A New Density-based Spatial Clustering Algorithm for Extracting Attractive Local Regions in Georeferenced Documents

Tatsuhiko Sakai, Keiichi Tamura, Member, IAENG, and Hajime Kitakami

Abstract—Nowadays, with the increasing attention being paid to social media, a huge number of georeferenced documents, which include location information, are posted on social media sites via the Internet. People have been transmitting and collecting information through these georeferenced documents. Georeferenced documents are usually related to not only personal topics but also local topics and events. Therefore, extracting “attractive” local regions associated with local topics from georeferenced documents is one of the most important challenges in different application domains. In this paper, a novel spatial clustering algorithm, called the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm, for extracting “attractive” local regions in georeferenced documents is proposed. We define a new type of spatial cluster called an \((\epsilon, \sigma)\)-density-based spatial cluster. The proposed clustering algorithm can recognize not only semantically-separated but also spatially-separated spatial clusters. To evaluate our proposed clustering algorithm, geo-tagged tweets posted on the Twitter site are used. The experimental results show that the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm can extract “attractive” local regions as \((\epsilon, \sigma)\)-density-based spatial clusters.

Index Terms—density-based clustering, spatial cluster, DBSCAN, social media, local topic extraction.

I. INTRODUCTION

In recent years, with widespread use of smart phones equipped with a GPS, as well as the increasing interest in social media, a huge number of georeferenced documents, which include location information, are posted on social media sites through the Internet. People have been transmitting and collecting information related to location through georeferenced documents [1], [2]. Georeferenced documents are usually closely related not only to personal topics but also to local topics and events. Therefore, extracting local topics and events from georeferenced documents [3] contribute to different geo-location application domains such as, local area marketing, tourism informatics, and local topic recommendation.

Researchers, who are interested in knowledge discovery on georeferenced documents posted on social media sites, have made a great effort to tackle the new challenges that extract local topics and events from georeferenced documents. Dense regions, in which many georeferenced documents including a keyword are posted, are the hot areas of local topics related to the keyword. For example, Crandall et al. [4] developed an algorithm for identifying hot sites and landmarks from geotagged photos posted on the Flickr site, one of the most famous photo-sharing sites. Sakaki et al. [5] focused on tweets posted on the Twitter site about typhoons and earthquakes to estimate a typhoon’s trajectory and an earthquake’s epicenter using dense regions.

We have been developing a new spatial clustering algorithm, which extracts “attractive” local regions that are dense regions in which many georeferenced relevant documents including some keywords relevant to local topics are posted. To extract “attractive” local regions, we define a new type of spatial cluster called a \((\epsilon, \sigma)\)-density-based spatial cluster. An \((\epsilon, \sigma)\)-density-based spatial cluster is not only spatially-separated but also semantically-separated from other spatial clusters. Thus, \((\epsilon, \sigma)\)-density-based spatial clusters are closely related to local topics and events.

The main contributions of this study are as follows:

- To extract \((\epsilon, \sigma)\)-density-based spatial clusters, we propose a new spatial clustering algorithm for georeferenced documents, called the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm, which is a natural extension of DBSCAN [6]. DBSCAN is a basic density-based spatial clustering algorithm and is based on neighborhood density and recognizes an area those density is higher than that of the other areas. However, it does not take account of similarities between the contents of georeferenced documents. The \((\epsilon, \sigma)\)-density-based spatial clustering algorithm can recognize \((\epsilon, \sigma)\)-density-based spatial clusters, which are both semantically-separated and spatially-separated from other spatial clusters.
- To recognize semantically/spatially-separated clusters as \((\epsilon, \sigma)\)-density-based spatial clusters, we define a new similarity measurement for georeferenced documents on social media sites. In social media sites, people usually post georeferenced documents that are short messages including a local topic and event. Therefore, if georeferenced documents include a same keyword, which are similar each other, the georeferenced documents are similar each other. On the basis of this concept, we define the new similarity measurement based on keyword-based Simpson’s coefficient.
- To evaluate the proposed spatial clustering algorithm, we performed evaluations using an actual data set consisting of 480,000 tweets from the Twitter site, which were posted from November 2011 to February 2012. We confirmed that the proposed spatial clustering algorithm can extract \((\epsilon, \sigma)\)-density-based spatial clusters that represent “attractive” local regions associated with...
local topics.

The remainder of this paper is organized as follows. In Section 2, related work is reviewed. In Section 3, the \((\epsilon, \sigma)\)-density-based spatial cluster is defined. In Section 4, the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm is described. In Section 5, the results of an evaluation using tweets posted on Twitter are presented. Finally, some concluding remarks are given in Section 6.

II. RELATED WORK

The popularization of smart phones equipped with a GPS has opened up entirely-new types of data on social media sites. That is georeferenced data, which includes its posted location (e.g., geo tag, address, and landmark name) as well as its posted time. People on social media sites are referred to as sensors that observe real world happening around them. In other words, considering people on social media sites as sensors, georeferenced data is like sensor data that observes topics and events in the real world [7].

Since the use of the Internet has become widespread, topic detection and tracking in documents on the Internet [8] has been one of the most attractive research topics in many kinds of application domain. Above all, in social media era, we face new types of documents, called georeferenced documents which are a kind of georeferenced data and include location information. For example, on the Twitter site, which is a micro-blogging service site, geo-tagged tweets are georeferenced documents.

The most significant impact on many studies related to our work is DBSCAN, a density-based spatial clustering algorithm [6], [9]. The shapes of spatial clusters in geo-spatial data usually vary in form. Even some spatial clusters are completely surrounded by (but not connected to) a different cluster. To extract arbitrarily shaped clusters, density-based spatial clustering algorithms focuses on high dense regions in data space, separated by regions of a lower density. DBSCAN and subsequent studies were applied to studies on extracting specific areas related to local topics and events from geo-spatial data.

Tamura et al. [10] proposed a novel density-based spatiotemporal clustering algorithm, which can extract spatially and temporally-separated clusters in georeferenced documents. Their proposed algorithm integrates spatiotemporal criteria into DBSCAN to separate spatial clusters temporally. Kisilevich et al. [11] also proposed P-DBSCAN, a new density-based spatial clustering algorithm based on DBSCAN, for analysis of attractive places and events using a collection of geo-tagged photos. They defined a new density according to the number of people in the neighborhood. Our work is close to these studies. However, P-DBSCAN and the density-based spatiotemporal clustering algorithm cannot recognize semantically-separated spatial clusters.

There are some studies on clustering techniques for extracting topics and events, which focused on geo-tagged tweets posted on the Twitter site and image-data posted on the Flickr site. Watanabe et al. [12] identified locations that are currently attracting attention. Lee et al. [13] developed a method of detecting local events using spatial partitions. They separate the entire area into sub-areas using a Voronoi diagram. Their method recognizes the sub-areas in which the number of posted tweets is increasing. Jaffe et al. [14] developed a spatial clustering algorithm for geo-tagged image data posted on the Flickr site. The spatial clustering algorithm is hierarchical and based on location information. Rattenbury et al. [15] also proposed an identification method of event places for geo-tagged image data posted on the Flickr site. Their method also can predict the contents of events using tag data. Yanai et al. [16] applied k-means to clustering geo-tagged image data. Kim et al. [17] introduced mTrend, which constructs and visualizes spatiotemporal trends of topics, named “topic movements.” These studies only focus on spatial clustering using location information, however our study focus not only spatially-separated statical clustering but also semantically-separated spatial clustering.

III. \((\epsilon, \sigma)\)-DENSITY-BASED SPATIAL CLUSTER

In this section, the definitions of \((\epsilon, \sigma)\)-density-based spatial criteria and \((\epsilon, \sigma)\)-density-based spatial cluster are presented.

A. Density-based Spatial Criteria

In the density-based spatial clustering algorithms, spatial clusters are dense regions separated from the regions of lower density. In other words, regions with a high density of data points are spatial clusters, whereas areas with a low density are not. The key idea of the density-based spatial clustering algorithms is that, for each data point of a spatial cluster, the neighborhood of a user-defined radius has to contain at least a minimum number of points; that is, the density in the neighborhood has to exceed some predefined threshold.

In DBSCAN, the \(\epsilon\)-neighborhood of a data point is defined as documents in the neighborhood of a user-defined given radius \(\epsilon\). In the \(\epsilon\)-neighborhood of a data point in a spatial cluster has to contain at least minimum number of data points. In this study, a data point is a georeferenced document and the definition of \(\epsilon\)-neighborhood of a georeferenced document is extended. We define the \((\epsilon, \sigma)\)-neighborhood of a georeferenced document to extract the semantically similar neighbors of a georeferenced document.

**Definition 1** \((\epsilon, \sigma)\)-neighborhood \(GN_{(\epsilon, \sigma)}(gdp)\) The \((\epsilon, \sigma)\)-neighborhood of a georeferenced document \(gdp\),
denoted by $GN_{(\epsilon, \sigma)}(gd)$, is defined as

$$GN_{(\epsilon, \sigma)}(gd) = \{ gdq \in GDS | dist(gdp, gdq) \leq \epsilon \text{ and } \sim(gdp, gdq) \geq \sigma \}, \quad (1)$$

where the function $dist$ returns the distance between georeferenced document $gdp$ and georeferenced document $gdq$, and the function $\sim$ returns the similarity between $gdp$ and $gdq$. The function $\sim$ is explained in the next section.

An example of the $\epsilon$-neighborhood of $gdp$ is shown on the left side of Fig. 1. The $\epsilon$-neighborhood of $gdp$ is a set of georeferenced documents that exist within $\epsilon$ from $gdp$. In this example, there are four georeferenced documents in the $\epsilon$-neighborhood of $gdp$. An example of the $(\epsilon, \sigma)$-neighborhood of $gdp$ is shown on the right side of Fig. 1. The $(\epsilon, \sigma)$-neighborhood of $gdp$ is a set of georeferenced documents existing within distance $\epsilon$ from $gdp$ and the similarity between each georeferenced document and $gdp$ is more than a value of $\sigma$. In this example, there are three georeferenced documents, $GN_{(\epsilon, \sigma)}(gdp) = \{ gdq_2, gdq_3, gdq_4 \}$. A georeferenced document $gdq_1$ is within $\epsilon$ from $gdp$; however, it is not in $GN_{(\epsilon, \sigma)}(gdp)$, because the similarity between $gdq_1$ and $gdp$ is less than than a value of $\sigma$.

**Definition 2 (Core/Border Georeferenced Document)** A document $gdp$ is called a core georeferenced document if there are at least a minimum number of georeferenced documents, $MinDoc$, in the $(\epsilon, \sigma)$-neighborhood $GN_{(\epsilon, \sigma)}(gdp)$ ($GN_{(\epsilon, \sigma)}(gdp) \geq MinDoc$). Otherwise, ($GN_{(\epsilon, \sigma)}(gdp) < MinDoc$), $gdp$ is called a border georeferenced document.

Suppose that $MinDoc$ is set to three. A georeferenced document $gdp$ in the left side of Fig. 2 is a core georeferenced document, because there are three documents in $GN_{(\epsilon, \sigma)}(gdp)$. A georeferenced document $gdp$ in the right side of Fig. 2 is a border georeferenced document because the number of documents in $GN_{(\epsilon, \sigma)}(gdp)$ is less than $MinDoc$.

**Definition 3 ($\epsilon, \sigma$)-density-based directly reachable**

Suppose that a georeferenced document $gd$ is the $(\epsilon, \sigma)$-neighborhood of $gdp$. If the number of georeferenced documents in the $(\epsilon, \sigma)$-neighborhood of $gdp$ is greater than or equal to $MinDoc$, i.e., $GN_{(\epsilon, \sigma)}(gdp) \geq MinDoc$, $gd$ is $(\epsilon, \sigma)$-density-based directly reachable from $gdp$. In other words, georeferenced documents in the $(\epsilon, \sigma)$-neighborhood of a core georeferenced document are $(\epsilon, \sigma)$-density-based directly reachable from the core georeferenced document.

On the left side of Fig. 2, document $gdp$ is a core georeferenced document, because $GN_{(\epsilon, \sigma)}(gdp) \geq MinDoc$. Georeferenced documents $gd_2, gd_3$ and $gd_4$ are in the $(\epsilon, \sigma)$-neighborhood of $gdp$. These three documents are $(\epsilon, \sigma)$-density-based directly reachable from $gdp$. On the other hand, on the right side of Fig. 2, document $gdp$ is a border georeferenced document, i.e., is not $GN_{(\epsilon, \sigma)}(gdp) \geq MinDoc$. These two georeferenced documents are not $(\epsilon, \sigma)$-density-based directly reachable from $gdp$ although georeferenced document $gd_2$ and $gd_3$ are in the $(\epsilon, \sigma)$-neighborhood of $gdp$.

**Definition 4 (($\epsilon, \sigma$)-density-based reachable)** Suppose that there is a georeferenced document sequence ($gd_1, gd_2, gd_3, \ldots, gd_n$) and the $(i+1)$-th georeferenced document $gd_{i+1}$ is $(\epsilon, \sigma$)-density-based directly reachable from the $i$-th georeferenced document $gd_i$. The georeferenced document $gd_n$ is $(\epsilon, \sigma$)-density-based reachable from $gd_1$.

An example of an $(\epsilon, \sigma$)-density-based reachable is shown Fig. 3. If $MinDoc = 3$, $gd_2$ is $(\epsilon, \sigma$)-density-based directly reachable from $gd_1$ and $gd_3$ is $(\epsilon, \sigma$)-density-based directly reachable from $gd_2$. The georeferenced document $gd_3$ is $(\epsilon, \sigma$)-density-based reachable from $gd_1$. On the other hand, $gd_3$ is not $(\epsilon, \sigma$)-density-based reachable from $gd_3$, i.e., $gd_2$ is not $(\epsilon, \sigma$)-density-based directly reachable from $gd_3$.

**Definition 5 (($\epsilon, \sigma$)-density-based connected**)

Suppose that georeferenced documents $gdp$ and $gdq$ are $(\epsilon, \sigma$)-density-based reachable from document $gdo$. If $ND_{(\epsilon, \sigma)}(gdp) \geq MinDoc$, we denote that $gdp$ is $(\epsilon, \sigma$)-density-based connected to $gdq$.

An example of an $(\epsilon, \sigma$)-density-based reachable is shown in Fig. 3. In this figure, $gd_3$ is $(\epsilon, \sigma$)-density-based reachable from $gd_1$ and $gd_3$ is $(\epsilon, \sigma$)-density-based reachable from $gd_1$. At this time, $gd_3$ is $(\epsilon, \sigma$)-density-based connected to $gd_3$.

**B. Definition of Cluster**

An $(\epsilon, \sigma$)-density-based spatial cluster consists of two types of document: core georeferenced documents, which are mutually $(\epsilon, \sigma$)-density-based reachable; and border georeferenced documents, which are $(\epsilon, \sigma$)-density-based directly reachable from a core georeferenced document.
reachable from the core georeferenced documents. A \((\epsilon, \sigma)\)-density-based spatial cluster is defined as follows.

**Definition 6 \((\epsilon, \sigma)\)-density-based spatial cluster**

An \((\epsilon, \sigma)\)-density-based spatial cluster (DSC) in a georeferenced document set \(GDS\) satisfies the following restrictions:

1. \(\forall gd_p, gd_q \in GDS\), if and only if \(gd_p\) is \((\epsilon, \sigma)\)-density-based reachable from \(gd_p, gd_q\) is also in \(DSC\).
2. \(\forall gd_p, gd_q \in DSC\), \(gd_p\) is \((\epsilon, \sigma)\)-density-based connected to \(gd_q\).

Even if \(gd_p\) and \(gd_q\) are border georeferenced documents, \(gd_p\) and \(gd_q\) are in a same \((\epsilon, \sigma)\)-density-based spatial cluster if \(gd_p\) is \((\epsilon, \sigma)\)-density-based connected to \(gd_q\).

**IV. \((\epsilon, \sigma)\)-Density-based Spatial Clustering Algorithm**

In this section, the proposed \((\epsilon, \sigma)\)-density-based spatial clustering algorithm is described.

**A. Data Model**

Let \(gd_i\) denote the \(i\)-th georeferenced document in \(GDS = \{gd_1, \cdots, gd_n\}\); then, \(gd_i\) consists of three items: \(gd_i = \langle text_i, pl_i, pt_i \rangle\), where \(text_i\) is the content (e.g., title, short text message, and tags), \(pl_i\) is the location when the geo-spatiotemporal document was posted, and \(pt_i\) is the time when the geo-spatiotemporal document was posted. Then, \(pt_i\) is located (e.g., latitude and longitude).

**B. Algorithm**

The algorithm of \((\epsilon, \sigma)\)-density-based spatial clustering is shown in Algorithm 1. In this algorithm, the function **IsClustered** checks whether document \(gd_p\) is already assigned to a spatial cluster. Then, the function **GetNeighborhood** returns the \((\epsilon, \sigma)\)-neighborhood of georeferenced document \(gd_p\). For each georeferenced document \(gd_p\) in \(GDS\), the following steps are executed. If \(gd_p\) is a core georeferenced document according to Definition 2, it is assigned to a new spatial cluster. Then, all the neighbors are queued to a candidate queue \(CQ\) for further processing. The function **MakeNewCluster** makes a new spatial cluster. The processing and assignment of georeferenced documents to the current spatial cluster continue until \(CQ\) is empty. The next georeferenced document is dequeued from \(CQ\). If the dequeued georeferenced document is not already assigned to the current spatial cluster, it is so assigned to the current spatial cluster. Then, if the \((\epsilon, \sigma)\)-neighborhood of the dequeued georeferenced document is queued to \(CQ\) using the function **EnQueue**, which puts input georeferenced documents into \(CQ\) if they are not already in \(CQ\).

**C. Keyword-based Similarity Function**

Let \(dt_i\) denote all words in \(text_i\) of \(i\)-th georeferenced document: \(dt_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,nw(i)}\}\), where \(w_{i,j} \in W, W\) is a set of all words including in \(\{text_1, text_2, \cdots, text_n\}\). In this study, morphological analysis extracts noun, verb and adjective phrase as words. Simpson’s coefficient has a feature of cosine similarity for similarity between sets. The word-based Simpson’s coefficient is defined as:

\[
wsim(gd_i, gd_j) = \frac{|dt_i \cap dt_j|}{|\min(dt_i, dt_j)|}
\]

The word-based Simpson’s coefficient has drawback, when the keywords are same but several words in georeferenced documents are different. For example, suppose that there are two georeferenced documents \(gd_1\) and \(gd_2\) that are related to “Itsukushima Shrine”. If \(dt_1 = \{‘Itsukushima Shrine’, ‘beautiful’, ‘historical’, ‘Hiroshima’\}\) and \(dt_2 = \{‘Itsukushima Shrine’, ‘wonderful’, ‘sea’, ‘clean’\}\), the similarity between two georeferenced documents is \(wsim(gd_1, gd_2) = 1/4 = 0.25\). The similarity between \(gd_1\) and \(gd_2\) is low, even though \(gd_1\) and \(gd_2\) cover the same topic “Itsukushima Shrine.”

If georeferenced documents include a same keyword, which are be located close to each other, the georeferenced documents are similar each other. On the basis of this concept, we define the new similarity measurement based on keyword-based Simpson’s coefficient. Let \(key_j\) denote all words in \(dt_i\) of \(i\)-th georeferenced document: \(key_j = \{k_{i,1}, k_{i,2}, \cdots, k_{i,nk(j)}\}\), where \(k_{i,j} \in wi, k_{i,j} \in K\) is a set of all keywords including in \(W\). The keyword-based Simpson’s coefficient is defined as:

\[
ksim(gd_i, gd_j) = \frac{|key_j \cap key_j|}{|\min(key_j, key_j)|}
\]

We define a new similarity function between georeferenced documents that is trade-off of the word-based Simp-
son’s coefficient and the keyword-based Simpson’s coefficient. The similarity function \(sim\) is defined as:

\[
sim(g_i, g_j) = w_1 \times usim(g_i, g_j) + w_2 \times ksim(g_i, g_j),
\]

(4)

where \(w_1 + w_2 = 1.0\). If \(w_1\) and \(w_2\) are set to 1.0 and 0.0 respectively, the keyword-based similarity function only use words similarities. On the other hand, If \(w_1\) and \(w_2\) are set to 0.0 and 1.0 respectively, the keyword-based similarity function only use keywords similarities.

In the example described above, suppose that \(w_1 = 0.5\) and \(w_2 = 0.5\). The return value of \(usim(g_i, g_j)\) is 0.25 and the return value of \(ksim(g_i, g_j)\) is 1.0. Thus, the return value of the keyword-based similarity function \(sim\) is 0.5 \(\times 0.25 + 0.5 \times 1.0 = 0.6125\). Georeferenced documents \(g_1\) and \(g_2\) including the local topic of “Itsukushima Shrine” are determined similar each by using a new similarity measurement.

V. Experimental Results

To evaluate the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm, we used an actual GDS that is composed of crawling geo-tagged tweets on the Twitter site. We collected geo-tagged tweets from the Twitter site using its API. The number of tweets is 480,000. The time period is from November 2011 to February 2012. In the experiments, we compare the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm with DBSCAN.

The parameters of DBSCAN were set to \(\epsilon=500m, MinDoc=5\). The parameters of the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm were set to \(\epsilon=500m, \sigma=0.7, MinDoc=5\). Moreover, we used two types of the keyword-based similarity functions. One is that weight parameters \(w_1\) and \(w_2\) are set to 1.0 and 0.0 respectively (called the words-based method). The other is that weight parameters \(w_1\) and \(w_2\) are set to 0.5 and 0.5 respectively (called the keywords-based method). We ranked the clusters on the basis of the number of tweets included in each cluster.

Table I, Table II and III show the details of extracted spatial cluster ranked in the number of tweets. These table show the number of tweets, the range of longitude and latitude of each cluster. Moreover, top 5 of frequent words in each cluster are shown, but words relevant to address such as Hiroshima and city is excluded.

Table I shows the details of extracted spatial cluster using DBSCAN. The region of cluster 1 covers the downtown of Hiroshima; however, there are many local topics in it. Fig. 4 shows the locations of tweets in clusters 1 on the geographical coordinate space. The density of posed tweets in the downtown of Hiroshima is high because there are many people there. Therefore, this region is extracted as one spatial cluster including several local topics. As a result, DBSCAN can not recognize semantically-separated spatial clusters.

Table II and III show the ranking of extracted spatial clusters using DBSCAN. The region of cluster 1 covers the downtown of Hiroshima; however, there are many local topics in it. Table II and III show the results of extracted spatial clusters using the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm. Table II shows the results of the proposed clustering algorithm using the words-based method. Table III shows the results of the proposed clustering algorithm using keywords-based method. In contrast to DBSCAN, the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm recognized multiple spatial clusters.

In Table II, the areas of cluster 1, cluster 3, cluster 4 are located in the downtown of Hiroshima. In Table III,
the areas of cluster 1, cluster 2, cluster 8 are located in the downtown of Hiroshima. Fig. 5 shows the locations of tweets in extracted spatial clusters located in the downtown of Hiroshima on the geographical coordinate space. The \((\epsilon, \sigma)\)-density-based spatial clustering algorithm can recognize semantically-separated spatial clusters; however cluster 1 in Table II includes local topics downtown in Hiroshima. There are many tweets related to “Okonomiyaki restaurant”, “streetcars” and “Hiroshima’s oyster.” These tweets include the same address. Table II shows the results of the words-based method. Therefore, the algorithm determined these tweets are similar. On the other hand, this cluster is not extracted in the keywords-based method.

The extracted spatial clusters clusters 4 of Table II and clusters 2 of Table III, although both clusters are “Atomic Bomb Dome”, the keywords-based method is six tweets less than the words-based method. We checked six tweets manually, the topic of these six tweets is “Atomic Bomb Dome Sta.” This result indicates that the keywords-based method can recognize accurate spatial cluster compared with the words-based method.

VI. CONCLUSION

In this paper, we propose a novel spatial clustering algorithm, called the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm, for extracting “attractive” local regions in georeferenced documents. The proposed spatial clustering algorithm can recognize not only spatially-separated but also semantically-separated spatial clusters. To evaluate our proposed clustering algorithm, geo-tagged tweets posted on the Twitter site are used. The experimental results show that the \((\epsilon, \sigma)\)-density-based spatial clustering algorithm can extract “attractive” local regions as \((\epsilon, \sigma)\)-density-based spatial clusters. In our future work, we are going to develop online algorithm to extract \((\epsilon, \sigma)\)-density-based spatial clusters.

REFERENCES