Sophy: a Morphological Framework for Structuring Geo-referenced Social Media

Kyoung-Sook Kim, Hirotaka Ogawa, Akihito Nakamura, Isao Kojima
Information Technology Research Institute
National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{ks.kim, h-ogawa, nakamura-akihito, isao.kojima}@aist.go.jp

ABSTRACT

Social networks have played a crucial role of information channels for understanding our daily lives beyond communication tools. In particular, their coupling with geographic location has boosted the worth of social media to detect, track, and predicate dynamic events and situations in the real world. While the amounts of geo-tagged social media are apparently increasing at every moment, we have few framework to handle spatiotemporal changes and analyze their relationships. In this paper, we propose a framework to understand dynamic social phenomena from the mountains of fragmented, noisy data flooding social media. First, we design a data model to describe morphological features of the populations of geo-location of social media and define a set of relationships by using differential measurements in spatial, temporal, and semantic dimensions. Then, we describe our real-time framework to extract morphometric features from streaming tweets, create the topological relationships, and store all features into a graph-based database. In the experiments, we show case studies related to two typhoons (Neoguri and Halong) and a landslide disaster (Hiroshima) with real tweet-sets in a visualization way.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial databases and GIS; H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Design, Human Factors

Keywords

Social geomorphology, Morphological features, Spatiotemporal phenomena, Movement analysis

1. INTRODUCTION

ACM SIGSPATIAL LBSN'14, November 4, 2014, Dallas, TX, USA

With coupling with location-based services and social networks, social media have become a typical type of user-generated geographic information. They have provided us considerable information to be able to detect, track, and predicate dynamic events and situations in the real world locally and globally. The geography of data shadows drawn from social media appears to be quite effective at reflecting experiences of real-world events[21]. However, they contain lots of noise with low precision and massive volumes comparing to traditional geographic information. The increasing rate of data volume is much faster than Web documents or blogs, and they are streaming into the server in real time. Therefore, it has difficulties to directly use data for decision making and crisis management and brings challenges to find meaningful information about human behaviours and understand social cognition about occurrences in the real world.

The emerging role of social media accompanying the physical space emphasizes the importance of spatiotemporal analysis over surging social messages and interactions. As shown in [24], understanding spatio-temporal processes of real-world phenomena and discovering pattern changes have been long-standing issues in many application domains. Numerous techniques have been developed including data structures, mining algorithms, stochastic models, and so on. The spatiotemporal locality, proximity, and relationships based on metric (like distance and direction) or non-metric (like shape and topology) criteria are central to extract knowledge about the dynamic complexity of geographic phenomena. In case of social sciences, the geo-spatial and temporal factors also have been dealt with to recognise the interactions between the environment and human activities [14]. We can roughly imagine the differences between cities and countryside lives of people. When we do, however, consider high volume and velocity of social media, we need a methodology to interpret how they are different or similar by computational models and operations. As being quoted by [4], Durkheim deployed the concept of social morphology to study the substratum of society by bridging geography and demography. Social morphology is based on how human population are geographically distributed and concentrated in space. Also, it usually focuses on the structural interrelations among phenomena, concepts, and ideas, such as structure of organisms in biology and structure of word morphemes for linguistics.

In this paper, we propose Sophy (Socio-geomorphological analysis) system, a framework to construct spatiotemporal relations among morphological features of the distribution of geo-tagged Twitter messages (for short, geo-tweets) regardless of the identification of users. The geomorphological

properties (such as geometric shape, size, slope, and curvature of landforms) and topological relations (such as intersection and disjoint) between two geo-spatial areas are foundational characteristic in spatiotemporal data analysis and prediction. Sophy extracts morphological features as units of analysis over the geographical population of geo-tweets in the real time. Then, it estimates basic relationships between features by using differential measurements in spatial, temporal, and semantic dimensions. Finally, the extracted features and relations are stored into a graph-based database. This paper exemplifies how to discover an interesting pattern of topic flocks through our framework by using three cases: two typhoons (Neoguri and Halong) and a landslide disaster (Hiroshima) in 2014. As considering a geomorphological approach, we can easily filter noise data and explore the spatiotemporal changes of location embedded social media data.

The remainder of the paper is organized as follows: Section 2 addresses our motivation on the basis of related work. Section 3 explains the data model with geometry and morphological characteristics in spatiotemporal semantic dimensions. In Section 4, we introduce the Sophy system of stream processing based on Storm and Neo4j and Section 5 provides a case study to analyze and visualize the spatiotemporal changes in a visual browser as our exploratory experimental results. Finally, we conclude this paper with a short description of future work in Section 6.

2. RELATED WORK

2.1 Location-based Twitter Analytics

Geo-referenced social media have significantly been applied to investigate human behaviour and understand social cognition about occurrences in the real world, especially Twitter has been a key resource of free and open data even though it has a limitation. TwitterStand [20] captures tweets related to breaking news from noise by online clustering method, and Sakaki et al. [19] propose the location/trajectory estimation methods such as Kalman filtering to estimate the locations of interesting events based on tweets– for examples, entertainment events such as sports games and concerts and natural/man-made disasters such as accidents, typhoons, and earthquakes. In [10], a dengue surveillance model using Twitter is proposed to perform spatio-temporal predictions based on volume, location, time, and population perception by the clustering algorithm. A probabilistic model for detecting flu outbreak based on a spatio-temporal Markov Network algorithm is shown in [17]. Lee et al. [16] present a geo-social event detection method and analyze urban characteristics with usual patterns of crowd behaviour. In [9], geo-tweets are used to find a hot spot for crime by means of kernel density estimation of past crime patterns. As considering location and time as well as topic analysis, we can expect the high potential of geo-tweets to detect, track, and predicate real-world events and situations both in local and global areas.

Despite many studies have focused on the event detection as shown in [6], few literature is proposed for dealing with spatiotemporal dynamics of topics. Pozdnoukhov et al. present the space-time dynamics of topics by using online topic model and kernel density estimation in [18]. They illustrate the changes of space-time hotspots of topics by physical movement of users and increased intensity of local communities. As developing online or offline methods of the event detection, we also need a quantitative and qualitative model to handle massive spatiotemporal features derived from the Twitter stream and understand the evolution in geometrical properties and topological relationships. For example, Figure 1 shows the change of spatial pattern of geo-tweets containing keyword 'typhoon' over time. The visualization is an efficient way for people to recognize their differences using shapes, sizes, and colors as shown as CartoDB[1] and TweetMap[2]. However, it is almost impossible to look at evolving patterns one by one when we face high-velocity and volume of population density generated by social media. In [23], six types of analysis for spatio-temporal changes are enumerated: (1) measuring time by fixing location and controlling the state of attributes, (2) measuring location by fixing time and controlling the state of attributes, (3) measuring the state of attributes by fixing time and controlling locations, (4) measuring time by fixing attributes and controlling locations, (5) measuring attributes by fixing location and controlling time, and (6) measuring locations by fixing attributes and controlling time. Based on those measurements, we propose a new method to manipulate dynamic social phenomena in geo-social media.

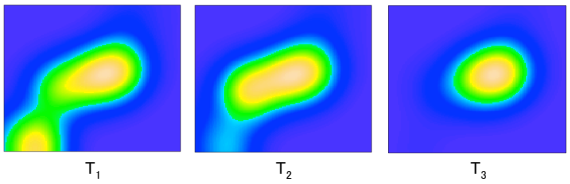


Figure 1: Spatiotemporal patterns

2.2 Movement Patterns

The dynamics of objects and phenomena have been typical issues in geographic information systems. While earlier models were concerned with the representation of changes in a discrete manner, moving-object models have tried to capture not only the discrete changes but also the continuous processes of the real world over time. A moving object is conceptually defined as a temporal function $mobject : time \rightarrow spatial - object$; this means that the model estimates the position of an object at any time during its lifetime [12]. We refer to this continuous representation of moving objects for our study. For example, a moving point is represented as a curve of successive locations over time in the three dimensional space. In [19], a typhoon trajectory is estimated by using tweets and compared with its real movement path. If we have a large amount of trajectory data of each topic or keyword, we can try to examine their mobilities and fine a certain pattern among a set of topic trajectories in space and time as shown in Figure 2. For example, the typhoon and heavy-rain trajectories of geo-tweets may stay together in a specific location during a certain time interval or they may show a similar movement with respect to the real one.

In order to represent a topic trajectory over geo-tweets, we first need to identify its spatiotemporal boundary over a set of data. We concern the spatiotemporal changes of top-

[1] http://cartodb.com
[2] http://worldmap.harvard.edu/tweetmap/

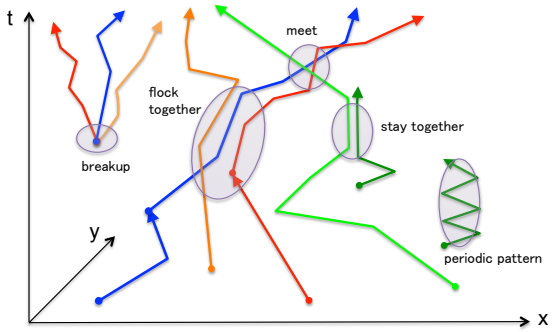


Figure 2: Trajectories that have movement patterns in spatiotemporal three dimensions: flock, stay, meet, breakup, and periodic visiting [11]

ics without considering user's identification and profiles in this work. For that clustering methods based on statistics, k-means, or kernel density will contribute to solve this problem. Second, the temporal mapping should be considered in the construction of the trajectory data. The moving-object models have provided basic data types and operations to analyze the continuous historical changes of the geographic phenomena with temporal function. While certain topics like a typhoon or a traffic jam would be refer to the movement information of real objects, not all topics have a good hypothesis. Finally, the real-time processing should be taken into account for handling the volume and velocity of geo-tweets. A complex model and algorithm requiring overhead costs are inappropriate for streaming geo-tweets. In this paper, we concentrate on real-time processing of geo-tweet streams to make spatiotemporal data structures for a pattern analysis.

3. MORPHOLOGY-BASED MODEL

Figure 3 shows the overview of Sophy (Social geomorphological analysis) framework that processes geo-tweet streams and constructs a spatiotemporal relations. It consists of three components of a distributed real-time processing engine, a database server, and a visual data browser. We will explain each of these in detain in the next section. This section first describes a geomorphology-based data model as a unit of processing and storing in the framework.

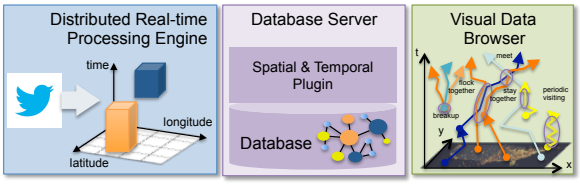


Figure 3: Overview of Sophy framework

3.1 Geomorphology Features

For our data model, we assume 1) certain topics would present a social pattern of geo-tweets' locations by local communities and have geomorphological features, such as peaks, pits, and passes; 2) local topics can propagate to near locations with spatiotemporal proximity, so geomorphological features may be observed at the different locations over time. First of all, we use the term '*observation*' to denote a geo-tweet formed as $ob = (u, s, t, msg)$, where the identifier u of a user, a spatial point of latitude and longitude $s = (x, y)$, a timestamp t, and a textual message msg. Basically, there are two spatial elements on the geo-tweets: place names of the user profile or on contents a user explicitly inputs and the user's geographical point location automatically tagged by GPS. We only perform the tweets tagged by the latitude and longitude coordinates rather than place names because the name of profile places rarely changes and the place name on contents is difficult to simply accept the presence of users at the place. After Part-of-Speech (POS) tagging of words on the message, an observation is transformed as several keyword occurrences to characterize patterns of spatiotemporal distribution of word-frequencies. Let a set of keyword occurrences O of $o_i = (s_i, t_i, w_i) : o_i \in O, i \in N$, where w is a word and N is the number of accumulated keyword occurrences observed on messages in a sliding window. We define the term '*locality*' as a measurement to reflect a local pattern of the keyword in geo-tweets within a pre-specified spatial region $S \subset \mathbb{R}^2$ and a bounded time-interval $T \subset \mathbb{R}$.

Definition 1. Locality
Given is a set of keyword occurrences O within a spatiotemporal domain $R = S \times T$, a locality l of keyword $w \in W$ is defined as a random real measure at a spatiotemporal point $st \in R$ given by

$$l(st, w) = \lambda(st, w)\kappa(st, w) \tag{1}$$

where W is a word bag appeared in O, λ is the intensity (expected number or density) of the spatiotemporal points st observed keyword w, and κ is the importance weight of keyword w at the point st.

This can be interpreted informally as the spatio-temporal-thematic distribution in the domain. There are many methods to estimate the intensity of a point process over an area considering applications as shown in [7]. Generally, the intensity function λ at location (s, t) is expressed as

$$\lambda(s, t) = \lim_{|ds|\to 0, |dt|\to 0} \frac{E[N(ds \times dt)]}{|ds| \cdot |dt|} \tag{2}$$

where ds is a small spatial region around spatial point s, dt is a small interval containing time instance t, $N(A)$ is the number of points within domain $A = ds \times dt$, and $|ds|$, $|dt|$ is the area of the region and the length of the time interval, respectively. In this paper we apply a quadrat method dividing a study area into subregions of equal size and count the frequency of points per unit area for the simplicity of stream processing. Therefore, we can substitute the Eq.(1) into Eq.(3) with a sequence of sub-regions $R = (r_i; i = 1, 2, ..., n)$.

$$l(r_i, w) = \lambda(r_i, w)\kappa(r_i, w) = \frac{N_w(r_i)}{|r_i|} \cdot \frac{N_w(r_i)}{N(r_i)} \tag{3}$$

where r_i is the i-th sub-region in spatiotemporal domain R, $|r_i|$ is the volume of sub-region r_i, $N(r_i)$ is the number of keyword occurrences within the sub-region, and $N_w(r_i)$ is the number of keyword occurrences whose important word is w in the subregion.

Next, we determine morphometric features (e.g., boundaries or shapes) in order to define topological relationships between topics of interests. We estimate time and location

of a social phenomenon by controlling the value of localities. This study focuses on two geomorphology types: peak (local maximum) and pit (local minimum) because they can evolve as a hot spot or outlier based on deviance from the normal situation. A peak is higher than all of its neighborhoods and mathematically described by the negative values of the second directional derivatives (i.e., $d^2f/dx^2 < 0$, $d^2f/dy^2 < 0$, and $d^2f/dt^2 < 0$). A pit is lower than all of its neighbors and the second directional derivatives are positive. In this paper, we apply the concept of neighborhood and open set in topological spaces. Let $R = \{r_1, r_2, ..., r_n\}$ be a set of n-regions in a topology space. We find a subset of R that satisfies the following definitions:

Definition 2. Peak and Pit
Given any keyword of w and a threshold value of θ, a subset P of set R is said to be $Peak_{w,\theta}$ and Pit_θ if it respectively satisfies as follows

$$Peak_{w,\theta} = \{r_i \in P \mid \max_{r_j \in NB(r_i), r_j \notin P} l(r_j, w) \leq l(r_i, w) \wedge \sum_i l(r_i, w) > \theta \ \wedge \ P \cap NB(P) = P\}$$

$$Pit_{w,\theta} = \{r_i \in P \mid \min_{r_j \in NB(r_i), r_j \notin P} l(r_j, w) \geq l(r_i, w) \wedge \sum_i l(r_i, w) < \theta \ \wedge \ P \cap NB(P) = P\}$$

where l is a locality value and $NB(r_i)$ is a set of neighborhood-regions of i-th subregion r_i in R, such that

$$NB(r_i) = \{r \in R \mid r_i \neq r, \partial r_i \cap \partial r \neq \varnothing\},$$

and

$$NB(P) = \bigcup_{r \in P} NB(r)$$

where ∂ is a set of boundary points (open sets) of a region.

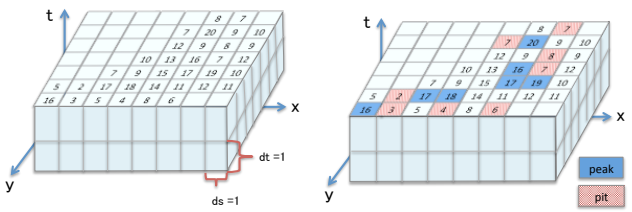
Figure 4 shows an example of the peaks and pits on the local patterns in a spatiotemporal space. If we capture locality values as shown in Figure 4 (a), then we can deduce geomorphology features as shown in Figure 4 (b). A peak/pit has the highest/lowest value than its 8 neighbours except the elements of the same group, e.g., any neighbour of the cell of 20 does not have a higher value. After selecting the geomorphology features, we put their information into instances of type *Form* as

Definition 3. Form
Given a peak of $P_w = \{r_i; i = 1, 2, ..., p\}$ of arbitrary word w (or a pit of P_w), *Form* type constructs a spatiotemporal continuum whose attributes are a word of w and a set of pairs of $(r, l(r, w))$ where r is a spatiotemporal region in P_w and l is a locality value of word w in region r.

Even though we here deal with only specific features such as peaks or pits, we can apply form type by using different methods like hot-spot and event detection if a spatiotemporal geometry can be composed of multiple sub-divisions.

3.2 Spatiotemporal Relationships

Now, we take into account spatiotemporal relationships based on the boundary information of form instances. For instance, a temporal relation such as *after*, *during*, and *overlap* in [5] or a spatial relation like *equal*, *overlap*, and *contain*



(a) Locality (b) Geomorphology Features

Figure 4: Geomorphology extraction based on quadrat analysis

as shown in [8] can be calculated by time interval and geometry operations. In this paper, we employ only a few relations to manage the spatiotemporal co-occurrences (which keywords get together) and movements of topics (which keywords flock together) over streaming geo-tweets. Table 1 shows two binary relations of form instances to be managed in our framework. First, we define *COINCIDENT* relation as pairs of forms that exist a spatial and temporal intersection and have different keywords such as 'heavy rain' and 'flood'. When two form instances are connected with each other by a *COINCIDENT* relationship, the relationship has the following attributes to measure proportion of co-occurrence: a type of Allen's interval relations [5], the duration of a intersected time interval, the degree and distance between two centroids, and the ratio of the intersection of each dimension. Through the *COINCIDENT* relationships, we can find a set of high-correlated keywords regarding to the spatiotemporal proximity of geomorphology features.

The second relation is *NEXT*, which represents a partial order of form instances in time. It has derivatives to measure the changes of spatio-temporal thematic states as well as a duration, a degree, and a distance attribute. If two forms given by the same word have fewer differences of the distance and the locality values within radius θ_d and threshold θ_l as defined in Table 1, they are connected by a *NEXT* relationship. However, we have a difficulty to decide the values of radius θ_d and threshold θ_l with respect to all keywords before starting the processing. Thus, we apply a nearest neighbor search problem for creating a *NEXT* relation of each keyword. Let f_{T_i} be a form instance within time interval T_i of the current window and $F_{T_{i-1}}$ be a set of forms that have the same keyword in time-interval T_{i-1} of the previous window (T_{i-1}<T_i). The *NEXT* relation can be obtained by the nearest neighbor query formulated as:

$$Next(F_{T_{i-1}}, f_{T_i}) = \underset{f \in F_{T_{i-1}}}{\operatorname{argmin}} \, d(f, f_{T_i}) \qquad (4)$$

where d is a function to measure a distance or difference from f_{T_i} to $f \in F_{T_{i-1}}$. The *NEXT* relation varies depending on the definition of function. In this study, we divide relationship *NEXT* into *NEXTBYDISTANCE* and *NEXTBYVALUE* by using the average of spatial distances and the average of locality differences of two forms, respectively. Figure 5 illustrates an example of basic relationships between form instances we have defined. By controlling the *NEXT* relationships, we can observe a phenomenon of changes of geomorphology features. For example, relationship $NEXT_{AC}$ between A and C can be interpreted by a diffusion phenomenon of keyword 'hurricane' because the locality of form

Table 1: Basic relationships between forms (a, a', and b are form types. S, T, L, and w denotes a set of spatial objects, time intervals, locality values, and a word of a form, respectively. $r = S \times T$ is a spatiotemporal region and $|\cdot|$ is the absolute value.)

Name	Spatial	Temporal	Thematic	Attributes
COINCIDENT(a, b)	$S_a \cap S_b \neq \phi$	$T_a \cap T_b \neq \phi$	$w_a \neq w_b$	type, duration, degree, distance, intersect-area ratio intersect-duration ratio
NEXT(a, a')	$distance(S_a, S_{a'}) \leq \theta_d$	$T_a < T_{a'}$	$w_a = w_{a'}$ and $\lvert L(r_a, w_a) - L(r_{a'}, w_{a'})\rvert \leq \theta_l$	duration, degree, distance, change of duration, change of area, change of measure, velocity

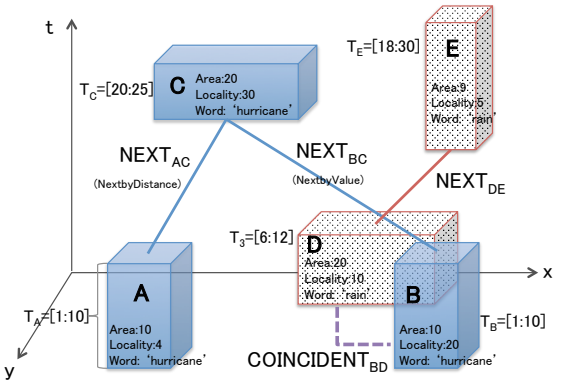


Figure 5: Example of relationships between forms (A, B, and C are derived from the geomorphology features of word 'hurricane', and D and E from 'rain').

C is much increased with a larger area during the less time. Relationship $NEXT_{DE}$ of keyword 'rain' shows an opposite case.

As shown in the above figure, a *COINCIDENT* relationship implies a spatiotemporal continuum like a form and a *NEXT* relationship can be considered as a trajectory segment between two successive time intervals. As time goes by, we generate a large number of trajectory segments through space and time by *NEXT* relationships, and mining spatiotemporal patterns in keyword trajectories is emerged in our framework. For the analysis of interesting spatiotemporal patterns of topics, we primarily add *Movement* type to draw a three-dimensional line segment in a spatiotemporal space. It is defined as a linear function to continuously map a time instance into a spatial point in a time interval such as a unit point in [12]:

Definition 4. Movement
Given a $NEXT$ relationship between two form instances f_i and f_j ($T_{f_i} < T_{f_j}$), *Movement* type constructs a following record of

$$Movement(f_i, f_j) = \{(t_i, t_j, x_i, x_j, y_i, y_j, l_i, l_j) \mid t_i < t_j, (x_i, y_i, t_i) \in S_{f_i} \times T_{f_i}, (x_j, y_j, t_j) \in S_{f_j} \times T_{f_j}\}$$

where t is a time instant, T and S is the temporal and spatial domain of a form, (x, y, t) is a coordinate in a spatiotemporal domain, and l denotes a real value of locality.

In our framework, we use the centroid of spatiotemporal coordinates and one of the summary functions such as *sum*, *min*, *max*, and *avg* to calculate a locality value of a form that may have several values of sub-regions.

Finally, we define two more relations of *SIMILAR* and *NEAR* of movement instances. Even though these relations represent a fragmentary information about movement patterns, we think they can help us to find a meaningful movement and extract spatio-temporal patterns such as tracks, flocks, and leaderships> as shown in [15]. Let M be a set of movements and $\theta \in \mathbb{R}$ is a threshold value to limit the number of relationships. A general definition of two relations is defined by pairs of movements that are satisfied as

$$Rel(M, \theta) = \{(m_i, m_j) \mid \forall m_i, m_j \in M \; : \; m_i \neq m_j \Rightarrow D(m_i, m_j) \leq \theta\}(i = 1, 2.., n) \quad (5)$$

where m is a movement instance and D is a function to return a distance measure between two movements. Depending on the distance function, we assign *Rel* into relation *SIMILAR* or *NEAR*.

For the *SIMILAR* relation, we adapt the Hausdorff distance metric into D function in Eq.(5) to measure a shape similarity [22]. The distance between two movements is

$$D(m_i, m_j) = \begin{cases} \infty & \text{if } T = T_{m_i} \cap T_{m_j} = \varnothing \\ Sim(ls_i, ls_j) & \text{otherwise} \end{cases}$$

where ls_i and ls_j are the line segments projected into the spatial domain with the common time interval T. The similarity measure is calculated by a shape similarity and direction similarity using

$$Sim(ls_i, ls_j) = \frac{H(ls_i, ls_j)}{diagonal} \cdot \cos(angle).$$

where *diagonal* is the max length of diagonal lines of the bounding rectangle containing ls_i and ls_j, *angle* is the angle between ls_i and ls_j, $H(ls_i, ls_j) = max(h(ls_i, ls_j), h(ls_j, ls_i))$ is the Hausdoroff distance and one-sided Hausdorff distance h is given by

$$h(ls_i, ls_j) = \max_{a \in ls_i} \min_{b \in ls_j} d(a, b)$$

(d is Euclidean distance between two spatial points a and b). Then, we normalize the distance with the max length of diagonal lines. Finally we multiply a cosine of the angle between two segments. If two movements have opposite directions to each other, they are not similar although their shapes are almost same. The function *Sim* returns a normalized real value from 0 (least similar) to 1 (most similar). A pair whose similarity measurement is greater than a threshold value becomes an instance of the *SIMILAR* relation.

In the case of *NEAR* relation, we use *MaxDist* function that returns the distance meter between two movements as

follows:

$$D(m_i, m_j) = \begin{cases} \infty & \text{if } T = \varnothing \\ MaxDist(m_i(T), m_j(T)) & \text{otherwise} \end{cases}$$

where $T = T_{m_i} \cap T_{m_j}$ is the common time interval between two movements and $m(T)$ is the slice of a movement within T interval. According to our definition, a continuous movement can be represented by linear functions of time t as $x = f_x(t) = at + b$ and $y = f_y(t) = ct + d$. Thus, we can evaluate the continuous distance between two slices by

$$MaxDist(m_i(T), m_j(T)) = \max_{t \in T} d(m_i(t), m_j(t))$$
$$= \max_{t \in T} \{\sqrt{(f_x(t) - g_x(t))^2 + (f_y(t) - g_y(t))^2}\}$$

where f and g is the linear function of movement m_i and m_j, respectively. After calculating the maximum distance between two movements, we select pairs within a threshold distance for the *NEAR* relation.

4. SOPHY FRAMEWORK

This section explains the system architecture of Sophy framework as shown in Figure 6. For the implementation of three components (a distributed real-time processing engine, a database server, and a visual data browser), the framework is based on Storm [3] system to handle real-time streaming geo-tweets, Neo4j graph database [1] to maintain geomorphology features and their relationships, and R Shiny [2] to present the spatiotemporal proximity and similarity of analysis results. The reason why we use a graph database is to provide a flexible way for handling associative data sets and basic operations for solving relationship-based problems, such as centrality analysis, path analysis, and isomorphism.

The real-time processing engine is composed of three modules: FormExtrator, FormRealtionConnector, and DerivedRelationHandler as follows:

FormExtrator The process starts by collecting from geotagged tweets via Twitter stream API in *TweetSpout*. In the *TweetSpout*, a geo-tweet is transformed into an observation instance with four properties: user, timestamp, geolocation, and content text except URL and user names. The observations are inserted into the database by *ObservationBolt* and passed to *LanguageMorphologyBolt* that generates keyword occurrences from one text message by assigning a category (e.g., noun, verb, adjective, etc.) to each word in the content of an observation. The emitted keyword occurrences from *LanguageMorphologyBolt* become input streams of two bolts that carry out a stream-to-relation operation with a sliding window: *GridIntensityBolt* and *SlidingSummaryBolt*. A window size can be set by the time duration or tuple capacity. In this study, we use a count-based sliding window that contains the last N items per each keyword in the *GridIntensityBolt*. By fixing the number of keyword occurrences, we can capture the burst with the information diffusion with shorter time duration. The *GridIntensityBolt* handles keyword occurrences with individual grid buffer per each keyword. The *SlidingSummaryBolt* is adapted to a time-based window that retains the last items observed during the last period of time T. Every time unit T the *SlidingSummaryBolt* refreshes term frequency in whole study area and inserts the summary of keyword frequencies per each cell into a database. The stored summaries are used

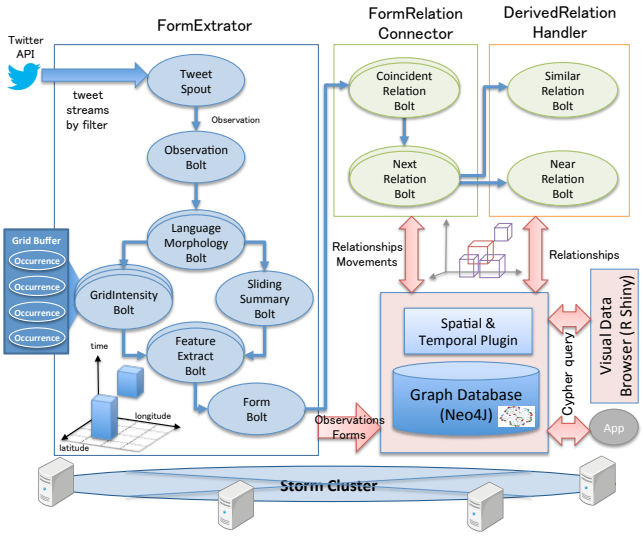


Figure 6: Sophy system architecture

for controlling different time windows of grid buffers. These two bolts aggregate keyword occurrences with sliding windows and measure each component of Eq.(4), i.e., a point intensity and an importance weight of each keyword with respect to the quadrats (grid cells) of the domain. After estimating each measure, they emit the values of cells into *FeatureExtractBolt*. The *FeatureExtractBolt* combines them into a locality value and extracts morphometric features as critical areas such as peaks and pits. The extracted features are transformed as instances of *Form* type with a pre-defined threshold and stored into the database in *FormBolt*. Finally, the features are going to *FormRelationConnector*.

FormRelationConnector It creates spatiotemporal relationships between *Form* instances. First, *CoincidentRelationBolt* selects candidates from the database by the spatial and temporal bounding box of an inserted instance and checks whether there is any intersection in both space and time. If there exists an intersection between two Form instances, they are connected by a *COINCIDENT* relationship with derived attributes such as a temporal relation pattern and the ratio of intersection as specified in Table 1. Then, the form instance is passed to *NextRelationBolt* that retrieves the nearest forms among the candidates inserted at the previous sliding window for the *NEXT* relation. If a *NEXT* relationship is defined, *NextRelationBolt* generates a movement instance and emits it to *DerivedRelationHandler*. All created relationships and movements are also stored in a database by each bolt.

DerivedRelationHandler There are two bolts to handle movement relations: *SimilarRelationBolt* and *NearRelationBolt*. According to our definition in Section 3, each bolt first searches a set of candidate movements that have a temporal intersection with respect to the new movement. Then they compute a measure between two movements by using their distance function. The *SimilarRelationBolt* has a predefined threshold when it is deployed in the Storm Cluster, and *NearRelationBolt* has an arbitrary number of k-nearest neighbors. Also the relationships have attributes about their measurements such as similarity measure and distance. We can use these attributes to make a ranked list of result sets without the recalculation. Two bolts also insert the rela-

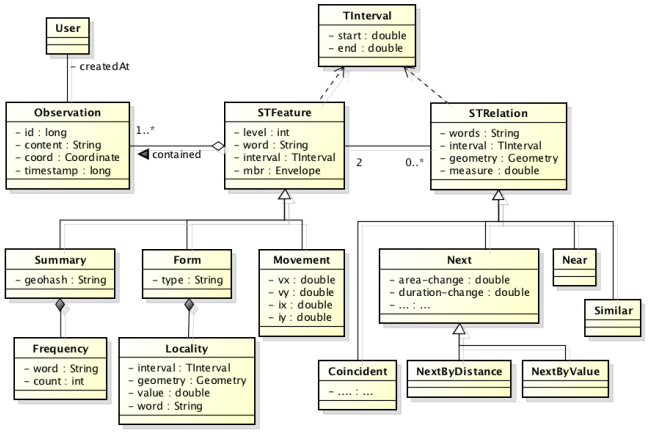


Figure 7: Data model of Sophy framework

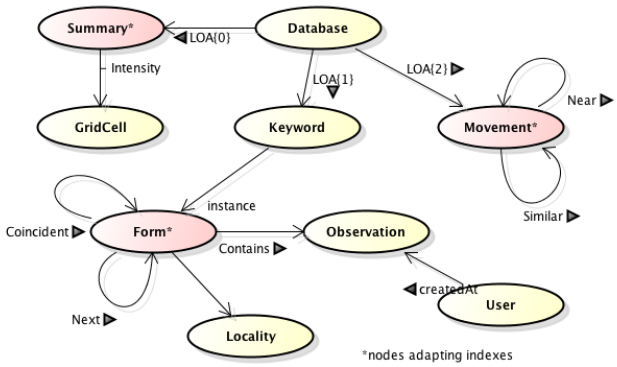


Figure 8: Graph structure in a database

tionships they found into the database through internal interfaces.

Figure 7 shows a diagram of the data model of the Sophy framework. As we above mentioned, we build a database on the basis of a graph database management system. Hence we represent data types by using nodes, edges, and their properties for graph data structures. Basically, type *STFeature* is mapped into a node and type *STRelation* is represented as an edge (relationship). Figure 8 shows an example of the database schema that contain meta-data nodes. The database system requires index structures to efficiently retrieve features and relationships without loading the whole data. The Sophy framework provides an interval tree to index time intervals for the query processing of temporal predicates as well as adapting a spatial plugin to handle geometry data. However, we found too many and complex indexes bring lots of overhead in the insertion operation against streaming data. The performance improvement of the database server will be one of our future challenges.

The last component is a three-dimensional (2D for space and 1D for time) visual data browser to review the processing results of the real-time engine. In order to display the changes of spatiotemporal information, we use the space-time cube presentation from time geography, which was developed as a model of society considering constraints on human behaviour by Hägerstand[13] and used for analysing and simulating human movements and individual activities in space and time. In this study, we are also interested in the movements of social phenomena that are implied from a large amount of geo-tweets. Thus, the spatiotemporal 3D visualization helps us to watch how they change and develop before starting deep analysis. The visual browser translates a user command into a Cypher[3] query, the graph query language of Neo4j database, and draws 3D geometries or tag clouds as the query results. For example, a query to retrieve topic movements of 'typhoon' within a time interval can be expressed by the following statement.

```
start n=node:ti_movements('between:[2014-08-05 TO
2014-08-09]') where n.word='typhoon' return n
```

The browser only supports a few types of queries for our experiments, but another application can access our database server by using Cypher language and RestAPI.

5. EXPERIMENTS

In the experiments, we try to verify functionalities of the framework with small data sets filtered by keyword of interest, even if it was designed to handle real-time streams via Twitter APIs. We prepared test files for the simulation in order to use a filespout that emits a record one by one from a file. Table 2 describes the data sets related to three disaster events: two typhoons ('Neoguri[4]' and 'Halong[5]') and a landslide ('Hiroshima[6]'). We can easily infer they are correlated with a heavy rain and bring small or big damages in physical and social infrastructures. For our experiments, we chose a few keywords such as 'heavy rain', 'flood', 'damage', and 'worry' to observe the following issues:

- Spatiotemporal proximity: The proposed measurement of locality represents the contribution of local communities against an event. Thus, we expect a certain geomorphology feature may reflect a local situation. For example, we can find more numbers or the highest/lowest of features near to physical events. We first look at how social communities act against disaster events that affect different geo-locations at different times and how they are connected to each other.
- Comparison of topic movements: We think that people publish their observations or opinions about their surrounding situation and these social phenomena follow the spatiotemporal changes of physical events. Consequently, movements of a topic in social media may have similar patterns to physical movements or other associated topics. We try to discover a movement pattern like flocks and breakups among the movements we generate.

For starting up the real-time processing, Sophy framework needs to be configured by parameter values such as the size/interval of sliding windows, a size of grid buffers, filtering thresholds, and so on. Depending on the parameter values, the relation structures and the system performance are influenced. However, this study remains an optimization problem of parameters for the future work and exemplifies processing results with arbitrary values as shown in Table 3. The following graphs grab snapshots of the processing results via our data browser.

[3] http://neo4j.com/guides/basic-cypher/
[4] http://en.wikipedia.org/wiki/Typhoon_Neoguri_(2014)
[5] http://en.wikipedia.org/wiki/Typhoon_Halong_(2014)
[6] http://en.wikipedia.org/wiki/2014_Hiroshima_landslides

Table 2: Data sets (Japanese tweets in Japan area)

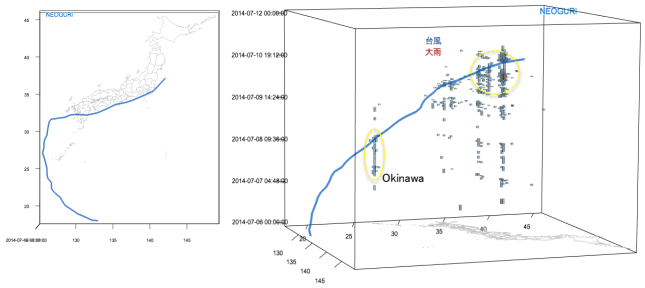
Description	NEOGURI	HALONG	HIROSIMA
disaster	typhoon	typhoon	landslide
physical peaks	2014/07/07-2014/07/09	2014/08/07-2014/08/09	2014/08/20 (4am-6am)
data time interval	2014/07/05-2014/07/13	2014/08/05-2014/08/13	2014/08/16-2014/08/30
common words	土砂災害(landslide),土砂崩れ(landslide) 大雨(heavy rain), 洪水(flood), 被害(damage), 心配(worry), 大変(tough)		
specific words	台風(typhoon)		広島 (Hiroshima)
# of tweets	45890	43945	41917

Table 3: Parameter setting of Sophy Framework

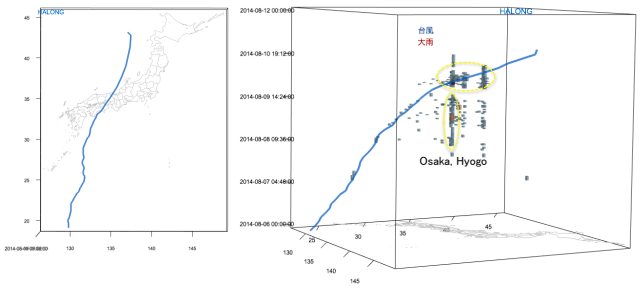
Description	Value
A size of grid buffer (km^2)	39.1x19.5
A size of count-based sliding window	300
A duration of time-based sliding window	3 hours
A threshold of similarity	0.4
A bounding distance (km)	100
POS tags of language morphology	noun
Types of geomorphology features	peak

Figure 9 shows the spatiotemporal proximity between physical typhoons and peak features of two datasets of geo-tweets. While typhoon 'Neoguri' whipped Okinawa with heavy wind and torrential rain, typhoon 'Halong' crossed the Japanese mainland and brought drenching rainfall and destructive wind to the country. In the figure, we found that social media react not only physical events but also the forecast information from the Japan Meteorological Agency (JMA). For example the development of peak clusters appears around Okinawa, Osaka and Hyogo areas before typhoon strikes. Even though not all keywords showed the physical proximity, some words to represent physical phenomena like 'typhoon' proved a strong spatiotemporal proximity with respect to real-world events. Table 4 lists the keyword occurrences of peak features derived from three data sets. We can see those terms sketch the spatiotemporal and thematic information, e.g., typhoon 'Halong' brought more wind damages ('竜巻' and '暴風') comparing to typhoon 'Neoguri'. Landslides ('土砂崩れ') happened and injured people ('被害者') in the Japanese city of Hiroshima without reading all messages.

Figure 10 shows COINCIDENT relationships among peaks of keyword 'damage(被害)', 'landslide(土砂災害)', and 'flood(洪水)'. The three-dimensional (3D) heat maps present the spatiotemporal intersections between two features and the strength of color indicates the ratio of the interaction portions. In our browser, a COINCIDENT relationship that has a larger spatial area and a smaller temporal interval of the intersection is presented by a stronger color. In addition, the browser helps us to explore which keywords are associated with each other in space and time. According to our results, the keywords of 'heavy rain(大雨)', 'landslide(土砂災害)', 'flood(洪水)', 'strong wind(強風)' and 'thunder(雷)' are usually accompanied by the watch(注意報) and warning(警報) information from JMA. As a result, they appear in close proximity. If an unfamiliar keyword is found in the word cloud, it may be considered as a special situation among



(a) NEOGURI [2014/07/05-2014/07/11]



(b) HALONG [2014/08/05-2014/08/11]

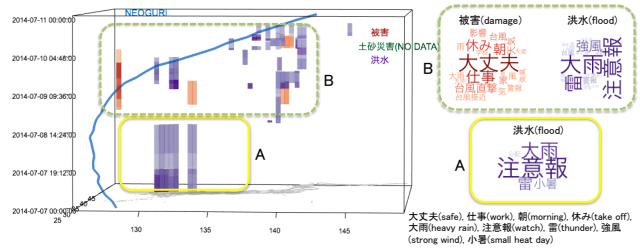
Figure 9: Spatiotemporal proximity between physical events and peak features ($locality \geq 0.02$) of keyword 'typhoon(台風)' and 'heavy rain(大雨)'

local communities. For example, keyword 'SORATOMO' shows up in Hiroshima data (Figure 10 (c)). This tag has been used for reporting weather and tourist information with geo-locations. After Hiroshima landslide, many communities use this tag to send a message of compassion for local people. Further, we found the spatiotemporal closeness among features of keyword 'damage(被害)', 'worry(心配)', 'tough(大変)' and 'safe(大丈夫)' during the disaster. It may testify the social media play an important role to check in with family and friends, seek emotional support and healing, and report situation about disaster areas by citizen.

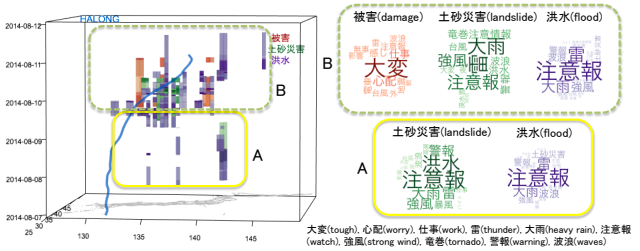
Next, we compare the movements of topics and attempt to discover a flock pattern among them. Figure 11 illustrates the trajectories of keyword 'typhoon(台風)' and 'worry(心

Table 4: Comparison of keyword occurrences of peak features

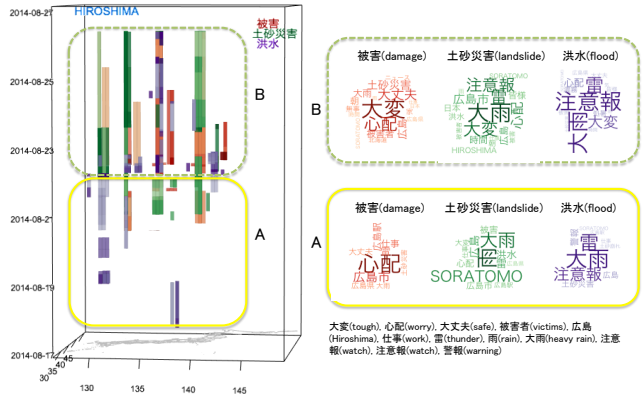
Description	Spatial terms	Temporal terms	Thematic terms
New common to appear	大阪, 東京, 日本, 家	朝, 夜	地震, 元気, 浸水害, 濃霧, 天気, 警報, 仕事, 気, 雨, 注意報, 雷
NEOGURI (only)	沖縄, 九州, 関東, 学校	夏, 小暑	普通, 期待, 接近, 目, 直撃, 休校, NEOGURI
HALONG (only)	愛知県, 大阪府, 青森県, 海	10日, 暇, 立秋	楽しみ, 台風情報, 高温注意情報, 竜巻注意情報, 暴風, HALONG, 予定, 中止
HIROSIMA (only)	廿日市市, 呉市, 京都, 福山市, HIROSHIMA	最後, 処暑	応援, 好き, 印象, 被害者, 心, 見舞い, SORATOMO, 土砂崩れ



(a) NEOGURI [2014/07/07-2014/07/11]



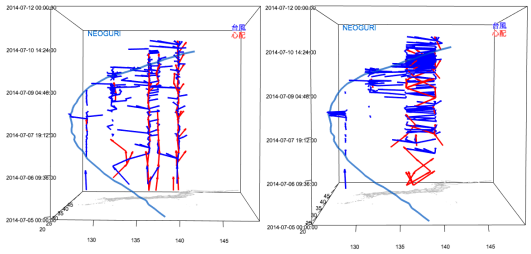
(b) HALONG [2014/08/07-2014/08/12]



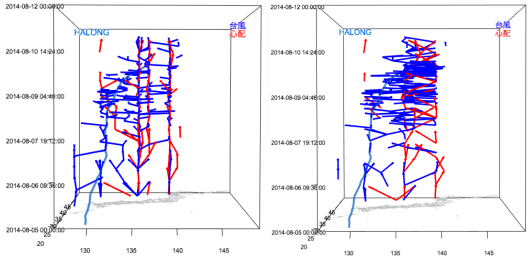
(c) HIROSHIMA [2014/08/17-2014/08/27]

Figure 10: Coincident keywords based on spatiotemporal intersections ($ratio_{interarea} \geq 0.5$) among peak features ($locality > 0$) of keyword 'damage(被害)', 'landslide(土砂災害)', and 'flood(洪水)'

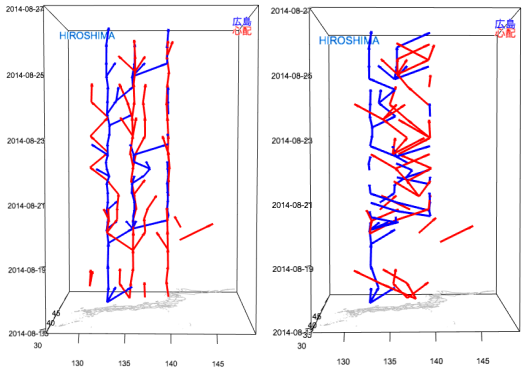
配)' generated by the NEXTBYDISTANCE and NEXTBYVALUE relationships. As shown in the figure we obtained different trends of trajectories regarding its measure function and some features have both relationships. Although a sophisticated method will be required to combine two measures for defining a NEXT relationship in future, we think the distance-based connection is more appropriate to guarantee the spatiotemporal proximity of movements. An interesting thing in Figure 11 (c) is to show a similar pattern between 'Hiroshima' and 'worry' movements at a certain times. In order to find a similar pattern, we finally look at the SIMILAR and NEAR relationship between generated movements. Figure 12 (a) shows the fragment information of movements that concurrently have both relationship SIMILAR and NEAR in Halong and Hiroshima datasets. We suppose these information can contribute to discover complex spatiotemporal patterns of movements such as flock patterns. In addition, the movement relationship helps us to track a continuous proximity in a spatiotemporal domain. In our result, the relationships between 'worry(心配)' and 'landslide(土砂災害)' movements are closer than 'worry' and 'tough' as shown by Figure 12 (b).



(a) NEOGURI: typhoon(台 風) and worry(心 配) [2014/07/05-2014/07/12]



(b) HALONG: typhoon(台 風) and worry(心 配) [2014/08/05-2014/08/12]

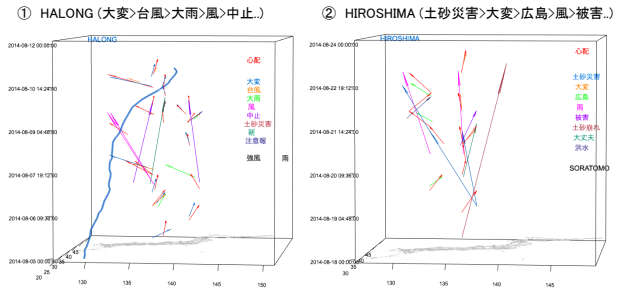


(c) HIROSHIMA: Hiroshima(広島) and worry(心配) [2014/08/17-2014/08/27]

Figure 11: Movements by NEXT relationships of peak features ($length_{segment} \leq 400km$): NEXTBYDISTANCE(left) and NEXTBYVALUE(right)

6. CONCLUSIONS

Location-based social media have provided us considerable information to be able to detect, track, and predicate dynamic events and situations in the real world; however, we have been struggled to understand the spatiotemporal dynamics from the mountains of fragmented, noisy data flooding today's social media. The main goal of this study is to construct spatiotemporal relations from geo-social media by using latent relationships in the real time. For that, we have proposed a geomorphology-based data model with the locality measurement and new spatiotemporal relationships. Moreover we have implemented Sophy framework to support our data model on the top of Storm platform and Neo4j graph database. In the experiments, we showed the functionalities of the framework with real tweet-sets related



(a) Movement relation

(b) Coincident relation

Figure 12: Comparison of keyword ranks with respect to keyword 'worry(心配)' based on spatiotemporal relationships: tough(大変), typhoon(台風), heavy rain(大雨), wind(風), stop(中止), landslide(土砂災害), Hiroshima(広島), damage(被害), safe(大丈夫)

to three disaster events in 2014. In particular, we investigated the spatiotemporal proximity of geomorphology features and the similarity of topic movements in social media. This work is just the first step to analyze complex spatiotemporal patterns such as flocks and breakups in social media. Through our experiments, we found the performance overheads of the database server and visualization of a large amount of data comparing to the processing engine. Moreover, there are many challenges to improve our framework such as parameter optimization, relationship computation, and pattern discovery for our future study.

Acknowledgements

This work is partly supported by JSPS KAKENHI Grant Number 24240015. The authors thank to Dr. Ryong Lee for his comments to improve this paper.

7. REFERENCES

[1] Neo4j - The World's Leading Graph Database. `http://www.neo4j.org/`.

[2] shiny: Easy web applications in R. `http://www.rstudio.com/shiny/`.

[3] Strom - Distributed and fault-tolenrant realtime computation. `https://storm.incubator.apache.org/`.

[4] L. Alex. *Key concepts in classical social theory*. SAGE Publications Ltd, London, 2011.

[5] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–848, 1983.

[6] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 2013.

[7] P. J. Diggle. Spatio-temporal point processes: Methods and applications. `http://biostats.bepress.com/jhubiostat/paper78`, June 2005.

[8] M. J. Egenhofer. Reasoning about binary topological relations. In *Proc. of the Second International Symposium on Advances in Spatial Databases*, pages 143–160, 1991.

[9] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61(0):115–125, 2014.

[10] J. Gomide, A. Veloso, W. Meira, Jr., V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, pages 3:1–3:8, 2011.

[11] J. Gudmundsson, P. Laube, and T. Wolle. Movement patterns in spatio-temporal data. pages 726 – 732, 2008.

[12] R. H. Güting, M. H. Böhlen, M. Erwig, C. S. Jensen, N. A. Lorentzos, M. Schneider, and M. Vazirgiannis. A foundation for representing and querying moving objects. *ACM Trans. Database Syst.*, 25(1):1–42, March 2000.

[13] T. Hagerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24:7–21, 1975.

[14] E. Jones and J. Eyles. *An Introduction to Social Geography*. Oxford University Press, 1977.

[15] P. Laube, M. van Kreveld, and S. Imfeld. Finding remo-detecting relative motion patterns in geospatial lifelines. In *Developments in Spatial Data Handling*, pages 201–215. Springer Berlin Heidelberg, 2005.

[16] R. Lee, S. Wakamiya, and K. Sumiya. Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal and Ubiquitous Computing*, 17(4):605–620, 2013.

[17] J. Li and C. Cardie. Early stage influenza detection from twitter. *CoRR*, abs/1309.7340, 2013.

[18] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proc. of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 1–8, 2011.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

[20] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in Tweets. In *Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, 2009.

[21] T. Shelton, A. Poorthuis, M. Graham, and M. Zook. Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum*, 52(0):167–179, 2014.

[22] X. Yu and M. K. Leung. Shape recognition using curve segment hausdorff distance. *Pattern Recognition, International Conference on*, 3:441–444, 2006.

[23] M. Yuan. Temporal gis and spatio-temporal modeling. `http://ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/sf_papers/yuan_may/may.html`, 2004.

[24] X. Zhou, S. Shekhar, and R. Y. Ali. Spatiotemporal change footprint pattern discovery: an inter-disciplinary survey. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, pages 1–23, 2014.