Portraying Collective Spatial Attention in Twitter

Émilien Antoine
Kyoto Sangyo University
emilien.antoine@cc.kyoto-su.ac.jp

Adam Jatowt
Kyoto University
adam@dl.kuis.kyoto-u.ac.jp

Shoko Wakamiya
Kyoto Sangyo University
shokow@cc.kyoto-su.ac.jp

Yukiko Kawai
Kyoto Sangyo University
kawai@cc.kyoto-su.ac.jp

Toyokazu Akiyama
Kyoto Sangyo University
akiyama@cc.kyoto-su.ac.jp

ABSTRACT

Microblogging platforms such as Twitter have been recently frequently used for detecting real-time events. The spatial component, as reflected by user location, usually plays a key role in such systems. However, an often neglected source of spatial information are location mentions expressed in tweet contents. In this paper we demonstrate a novel visualization system for analyzing how Twitter users collectively talk about space and for uncovering correlations between geographical locations of Twitter users and the locations they tweet about. Our exploratory analysis is based on the development of a model of spatial information extraction and representation that allows building effective visual analytics framework for large scale datasets. We show visualization results based on half a year long dataset of Japanese tweets and a four months long collection of tweets from USA. The proposed system allows observing many space related aspects of tweet messages including the average scope of *spatial attention* of social media users and variances in spatial interest over time. The analytical framework we provide and the findings we outline can be valuable for scientists from diverse research areas and for any users interested in geographical and social aspects of shared online data.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous

Keywords

Location Mention; Twitter; Social Network; Spatial Analysis; Visualization

1. INTRODUCTION

Twitter and other social media are frequently used to express opinions or to share reports of daily life activities. They have been also a common target for various kinds of analyses including detecting real-world events [32] or investigating the process of information spread among users [7]. Utilizing spatial information has been of particular interest in these studies [2, 17] as it is possible to estimate locations from where tweets originate. Due to high immediacy of tweets, timely detection of local events has become feasible by detecting spikes of similar tweets coming from same neighborhoods [33]. What has been however often neglected in spatial-focused analysis of microblogging, is the usage of location mentions within tweet content. These may either relate to an area where some event happens or to the locality of personal concern of a Twitter user. A user may for example mention location names in tweets to express her travel plans, to reminisce places she visited, to discuss current events or just to refer to some interesting places.

In this paper we introduce the concept of *collective spatial attention*. We define *spatial attention* as the geographic area of interest and focus of an online user. When treated collectively, the spatial attention of many interacting users forms collective signal that can assume different types of patterns. Users at the same time may *align* their attention towards the same locations due to the occurence of sudden events, shifts in calendar seasons or due to population migrations. Note that the concept of collective spatial attention is orthogonal to the topical and temporal attention [15] that have been recently studied on large collections of user-generated content such as Twitter datasets. Analyzing collective spatial attention offers information complementary to the standard social media analysis and should help us in better characterizing spatial aspects in online media and its dynamics.

We then propose a dedicated visual analytics system that extracts location references from large datasets of messages and portrays them en masse to investigate and attract collective spatial attention. For example, it is possible to compare opinions about particular place expressed by nearby or far away users. Or, in another example, one can find location pairs connected by common spatial attention of tweeting users. In the design process, we have also considered the temporal factor with the aim to track and understand variations of spatial attention over time. Thus, the visualization framework we create allows not only drawing horizons of spatial references but also empowers users to detect temporal variances of interesting patterns related to places.

Since spatial perception and thinking are strongly related to country geography and particular culture we study two

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783418.

countries that differ significantly in terms of size, shape, population and culture: USA and Japan. Our field of study is the portion of USA and Japanese tweets collected over the time frame of, respectively, 6 and 4 months in 2013. Using the visualizations on the provided datasets we exhibit several interesting findings that shed new light on the characteristics of spatial attention of social media users in both countries. Although many of the findings are common to both datasets, we notice several differences. The system is available online [37, 38] for anyone interested in quick overview or in detailed exploration of spatial aspects in shared messages.

We would like to emphasize that tweets tend to be noisy and lack comprehensive context for small scale analysis. Even if nowadays natural language processing techniques allow tagging spatial expressions, still, extracting information of location mentions, disambiguating and mapping them precisely are difficult and prone to errors. On the other hand, aggregating multiple spatial expressions from numerous tweets should offer more trustworthy view of common spatial thinking of online users. The insights about the character of collective spatial attention could then assist us in better understanding spatial references within individual tweets.

The reminder of the paper is structured as follows. After reviewing the prior literature in Section 2, we introduce our datasets and outline data models in Section 3. This section also contains the general overview of the visualization systems we use, although, some of their technical details are deferred to Section 5. Section 4 is the key part of this paper and contains the overview of our findings. We conclude the paper in Section 6.

2. RELATED WORK

The rise of online social networks makes it easy to collect large amounts of data on human behavior, characteristics and social connections. Twitter data analytics is then frequently used in social sciences utilizing numerous traces of daily life activities left by users [34, 17]. Extensive work has been carried so far to exploit large collections of personal messages crawled from Twitter as shown in this survey [12]. The heterogeneous nature of topics discussed in Twitter provides valuable data to perform large scale analysis of societal interests, particularly, from spatial viewpoint which is the focus of this paper.

In this work we focus on spatial factors in microblogging. Studying spatial aspects is quite important. In fact, daily communication about the world seems to revolve around the *space*. This becomes understandable when considering that the global news media mentions any location every 200-300 words, thus, more than any other information type [18]. Even access to information is largely associated with spatial attributes as over a quarter of web searches contain geographic terms and 13% of all web searches have geographic character [11].

Twitter presents an effective playground for socio-spatial studies. Since early 2010 explicit GPS tags can be specified for tweets, thus resulting in a large dataset of messages with location stamps. However, not every message has an explicit location stamp. Hence, location inference has become one of the most prolific domains of study, either done by mining the social graph [3, 9] to utilize the location of friends, or by parsing textual tweet content as in [8]. The latter work introduced a concept of *local words*, later extended in [16, 6], and used them to infer home locations of Twitter users.

This last approach also leads to the application of inference methods based on topic models in [10, 14, 36]. Statistical models allow inferring home location from the user history, tags and other attributes as shown in [13, 27] and can even serve for predicting user movements [9]. In [21, 20, 31] both social graph and content processing are used to infer fine-grained user locations, even when the users choose to keep their location information private. All these works demonstrate that it is possible to determine user location using public data shared on social networks, usually, on Twitter.

Prior research has also emphasized the usefulness of social networks and especially Twitter for extraction of real-time information by detecting events [35, 30, 19]. Also, studies in [23, 1, 32] provide a suite of visualization capabilities to explore tweets from different dimensions relating to specific application like fire combat and earthquake detection. More generic platforms have been designed with visualization tools to analyze data in spatial perspectives [26, 24, 25].

To the best of our knowledge, there are no similar researches nor visualization systems that would collectively portray location mentions in microblogs, analyze their usage, characteristics, relations and dynamics in order to provide new kind of spatial knowledge. The current work thus provides a novel type of spatial analysis, which, as we believe, can complement the above mentioned studies.

3. DATA MODEL

We are particularly interested in space-referring tweet messages from which we can extract spatial expressions. The spatial expressions allow us to categorize tweets into those about far away locations and those about close locations as well as to locate them on a map after prior disambiguation. We will describe the datasets we use in Section 4.1. The following details our data model based on space-related information extraction from tweets.

3.1 Data Structure

Each tweet is first represented as a tuple of 4 attributes as follows:

$< user\ id,\ tweet\ content,\ timestamp,\ location\ stamp >$

The *user id* is the Twitter identification number, the *tweet content* is the text of the tweet, the *timestamp* is the time when the tweet has been published on Twitter, and the *location stamp* is the GPS coordinates from where the tweet has been posted. All the tweets crawled to build our datasets have *location stamps* provided by tweeting users. Note that although the GPS coordinates might be accurate when someone posts on Twitter via a smartphone using the official Twitter application, they might be rather approximate for a traditional web user posting from a PC. That is why in this paper we are interested only in the distances greater than 1km. Based on the above listed attributes, the information on location is extracted and represented by the two additional attributes computed from the basic attributes:

$< location\ mention,\ location\ diff >$

The *location mentions* are extracted and disambiguated from the *tweet content* by applying entity-recognition techniques (described in Section 5.3). The *location diff* is computed for each *location mention* as the euclidean distance between the *location mention* and the *location stamp*. In the rest of the paper, we will use the term *spatial expression* to name any spatial annotation given by the NLP parser

containing one or few consecutive words. From the *spatial expressions* our system infers only one *location mention* for each tweet. This choice is discussed in Section 5.3.

The above data model is used to contrast two key spatial attributes which characterize tweets in order to represent novel kind of information about the global and per user behavior of microblogging users. Intuitively, the attribute sextuple should allow answering simple but fundamental questions of the following type:

- What do people at the same places tend to say? (*location stamp*)
- What do people tend to say about the same places? (*location mention*)
- What do people tend to say about places located at the same distance from them? (*location diff*)

3.2 Space Mention Extraction

We consider only tweets with *spatial expressions* that could have been extracted. Note that while there are ready spatial taggers for English, we are unaware of any credible tool for extracting and mapping *spatial expressions* in Japanese language. Therefore, for English we use the Stanford CoreNLP tagger [22], while for the Japanese dataset we do the locations' extraction from text by ourselves.

We consider only *location mentions* that correspond to a unique specific place or area identified and localizable in space by GPS coordinates. For instance, the *spatial expression* "at the University" without any further context cannot be disambiguated and thus no *location mention* will be found, whereas the expression "at Kyoto University" will be disambiguated.

3.3 Snapshots of Visualizations

The objective is to enable users to see and reason about the *spatial attention* of tweeting users based on:

- the *location diff* and *location stamp* (Figure 1)
- the *location diff* and *time stamp* (Figure 2 and 3)
- the *location mention* and *location stamp* (Figure 4 and 5)
- the *location stamp* and *location mention* on the USA geographic map (Figure 6(a) and 6(b))

The following visualizations are 2D plots with colored cells in the form of a heat map. The cell color represents the *intensity* with which tweets in our dataset refer to that cell. According to the location attributes of a tweet t, the cell $C_{i,j}$ has a probability $P(C_{i,j}|t)$ such that $t \in C_{i,j}$. Intuitively, $P(C_{i,j}|t)$ is the amount of its probability mass that the tweet assigns to the cell. The probability depends on the mapping of the *location mention* from natural language to a computable value, usually, GPS coordinates of an area or a point. The *intensity* of a cell $C_{i,j}$ is the sum of probability of all the tweets in the dataset T that refer to $C_{i,j}$, that is, $I(C_{i,j}) = \sum P(C_{i,j}|t) : \forall t \in T$.

4. VISUAL DATA ANALYTICS

This section first describes our visualizations and design choices we made. We then discuss findings we could obtain.

All the visualizations we used are shown in 2D panes in the form of heat maps with colored cells. Cells have different meaning in different graphs as it will be discussed in the following subsections. The color of each cell is selected based on the *intensity* with which tweets in our dataset refer to the particular cell. Below each graph, we display the color scale ranging from dark blue (the lowest *intensity*) to dark red (the highest *intensity*). From the set of tweets associated to each cell, the system computes the most representative words among the tweet contents. Computation is described in Section 5.2. The top-30 representative words are displayed in pop-up window when a mouse passes over any segment. Moreover, for receiving more details the user can click on any cell or segment to open a separate window with the top 100 representative words with their scores and the exhaustive list of tweets in that cell or segment. In Figure 1, 2 and 3, the gray buttons on top and right sides of the graphs relate to the column and line aggregates of cells, respectively. All the graphs shown in this section are built from the same datasets described in Section 4.1.

4.1 Experimental Datasets

We first report the overall statistics of the data we use. The Japanese dataset has been built retrieving 31.6M (millions) tweets posted from Japan between July 21, 2013 and January 12, 2014. The USA dataset contains 198M tweets created between September 25, 2013 and January 17, 2014. Unfortunately, due to technical limitations the data crawling was disrupted at certain times. This explains the blank sections in Figure 2(a) from August 11 to 14, October 13 to 17, December 13 to 15 and December 25 to 28, and the two blank sections in Figure 2(b) from October 11 to 29 and December 13 to 27.

We applied a preprocessing step that removed all the tweets not in Japanese or English using the language detection method based on Naive Bayesian filter (found to work with nearly 99% precision[1]). As a result, the dataset contains roughly 25M tweets written in Japanese in the Japanese dataset and 158M written in English from the USA dataset. Next, from these data we singled out all those tweets that mention geographic locations. For the USA dataset 30M tweets were annotated with spatial expressions by Stanford CoreNLP tagger and among them we consider 4.3M that where successfully disambiguated by our system and situated in the USA. For the Japanese dataset 684K tweets have been found to contain location mention based on the list of the names of the 47 prefectures. The users tweeting with location mentions that were considered in our study represent 22% of the total number of Japanese users and 28% of the USA users within the dataset.

4.2 Distribution of Spatial References

The first issue we investigate is the horizon of spatial attention of microblogging users. Since there are many locations from which tweets could be written and/or to which they could refer, plotting all the combinations in one graph would not be very efficient. We then visualize spatial attention through aggregating tweets based on the differences of their location mentions to location stamps.

As shown in [4, 9] users are mainly interested in close locations around their home and work. For an average user, the majority of her daily life events or even weekend trips is expected to be within rather small distance from her usual place of accommodation. Even during the trips to far away

[1] `http://code.google.com/p/language-detection`

locations users are expected to tweet more about places they are actually visiting rather than some other places. This would then suggest rather short span of spatial attention. However, in the globalized world, information on events occurring in far places is easily and quickly reachable. In addition, it is relatively easy now to travel and migrate these days. Thus, we could expect that these factors would rather stretch the average distance of the spatial attention of users.

To shed more light on these aspects, we provide our first visualization demonstrated in Figure 1. It portrays the average *location diff* of tweets depending on their *location stamps*. For the ease of analysis, the tweets are naturally clustered by the prefectures for Japan or states for USA, from where they originate. Therefore, the horizontal axis represents the prefectures/states which include disambiguated *location stamp* attributes.

Note that it would be difficult to visualize the spatial attention using the linear scale. To show all possible distances including far away ones the vertical axis would have to be stretched thus making the data about close neighborhoods difficult to see and analyze. Therefore, the *location diff* is given in logarithmic scale in ordinate in order to visually portray data over wide range of space. The choice of logarithmic scale was also driven by the study [5] showing that users tend to switch to a larger spatial granularity when referring to far locations.

At the bottom of each graph, the blue histogram shows the numbers of tweets issued from each prefecture/state. As the numbers are used for normalizing data we display them for reference in each corresponding column. In addition, for the ease of analysis we show the blue lines on the right hand sides of both graphs in Figure 1. They correspond to the aggregate location differences over the total amount of tweets. These lines thus display the aggregate distribution of location differences within the whole dataset.

Another design issue relates to ordering of states or prefectures on the horizontal axis. These should be arranged in order to better capture proximity-driven spatial patterns following the intuitive reasoning that nearby locations exhibit many commonalities. We use here the standard region division for both countries to group the corresponding administrative regions (i.e., prefectures or states).

As mentioned above, we consider the data below 1 km for *location diff* as not relevant due to the inherent imprecision of the GPS coordinates retrieved via Twitter. All such tweets are then collapsed and displayed as a single cell in each column right above the horizontal axis. Above the 1km limit we can observe that, on average, the attention of users seems to have some correlation with geographical distance. For the Japanese dataset (Fig. 1(a)), the *spatial attention* falls within the range of 50 to 2000 kilometers. Note that the strict lower bound in the spatial attention in between 10 to 50 km for Japan is a consequence of the prefecture granularity we adopted. Comparatively, in the USA dataset in Figure 1(b), the *spatial attention* covers the whole range with a strong red bottom line, indicating that much of the Twitter attention is local, following the assumption that people tweet about their direct environment as shown in [32]. Next we can observe two main landmarks, the first between 5 to 10km, and the second between 500 to 3000km (see the blue line on the right hand side of Figure 1(b) as evidence). By analyzing the top-words in the corresponding horizontal segments, we found that the first is related to daily life activities as evidenced by representative words such as restaurant, grill, home, etc. The second seems to be related to national events with a clear domination of sport events and, in many cases, American Football. This confirms the notion of *local words* introduced in [8]. Our proposed visualization system can then allow studying, enriching and validating such hypotheses.

Next, we look for any isolation effect in our datasets. In the Japanese dataset only the two most distant prefectures, Hokkaido (the farthest north in Japan) and Okinawa (the farthest in the south) have any tweets about them that come from further away than 2000km (see the first and the last columns in Fig. 1(a)). Both are actually occupying whole islands. Similarly, the states of Hawaii and Alaska (see the two last columns in Fig. 1(b)) have a lower bound above 2000km. In addition, both these states have few or no tweets coming from less than 1000km away from them. The existence of isolated regions conforms to the topology of Japan and USA where these prefectures and states are separated and lie furthest from the main land mass of the countries.

4.3 Referring to Space over Time

The purpose of the next data view is to analyze how the *spatial attention* changes and evolves over time. We would expect some kind of calendar effect due to different activities and more free time that users have on average on weekends. Also, one could imagine the possibility of certain seasonal effect such as summer season being visibly different from autumn and winter.

To study this we use another visualization to display the changes in aggregated spatial references over time. The graphs in Figure 2, show the value of the *location diff* attribute for tweets arranged over time with daily granularity. The *location diff* values are given in ordinate in logarithmic scale and *time stamp* values are shown on the abscissa. Below the graph we also display the curve of the total number of tweets crawled (in blue) and the percentage of the tweets with location mentions (in orange) to compare the visualization on particular day with the total amount of data used to visualize that day. Note that the data are now portrayed on aggregate over all the states and prefectures unlike in Figure 1 where it was grouped according to country administrative divisions.

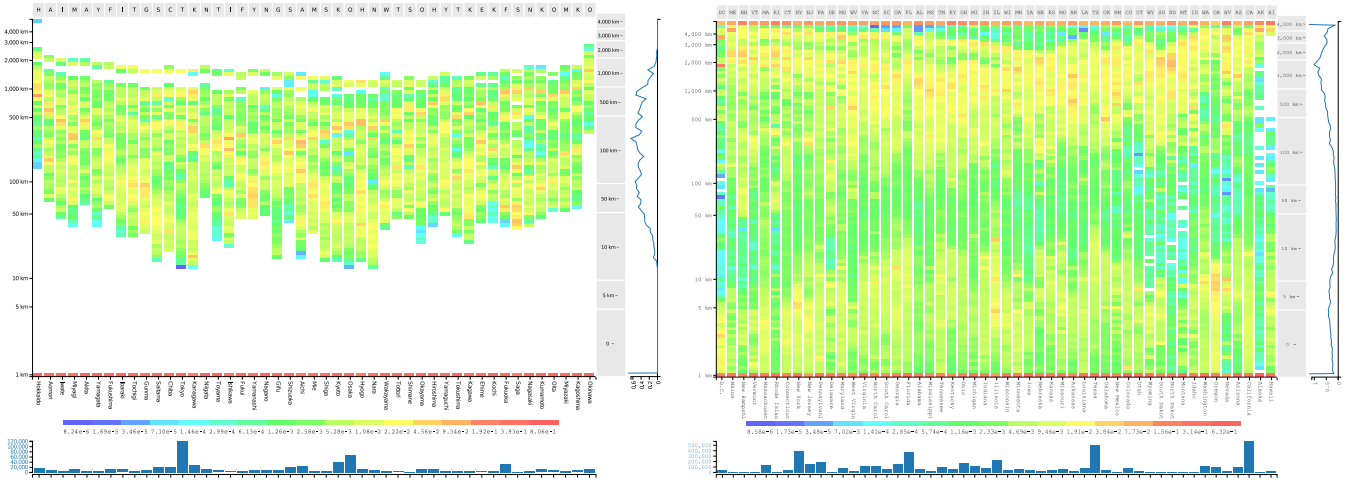
Metropolis effect.

The most striking pattern for the graphs in Figure 2 are the horizontal red lines all along the time period, especially, obvious in the Japanese graph. This is a consequence of the *spatial attention* between the populated metropolises.

For the Japanese dataset in Figure 2(a), three main lines are seen:

- around 390km for Tokyo↔Osaka, Kyoto, Kobe
- around 850 for Tokyo↔Fukuoka
- around 1,500km for Tokyo↔Okinawa

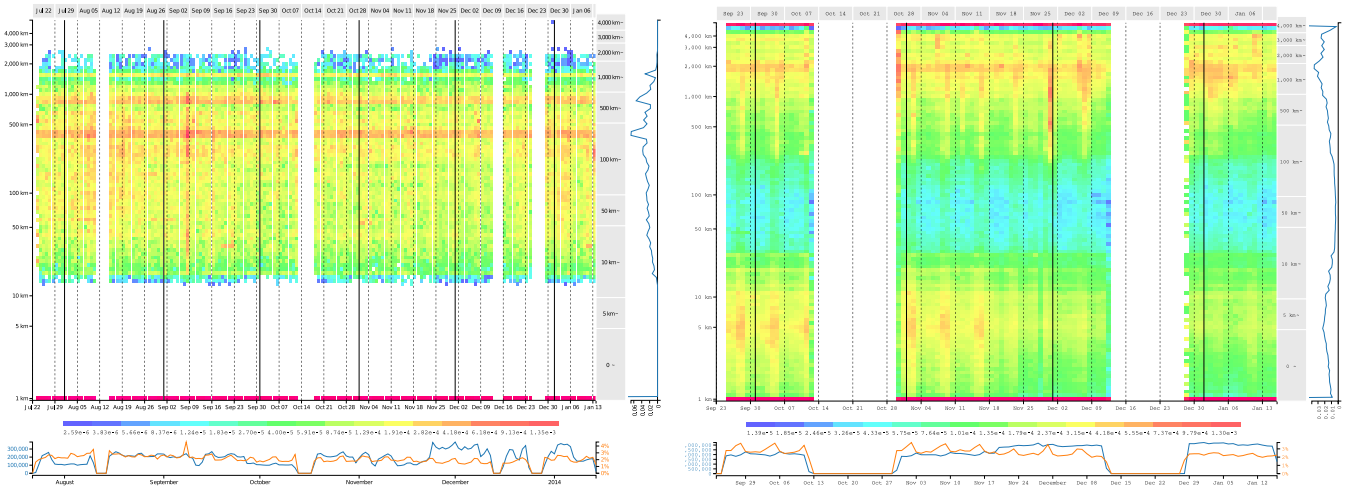
That is confirmed by the top words found in their respective horizontal segments. For the USA dataset in Figure 2(b), there is only one obvious horizontal line, which in fact is due to our mapping of the *location mention* of USA as a country that is mapped to the GPS coordinates of the middle point of continental USA.



(a) Snapshot of Japanese dataset available online [39]

(b) Snapshot of USA dataset available online [40]

Figure 1: Heat map of location differences per prefectures or states



(a) Snapshot of Japanese dataset available online [41]

(b) Snapshot of USA dataset available online [42]

Figure 2: Heat map of location differences over time with $1.5e^{-3}$ limit z-value

Weekend Pattern.

The *weekend pattern* (especially visible in the USA dataset) is characterized by the vertical lines of stronger attention in the two previous days just before the vertical dashed lines displayed to indicate week boundaries. These two days correspond to Saturdays and Sundays that are characterized by a periodic increase in the range of the *spatial attention*. The top representative words that we examined actually reveal the tendency of users to speak about their trips or travel plans during weekends, which explains why the *spatial attention* focused more on farther distance than that of the weekday activities (notice the longer horizontal yellow color lines on weekends).

Summer holiday pattern.

Especially visible in the Japanese dataset, during the summer period in August, the high ratio of tweets mentioning events between 10 to 900 kilometers is more constant throughout the weeks. The previously described weekend effect seems to fade out in this period since August is an important ceremonial and vacation period in Japan called "Obon" during which people often travel throughout the country and families tend to reunite. From the top words we confirm that users tend to talk about their vacation destinations or plans independently of the weekend period.

Place-based Filtering.

The visualizations shown in the graphs in Figure 2 allow also users to render the views based on the name of particular state or prefecture. Correspondingly, Figure 3 is the result of such selection that keeps only the tweets either originating from Tokyo or those mentioning Tokyo.

This allows spotting particular events linked to this location, such as the earthquake that occurred on August 8, 2013 in Nara and which caused concern among the people of Tokyo.

Figure 3: Heat map of location differences over time for Tokyo available online [39]

4.4 Origin and Destination of Space Mentions

The two Figures 4(a) and 4(b), available online at [43], contrast the origin (*location stamp*) of tweets with their spatial focuses (*location mention*). They show on the abscissa the proportion of tweets issued from a given place that mentions places as arranged in ordinate.

The data shown in Figure 4(a) and 5(a) are row-normalized. Thus, when looking at a given line l, one can see the probabilities that a tweet mentioning the place l may come from a place listed in column c. Intuitively, this shows the interest of users from a prefecture or a state c in the prefecture or a state l. Figure 4(b) and 5(b) portray the same data but in the column-normalized way, meaning that they visualizes the probabilities that a tweet from a place in column c mentions place at any line l. Intuitively, this data could be interpreted as the popularity of a prefecture or state l among users from a given prefecture or state c. The prefectures or states are ranked from left to right and top to down, following the order described previously in Section 4.2 to exhibit the effect of vicinity.

Diagonal. An immediate observation from both figures is the pronounced diagonal on the graphs which highlights the fact that people tweeting from a given prefecture or state are most likely to tweet about the same prefecture or state. This observation supports the concept of "here" dominating the spatial attention, similar to the concept of "now" in time-based analysis [15]. "Here and now" would then seem to be what Twitter users mainly care about.

Influential places. We can observe few places with a wide range of influence and high popularity. Tokyo and Osaka appear to receive relatively high amount of *spatial attention* due to their role as economical, transportation and political hubs. Figure 4(a) shows that tweets from Tokyo and Osaka significantly mention every prefecture. Conversely, Figure 4(b) demonstrates that every prefecture significantly mentions Tokyo and Osaka. The situation is similar in the USA dataset for New York, Florida, Texas and California as shown in Figure 5(a) and 5(b).

Attractive places. Some places continue to be mentioned despite having rather moderate or low population count and economical role in the country. Okinawa (pop: 1.3M, rank: 32th) and Hokkaido (pop: 5.7M, rank: 7th) are such popular prefectures about which people tend to tweet from different places, even from far away ones. This is because both islands actually belong to the most popular touristic places in Japan. This is shown in Figure 4(b) where both Okinawa and Hokkaido related lines are featured by numerous green cells. Note that in Figure 4(a) the columns of both Okinawa and Hokkaido are mostly blue showing a small interest in other prefectures than themselves.

Isolated places. On the other hand, isolated places are mentioned only by local or nearby located users. For example, looking at the line of Shimane on Figure 4(b), one can see that only Tottori and Yamaguchi, which are close neighbors, significantly mention Shimane prefecture. Similarly, USA states such as Mississippi, West Virginia, North Dakota, Rhode Island, and Delaware exhibit the same pattern.

Clusters of interconnected places. The last observation is the effect of geographical proximity and economical dependence on the *spatial attention* depicted by square of mutual attention in both column and row normalized graphs. For example, there are many tweets from and to Saitama, Chiba, Tokyo and Kanagawa since they are adjacent prefectures in the same metropolis. Also, a cluster such as Tokushma, Kagawa, Ehime and Kochi corresponds to the four prefecture constituting the Japanese island of Shikoku.

4.5 Geographic Distribution of Location Mentions

Lastly, we show the two Figures 6(a) and 6(b), available online at [45] and [46] in order to portray spatial attention on the actual USA map. Figure 6(a) colors cells depending on the amount of tweets with any location mentions that are issued from those cells. On the other hand, Figure 6(b) colors cells depending on how often they are mentioned in our datasets. We can notice that the two graphs are not equal. Location mentions are actually more clustered suggesting that while people tweet from large variety of places the amount of locations they mention is more limited than the places they tweet from. Note however that the red cells in the middle of each state correspond to the state granularity spatial references due to our disambiguation settings. The strong intensity in these cells could be actually distributed uniformly among all cells covering the given state area if more precise mapping technique is adopted. Note also that the graphs can be time adjusted so it is possible to see the visualizations for a given selected week.

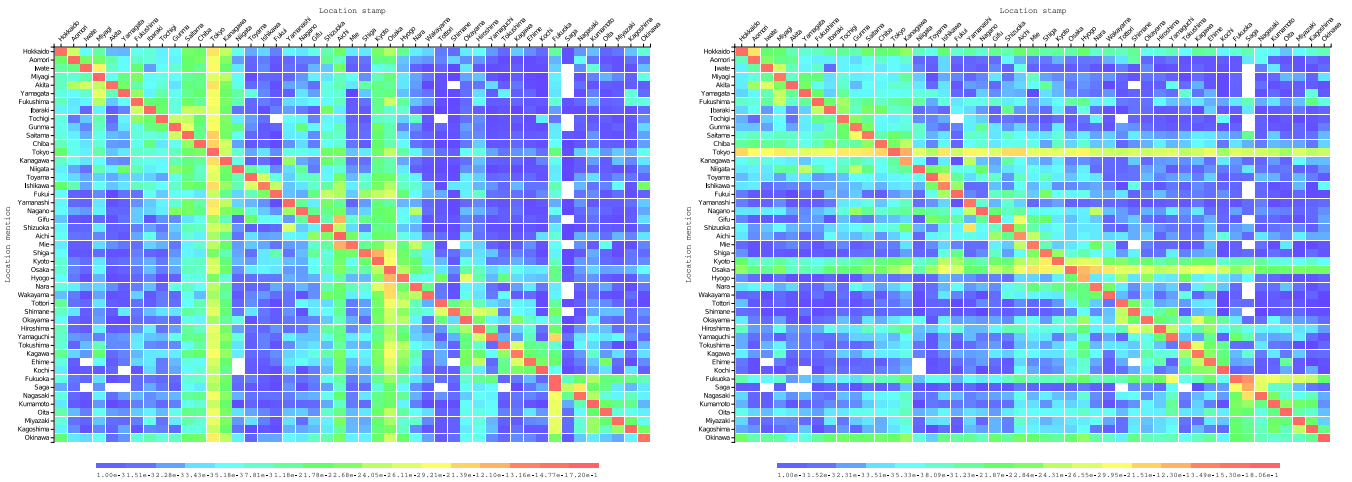
5. METHODOLOGY

Our system has been implemented in SCALA programming language for the back-end. We have used D3 graphical library for the front-end. To handle Japanese text processing tasks we used Kuromoji [2] Japanese morphological libraries, while Stanford CoreNLP [22] was used to parse English texts and location detection. We are aware that there might be better name entity recognizers specially designed to deal with the noisy nature of Twitter [29] (their adoption is left for future work). We give the formal definitions and explain the key computations of our visualizations in the following sections.

5.1 Probability Mass Function

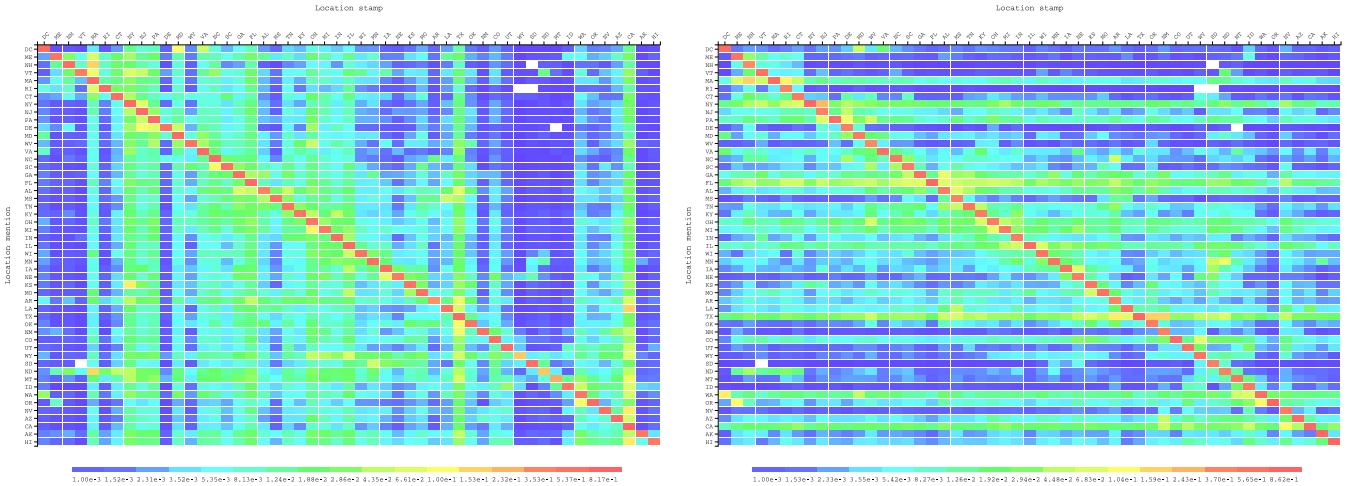
According to *location mention* in a tweet t the cell $C_{i,j}$ is associated with a probability $P(C_{i,j}|t)$ such that $t \in C_{i,j}$. When the *spatial expression* is a point in space, $P(C_{i,j}|t) \rightarrow$

[2] `http://www.atilika.org/`



(a) Interest of X in Y (row-normalized) (b) Popularity of Y among X (column-normalized)

Figure 4: Heat map origin-target of tweets from/about prefectures in Japan available online [43]



(a) Interest of X in Y (row-normalized) (b) Popularity of Y among X (column-normalized)

Figure 5: Heat map origin-target of tweets from states and about states in USA available online [44]

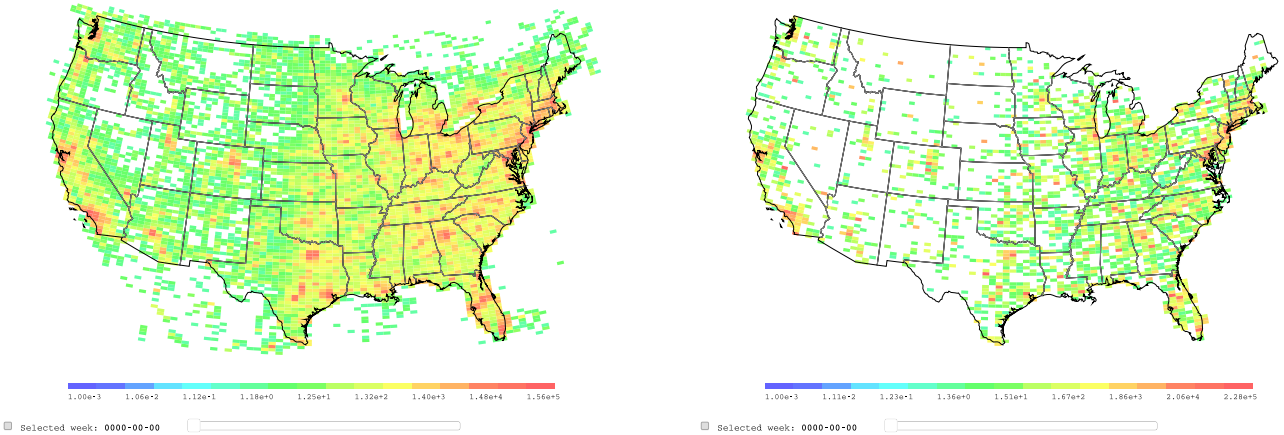
$\{0, 1\}$. In the visualizations proposed in this paper, the *location mention* is always one specific point in space due to our disambiguation mechanism. Mentions of states, prefectures, cities and other kinds of *spatial* areas that we retrieve are then reduced to their central points. That choice is of course not optimal and has been decided due to efficiency reasons. We note however that our model can be easily extended to use probabilities. For example, to map an area such as a state that can span multiple cells, the weight of a tweet can be distributed following a probability mass function P over several cells covered by the state area.

5.2 Ranking Top Words

For each visualization, the system provides a ranking of the words for any *cell* on the heat map. In addition, the system provides a list of top-words for group of lines or group or rows in a heat map called *segment*. The goal is to return a top-k list of the most characteristic words for the area of given *spatial attention* delimited by the segment's boundaries. We compute term scores by adapting TF-iDF weighting scheme. Intuitively, TF-iDF would assign high scores to terms that often appear in tweets associated with a given cell (or segment), while appearing infrequently in other cells (or segments). We modify TF-iDF using probabilities instead of counts and we introduce two hierarchical levels defined for *cell* and *segment*. We describe them in the subsequent paragraphs.

Cell.

Formally, $t \in T$ is a tweet characterized by a bag of words W_t and T is the set of all the tweets in the dataset. Our visualizations are heat maps divided into cells delimited by an interval of values on the x-axis and on the y-axis. Let *CELLS* be the set of all the cells in a heat map and $C \in CELLS$ be one cell in the heat map. Then we use $C_{i,j}$ to denote a cell in position i, j in a heat map.



(a) Tweets mapped to their location stamp [45]

(b) Tweets mapped to their location mention [46]

Figure 6: Geographic distribution of tweets according to their location stamp and mention

The score to rank a word w in a cell $C_{i,j}$ based on the whole set of tweets T is:

$$\text{Score}(w, C_{i,j}, T) = \frac{\sum_{t \in T} P(C_{i,j}|t) : w \in W_t}{\sum_{t \in T} P(C_{i,j}|t)} \times \log \frac{|CELLS|}{|C \in CELLS : \exists t \in C : w \in W_t|}$$

When we rewrite the previous equation in natural language it would be:

$$\text{Score}(w, C_{i,j}, T) = \frac{A}{B} \times \log \frac{C}{D}$$

where:

A is the sum of the weights of the tweets containing the word w in the cell i, j

B is the sum of the weights of all the tweets in the cell i, j

C is the number of cells in the graph

D is the number of cells in the graph with a tweet containing the word w

Segment.

In the visualization displayed in Figure 2, the gray buttons on the top and on the right-hand side represent *segments*, which are regarded as virtual documents. The *segments* are either row-wise or column-wise. They are sets of cells that span respectively, all columns covering the width of the selected rows, or all rows covering the width of the selected columns. For instance, the top left gray cell in Figure 2 with label "Jul 22" gathers all the cells between column July 22 to 28 on every row. We compute a second level TF-iDF score to rank top words in each *segment*. For this a segment is considered as a virtual document. The content of a virtual document is the set of words W_{S_i} built from all the words of the cells in that *segment*.

Let SEG be the set of all the horizontal segments in a heat map and $S \in SEG$ be one horizontal segment in the heat map. The horizontal segment is a set of all the tweets appearing in all the cells in several adjacent lines. Then, S_i is a horizontal segment in position i in a heat map. We compute the weight of a word w in a segment as the sum of the weights of the tweets in those segments in which the word w appears.

$$\text{Score}(w, S_i, T) = \frac{\sum_{t \in T} P(C|t) : \forall C \in S_i : w \in W_C}{\sum_{t \in T} P(C|t) : \forall C \in S_i} \times \log \frac{|SEG|}{|S \in SEG : \exists t \in S : w \in W_t|}$$

The vertical segments are computed in the same way with SEG denoting in this case the set of all the vertical segments and $S \in SEG$ being one vertical segment.

5.3 Location Disambiguation

For the reasons of simplicity, efficiency and to maintain good precision, we assume in our model that a tweet can point to only one *location mention*. Therefore, if the name entity recognizer annotates several location expressions within the same tweet, we assume that they refer to the same real world location. First, it is reasonable since Twitter messages are short. Second, the disambiguation is costly. Thanks to this simplification only one disambiguation per tweet is required exploiting all the *spatial expressions* at once. Otherwise, the system needs to start as many disambiguation processes as the number of elements in the power set of *spatial expressions* of the tweet, which becomes inapplicable in the context of large-scale data. Third, even disregarding the efficiency problem, choosing among the results of each disambiguation is not trivial. For example, consider a tweet with three *spatial expressions* and two real geographical locations mentioned by the user as below:

"I am moving from $\underbrace{\text{Paris}}_{1}$, $\underbrace{\text{TX}}_{2}$ to $\underbrace{\text{Tokyo}}_{3}$."

There would be 8 possible sets of parameters for the disambiguation. For example <Paris> returns Paris in France, <Paris, TX> returns Paris in Texas, <Tokyo> returns Tokyo

in Japan and for <Paris, TX, Tokyo> the system would make an arbitrary choice between the preceding possibilities depending on the implementation.

Disambiguation workflow.

Our work benefits from the previous study [28], in which the authors used a gazetteer lookup and the disambiguation of locations based on the closest-country and the level of place importance. With these settings they could achieve 77% precision suggesting that such simple techniques work reasonably well. For the USA dataset, we then use the same techniques as the part of our disambiguation mechanism for the reasons of efficiency in the large-scale data processing. We manage to disambiguate 62% of the tweets that contain spatial expressions using GeoNames[3] as our gazetteer. In addition, to increase the coverage, we perform another lookup in a location index ranked by popularity. We have built this index by finding and extracting any geotags appearing in tweets and ranking them by their popularity. The application of this additional index allows us to disambiguate 8% more tweets which were not found by Geonames. Note that geotags within tweets are typically added by check-in applications such as Foursquare, and are quite reliable. Whenever a tweet contains a geotag in its content, we can assume that its location mention can be simply mapped to the GPS coordinates of the tweet (i.e. its location stamp). The additional index with location mention-GPS pairs found using the geotags harvested from our dataset complements the main geographical index with smaller yet popular locations that cannot be found by gazetteers.

Figure 7 shows the workflow to map a tweet with *location expressions* to a disambiguated *location mention* associated with GPS coordinates. First, in the same fashion as in [28] we filter out noise and apply heuristics to detect standard patterns used when humans write about locations like a city name followed by a comma and two capital letters denoting state name, to refine the grouping of words in *locations expressions*. Then we query our gazetteer, Geonames, with *location expressions* and order the returned results by their population counts. Depending on the distance from the location stamp (in the USA or close to the border in Mexico and Canada), and the nature of the location disambiguated, which can be either a space point or a large area, we consider three cases from left to right (see Figure 7). First, if the location mention is an area (very often a state) in North America, and if the tweet originates from this area, the tweet is considered local and the *location mention* GPS will be the *location stamp* GPS. Otherwise, that is, if the tweet originates from outside of the area (state), the *location mention* is mapped to the GPS coordinate of the midpoint of the state. Second, if the location mention is outside of North America, the *location mention* will be the middle point of the location disambiguated (country, mountain, etc.). Third, if the found location is a specific point in USA, we use our geotag dictionary ranked by popularity to refine ranking.

For the Japanese dataset, we built our own parser because of the lack of name entity recognizer to extract *location expressions* from Japanese language. We limit the *location mentions* to the names of prefectures in Japan for simplicity and for the ease of visualization keeping manageably low granularity level. Prefecture names are commonly used in Japan when referring to space and they are, on average, relatively small in terms of geographic size. They are also easy to be spotted and disambiguated.

[3]http://www.geonames.org/

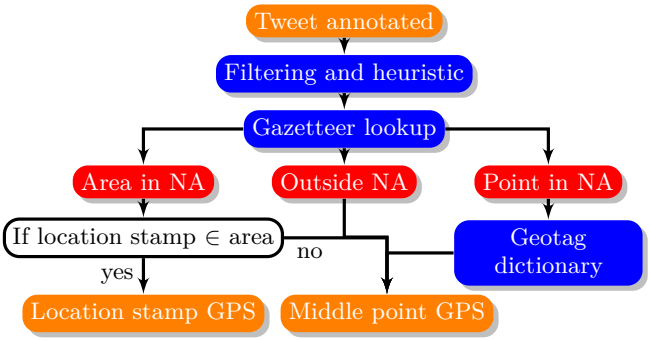


Figure 7: Flow chart of disambiguation workflow

6. CONCLUSIONS

In this paper we demonstrate a novel visualization system and a data model for studying *location mentions* in microblogging. Based on the subset of Japanese tweets collected over half a year and USA tweets collected over 4 months of 2013 we present several visualizations that enable novel data analytics. We have found several interesting observations that emphasize the analytical capability of our visualization frameworks. The observations we make and the proposed systems help us to better understand the way in which users refer about the space and should open way for further, more refined analytical models.

We are aware of limitations of this work due the exploratory character of our analysis. In case of the Japanese dataset our observations are currently done only on the mentions of Japanese prefecture names. Although this makes it easier to find relevant tweets, other spatial expressions are missed.

In the future we will search for lexical field associated with particular *location diff* such as ones related, for example, to traveling activity which may be more linked to long distances when compared, for example, to shopping activity that would likely be associated with closer range distances. Once we build vocabularies associated with particular distances and with particular locations, we plan to automatically infer location information from messages that lack any explicit *spatial expressions*.

7. ACKNOWLEDGMENTS

This work was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan and the JSPS KAKENHI Grants (Nos. 26280042, 15K00162).

References

[1] Fabian Abel, Claudia Hauff, Geert-Jan Houben, et al. “Twitcident”. In: *WWW Companion*. 2012.

[2] Sebastien Ardon, Amitabha Bagchi, Anirban Mahanti, et al. “Spatio-temporal and Events Based Analysis of Topic Popularity in Twitter”. In: *CIKM*. 2013.

[3] Lars Backstrom, Eric Sun, and Cameron Marlow. “Find Me if You Can”. In: *WWW*. 2010.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. "The scaling laws of human travel". In: *Nature* 439.7075 (Jan. 26, 2006), pp. 462–465.

[5] Elena Camossi, Michela Bertolotto, Elisa Bertino, and Giovanna Guerrini. "A Multigranular Spatiotemporal Data Model". In: *GIS*. 2003.

[6] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. "@Phillies Tweeting from Philly?" In: *ASONAM*. 2012.

[7] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. "Emerging Topic Detection for Organizations from Microblogs". In: *SIGIR*. 2013.

[8] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. "You Are Where You Tweet". In: *CIKM*. 2010.

[9] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. "Friendship and Mobility". In: *SIGKDD*. 2011.

[10] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. "A Latent Variable Model for Geographic Lexical Variation". In: *EMNLP*. 2010.

[11] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. "Analysis of Geographic Queries in a Search Engine Log". In: *LOCWEB*. 2008.

[12] Oshini Goonetilleke, Timos Sellis, Xiuzhen Zhang, and Saket Sathe. "Twitter Analytics". In: *SIGKDD* 16.1 (Sept. 2014), pp. 11–20.

[13] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. "Tweets from Justin Bieber's Heart". In: *CHI*. 2011.

[14] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, et al. "Discovering Geographical Topics in the Twitter Stream". In: *WWW*. 2012.

[15] Adam Jatowt, Émilien Antoine, Yukiko Kawai, and Toyokazu Akiyama. "Mapping Temporal Horizons: Analysis of Collective Future and Past related Attention in Microblogging". In: *WWW*. 2015.

[16] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. ""I'M Eating a Sandwich in Glasgow"". In: *SMUC*. 2011.

[17] David Lazer, Alex (Sandy) Pentland, Lada Adamic, et al. "Life in the network". In: *Science* 323.5915 (Feb. 6, 2009), pp. 721–723.

[18] Kalev Leetaru. "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space". In: *First Monday* 16.9 (2011).

[19] Chenliang Li, Aixin Sun, and Anwitaman Datta. "Twevent: Segment-based Event Detection from Tweets". In: *CIKM*. 2012.

[20] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. "Multiple Location Profiling for Users and Relationships from Social Network and Content". In: *VLDB* 5.11 (July 2012), pp. 1603–1614.

[21] Rui Li, Shengjie Wang, Hongbo Deng, et al. "Towards Social User Profiling". In: *SIGKDD*. 2012.

[22] Christopher D. Manning, Mihai Surdeanu, John Bauer, et al. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *ACL*. 2014.

[23] Adam Marcus, Michael S. Bernstein, Osama Badar, et al. "Twitinfo". In: *CHI*. 2011.

[24] Andrew J. McMinn, Daniel Tsvetkov, Tsvetan Yordanov, et al. "An Interactive Interface for Visualizing Events on Twitter". In: *SIGIR*. 2014.

[25] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. "Understanding Twitter Data with TweetXplorer". In: *SIGKDD*. 2013.

[26] Mashaal Musleh. "Spatio-temporal Visual Analysis for Event-specific Tweets". In: *SIGMOD*. 2014.

[27] Tatiana Pontes, Gabriel Magno, Marisa Vasconcelos, et al. "Beware of What You Share". In: *ICDMW*. 2012.

[28] Bruno Pouliquen, Marco Kimler, Ralf Steinberger, et al. "Geocoding multilingual texts: Recognition, disambiguation and visualisation". In: *LREC* (2006).

[29] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. "Named Entity Recognition in Tweets". In: *EMNLP*. 2011.

[30] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. "Open Domain Event Extraction from Twitter". In: *SIGKDD*. 2012.

[31] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. "Finding Your Friends and Following Them to Where You Are". In: *WSDM*. 2012.

[32] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users". In: *WWW*. 2010.

[33] George Valkanas and Dimitrios Gunopulos. "How the Live Web Feels About Events". In: *CIKM*. 2013.

[34] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. "Microblogging During Two Natural Hazards Events". In: *CHI*. 2010.

[35] Jianshu Weng and Bu-Sung Lee. "Event Detection in Twitter." In: *ICWSM*. 2011.

[36] Quan Yuan, Gao Cong, Zongyang Ma, et al. "Who, Where, when and What". In: *SIGKDD*. 2013.

Online Resources

[37] URL: `http://scope13.cse.kyoto-su.ac.jp/JP4/chartlist.html`.

[38] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/chartlist.html`.

[39] URL: `http://scope13.cse.kyoto-su.ac.jp/JP4/locationcity.html`.

[40] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/locationcity.html`.

[41] URL: `http://scope13.cse.kyoto-su.ac.jp/JP4/location.html`.

[42] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/location.html`.

[43] URL: `http://scope13.cse.kyoto-su.ac.jp/JP4/cityMatrix.html`.

[44] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/cityMatrix.html`.

[45] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/locationStamp.html`.

[46] URL: `http://scope13.cse.kyoto-su.ac.jp/US7/locationMention.html`.