Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I,
IMECS 2014, March 12 - 14, 2014, Hong Kong

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

Tatsuhiro 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 (ϵ, σ)-density-based spatial clustering algorithm, for extracting "attractive" local regions in georeferenced documents is proposed. We defined a new type of spatial cluster called an (ϵ, σ)-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 (ϵ, σ)-density-based spatial clustering algorithm can extract "attractive" local regions as (ϵ, σ)-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 (ϵ, σ)-density-based spatial cluster. An (ϵ, σ)-density-based spatial cluster is not only spatially-separated but also semantically-separated from other spatial clusters. Thus, (ϵ, σ)-density-based spatial clusters are closely related to local topics and events.

The main contributions of this study are as follows:

- To extract (ϵ, σ)-density-based spatial clusters, we propose a new spatial clustering algorithm for georeferenced documents, called the (ϵ, σ)-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 (ϵ, σ)-density-based spatial clustering algorithm can recognize (ϵ, σ)-density-based spatial clusters, which are both semantically-separated and spatially-separated from other spatial clusters.
- To recognize semantically-/spatially-separated clusters as (ϵ, σ)-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 (ϵ, σ)-density-based spatial clusters that represent "attractive" local regions associated with

This work was supported in part by Hiroshima City University Grant for Special Academic Research (General Studies).

T.Sakai is with Faculty of Information Sciences, Hiroshima City University, 3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima 731-3194 Japan, corresponding e-mail: sakai.hcu@gmail.com

K.Tamura and H.Kitakami are with Graduate School of Information Sciences, Hiroshima City University, 3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima 731-3194 Japan, e-mail: {ktamura, kitakami}@hiroshima-cu.ac.jp

local topics.

The remainder of this paper is organized as follows. In Section 2, related work is reviewed. In Section 3, the (ϵ, σ)-density-based spatial cluster is defined. In Section 4, the (ϵ, σ)-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 sites. 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

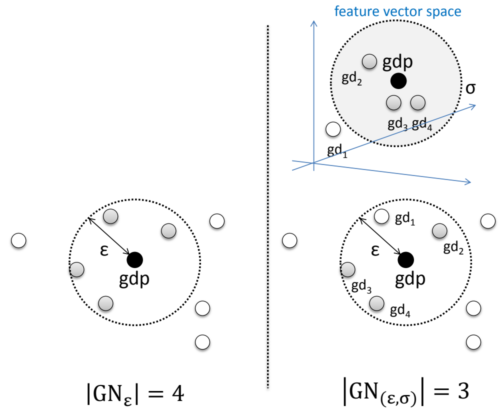

Fig. 1. Example of definition 1.

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. (ϵ, σ)-Density-based Spatial Cluster

In this section, the definitions of (ϵ, σ)-density-based spatial criteria and (ϵ, σ)-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 are 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 ϵ-neighborhood of a data point is defined as documents in the neighborhood of a user-defined given radius ϵ. In the ϵ-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 ϵ-neighborhood of a georeferenced document is extended. We define the (ϵ, σ)-neighborhood of a georeferenced document to extract the semantically similar neighbors of a georeferenced document.

Definition 1 **((ϵ, σ)-neighborhood $GN_{(\epsilon,\sigma)}(gdp)$)** The (ϵ, σ)-neighborhood of a georeferenced document gdp,

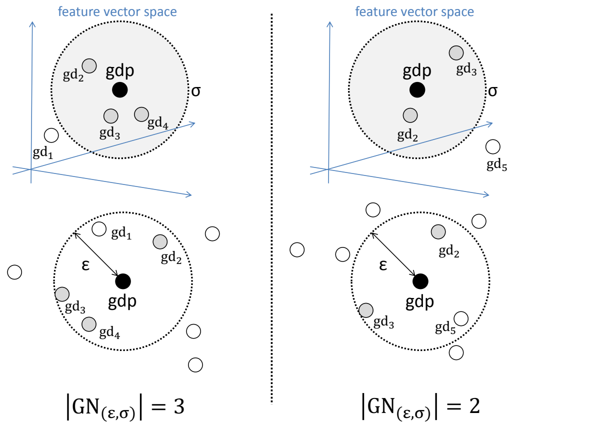

Fig. 2. Example of definition 2 and definition 3.

denoted by $GN_{(\epsilon,\sigma)}(gdp)$, is defined as

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

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 ϵ-neighborhood of gdp is shown on the left side of Fig. 1. The ϵ-neighborhood of gdp is a set of georeferenced documents that exist within ϵ from gdp. In this example, there are four georeferenced documents in the ϵ-neighborhood of gdp. An example of the (ϵ, σ)-neighborhood of gdp is shown on the right side of Fig. 1. The (ϵ, σ)-neighborhood of gdp is a set of georeferenced documents existing within distance ϵ from gdp and the similarity between each georeferenced document and gdp is more than a value of σ. In this example, there are three georeferenced documents, $GN_{(\epsilon,\sigma)}(gdp) = \{gd_2, gd_3, gd_4\}$. A georeferenced document gd_1 is within ϵ from gdp; however, it is not in $GN_{(\epsilon,\sigma)}(gdp)$, because the similarity between gd_1 and gdp is less than than a value of σ.

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

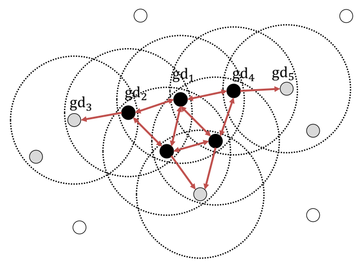

Fig. 3. Example of definition 4 and definition 5.

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 (ϵ, σ)-neighborhood of gdp. These three documents are (ϵ, σ)-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 (ϵ, σ)-density-based directly reachable from gdp although georeferenced document gd_2 and gd_3 are in the (ϵ, σ)-neighborhood of gdp.

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

An example of an (ϵ, σ)-density-based reachable is shown Fig. 3. If $MinDoc = 3$, gd_2 is (ϵ, σ)-density-based directly reachable from gd_1 and gd_3 is (ϵ, σ)-density-based directly reachable from gd_2. The georeferenced document gd_3 is (ϵ, σ)-density-based reachable from gd_1. On the other hand, gd_5 is not (ϵ, σ)-density-based reachable from gd_3, i.e., gd_2 is not (ϵ, σ)-density-based directly reachable from gd_3.

Definition 5 ((ϵ, σ)-density-based connected) Suppose that georeferenced documents gdp and gdq are (ϵ, σ)-density-based reachable from document gdo. If $ND_{(\epsilon,\sigma)}(gdp) \geq MinDoc$, we denote that gdp is (ϵ, σ)-density-based connected to gdq.

An example of an (ϵ, σ)-density-based reachable is shown in Fig. 3. In this figure, gd_3 is (ϵ, σ)-density-based reachable from gd_1 and gd_5 is (ϵ, σ)-density-based reachable from gd_1. At this time, gd_3 is (ϵ, σ)-density-based connected to gd_5.

B. Definition of Cluster

An (ϵ, σ)-density-based spatial cluster consists of two types of document: core georeferenced documents, which are mutually (ϵ, σ)-density-based reachable; and border georeferenced documents, which are (ϵ, σ)-density-based directly

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

reachable from the core georeferenced documents. A (ϵ, σ)-density-based spatial cluster is defined as follows.

Definition 6 ((ϵ, σ)-density-based spatial cluster)
An (ϵ, σ)-density-based spatial cluster (DSC) in a georeferenced document set GDS satisfies the following restrictions:

(1) $\forall gdp$, $gdq \in GDS$, if and only if gdq is (ϵ, σ)-density-based reachable from gdp, gdq is also in DSC.
(2) $\forall gdp$, $gdq \in DSC$, gdp is (ϵ, σ)-density-based connected to gdq.

Even if gdp and gdq are border georeferenced documents, gdp and gdq are in a same (ϵ, σ)-density-based spatial cluster if gdp is (ϵ, σ)-density-based connected to gdq.

IV. (ϵ, σ)-Density-based Spatial Clustering Algorithm

In this section, the proposed (ϵ, σ)-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: $gd_i =< text_i, pt_i, pl_i >$, where $text_i$ is the content (e.g., title, short text message, and tags), pt_i is the time when the geo-spatiotemporal document was posted, and pl_i is the location where gd_i was posted or is located (e.g., latitude and longitude).

B. Algorithm

The algorithm of (ϵ, σ)-density-based spatial clustering is shown in Algorithm 1. In this algorithm, the function **IsClustered** checks whether document gdp is already assigned to a spatial cluster. Then, the function **GetNeighborhood** returns the (ϵ, σ)-neighborhood of georeferenced document gdp. For each georeferenced document gdp in GDS, the following steps are executed. If gdp is a core georeferenced document according to Definition 2, it is assigned to a new spatial cluster, and 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 (ϵ, σ)-neighborhood of the dequeued georeferenced document are queued to CQ using the function **EnNniqueQueue**, which puts input georeferenced documents into CQ if they are not already in CQ.

```
input : GDS - georeferenced document set, ε -
        neighborhood radius, σ - similarity rate,
        MinDoc - threshold value
output: SC - set of clusters

cid ← 1;
SC ← φ;
for i ← 1 to |GDS| do
    gdp ← gd_i ∈ GDS;
    if IsClustered(gdp) == false then
        GN ← GetNeighbors(gdp,ε,σ);
        if |GN| ≥ MinDoc then
            stc_cid ←MakeNewCluster(cid,gdp);
            cid ← cid + 1;
            EnQueue(CQ,GN);
            while CQ is not empty do
                gdp ← DeQueue(CQ);
                GN ← GetNeighbors(gdp,ε,σ);
                if |GN| ≥ MinDoc then
                    EnNniqueQueue(CQ,GN);
                end
                stc_cid ← stc_cid ∪ gdp
            end
            SC ← SC ∪ stc_cid;
        end
    end
end
return SC;
```

Algorithm 1: (ϵ, σ)-Density-based Spatial Clustering Algorithm

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:

$$wsim(gd_i, gd_j) = \frac{|dt_i \cap dt_j|}{|min(dt_i, dt_j)|}. \tag{2}$$

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 document 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_i denote all words in dt_i of i-th georeferenced document: $key_i = \{k_{i,1}, k_{i,2}, \cdots, k_{i,nk(i)}\}$, where $k_i \in wi$, $k_{i,j} \in K$, K is a set of all keywords including in W. The keyword-based Simpson's coefficient is defined as:

$$ksim(gd_i, gd_j) = \frac{|key_i \cap key_j|}{|min(key_i, key_j)|}. \tag{3}$$

We define a new similarity function between georeferenced documents that is trade-off of the word-based Simp-

ISBN: 978-988-19252-5-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
CLUSTERING RESULTS OF DBSCAN

No	Number of Tweets	Range (longitude)	Range (latitude)	Top-5 Frequent Words
1	2173	132.34259769 - 132.5139095	34.34225649 - 34.41800308	shop, inside, today, station, come
2	288	132.301779 - 132.32664956	34.291072 - 34.317351	Miyajima, Itsukushima Shrine, Miyajimaguchi, oyster, ferry
3	170	132.4580275 - 132.4968043	34.43618755 - 34.48192577	shop, day, lunch, AEON MALL Hiroshima Gion, come
4	128	132.90427752 - 132.91733343	34.331726 - 34.348506	Tamayura, station, cat, Mr/Ms, Okonomiyaki
5	97	132.54589487 - 132.57154524	34.2343527 - 34.25657546	Yamato, museum, center, shop, noodle
6	96	132.7203672 - 132.75817651	34.4141014 - 34.43534496	Geso, person, today, set menu, shop
7	86	132.5285826 - 132.54099838	34.3442324 - 34.3628074	Mr/Ms, senaponcoro, shop, buy, seem
8	67	132.30352202 - 132.31108951	34.35173988 - 34.35770497	octopus, ball, while, open, today

TABLE II
CLUSTERING RESULTS OF THE PROPOSED SPATIAL CLUSTERING ALGORITHM ($w_1 = 1.0$ AND $w_2 = 0.0$)

No	Number of Tweets	Range (longitude)	Range (latitude)	Top-5 Frequent Words
1	97	132.4572834 - 132.46863105	34.389778 - 34.398638	shop, inside, Okonomiyaki, the head shop, Hondori
2	91	132.3154613 - 132.323433	34.2952182 - 34.304972	Miyajima, Itsukushima Shrine, Otorii, Itsukushima, Shrine
3	89	132.47242982 - 132.478453	34.39267358 - 34.401398	station, JR, Sta, Shinkansen, shop
4	47	132.4516591 - 132.45680987	34.39113274 - 34.39614078	Atomic Bomb Dome, Dome, bomb, Atomic, inside
5	32	132.9155353 - 132.919807	34.4374464 - 34.44173556	Hiroshima airport, HIJ, RJOA, lounge, ANA
6	18	132.177305 - 132.179825	34.16595235 - 34.169017	Kintaikyo, Yokoyama, the foot of the bridge, back side, cross
7	18	132.303433 - 132.310635	34.30675418 - 34.311843	Miyajima, ferry, Miyajimaguchi, JR West Japan, conger
8	15	132.31584043 - 132.31844813	34.36297389 - 34.36718941	Miyajima SA, outbound, San'you Expressway, Starbucks, coffee

TABLE III
CLUSTERING RESULTS OF THE PROPOSED SPATIAL CLUSTERING ALGORITHM ($w_1 = 0.5$ AND $w_2 = 0.5$)

No	Number of Tweets	Range (longitude)	Range (latitude)	Top-5 Frequent Words
1	58	132.47208448 - 132.47934873	34.39384782 - 34.40005438	Station, JR, Sta, Shinkansen, platform
2	41	132.4522132 - 132.45680987	34.39113274 - 34.395784	Atomic Bomb Dome, Atomic, Dome, Bomb, inside
3	34	132.3154613 - 132.32271635	34.295341 - 34.3043505	Miyajima, Otorii, Itsukushima, oyster, do
4	25	132.31876669 - 132.32147207	34.2958401 - 34.30074774	Itsukushima Shrine, Itsukushima, Shrine, Shrine, Itsukushima
5	17	132.177305 - 132.179825	34.16595235 - 34.169017	Kintaikyo, Yokoyama, the foot of the bridge, back side, Cross
6	15	132.9155353 - 132.91950762	34.4374464 - 34.44173556	Hiroshima Airport, HIJ, RJOA, Arrival, B787
7	13	132.42671107 - 132.42702243	34.37271835 - 34.37327164	SemiHard Toast, baked, one down, favor, today
8	12	132.45691723 - 132.45915413	34.40035934 - 34.40379812	Castle, Castle, beautiful, huge castle, Mizuhori

son's coefficient and the keyword-based Simpson's coefficient. The similarity function sim is defined as:

$$sim(gd_i, gd_j) = w_1 \times wsim(gd_i, gd_j) + w_2 \times ksim(gd_i, gd_j), \quad (4)$$

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 $wsim(gd_i, gd_j)$ is 0.25 and the return value of $ksim(gd_i, gd_j)$ 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 gd_1 and gd_2 including the local topic of "Itsukushima Shrine" are determined be similar each other by using a new similarity measurement.

V. EXPERIMENTAL RESULTS

To evaluate the (ϵ, σ)-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 (ϵ, σ)-density-based spatial clustering algorithm with DBSCAN.

The parameters of DBSCAN were set to ϵ=500m, $MinDoc$=5. The parameters of the (ϵ, σ)-density-based spatial clustering algorithm were set to ϵ=500m, σ=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 the (ϵ, σ)-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 (ϵ, σ)-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,

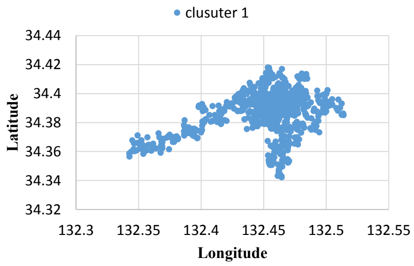

Fig. 4. Data plots in downtown of Hiroshima using DBSCAN.

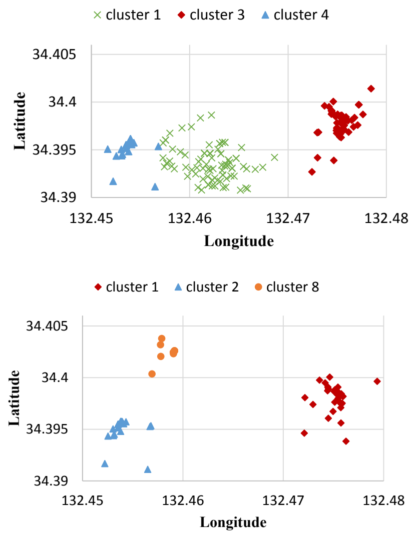

Fig. 5. Data plots in downtown of Hiroshima using the proposed spatial clustering algorithm (the upper figure is the words-based method and the lower figure is the keywords-based method).

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 (ϵ, σ)-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 (ϵ, σ)-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 (ϵ, σ)-density-based spatial clustering algorithm can extract "attractive" local regions as (ϵ, σ)-density-based spatial clusters. In our future work, we are going to develop online algorithm to extract (ϵ, σ)-density-based spatial clusters.

References

[1] M. Naaman, "Geographic information from georeferenced social media data," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 54–61, jul 2011.
[2] S. Van Canneyt, S. Schockaert, O. Van Laere, and B. Dhoedt, "Detecting places of interest using social media," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '12, 2012, pp. 447–451.
[3] H. Yang, S. Chen, M. R. Lyu, and I. King, "Location-based topic evolution," in *Proceedings of the 1st international workshop on Mobile location-based service*, ser. MLBS '11, 2011, pp. 89–98.
[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th international conference on World wide web*, ser. WWW '09, 2009, pp. 761–770.
[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10, 2010, pp. 851–860.
[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, E. Simoudis, J. Han, and U. M. Fayyad, Eds. AAAI Press, 1996, pp. 226–231.
[7] M. F. Goodchild, "Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0," *International Journal of Spatial Data Infrastructures Research*, vol. 2, pp. 24–32, 2007.
[8] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 37–45.
[9] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, jun 1998.
[10] K. Tamura and T. Ichimura, "Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents," in *Proceedings of the IEEE International Conference on System, Man, and Cybernetics, SMC 2013*, 2013, pp. 2079-2084.
[11] S. Kisilevich, F. Mansmann, and D. Keim, "P-dbscan: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, ser. COM.Geo '10, 2010, pp. 38:1–38:4.
[12] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, ser. CIKM '11, 2011, pp. 2541–2544.
[13] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ser. LBSN '10, 2010, pp. 1–10.
[14] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, ser. MIR '06, pp. 89–98.
[15] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '07, pp. 103–110.
[16] K. Yanai, K. Yaegashi, and B. Qiu, "Detecting cultural differences using consumer-generated geotagged photos," in *Proceedings of the 2nd International Workshop on Location and the Web*, ser. LOCWEB '09, 2009, pp. 12:1–12:4.
[17] K.-S. Kim, R. Lee, and K. Zettsu, "mtrend: discovery of topic movements on geo-microblogging messages," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '11, 2011, pp. 529–532.