Journal of Universal Computer Science, vol. 24, no. 3 (2018), 322-337
submitted: 30/9/17, accepted: 5/3/18, appeared: 28/3/18 © J.UCS

Multi-scaled Spatial Analytics on Discovering Latent Social Events for Smart Urban Services

O-Joun Lee
(Department of Computer Engineering
Chung-Ang University, Seoul, Korea
concerto34@cau.ac.kr)

Yunhu Kim
(Department of Computer Engineering
Chung-Ang University, Seoul, Korea
yunhu0110@gmail.com)

Hoang Long Nguyen
(Department of Computer Engineering
Chung-Ang University, Seoul, Korea
longnh238@gmail.com)

Jai E. Jung[1]
(Department of Computer Engineering
Chung-Ang University, Seoul, Korea
j3ung@cau.ac.kr)

Abstract: The goal of this paper is to discover latent social events from social media for sensitively understanding social opinions that appeared within a city. The latent social event indicates a regional and inconspicuous social event which is mostly buried under macroscopic trends or issues. To detect the latent social event, we propose three methods: i) discovering areas-of-interest (AOIs), ii) allocating social texts to the AOIs, and iii) detecting social events in each AOI. The AOIs can be composed by grouping social texts which are topically and spatially homogeneous. To make the AOIs dynamic and incremental, we use windows for allocating a social text to an adequate AOI. Lastly, the latent social events are detected from the AOI on the basis of keywords and temporal distribution of the social texts. Although, in this study, we limited the proposed method into analyzing social media, it could be extended to detecting events among agents/things/sensors.

Key Words: Social event detection, Area-of-interest, Social opinion mining, Spatio-temporal analysis.
Category: I.2.8, J.4, I.m

1 Introduction

Traditionally, a city has been simply regarded as a physical space where people live, work, and move. It has been important to collect and analyze all possible data for

[1] Corresponding author.

implementing 'smart city'. Various studies have been conducted, and most of them have focused on designing infrastructures (e.g., internet of agents [Bui and Jung, 2018], smart gird for energy services [Karnouskos et al., 2012], and effective water management [Dickey, 2018]).

More importantly, a city is considered as a social space among citizens, and social data from the citizens of the city is a valuable resource for constructing smart city [Cranshaw et al., 2012, Mainka et al., 2014, Balduini et al., 2014]. In this study, we focus on detecting regional and inconspicuous social events [Nguyen et al., 2017] which happened in the city by using geo-tagged data [Pham et al., 2014, Nguyen et al., 2014, Tri and Jung, 2015, Jung, 2016].

Various studies for detecting regional/local/city-scale social events have been conducted. Lee et al. proposed a method for detecting geo-social events by defining them as a unusual status of a region which is exposed on the social media [Lee et al., 2011]. However, social events which are required for realizing smart city as a service are not only abnormal or unusual but also periodic or ever-present. If a graduation ceremony which is hold in a university causes traffic congestion, we consider it as a usual event. Nevertheless, graduation ceremony is an annual event and it also brings significant contribution for smart city services. Besides, most of methods for discovering social trends or events have been focused on macroscopic events [Puiu et al., 2016]. The authors extracted the human flow dynamics from taxi traces. In another research, authors discovered the event's time, venue, and scale through detecting abnormal social activeness. In addition, they measured impact of the event by detecting traffic congestion and its change. This study is on the basis of an assumption: the social events are outstanding occurrences that involve a large number of people but it limited scope of the study into discovering abnormal events that happened in a city [Zhang et al., 2015]. Moreover, in order to realize the smart city as an application/service, we have to deal with social events more minutely and finely. Trivial, ordinary events also provide meaningful information, as much as abnormal and eccentric events. For example, it is obvious that lots of citizens feel cold during winter. However, an event that particular citizens feel cold provides us useful information like which district of the city is extraordinarily cold. This information could be used for smart grid or smart home.

Although the macroscopic events affect residents of the city, either, in order to detect and deal with social events which are more intimate with the citizens, we have to discover the social events spatial-specifically. For example, when we operate a smart grid service, social opinions about climates could be useful information to predict power consumption of the city for air-conditioning. For small cities, we might be able to estimate the power consumption by using weather forecasts. Also, we could predict it based on frequency of social texts that talk about too hot/cold weather.

In order to deal with big city/metropolitan, we should not uniformly handle all the districts of the city, since each district has its own characteristics and contexts. In Fig. 1, let suppose that 'District A' and 'District B' are downtown and uptown, re-

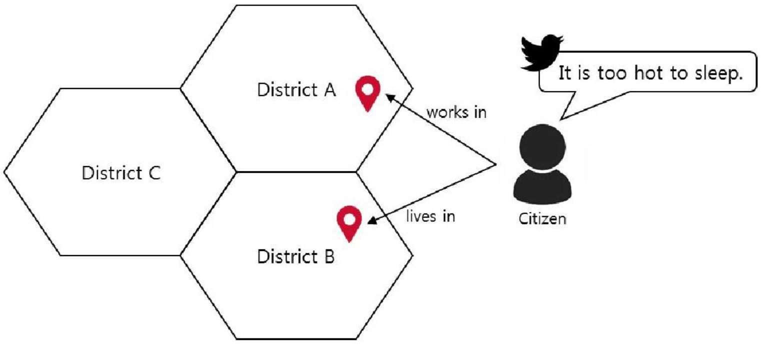

Figure 1: An example of a social text that contains the semantic ambiguity.

spectively. If we predict the power consumption during the night time based on the frequency of social texts for hot/cold weather, it will be overestimated for the 'District A' and underestimated for the 'District B'. Therefore, we have to detect the social events, spatial-specifically. Concept of Point Of Interest (POI) and Area Of Interest (AOI) are very effective in this circumstance [Vu and Shin, 2015, Vu et al., 2016, Nguyen and Shin, 2017].

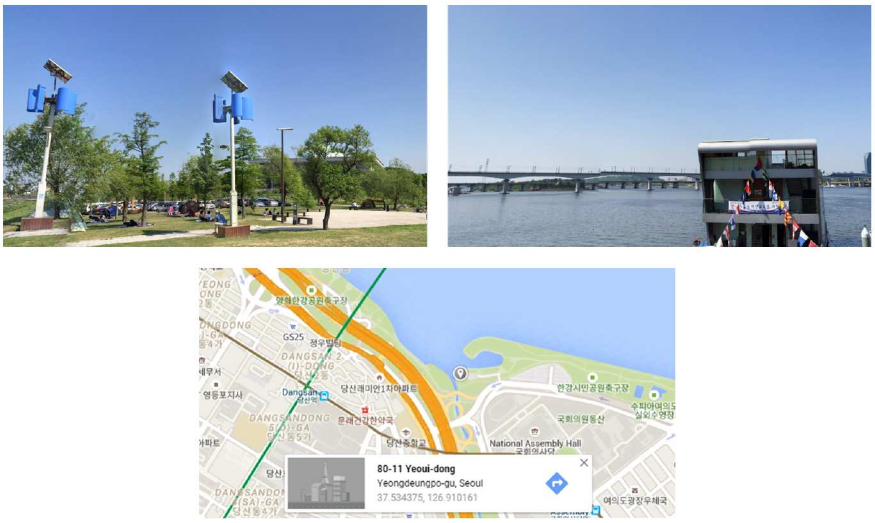

Figure 2: An example of social texts discussing different topics and sharing a common location.

Second, as shown in Fig. 2, even if two arbitrary citizens posted social texts at a common location, they could talk about totally different topics from each other. Let suppose a location A is near by a riverside park, a citizen U is a visitor for the park, and another citizen V works in a quay around location B. Although they commonly posted as "It is too crowded." or "It is too noisy.", however, these social text might have completely different meanings from each other.

To overcome aforementioned problems, we propose an adaptive method to detect latent social events for smart urban services in this paper. We discover regional and inconspicuous social events which are buried under massive social trends or issues and call them as latent social events. A latent social event accompanies a set of social texts that are topically and spatio-temporally homogeneous. To detect the latent social event from a social data stream, first, we compose areas-of-interest (AOIs) which are spatial regions discussing common topics. For incrementally updating the AOIs, novel social texts are allocated into the AOIs and outdated social texts are discarded from the AOIs, based on fixed-size windows. Finally, we discover the latent social events within the AOIs by using occurrence time points and keywords of social texts. In this paper, we does not cover processing the detected events since various methods already have been proposed to analyze the events and apply for useful applications and services [Lee and Jung, 2017].

The remainder of this paper is organized in the following manners. In Sect. 2 we define main problems in this paper. Further, Sect. 3 mentions about an adaptive method to discover latent social event. Next, we discuss about a case study in Sect. 4 for expressing the effective of this study. Finally, we conclude and state some future works in Sect. 5.

2 Problem Definition

In most of the previous studies, the social event was defined as an abnormal or unusual occurrence. However, in order to provide pervasive and spontaneous urban services, we have to collect social events related with citizens' life, which are small-scaled, periodic, or inconspicuous.

This study attempts to discover the social events from social data which is posted on the social networking services by the citizens. We got advantages from study of Nguyen et al. [Nguyen and Jung, 2018] for the definition of social data, as follows:

Definition 1 (Social Data). Social data $\mathfrak{D}$ represents facts about our society that its formula is as follows

$$\mathfrak{D} \equiv \langle\, \mathcal{U}, \mathcal{S}, \mathcal{A}\,\rangle \tag{1}$$

where $\mathcal{U}$ is a citizen which creates data, $\mathcal{S}$ is source of social data (e.g., Facebook, Twitter, and Instagram), and $\mathcal{A}$ is social attributes of data.

We investigated on different data sources to obtain social attributes of data. There are lots of properties of data, however, topic, location, and time are three main attributes for discovering useful and understandable patterns to help us to discover our social activities.

Definition 2 (Social Attribute of Data). Social attributes of data includes these three components: ϑ, ζ, and τ. They are topical, spatial, and temporal attribute respectively.

$$\mathcal{A} \equiv \langle\ \vartheta, \zeta, \tau\ \rangle \tag{2}$$

Tacit and explicit data are two types of social data [Smith, 2001]. Tacit data is learn from experience and is obtained in human mind (i.e., competence, deed, experience, and thinking) and explicit data consists of visible things such as printed and electronic materials. In this paper, we consider data as explicit data, especially social text data because it has high significance for discovering social event.

Definition 3 (Social Text). A social text denotes a relatively short text shared through social media by citizens. A social text t_i consists of a sequence of words, W_i written by a citizen $\mathcal{U}_i$ from source $\mathcal{S}_i$ with attribute $\mathcal{A}_i$. It can be formulated as:

$$t_i = \langle \mathfrak{D}_i, W_i \rangle, \tag{3}$$

$$W_i = \langle w_{i,1}, \cdots, w_{i,k}, \cdots, w_{i,K} \rangle, \tag{4}$$

where t_i denotes an i-th social text and $w_{i,k}$ indicates a k-th word that appeared in t_i.

Each social text reflects an opinion of whom posts the social text. If there are multiple social texts that handle a common topic with similar opinions, we can say this set of social texts represents a social event. We defined the social event, as follows;

Definition 4 (Social Event). A social event indicates an occurrence that involves multiple people. The social event accompanies a meaningful number of social texts that share similar opinions for a common topic. Also, the number of social texts increases and decreases according to a development of the social event. Therefore, we can reflect the social event by using a set of social texts that share a topic, an opinion for the topic, and a time duration. It can be formulated as:

$$\mathbb{E}_e = \{t_{i,u_j} | \tau_i \in [t_s, t_e], W_i \ni \vartheta_e, W_i \cap K_e \neq \emptyset\}, \tag{5}$$

where t_s and t_e denote a starting point and an ending point, respectively, ϑ_e refers to a topic which $\mathbb{E}_e$ addresses, K_e indicates a set of keywords what majority of $\mathbb{E}_e$'s elements include, and $W_i \cap K_e \neq \emptyset$ means that t_i contains a part of keywords in K_e. $W_i \cap K_e \neq \emptyset$ implicitly refers to that t_i represents a similar opinion with other members of $\mathbb{E}_e$ for a topic ϑ_e.

As displayed in Fig. 3, small-scaled social events that reflect quotidian opinions of the citizens are buried by large-scaled trends or issues. In here, a problem is that these small-scaled events are intimate with the citizens' life; e.g., complains for urban services or governance policies. We call these small-scaled and inconspicuous social events 'latent social events'. We defined the latent social event, as follows;

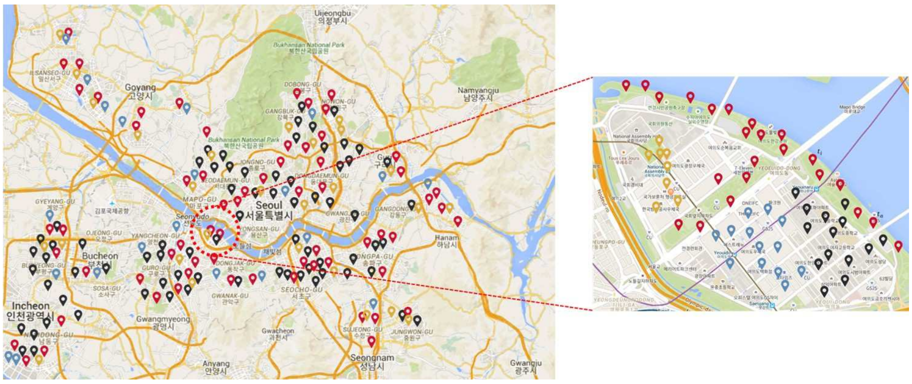

Figure 3: An example of latent social events buried by massive social data.

Definition 5 (Latent Social Event). A latent social event indicates a social event, which is regional and relatively small-scaled. Thus, we define the latent social event by limiting the social events into a particular spatial region. It is formulated as:

$$\mathbb{L}_e = \{t_i | \zeta_i \in \mathcal{R}_e, \tau_i \in [t_s, t_e], W_i \ni \vartheta_e, W_i \cap K_e \neq \emptyset\}, \quad (6)$$

where $\mathcal{R}_e$ refers to a region where $\mathbb{L}_e$ occurred.

In order to detect the latent social event, the most important requirement is we have to observe the social data stream with multiple scopes. Thus, this study proposes methods for i) discovering areas-of-interest (AOIs) (in Sect. 3.1), ii) allocating social texts to the AOIs (in Sect. 3.2), and iii) detecting social events in each AOI (in Sect. 3.3). The AOI indicates a spatial area related with a particular social issue.

3 Latent Social Event Detection

When we collect social data for all over the city, if we use an uniform scope (e.g., city and district), small-scaled social events will be buried under macroscopic trends or issues. Nevertheless, the small-scaled events are the things what we have to discover, since they are intimate with citizens' life and indistinct.

We are referred to the small and indistinct events as the latent social events. A method for detecting the latent social events consists of three steps: i) discovering areas-of-interest (AOIs), ii) allocating social texts to the AOIs, and iii) detecting social events in each AOI. In following sections, we introduce methods for conducting these steps based on topical and spatio-temporal homogeneity among the social texts.

3.1 Discovering Area of Interest

The area-of-interest (AOI) does not indicate only a certain spatial region, but also a region corresponds a social issue. In other words, an AOI is a set of social texts, which

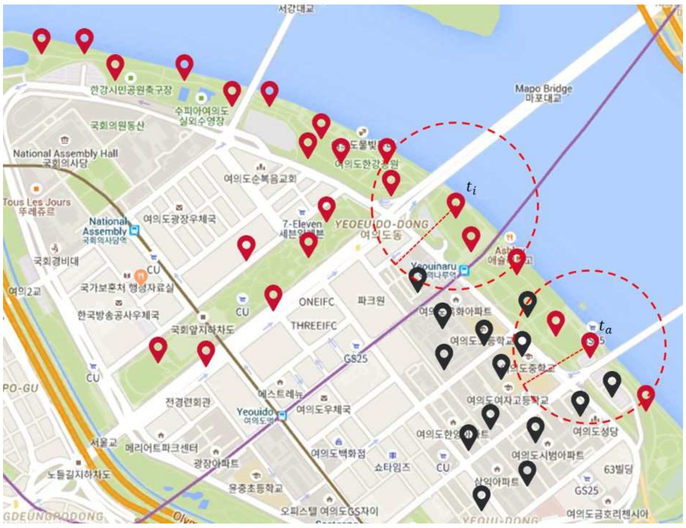

Figure 4: An example of grouping the social texts by using their N-nearest neighborhoods, where the geo-markers denote social texts and their location and colors of the markers indicate topics that are discussed by the social texts.

share a common spatial region and a social issue. Therefore, the AOIs could be overlapped with each other.

Definition 6 (Area-of-Interest). Let suppose $\mathbb{A}_a$ indicates an a-th AOI in the city. $\mathbb{A}_a$ is a set of social texts that were posted within a certain spatial region, $\mathcal{R}_a$ and addressing a particular topic, ϑ_a. It is formulated as:

$$\mathbb{A}_a = \{t_i | \zeta_i \in \mathcal{R}_a, W_i \ni \vartheta_a\}. \quad (7)$$

For composing the AOI, [Vu et al., 2016] used landmarks and radius. It attempted to find an optimal radius from a landmark, which includes social texts discussing the landmark, as many as possible. Lastly, a contour that connects the most outer geotags was determined as a border of the AOI. However, not all the social issues have spatial pivot points like landmarks.

To improve this problem, we apply neighborhoods of social texts for determining whether they address local issues or not. It is based on the following assumption.

Assumption 1 *If N-nearest neighborhoods of t_i are discussing a common topic with t_i, the topic, ϑ_a represents a social issue. Also, when a spatial distribution of social texts that are addressing ϑ_a and adjacent with t_i is $\mathcal{R}_a$, we can figure out an AOI, $\mathbb{A}_a = \{t_i | \zeta_i \in \mathcal{R}_a, W_i \ni \vartheta_a\}$.*

As displayed in Fig. 4, t_i and its neighborhoods are sharing a common topic. However, t_a addresses a different topic from most of its neighborhoods. From this observation, we can determine a topic that is discussed by t_i and its neighborhoods is a social issue. Moreover, if neighboring social texts of t_i are also adjacent with other social texts that discuss the common topic, we can compose a spatial region where the topic of t_i holds a majority.

Thus, in order to discover AOIs, we have to find i) which topics were discussed by the citizens, ii) whether the topics are local issues or not, and iii) where AOIs' borders are. First, we discover and measure a relationship between ϑ_a and W_i, $\mu_{\vartheta_a}(W_i)$. In order to measure $\mu_{\vartheta_a}(W_i)$, we use the method proposed in [Lee et al., 2017].

It is on the basis of a well-known information retrieval (IR) model, Bag-Of-Words (BOF). We define a words-social texts matrix, $\mathcal{M} \in \mathbb{R}^{K \times I}$, where K and I are the total number of words and social texts, respectively. In order to deal with the semantic ambiguity of words, first, we conduct TF-IDF (Term Frequency-Inverse Term Frequency) and LSA (Latent Semantic Allocation) for $\mathcal{M}$. Then, we compose feature vectors of the words by using SVD (Singular Value Decomposition). $\mathcal{M}$ is decomposed as $\mathcal{M} = W \Sigma T^T$, where W and T are $K \times K$ and $I \times I$ unitary matrices, respectively, and Σ is a $K \times I$ diagonal matrix that consists of eigen values. By reordering columns and rows of W, Σ, and T in descending order of the eigen values, we can recognize which latent feature of words is important than others. Finally, by reducing a size of W into $K \times \theta_K$, we obtain a matrix that consists of the words' feature vectors, where θ_K is a user-defined variable.

To reveal the topics of social texts, we apply the fuzzy c-means clustering, since any words do not correspond only one topic. The clustering is conducted to minimize an objective function:

$$\underset{\vartheta}{\operatorname{argmin}} \sum_{\forall \vartheta_\alpha} \sum_{\forall w_k} \mu_{\vartheta_\alpha}(w_k)^{-m} \mathcal{D}(w_k, C_{\vartheta_\alpha}), \tag{8}$$

$$\mu_{\vartheta_\alpha}(w_k) = \left[\sum_{\forall \vartheta_\beta} \left(\frac{\mathcal{D}(w_k, C_{\vartheta_\alpha})}{\mathcal{D}(w_k, C_{\vartheta_\beta})} \right)^{\frac{2}{m-1}} \right]^{-1}, \tag{9}$$

where ϑ is the topic model, ϑ_α denotes an α-th topic, and C_{ϑ_α} indicates a center of ϑ_α. A distance between a k-th word, w_k and C_{ϑ_α} is measured by the Euclidean distance between their feature vectors. Also, C_{ϑ_α} is estimated by a weighted average of ϑ_α's elements as:

$$C_{\vartheta_\alpha} = \frac{\sum_{\forall w_k \in \vartheta_\alpha} \mu_{\vartheta_\alpha}(w_k)^m \times w_k}{\sum_{\forall w_k \in \vartheta_\alpha} \mu_{\vartheta_\alpha}(w_k)^m}. \tag{10}$$

From Eq. 4, content of a social text, t_i is represented by a sequence of words included in t_i, W_i. Thus, in order to estimate a membership degree of t_i for ϑ_α, we combine $\mu_{\vartheta_\alpha}(w_{i,k}), \forall w_{i,k} \in W_i$. It can be formulated as:

$$\mu_{\vartheta_\alpha}(W_i) = \bigoplus_{\forall w_{i,k} \in W_i} \mu_{\vartheta_\alpha}(w_{i,k}), \tag{11}$$

where $oplus$ indicates the triangular norm, which is a maximum operation in this study.

By using the $\mu_{\vartheta_\alpha}(W_i)$, we can measure one of two major elements that define the AOI. To build sets of social texts that share both of topics and spatial regions, we apply a clustering approach for them based on their (i) topics, (ii) N-nearest neighborhoods, and (iii) locations.

The proposed clustering method aims to compose groups of social texts which contain a common topic and are closely located. An objective function of the clustering algorithm can be formulated as:

$$\underset{\mathbb{A}}{\operatorname{argmin}} \sum_{\forall \mathbb{A}_a} \sum_{\forall t_{i,u_j} \in \mathbb{A}_a} \mu_{\vartheta_a}(W_i)^{-m} \times \mathcal{D}(\zeta_i, \mathbb{A}_a) + \mathcal{D}_N(\zeta_i, \mathbb{A}_a), \tag{12}$$

$$\mathcal{D}(\zeta_i, \mathbb{A}_a) = \min_{\forall C_{c,\mathbb{A}_a}} \mathcal{D}(\zeta_i, C_{c,\mathbb{A}_a}), \tag{13}$$

$$\mathcal{D}_N(\zeta_i, \mathbb{A}_a) = \frac{1}{N} \sum_{\forall \mathcal{N}_n} \mu_{\vartheta_a}(W_n)^{-m} \times \mathcal{D}(\zeta_i, \zeta_n), \tag{14}$$

where $\mathbb{A}$ indicates all the AOIs that currently exist in the city, $\mathcal{D}(\zeta_i, \mathbb{A}_a)$ refers to a distance between ζ_i and $\mathbb{A}_a$'s center, $C_{c,\mathbb{A}_a}$ indicates a c-th center of $\mathbb{A}_a$, $\mathcal{D}_N(\zeta_i, \mathbb{A}_a)$ denotes a distance between t_i and t_i's neighborhoods within $\mathbb{A}_a$, $\mathcal{N}_n$ refers to a n-th nearest neighborhood of t_i, and W_n and ζ_n are $\mathcal{N}_n$'s contents and spatial location, respectively.

Since shape of the AOI is informal, it is difficult to fix a center of the AOI as a point. Therefore, we postulate multiple centers for an AOI. And we also use t_i's N-nearest neighborhoods for estimating the intimacy between t_i and $\mathbb{A}_a$. In here, the topical distance, $\mu_{\vartheta_a}(W_i)^{-m}$ works as a weighting factor for the two spatial distances: $\mathcal{D}(\zeta_i, \mathbb{A}_a)$ and $\mathcal{D}_N(\zeta_i, \mathbb{A}_a)$.

As the same with the objective function, the first term of $\mathcal{D}_N(\zeta_i, \mathbb{A}_a)$ estimates whether the neighborhoods of t_i are also discussing the same topic with $\mathbb{A}_a$. The second term measures the spatial distance between t_i and its neighborhoods. We used the Euclidean distance for calculating the spatial distance. Thereby, $\mathcal{D}_N(\zeta_i, \mathbb{A}_a)$ has a low value, when (i) t_i's neighborhoods address a common topic with $\mathbb{A}_a$ and (ii) the neighborhoods are located closely with t_i.

In order to discover AOIs' centers, we apply density of the social texts. The higher density of the social texts indicates the shorter distances between the neighborhoods. Therefore, if t_i is a center of $\mathbb{A}_a$, it has to satisfy a following condition:

$$\mathcal{D}_N(\zeta_i, \mathbb{A}_a) \geq \max_{\forall \mathcal{N}_n} \mathcal{D}_N(\zeta_n, \mathbb{A}_a). \tag{15}$$

Owing to usage of the clustering approach, we have to determine the number of clusters. We measured the quality of the total cluster model, as the number of clusters increases one by one. The benefit of each increment is estimated by:

$$\mathcal{B}_{|\mathbb{A}|} = (1-\theta_Q) \times \Delta\mathcal{Q}_{|\mathbb{A}|} + \theta_Q \times \Delta\mathcal{Q}_{|\mathbb{A}|-1}, \tag{16}$$

$$\Delta\mathcal{Q}_{|\mathbb{A}|} = \mathcal{Q}_{|\mathbb{A}|} - \mathcal{Q}_{|\mathbb{A}|-1}, \tag{17}$$

where $|\mathbb{A}|$ indicates the number of clusters in the current cluster model and θ_Q denotes a user-defined parameter that represents the momentum of the cluster model's quality. When the number of clusters increases to $|\mathbb{A}|$, $\mathcal{Q}_{|\mathbb{A}|}$ refers to the quality of the cluster model, $\Delta\mathcal{Q}_{|\mathbb{A}|}$ denotes the amount of changes in the quality, and $\mathcal{B}_{|\mathbb{A}|}$ indicates to the gain from the increment of the number of clusters.

If the $\mathcal{B}_{|\mathbb{A}|}$ had a positive value, the proposed method went on to the next iteration by $|\mathbb{A}| := |\mathbb{A}| + 1$. Otherwise, it determined the optimal number of clusters as $|\mathbb{A}|$.

The quality of the total cluster model, $\mathcal{Q}_{|\mathbb{A}|}$ was estimated by the internal compactness and the external adjacency of clusters, $\forall\mathbb{A}_a \in \mathbb{A}$. It is formulated as:

$$\mathcal{Q}_{|\mathbb{A}|} = \sum_{\forall\mathbb{A}_a}\Bigg[\sum_{\forall t_i\in\mathbb{A}_a} \mu_{\mathbb{A}_a}(t_i)^{-m} \times \mathcal{D}_N(\zeta_i, \mathbb{A}_a) - \sum_{\forall t_b\notin\mathbb{A}_a} \mu_{\mathbb{A}_a}(t_b)^{-m} \times \mathcal{D}_N(\zeta_b, \mathbb{A}_a)\Bigg]. \tag{18}$$

Thereby, the first term of Eq. 18 measures the compactness of each cluster, the second term indicates the adjacency among the clusters, and $\mathcal{Q}_{|\mathbb{A}|}$ indicates how well-constructed the $\mathbb{A}$ is. If each AOI in the model properly represents latent social opinions, $\mathcal{Q}_{|\mathbb{A}|}$ might have a low value.

In addition, m, which is used as exponent of the membership functions, is a user-defined parameter. As m becomes higher, the membership degree of the movies gets more consideration. In this study, m is equal to 2.

3.2 Allocating Social Texts to AOIs

The social data is a kind of data stream rather than a static dataset. Therefore, in order to analyze the social data, the proposed model has to be incremental. For a newly appeared social text, the proposed model is easily able to add the new one into an adequate AOI.

In order to incrementally add and delete a new social text to and from the AOI, we propose two naive methods using fixed-size windows. First, we can use a window defined by a time duration, Δt. This method simply adds the newly occurred social text into an adequate AOI, and discards social texts that are older than Δt. Therefore, the AOI is dynamically changed for representing social texts that occurred during a time period, $[t_0 - \Delta t, t_0]$, where t_0 indicates the current time. Thereby, the definition of AOI in Eq. 7 is changed into:

$$\mathbb{A}_a = \{t_i | \zeta_i \in \mathcal{R}_a, \tau_i \in [t_0 - \Delta t, t_0], W_i \ni \vartheta_a\}. \tag{19}$$

On the other hand, we are able to consider a window defined by the number of social texts within the AOI. This method limits size of of the AOI, $|\mathbb{A}_a|$ lower than a user-defined parameter, $\theta_{\mathbb{A}}$. It works as a queue (i.e., first in, first out). Thus, Eq. 7 is re-defined by adding following constraints:

$$\underset{\mathbb{A}_a}{\operatorname{argmin}} \sum_{\forall t_i \in \mathbb{A}_a} \|\tau_i - t_0\|, \; |\mathbb{A}_a| \leq \theta_{\mathbb{A}}. \tag{20}$$

Above two methods have their own merits and demerits. When social texts within an AOI sparsely occur or a temporal distribution of the social texts is not uniform, size of the AOI will be too drastically changed, if we use the time-based window. However, when we attempt to detect abnormal and abrupt occurrence of social events, the drastic changes provide us meaningful information.

However, in order to update the AOI model, $\mathbb{A}$, we have to conduct the whole clustering process, which is computationally expensive. Therefore, the proposed method incrementally handles new social texts, and periodically re-builds the AOI model.

3.3 Detecting Latent Social Event

In Sect. 3.1, we proposed the novel method for building the AOI, and also introduced the incremental approaches for dynamically extending the AOI in Sect. 3.2. Finally, in this section, we propose a method for detecting the latent social event from the AOI.

In Def. 5, we defined the latent social event as a set of social texts that share a spatio-temporal region, a topic, and a similar opinion. To expose the opinions within social texts, we use a set of keywords, K_e that corresponds a latent social event, $\mathbb{L}_e$. Also, since $\mathbb{L}_e$ is a subset of $\mathbb{A}_a$ when $\mathbb{L}_e$ is detected in $\mathbb{A}_a$, Eq. 6 can be simplified as:

$$\mathbb{A}_a \supset \mathbb{L}_e = \{t_i | t_i \in \mathbb{A}_a, \tau_i \in [t_s, t_e], W_i \cap K_e \neq \emptyset\}. \tag{21}$$

Therefore, a latent social event is a set of social texts within an AOI that share a temporal region and keywords. In order to detect it, we use a similar method with Eq. 12 by using the Euclidean distance and the Jaccard distance. Our objective function can be defined as:

$$\underset{\mathbb{L} \subset P(\mathbb{A}_a)}{\operatorname{argmin}} \sum_{\forall \mathbb{L}_e \subset \mathbb{A}_a} \sum_{\forall t_i \in \mathbb{L}_e} \mathcal{D}\left(\tau_i, C_{\mathbb{L}_e}\right) \times \mathcal{D}_J\left(W_i, K_e\right), \tag{22}$$

where $\mathcal{D}\left(\tau_i, C_{\mathbb{L}_e}\right)$ denotes an Euclidean distance between τ_i and a center of $\mathbb{L}_e$ and $\mathcal{D}_J\left(W_i, K_e\right)$ indicates a Jaccard distance between W_i and K_e. K_e is composed by:

$$K_e = \left\{ w_k | w_k \in K, f_{w_k} > \frac{1}{|K|} \sum_{\forall w_j \in K} f_{w_j} \right\}, \; K = \bigcup_{\forall t_i \in \mathbb{L}_e} W_i, \tag{23}$$

where f_{w_k} refers to the frequency of w_k that occurred within social texts included in $\mathbb{L}_e$. Thereby, K_e is a set of words that frequently appeared in $\mathbb{L}_e$.

Nevertheless, if we use conventional clustering methods, it is inappropriate for processing social data, which is a massive data stream. Therefore, we propose an incremental algorithm based on Eq. 22. A procedure of this algorithm can be enumerated, as follows:

1. When a new social text, t_i appears in $\mathbb{A}_a$, estimate the membership of t_i for existing latent social events.
2. Add t_i into a latent social event which t_i has the highest membership degree for.
3. Update features of the latent social event where t_i is added to.
4. If the membership degree of t_i is lower than a minimum threshold for all the existing latent social events, make a novel latent social event.
5. Discard latent social events which can not embrace new social texts, anymore.

In order to dynamically model the latent social event, we manage its temporal region by the Gaussian distribution, which is defined by the mean and the standard deviation. Thus, a membership degree of t_i for $\mathbb{L}_e$ can be formulated as:

$$\mu_{\mathbb{L}_e}(t_i) = f\left(\tau_i, \overline{\tau_{\mathbb{L}_e}}, \sigma_{\mathbb{L}_e}\right) \times \mathcal{D}_J\left(W_i, K_e\right)^{-1}, \tag{24}$$

where $f\left(\tau_i, \overline{\tau_{\mathbb{L}_e}}, \sigma_{\mathbb{L}_e}\right)$ is a probability density function for temporal distribution of social texts included in $\mathbb{L}_e$. Also, $\overline{\tau_{\mathbb{L}_e}}$ and $\sigma_{\mathbb{L}_e}$ are an average and a standard deviation of occurrence time of social texts included in $\mathbb{L}_e$, respectively.

On the social data stream, we can not iteratively access all the social texts. However, it is relatively easy to incrementally update $\overline{\tau_{\mathbb{L}_e}}$, $\sigma_{\mathbb{L}_e}$, and K_e, if we store only $|\mathbb{L}_e|$, $|K|$, and $f_{w_k}, \forall w_k \in K$.

It is difficult to foreknow an occurrence of a novel latent social event. Thus, we use the concept drift which is an appearance of a social text that does not follow models of existing latent social event. If $\mu_{\mathbb{L}_e}(t_i) < \theta_{\mathbb{L}}, \forall \mathbb{L}_e \subset \mathbb{L}$, we construct a new latent social event that has t_{i,u_j} as an only element, where $\theta_{\mathbb{L}}$ is a user-defined minimum threshold.

Also, when $f\left(\tau_i, \overline{\tau_{\mathbb{L}_e}}, \sigma_{\mathbb{L}_e}\right) < \theta_{\mathbb{L}}$, $\mathbb{L}_e$ is discarded from active latent social events, since $\mathcal{D}_J\left(W_i, K_e\right)^{-1} \in [0, 1]$.

4 Utilizing Latent Social Event for Transportation Problem: a Case Study in Seoul

In order to examine the feasibility of our idea, we select Seoul as a case study with transportation problem. Keyword "road" and geo-tagged "Seoul" is used to collect social text data about transportation on Twitter in 20 May 2017. There are more than 1,000 tweets was posted on that day, however, it only shows some abnormal signals at several time points as in Fig. 5. In this paper, we do not focus on proposing algorithms to detect social events because various of studies solved this issue in novel ways. Assuming that we obtained 5 social events at different time points (i.e., 7:00, 7:30, 9:30, 12:15, and 17:30). We consider tweets which are posted at 17:30 as a case study to investigate which happens at this time. Our algorithm which is proposed in Sect. 3 is used to discover latent social events for the aforementioned data. The result shows that there are two main social events at two different regions (i.e., Gwangjin-gu and Gangnam-gu) because almost tweets come from there. At other regions the number of tweets is trivial, therefore, we ignored tweets at these places.

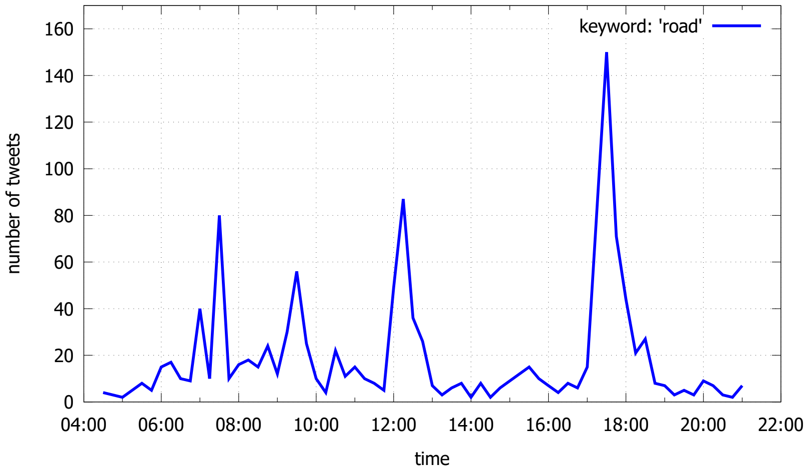

Figure 5: Distribution of tweets which was posted on 20-May-2017 at Seoul.

Only tweets whose geo-tagged from Gwangjin-gu and Gangnam-gu are considered. They are divided into two groups in order for us to discover specific patterns. We find out that almost tweets from Gangnam-gu are mentioned about traffic jam while a sudden sinkhole get lots of attention in Gwangjin-gu. Fig. 6 expresses latent social events at these locations. This result seems reasonable because Gangnam-gu is a business city, hence, traffic jam is a serious problem here. On the other way, transportation in Gwangjin-gu is effect by a sinkhole instead of traffic congestion. Taking immediate and precise action with a problem is very important towards a smart city. By obtaining latent social events, government can have promptly process for assisting citizens (e.g., setting up adaptive traffic signals, spreading real-time traffic feedback for solving traffic jam at Gangnam-gu and repairing sinkhole to resurface the road in Gwangjin-gu).

However, there are also some limitations of our work. First of all, almost data is

Figure 6: Latent social events on 17:30 20-May-2017 at Seoul.

lack of standard quantitative metrics because it depends on social media users' behavior. Therefore, data may contain inaccurate sentiment; written in short, unstructured, informal way; and comprise lots of spam message. Besides, we must collect social media data from different sources instead of using a specific one due to imbalanced data problem [Long and Jung, 2015]. Therefore, a framework which covers different social media sources is necessary.

5 Conclusion and Future Work

In this paper, we proposed an adaptive method to discover latent social events by using social data. We believe that our work is very essential for assisting citizens within smart cities. For further researches, we aim to integrate our proposing idea into SocioScope system [Nguyen and Jung, 2018] to provide a completed framework for smart cities including collecting social data, discovering latent social events, and supplying utilized application for residents. Moreover, we intend to extend our social data sources by collecting data from sensors (e.g., CCTV and wearable devices) and mass media (e.g., newspaper and television). Sensor data can be utilized for solving data delay problem and mass media data is effective for us to extract topics. Using social data from various sources can bring us an overview of every events which happen in our society.

Acknowledgment

This study was supported by the Chung-Ang University Research Scholarship Grants in 2017. Also, this research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A41015675).

References

[Balduini et al., 2014] Balduini, M., Bocconi, S., Bozzon, A., Valle, E. D., Huang, Y., Oosterman, J., Palpanas, T., and Tsytsarau, M. (2014). A case study of active, continuous and predictive social media analytics for smart city. In *Proceedings of the 5th Workshop on Semantics for Smarter Cities, a Workshop at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014*, pages 31–46.

[Bui and Jung, 2018] Bui, K.-H. N. and Jung, J. J. (2018). Internet of agents framework for connected vehicles: A case study on distributed traffic control system. *Journal of Parallel and Distributed Computing*. (To appear).

[Cranshaw et al., 2012] Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. M. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM 2012), Dublin, Ireland, June 4-7, 2012*.

[Dickey, 2018] Dickey, T. (2018). Smart water solutions for smart cities. In *Smart Cities*, pages 197–207. Springer.

[Jung, 2016] Jung, J. J. (2016). Exploiting geotagged resources for spatial clustering on social network services. *Concurrency and Computation: Practice and Experience*, 28(4):1356–1367.

[Karnouskos et al., 2012] Karnouskos, S., da Silva, P. G., and Ilic, D. (2012). Energy services for the smart grid city. In *Proceedings of the 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST 2012), Campione d'Italia, Italy, June 18-20, 2012*, pages 1–6.

[Lee et al., 2017] Lee, O., Hoang, L. N., Jung, J. E., Um, T., and Lee, H. (2017). Towards ontological approach on trust-aware ambient services. *IEEE Access*, 5:1589–1599.

[Lee and Jung, 2017] Lee, O. and Jung, J. E. (2017). Sequence clustering-based automated rule generation for adaptive complex event processing. *Future Generation Computer Systems*, 66:100–109.

[Lee et al., 2011] Lee, R., Wakamiya, S., and Sumiya, K. (2011). Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349.

[Long and Jung, 2015] Long, H. N. and Jung, J. J. (2015). Privacy-aware framework for matching online social identities in multiple social networking services. *Cybernetics and Systems*, 46(1-2):69–83.

[Mainka et al., 2014] Mainka, A., Hartmann, S., Stock, W. G., and Peters, I. (2014). Government and social media: A case study of 31 informational world cities. In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS 2014), Waikoloa, HI, USA, January 6-9, 2014*, pages 1715–1724.

[Nguyen and Jung, 2018] Nguyen, H. L. and Jung, J. E. (2018). Socioscope: A framework for understanding internet of social knowledge. *Future Generation Computer Systems*, 83:358–365.

[Nguyen and Shin, 2017] Nguyen, M. D. and Shin, W. (2017). Dbstexc: Density-based spatio-textual clustering on twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2017), Sydney, Australia, July 31 - August 03, 2017*, pages 23–26.

[Nguyen et al., 2017] Nguyen, T. T., Camacho, D., and Jung, J. E. (2017). Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services. *Personal and Ubiquitous Computing*, 21(2):267–279.

[Nguyen et al., 2014] Nguyen, T. T., Hwang, D., and Jung, J. J. (2014). Social tagging analytics for processing unlabeled resources: A case study on non-geotagged photos. In *Proceedings of the 8th International Symposium on Intelligent Distributed Computing (IDC 2014), Madrid, Spain, September 3-5, 2014*, pages 357–367.

[Pham et al., 2014] Pham, X. H., Nguyen, T. T., Jung, J. J., and Hwang, D. (2014). Extending HITS algorithm for ranking locations by using geotagged resources. In *Proceedings of the 6th International Conference on Computational Collective Intelligence. Technologies and Applications (ICCCI 2014), Seoul, Korea, September 24-26, 2014*, pages 332–341.

[Puiu et al., 2016] Puiu, D., Barnaghi, P., Tönjes, R., Kümper, D., Ali, M. I., Mileo, A., Parreira, J. X., Fischer, M., Kolozali, S., Farajidavar, N., et al. (2016). Citypulse: Large scale data analytics framework for smart cities. *IEEE Access*, 4:1086–1108.

[Smith, 2001] Smith, E. A. (2001). The role of tacit and explicit knowledge in the workplace. *Journal of Knowledge Management*, 5(4):311–321.

[Tri and Jung, 2015] Tri, N. T. and Jung, J. J. (2015). Exploiting geotagged resources to spatial ranking by extending hits algorithm. *Computer Science and Information Systems*, 12(1):185–201.

[Vu and Shin, 2015] Vu, D. D. and Shin, W. (2015). Low-complexity detection of POI boundaries using geo-tagged tweets: A geographic proximity based approach. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN 2015), Bellevue, Washington, USA, November 3-6, 2015*, pages 5:1–5:4.

[Vu et al., 2016] Vu, D. D., To, H., Shin, W., and Shahabi, C. (2016). Geosocialbound: an efficient framework for estimating social POI boundaries using spatio-textual information. In *Proceedings of the 3rd International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data (GeoRich 2016), San Francisco, California, USA, June 26 - July 1, 2016*, pages 3:1–3:6.

[Zhang et al., 2015] Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., and Wu, Z. (2015). City-scale social event detection and evaluation with taxi traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):40.