitoring

First of all, we need to gather geo-tagged tweets from Twitter to observe crowd activities in the real world as depicted in Fig. 3a. However, it takes a considerable amount of efforts to acquire a significant number of geo-tagged tweets because of certain practical limitations: In fact, Twitter presents open API [26] that solely supports the simplest near-by search based on a specified center location and a radius. Furthermore, each query can only obtain up to 1,500 tweets for past one week. Therefore, in order to overcome these restrictions and perform periodic monitoring of any size of user-specified regions, we developed a geographic microblog monitoring system [18, 19] that can monitor crowd behavior for a specific region of any size depending on the density of massive geographic microblogs as shown in Fig. 4a. Figure 4b shows the quad-tree-based geographic distribution of geo-tagged tweets from crowds in a part of area surrounding Osaka prefecture in Japan. The location information of geo-tagged tweets can be received either in a raw text form or in very precise location coordinates. Hence, in the case of the former

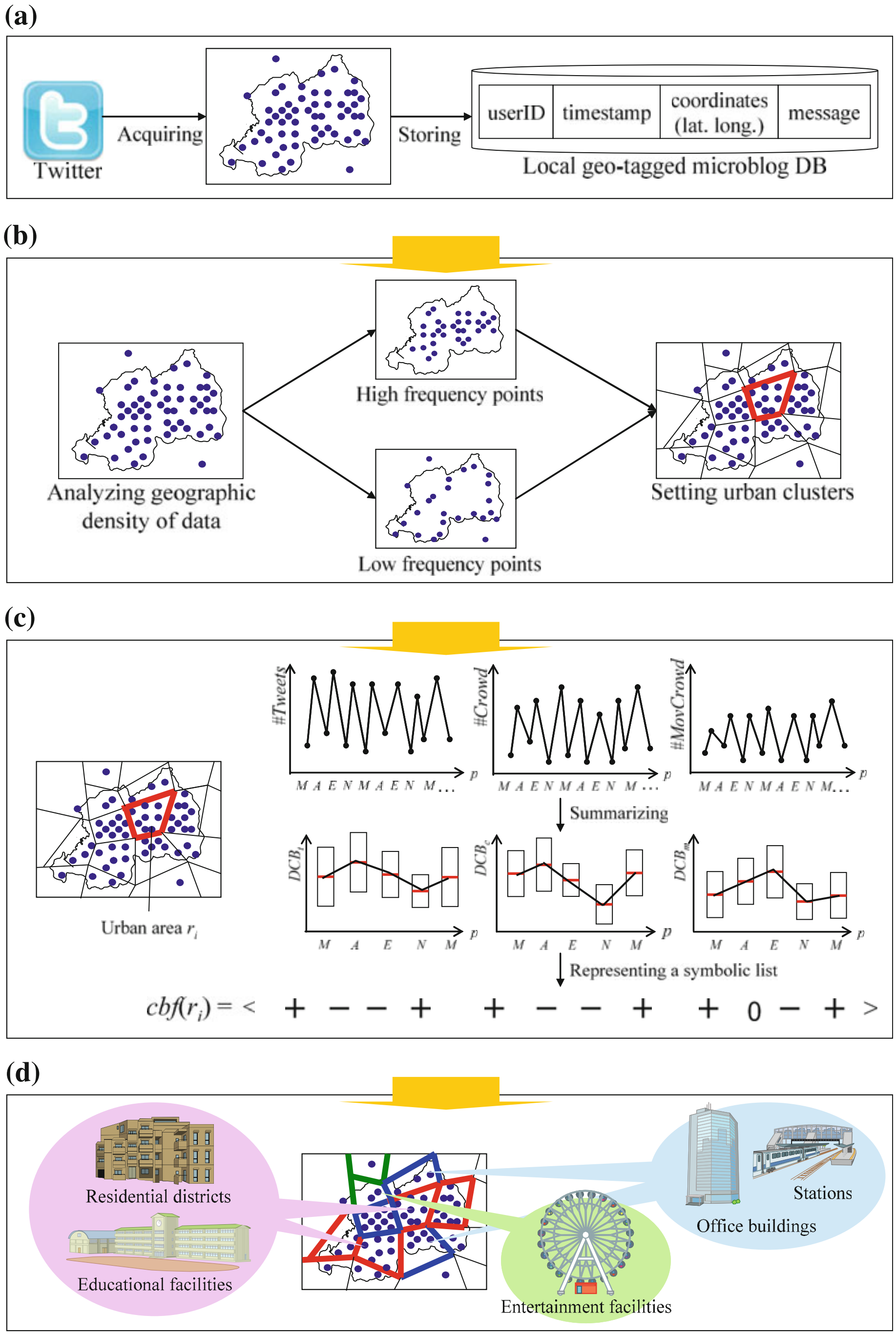

Fig. 3 Overview of crowd-based urban characterization. **a** Crawling geo-tagged microblogs from social networking sites and archiving them to the database. **b** Establishing socio-geographic boundaries. **c** Modeling crowd behavioral features in urban areas. **d** Characterizing urban areas by exploiting significant crowd behavioral patterns

textual style, we needed to perform geo-coding to identify the exact coordinates by translating place names into the corresponding exact locations. We were able to solve the problem easily by using another mash-up service with Google Map's geo-coding service [27]. We directly transferred the place names to this conversion service and received the precise coordinates. Subsequently, we could accurately determine when and where each tweet was written.

4.2 Socio-geographic boundaries of crowd activities

Next, in order to monitoring crowd behavior in urban areas for extracting urban characteristics, we need to define urban areas by partitioning a given region into sub-areas. As for how to partition the region, there are several different space partitioning, for instance, administrative area, grid, and cluster as shown in Fig. 5. Hence, we should select the most appropriate method for our goal,

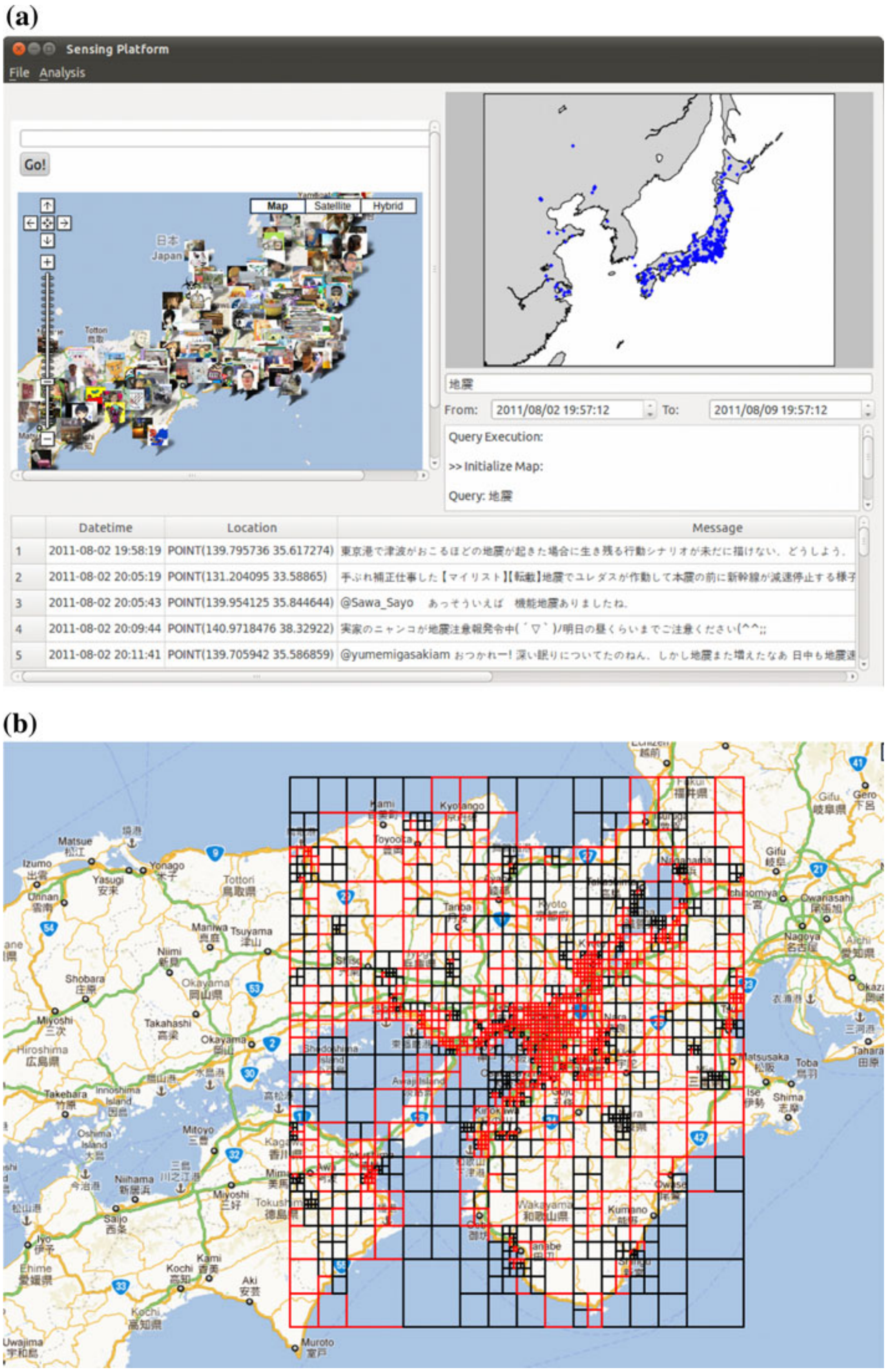

Fig. 4 Geo-tagged lifelogs acquired by the geographic microblog monitoring system. **a** Geographic microblog monitoring system. **b** Quad-tree-based geographic distribution of tweets around Osaka, Japan

地震

From: 2011/08/02 19:57:12 To: 2011/08/09 19:57:12

Query Execution:

>> Initialize Map:

Query: 地震

	Datetime	Location	Message
1	2011-08-02 19:58:19	POINT(139.795736 35.617274)	東京港で津波がおこるほどの地震が起きた場合に生き残る行動シナリオが未だに描けない．どうしよう．
2	2011-08-02 20:05:19	POINT(131.204095 33.58865)	手ぶれ補正仕事した【マイリスト】【転載】地震でユレダスが作動して本震の前に新幹線が減速停止する様子
3	2011-08-02 20:05:43	POINT(139.954125 35.844644)	@Sawa_Sayo　あっそういえば　機能地震ありましたね。
4	2011-08-02 20:09:44	POINT(140.9718476 38.32922)	実家のニャンコが地震注意報発令中(´▽`)/明日の昼くらいまでご注意ください(^^;;
5	2011-08-02 20:11:41	POINT(139.705942 35.586859)	@yumemigasakiam おつかれー! 深い眠りについてたのねん．しかし地震また増えたなあ 日中も地震速

characterizing urban areas based on crowd behavior there. Administrative areas are formed by splitting a target region into prefectures and municipalities based on administrative boundaries: limits or borders of a geographic area under the jurisdiction of some governmental or managerial entity as shown in Fig. 5a. Therefore, it is difficult to figure out whether crowd behavior areas are relevant or almost dependent on administrative areas, because crowd behavior often easily cross over the administrative boundaries. Accordingly, we consider that this method might be inappropriate for examining crowd behavior. Next, as for the grid, it is difficult to decide the adequate cell size because a grid is formed by a lot of equal-sized cells as shown in Fig. 5b. In addition, it would consume considerable unnecessary costs for observing crowd behavior since it does not consider non-uniform distribution of crowds. On the other hand, in case of the clustering, it can reflect the geographic distribution of crowds based on location

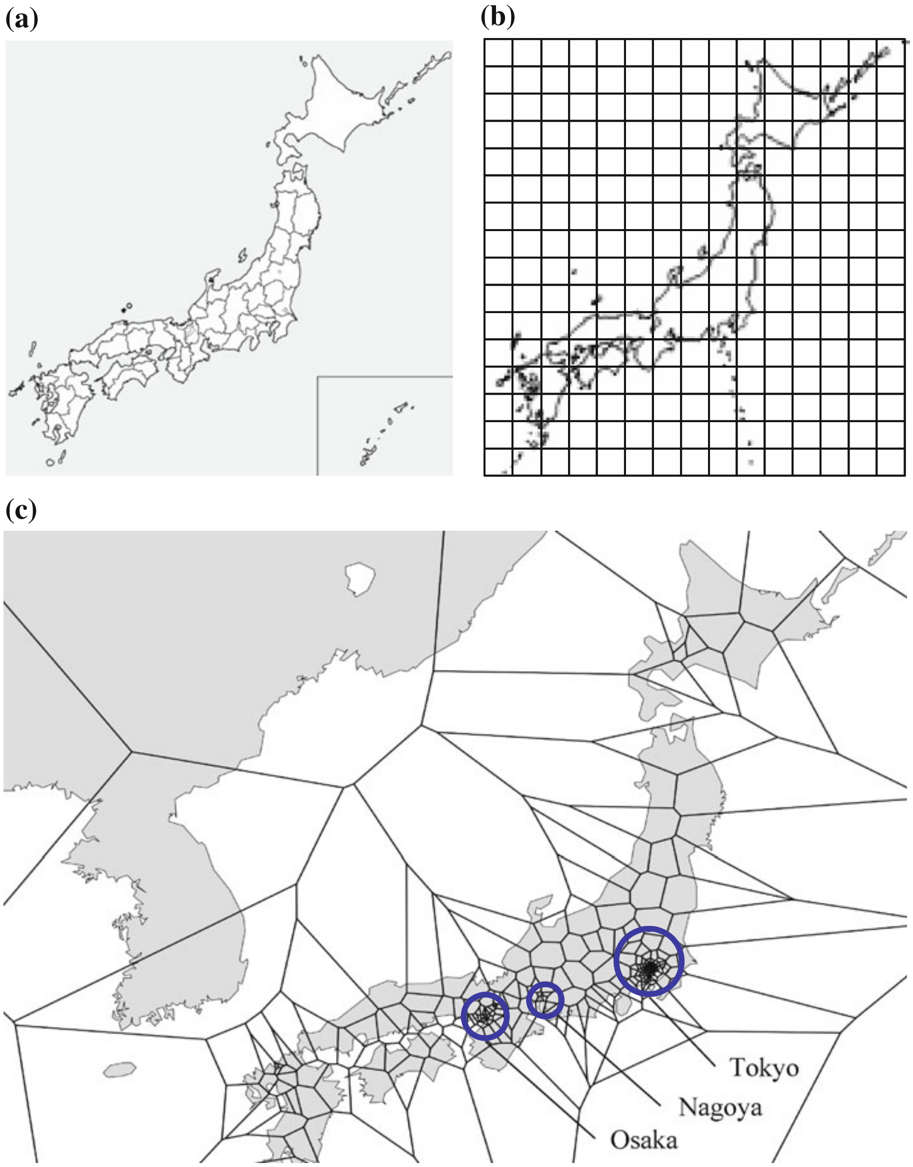

Fig. 5 Methodologies of space partitioning to monitor crowd behavior. **a** Administrative area. **b** Grid. **c** Cluster

information of geo-tagged tweets and deal with heterogeneous regions differently. As a result of this approach, small urban areas of densely populated areas appeared around major cities such as Tokyo, Nagoya, and Osaka in Japan as shown in Fig. 5c. In contrast, large urban areas of sparsely populated areas are spread over the other suburban areas or surrounding sea. Thus, we can effectively establish the appropriate socio-geographic boundaries for the target region and partition into urban areas by referring to the actual geographic crowd behavior.

On the basis of these reasons, in this work, we selected the cluster-based space-partitioning method, which can take into consideration geographic distribution of crowds. Especially, in our experiment described in Sect. 5, since we deal with millions of locational data of crowds obtained from Twitter as shown in Fig. 6a, it requires enormous computational effort. Therefore, we have to reduce the data size in much smaller and computable size without lack of essential quality of the original data. For this, we adopted the NNClean algorithm [28, 29] to split the data into two groups of high-frequency and low-frequency parts as shown in Fig. 6b and c. In many cases, the algorithm is used for distinguishing noise from a given data—low-frequency part. However, in the case of crowd-sourced data over social networks, high-frequency points are naturally observed around high-populated areas. Therefore, using only high-frequency points works out to ignore the suburban areas. In order to solve this problem, we also utilize the low-frequency points. Consequently, we generate clusters from these two different sources respectively by applying EM algorithm [30] as shown in Fig. 6d and e. After that, we depicted a Voronoi diagram [31, 32] using

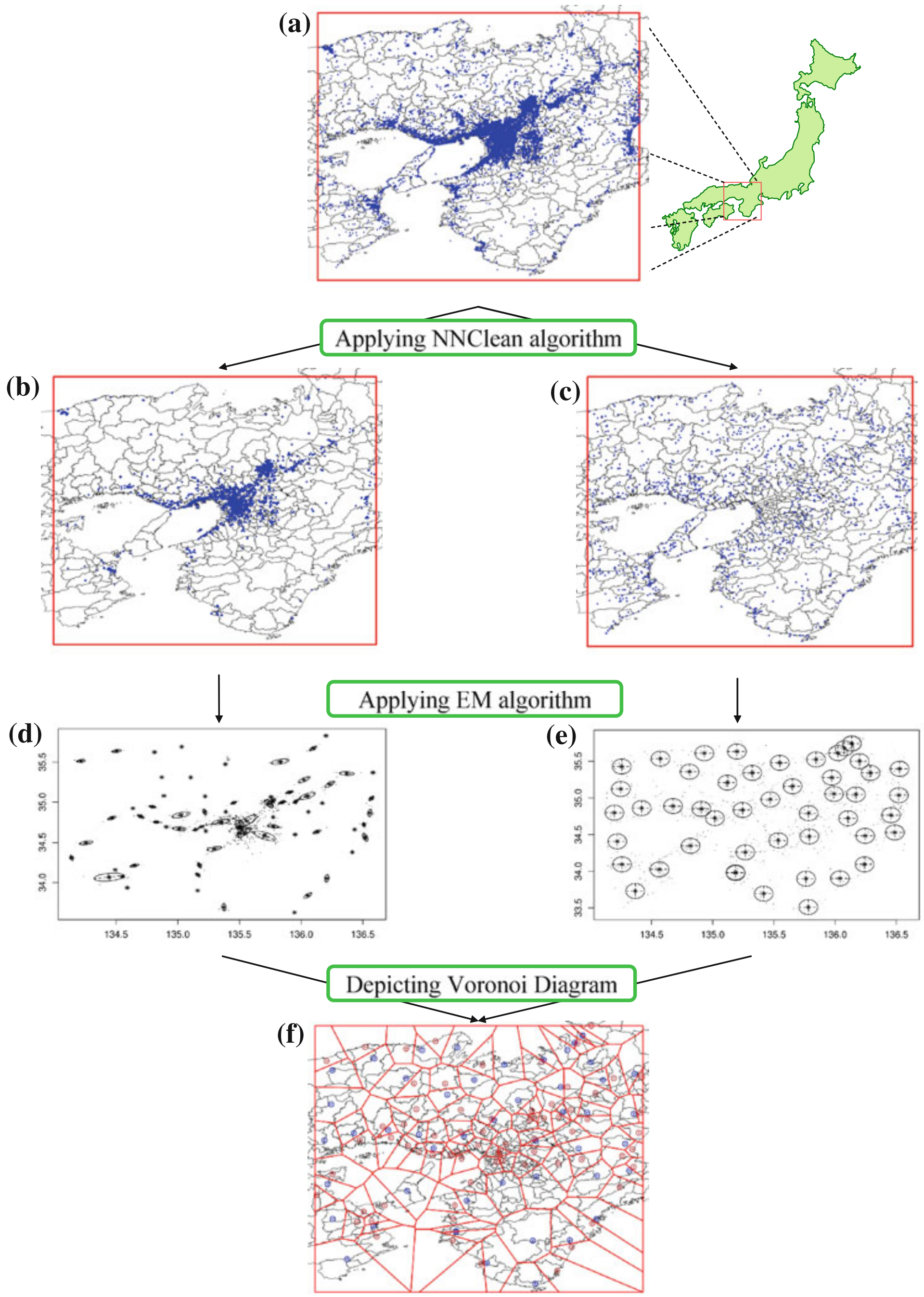

Fig. 6 Process of constructing socio-geographic boundaries based on EM algorithm and Voronoi diagram. **a** Original data distribution. **b** High frequency points. **c** Low frequency points. **d** High frequency clusters. **e** Low frequency clusters. **f** Socio-geographic boundaries

the center points (latitude, longitude) of all clusters and defined the formed polygonal regions as urban clusters as shown in Fig. 6f.

4.3 Extracting crowd behavioral features

In order to grasp crowd lifestyles in urban areas, we monitor crowd behavior through their lifelogs over social networks. Specifically, we compute crowd behavioral feature based on parameters computable using primitive metadata of geo-tagged tweets: user ID, timestamp, and location information. In addition, since crowd behavior would vary depending on certain time slots of a day, we should periodically monitor crowd behavior by splitting a day into certain time period p_j. In this paper, we empirically set 6-h by splitting a day into four equal time slots: morning (M, 06:00–12:00), afternoon (A, 12:00–18:00), evening (E, 18:00–24:00), and night (N, 24:00–06:00), and model crowd behavior in terms of three parameters defined as follows:

$\#Tweets|_{r_i,p_j}$: The total number of tweets occurring inside an urban area r_i in a time period p_j.

$\#Crowd|_{r_i,p_j}$: The number of distinct users observable in an urban area r_i during a specific period of time p_j. Naturally, the in-equality $\#Crowd \leq \#Tweets$ is valid since any individual user can write one or more tweets.

$\#MovCrowd|_{r_i,p_j}$: The number of mobile users in an urban area r_i in a time period p_j. To be more precise, this is the number of users posting two or more tweets at different locations in the area. Obviously, the in-equality $\#MovCrowd \leq \#Crowd$ is valid. This parameter is the most dynamic one, which can explicitly reflect the temporal usage of real space by crowds.

Crowd behavior feature for an urban area r_i is represented based on these three features: *#Tweets*, *#Crowd*, *#MovCrowd*. Although the scale of crowd behavioral feature would be different depending on crowd in each urban cluster, while periodic occurrence of geo-tagged tweets or the absolute number of crowds are different in each cluster, two or more regions can be similar in terms of an increasing and decreasing tendency, for example, there can be increasingly crowded places such as a large railway station in the morning and the evening. Therefore, in order to explore such significant patterns, we represent crowd behavioral features based on relative temporal changes of the degrees of crowd behavior as shown in Fig. 7a, where DCB_x refers to a sequence of temporal change between the time slots about the parameter $x \in \{\#Tweets, \#Crowd, \#MovCrowd\}$. Specifically, we compute the differences between the degrees of crowd behavior at two consecutive time slots; $t1$, $t2$, $t3$, and $t4$ from $DCB_t(r_i)$, $c1$, $c2$, $c3$, and $c4$ from $DCB_c(r_i)$, and $m1$, $m2$, $m3$, and $m4$ from $DCB_m(r_i)$ in Fig. 7b. Here, the four suffixes, 1, 2, 3, and 4, represent the differences from morning to afternoon (*M–A*), from afternoon to evening (*A–E*), from evening to night (*E–N*), and from night to morning (*N–M*). For the simplification, in this paper, we kept crowd behavioral features by only looking up the differences by transforming the representation to a symbol list such as $\langle +, -, 0, + \rangle$, which respectively means the increase from *M* to *A*, the decrease from *A* to *E*, no change from *E* to *N*, and the increase from *N* to *M*.

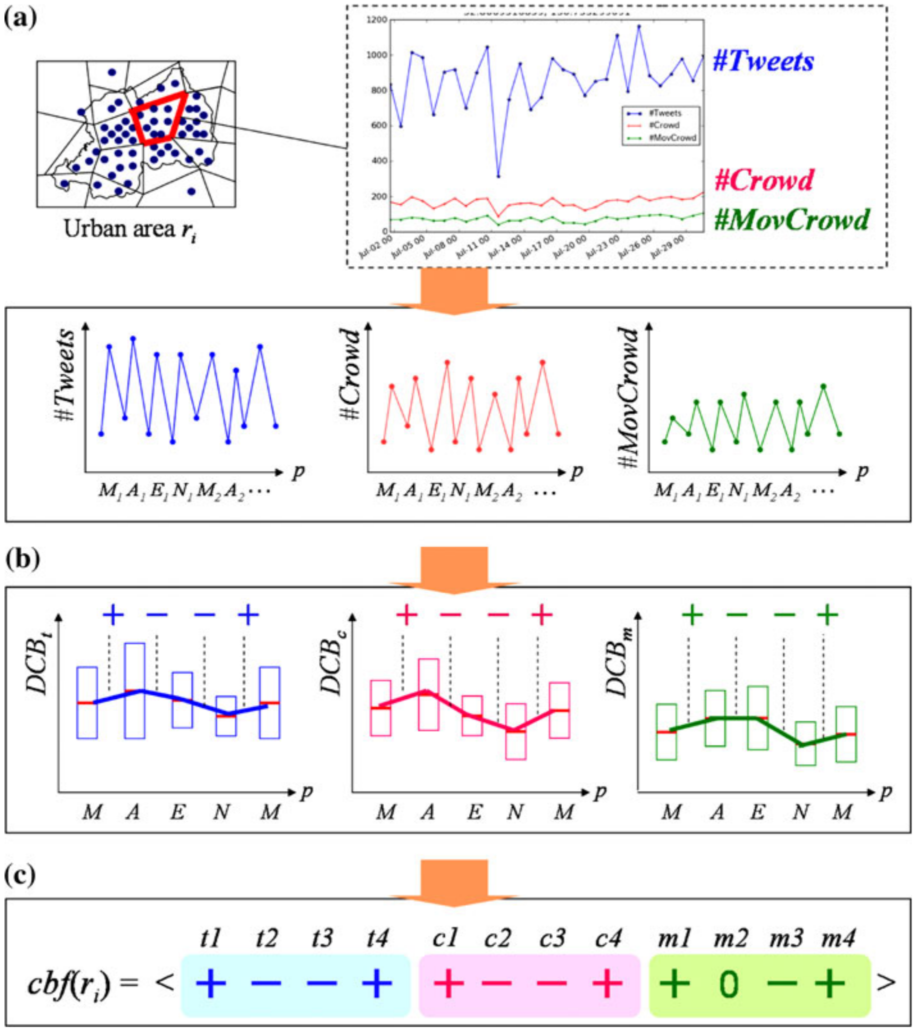

Fig. 7 Process of extracting crowd behavioral features. **a** Aggregating crowd monitored data in an urban area. **b** Extracting crowd behavior feature in the urban area. **c** Representing crowd behavioral feature with symbols

Indeed, the temporal changes can signify the dynamicity of crowd behavior so that we can obtain many possible combinations such as $\langle +, -, -, + \rangle$, $\langle +, 0, -, + \rangle$, etc. Finally, based on these symbolic patterns, we define crowd behavior feature, $cbf(r_i) = \langle DCB_t(r_i),\ DCB_c(r_i),\ DCB_m(r_i) \rangle$ as illustrated in Fig. 7c. This representation will be used as a summary of crowd behavior in a region r_i.

4.4 Exploiting crowd behavioral patterns

In the above, we described the method to extract crowd behavioral features in urban areas based on crowd-sourced data. Then, we need to find out significant crowd behavioral patterns monitored in multiple clusters by mining crowd behavioral features. Since each crowd behavioral feature consists of a series of the temporal changes of the degrees of crowd behavior *cbf* represented by 12-dimensional symbols as depicted in Fig. 7c, there will be numerous combinations up to the maximum 3^{12}; hence, the computational cost to examine the common full-size or partial characteristic patterns would be unbearable. In order to simplify the problem, we adopted a frequent itemset mining algorithm [33] for statistically looking for common frequent patterns from the item sets having a huge size of combinations. Ultimately, we can extract the common characteristic crowd behavioral patterns that can characterize many urban areas in common. In fact, each feature is described by a unique representation, such as $cbf(r_i) = \langle t1+,\ t2-,\ t3-,\ t4+,\ c1+,\ c2-,\ c3-,\ c4+,\ m1+,\ m20,\ m3-,\ m4+ \rangle$ which comes from the original form $cbf(r_i) = \langle +, -, -, +, +, -, -, +, +, 0, -, + \rangle$, and $t1+$ is a symbol consisting of three factors ('*t*', '1', '+'); here, we explain the extended symbols. '*t*' means that it is a parameter of *#Tweets*. Next, '1' is an index to distinguish from the others. Finally, '+' is a symbolic representation of the relative change of degrees of crowd behavior during two periods. On behalf of this translation, each symbol can be seen as a unique item, which can be utilized in the frequent itemset mining. Thus, based on this method, we can extract the common partial patterns effectively.

5 Experiment

In this section, we describe our experiment to extract significant patterns of crowd behavior that was described in Sect. 3. We were able to gather a great deal of geo-tagged tweets from Twitter in the geographic range of Japan. We computed the aforementioned crowd behavioral features based on three parameters using metadata of geo-tagged tweets and extracted urban characteristics by examining their common changing patterns. Lastly, we successfully confirmed that our experimentation to utilize the social

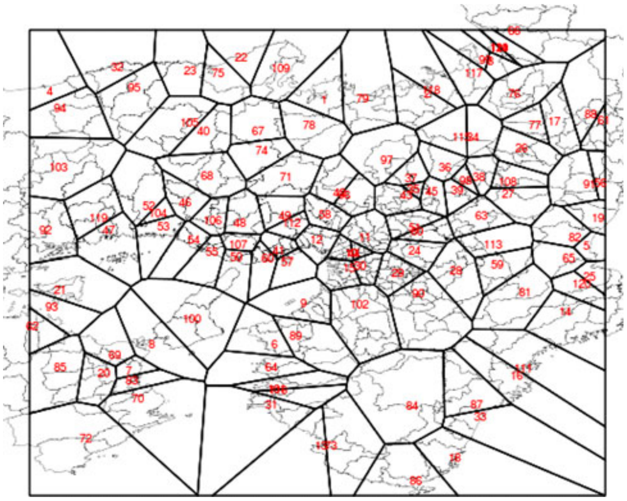

Fig. 8 Urban areas partitioned by socio-geographic boundaries

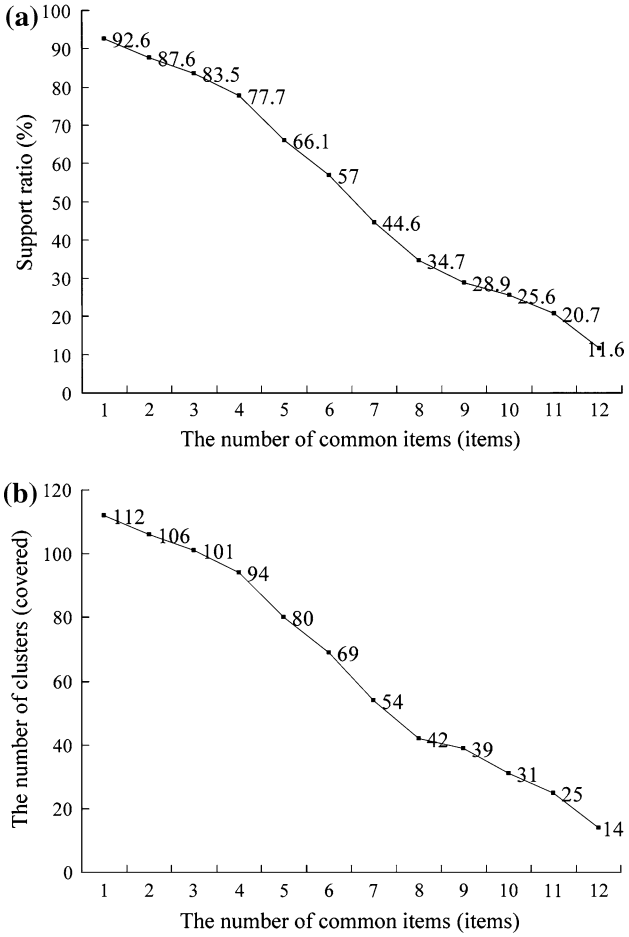

Fig. 9 Effects of common item control on support and covered clusters. **a** Relation between support ratio and the number of common items. **b** Relation between the number of clusters and the one of common items

network–based crowd behavior as an estimator for urban characterization was achieved by investigating categories of major facilities in each clustered region.

5.1 Experimental setting

First, we collected geo-tagged tweets from Twitter for one month between June 5, 2010 and July 5, 2010 around part of Japan with the latitude range [33.384555–35.839419] and the longitude range [134.126551–136.58890] using our geographic microblog monitoring system as shown in Fig. 4a. As a result, we could acquire 1,891,186 geo-tagged tweets from 39,898 distinct users as shown in Fig. 4b. Next, we constructed socio-geographic boundaries considering the density of crowd behavior and partitioned the target space into 121 urban clusters by the method described in Sect. 4 as shown in Fig. 8.

5.2 Exploring significant crowd behavioral patterns

Next, in each cluster, we computed crowd behavioral feature based on three parameters, for every 6-h time slot during the training period. Then, we extracted significant crowd behavioral patterns by analyzing the crowd behavior features for all clusters. In this experiment, we applied a frequent itemset mining algorithm [33] that can statistically mine common features to full size: 12 items. Specifically, we tried to extract significant features by examining partial patterns from the 12-dimensional crowd behavioral features. By applying the pattern extraction method described

Table 1 Crowd behavioral patterns extracted based on the frequent Itemset mining algorithm (12 items)

Pattern	DCB_t				DCB_c				DCB_m				Occurrence ratio (%)
	*t*1	*t*2	*t*3	*t*4	*c*1	*c*2	*c*3	*c*4	*m*1	*m*2	*m*3	*m*4	
Pattern 1	+	−	+	+	−	−	+	+	0	0	0	0	11.6
Pattern 2	−	−	+	+	−	−	+	+	0	0	0	0	7.4
Pattern 3	+	−	+	+	−	−	+	+	−	−	+	+	5.0
Pattern 4	+	−	+	+	+	−	+	+	0	0	0	0	3.3

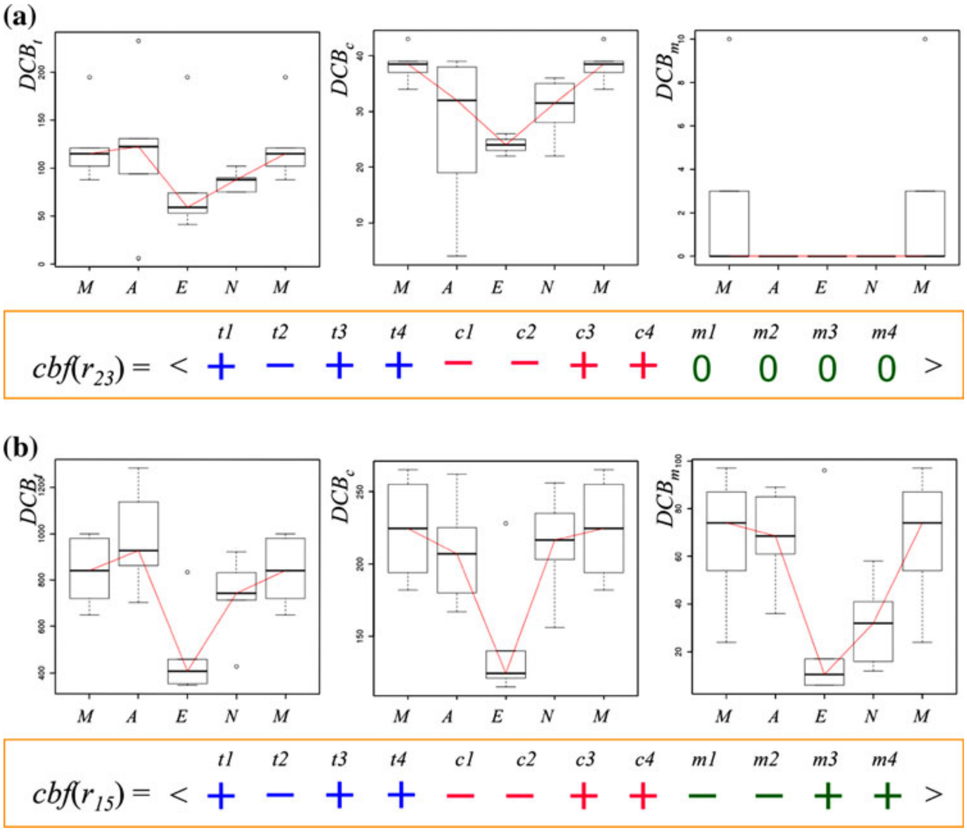

Fig. 10 Examples of crowd behavioral patterns. **a** A cluster of *pattern 1*. **b** A cluster of *pattern 3*

in Sect. 4.4, we were able to extract characteristic and frequent crowd behavioral patterns. Here, we can consider two important parameters for the pattern extraction. First, each pattern can have a support value that says how many clusters actually support the correspondent patterns. By specifying the support value, we can control the resulting number of patterns. Furthermore, we can consider the size of pattern, that is, how many number of each cluster should meet the patterns. For instance, for a pattern $\langle t1+, t2*, t3+, t4-, c1*, c2+, c3*, c4-, m10, m2+, m3-, m4-\rangle$, three parts of '$*$' symbol can include any cases of '+,' '−' or '0,' we also specified the preferable size of patterns from 12 down to only 1. Consequently, the effect of the support value and the common item size was examined as illustrated in Fig. 9a. Another important aspect is that depending on the size of common items, the number of clusters covered by the extracted patterns is decided as drawn in Fig. 9b. In other words, when we set the size of pattern shorter, many clusters can be included in the final results. Finally, with a setting with the preferable size of frequent itemset (=12) and minimum support (=3.0 %), we could obtain four different patterns as depicted in Table 1. Among these patterns, we examined two interesting patterns of *pattern* 1 and *pattern* 3 as shown in Fig. 10a and b, respectively.

5.3 Reasoning urban characteristics

In order to reason the extracted patterns and characterize urban clusters of the same patterns, we investigated what kinds of local facilities can be found in the clusters respectively. In order to obtain a baseline dataset, we

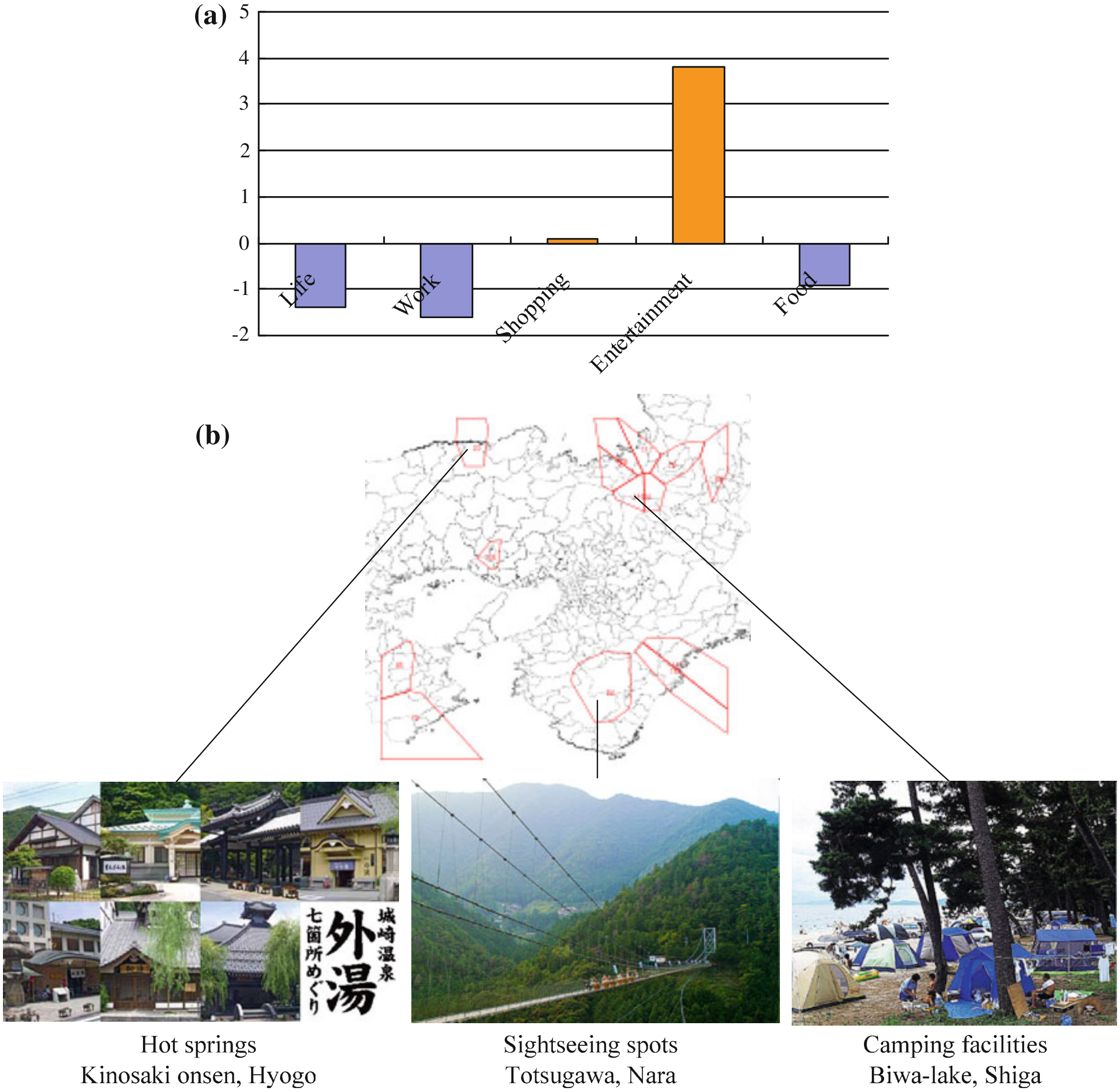

Fig. 11 Reasoning characteristic of urban clusters of *pattern* 1. **a** Facility genre-based reasoning of crowd behavioral *pattern 1*. **b** 14 Clusters of crowd behavioral *pattern 1*

prepared a database of local facilities by referring to geographic facilities information of the local search provided by Yahoo! Japan (Yahoo! Loco) [34] which answers surrounding facilities' information with precise location information in Japan. Most facilities have a hierarchy of three levels, respectively, for instance, 'Food' (1st), 'Chinese' (2nd), 'Beijing Food' (3rd) and can be eventually aggregated into four genres at the most upper level; 'Food,' 'Shopping,' 'Entertainment,' and 'Life.' However, there were some facilities that are not allocated yet to any category above such as factory and distribution center. We found that these facilities are mostly relevant to industrials; thus, we manually added a new genre 'Work' and subcategories by referring to Japan Standard Industry Classification [35].

Based on this local facilities database, we looked into the significant facilities of each region in terms of five genres; 'Life,' 'Work,' 'Shopping,' 'Entertainment,' and 'Food' as shown in Figs. 11a and 12a. Here, the Y-axis in the graphs of Figs. 11a and 12a represents a relative significance compared to the average over all clusters. In other words, if the value of each genre is under zero, it represents that the genre is weaker than the average. For example, the clusters grouped by *pattern* 1 are commonly characterized by the genre feature as shown in Fig. 11a. That is, the type of regions relatively have many entertainment facilities such as hot springs, camping parks, and sightseeing spots. In fact, we can find entertainment facilities in urban clusters of *pattern* 1 as shown in Fig. 11b. On the other hand, the clusters grouped by *pattern* 3 are characterized as

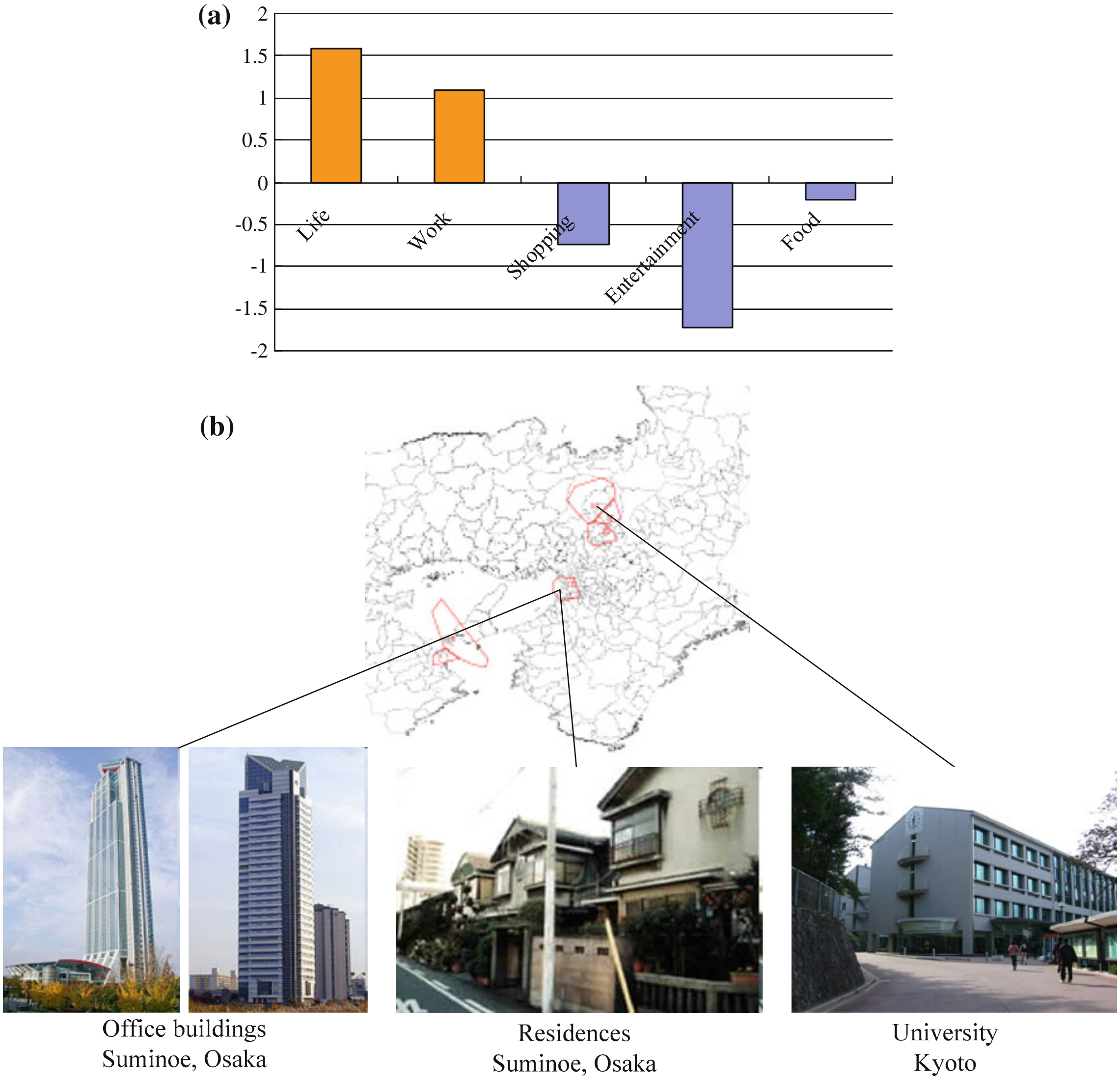

Fig. 12 Reasoning characteristic of urban clusters of *pattern* 3. **a** Facility genre-based reasoning of crowd behavioral *pattern 3*. **b** 6 Custers of crowd behavioral *pattern 3*

relatively many life and work facilities such as residences, educational facilities, and office buildings as illustrated in Fig. 12a. We can actually confirm residential towns, universities, and high-rise buildings for business in urban clusters of *pattern* 3 as shown in Fig. 12b. In fact, each cluster includes a variety of facilities eventually showing mixed crowd behavior. In other words, it is hard to say that a town only has a specific characteristic such as only shopping town. In our work, instead, we were able to find out the distinctive and significant configurations of local facilities in the regions clustered by crowd behavioral patterns to understand the characteristics of urban regions.

6 Conclusion

In this paper, we proposed a crowd-based urban characterization method based on crowd-sourced data obtained from social networking sites. In our proposed method, we collected geo-tagged tweets from Twitter as a source of monitoring crowd behavior in the real world and modeled crowd behavior in terms of primitive parameters extracted from the tweets. Then, we extracted crowd behavioral features for urban areas and investigated urban characteristics by exploiting common behavioral patterns. Furthermore, on the basis of our proposed method, we conducted the experiment using massive geo-tagged tweets and clustered urban areas based on crowd behavioral patterns, and reasoned each area utilizing the genre information of local facilities.

In the future work, we will extend our work by exploiting much deeper crowd's minds and complicated crowd behavior by improving our approach with the analysis of textual messages of tweets to strengthen the validity of our proposed method. Furthermore, we will investigate further crowd behaviors affected by various social and natural events.

Acknowledgments This research was supported in part by the 7th Microsoft Research IJARC Core Project.

References

1. Mann S. wearcam.org: http://wearcam.org/eastcampusfire.htm
2. Bell G, Gemmell J (2009) Total recall: how the e-memory revolution will change everything. Penguin Group, USA
3. Twitter: http://twitter.com/
4. Foursquare: http://foursquare.com/
5. Gowalla: http://gowalla.com/
6. Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in rome. IEEE Trans Intell Transp Syst 12(1):141–151
7. Jiang B, Claramunt C, Batty M (2000) An integration of space syntax into GIS for modelling urban spaces. Int J Appl Earth Obs Geoinf 2(3–4):161–171
8. Social Graph: Concepts and Issues: http://www.readwriteweb.com/archives/social_graph_concepts_and_issues.php
9. Cudré-Mauroux P, Elnikety S (2011) Graph data management systems for new application domains. PVLDB 4(12):1510–1511
10. Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10, pp 759–768
11. Sysomos: Twitter Statistics for 2010: an in-depth report at Twitter's Growth 2010, compared with 2009: http://www.sysomos.com/insidetwitter/twitter-stats-2010/
12. Barkhuus L, Brown B, Bell M, Sherwood S, Hall M, Chalmers M (2008) From awareness to repartee: sharing location within social groups. In: Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems, CHI '08, pp 497–506
13. Hightower J (2003) From position to place. In: Proceedings of the 2003 workshop on location-aware computing, part of the 2003 ubiquitous computing conference
14. Zheng VW, Zheng Y, Xie X, Yang Q (2010) Collaborative location and activity recommendations with GPS history data. In: Proceedings of the 19th international conference on World wide web, WWW '10, pp 1029–1038
15. De Longueville B, Smith RS, Luraschi G (2009) "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: Proceedings of the 2009 i nternational workshop on location based social networks, LBSN '09, pp 73–80
16. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, WWW '10, pp 851–860
17. Cheng Z, Caverlee J, Lee K, Sui DZ (2011) Exploring millions of footprints in location sharing services. In: Proceedings of fifth international AAAI conference on weblogs and social media, ICWSM '11, pp 81–88
18. Lee R, Sumiya K (2010) Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL international workshop on location bsased social networks, LBSN '10, pp 1–10
19. Lee R, Wakamiya S, Sumiya K (2011) Discovery of unusual regional social activities using geo-tagged microblogs. World Wide Web (WWW) Special Issue on Mobile Services on the Web 14(4):321–349
20. Wakamiya S, Lee R, Sumiya K (2011) Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs. In: Proceedings of the 2nd international workshop on social networks and social media mining on the web, SNSMW'11, pp 390–401
21. Wakamiya S, Lee R, Sumiya K (2011) Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating. In: Proceedings of the 5th international conference on ubiquitous information management and communication, ICUIMC '11, pp 39:1–39:10
22. Lynch K (1960) The image of the city. The MIT Press, Cambridge
23. Tezuka T, Lee R, Takakura H, Kambayashi Y (2001) Models for conceptual geographical prepositions based on web resources. J Geogr Inf Decis Anal 5(2):83–94
24. Tezuka T, Lee R, Takakura H, Kambayashi Y (2003) Cognitive characterization of geographic objects based on spatial descriptions in web resources. In: Proceedings of the workshop on spatial data and geographic information systems (SpaDaGIS)
25. Facebook: http://www.facebook.com/
26. Twitter Open API: http://apiwiki.twitter.com/Twitter-Search-API-Method%3A-search

27. Google Geocoding API: http://code.google.com/intl/ja/apis/maps/documentation/geocoding/
28. Baddeley A (2010) Analysing spatial point patterns in R. CSIRO
29. Byers S, Raftery A (1998) Nearest-neighbour clutter removal for estimating features in spatial point processes. J Am Stat Assoc 93:577–584
30. Dempster AP, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser R 39(1):1–38
31. Berg MD, Cheong O, Kreveld MV, Overmars M (2008) Computational geometry: algorithms and applications, third edition. Springer, Berlin
32. Lloyd SP (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137
33. Yu JX, Li Z, Liu G (2008) A data mining proxy approach for efficient frequent itemset mining. VLDB J 17:947–970
34. Yahoo! Local Search Web API: http://developer.yahoo.co.jp/webapi/map/localsearch/v1/localsearch.html
35. Japan Standard Industry Classification: http://www.stat.go.jp/english/index/seido/sangyo/san07-2.htm