Analyzing Global and Pairwise Collective Spatial Attention for Geo-social Event Detection in Microblogs

Shoko Wakamiya
Nara Institute of Science and Technology
Nara, Japan
wakamiya@is.naist.jp

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

Yukiko Kawai, Toyokazu Akiyama
Kyoto Sangyo University
Kyoto, Japan
{kawai,akiyama}@cc.kyoto-su.ac.jp

ABSTRACT

Microblogging has been recently used for detecting common opinions of users at di erent geographic places. In this paper we propose a novel spatial visualization system for uncovering collective spatial attention and interest of users not at but rather towards di erent locations. In other words, we aim to answer questions of the type: what do users collectively talk about when they refer to certain geographical places? In addition, we analyze relations between geographical locations from where Twitter users issue messages and the locations they tweet about. This allows answering questions such as: what do users at a certain place commonly talk about when they refer to another geographical place? We demonstrate an online visualization system that supports the interactive analysis of collective spatial attention over time using 4 months' long collection of tweets in USA.

Keywords

Microblogs; Spatial Analysis; Visualization

1. INTRODUCTION

By aggregating large numbers of microblog messages such as tweets we can detect topics commonly discussed by multiple users (e.g., [4]). Furthermore, associated GPS data o ers a rich medium for various geo-social studies making it possible to detect opinions, topics and sentiment shared by users at the same geographical areas [2, 3]. Many researches have been recently undertaken to track diverse quantities over space such as earthquakes [7], user locations [6], etc.

However, the spatial-focused analysis of microblogs (e.g., Twitter) has been mainly limited to the analysis of tweets based on their location stamps, as given by GPS coordinates. Few approaches tried to utilize another important source of spatial information - location mentions expressed in tweets, despite the fact that users often refer to various geographical places in their messages [1]. An aggregate of multiple tweets referring to spatial locations around the same time can then o er an interesting signal to study, which we call a Collective Spatial Attention (CSA) . In this work we propose to track such collective spatial signal over time and to present it on a geographical map for detailed analysis. In particular, we detect spatial areas to which users from various places collectively refer at the same time and we show the topics associated with such references. We then contrast such collective spatial attention with its aligned version called Pairwise Collective Spatial Attention (Pairwise CSA) following the well-known Focus+Context visualization style [5]. Pairwise CSA is de ned as the common focus of users from a given spatial area on another spatial area.

We demonstrate an interactive system available online[1] that allows investigating collective spatial attention and its pairwise version from diverse angles as well as across time. As an underlying dataset, we utilize tweets issued during 4 months in 2013/2014 in USA. We rst detect and map spatial references in tweet content. Then we visualize collective spatial attention over time by combining the information from the detected location mentions with the GPS coordinates of tweets, and by considering also tweet timestamps as well as their content. We believe that the data processing system and the visualizations we propose could provide complementary knowledge to many social media studies interested in location-based analysis of user activities or in geo-social event detection.

2. SYSTEM OVERVIEW

2.1 Contextual Data Views

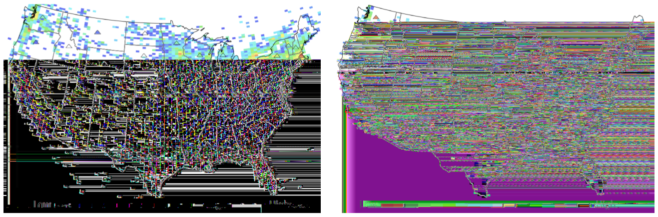
The proposed system has four basic contextual views in the form of heatmaps superimposed on a geographical map:

- Intensity View of CSA:
  - based on location stamps (Fig. 1(a))
  - based on location mentions (Fig. 1(b))
- Distance View of CSA:
  - based on location stamps (Fig. 2(a))
  - based on location mentions (Fig. 2(b))

The rst intensity view displays the intensity of CSA originating from given places by aggregating tweets based on their location stamps (i.e., GPS coordinates). It thus allows investigating what users at a certain place collectively talk about? The second view portrays the intensity of CSA to any given place by aggregating tweets containing mentions of

[1] ht tp://goo.gl/W3lOqN

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2890551.



(a) Location stamps (b) Location mentions

Figure 1: Intensity views of CSA for the entire period. The cells are coloured based on log scale of the values.

this place, thus enabling to reason on what users collectively talk about when they refer to a given geographical place?

In contrast, the distance views show the average distances of CSA based on location stamps (how far from a given location does collective spatial attention reach?) and location mentions (from how far does collective spatial attention directed to a given location come?)

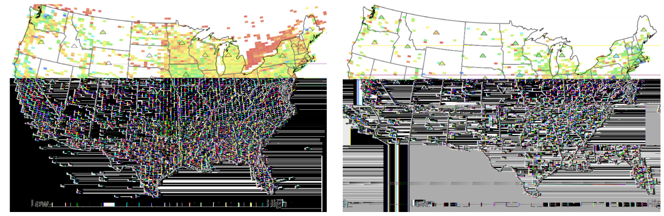
2.2 Data Model

A tweet consists of basic sextuplet of attributes: $tweetID$, $date$, lat, lng, $text$, $userID$. $tweetID$ is the unique number of a tweet. $date$ is the date when a tweet was issued. lat and lng are location stamp (GPS coordinates), $text$ is the textual message and $userID$ is the unique number of a user writing a tweet. The intensity view based on the location stamps can be already directly drawn using these basic attributes. In order to construct the remaining data views, we need to extract and compute other location-related attributes (i.e., location mentions and distances). Specifically, the location mentions are extracted from $text$ by applying morphological analysis and by consulting a geographical dictionary. We use here GeoNames which is a well-known geographical dictionary. For simplicity, in the current implementation, we mainly focus on village, city, state and country mentions, leaving others (e.g., buildings or city districts) as a future work. Distances (represented as delta in the system) are calculated as Euclidean distances between the GPS coordinates of tweets and the ones of their disambiguated location mentions.

2.3 Collective Spatial Attention

To create each of the views listed in Sec. 2.1 we set rectangle cells on a geographical map. They aggregate tweets based on their location stamps or location mentions (Figs. 1 and Figs. 2). Depending on these view types tweets that either originate from a place located in a given cell or which contain location mentions that point to a place in the cell are automatically allocated to that cell. In addition, triangle cells are placed in the centers of states and an inverted triangle cell is set in the center of the country. These indicate the degree of CSA based on the coarser granularity levels of location mentions such as the state and country levels.

The cells are coloured ranging from blue to red according to the intensity of CSA (Figs. 1) or its average distance (Figs. 2). In the intensity views, the colors are allocated based on the frequency of tweets mapped to cells, while, in the distance views, the colors are decided according to the average distance of CSA from/to cells. The colors of the triangle cells and the inverted triangle cell are assigned in a similar way.



(a) Location stamps (b) Location mentions

Figure 2: Distance views of CSA for the whole period. The cells are coloured based on linear scale of the values.

2.3.1 *Explaining Collective Spatial Attention*

Displaying the intensity is not enough to understand the reasons behind the collective spatial attention. The system then displays representative keywords to summarize tweets associated with any cell when clicking on the cell. The keywords are ranked based on their TF-iCF (Term Frequency - inverse Cell Frequency) values and are presented in a new window along with their scores and raw counts. A cell is treated here as a virtual document that contains the combined text of tweets associated with the cell.

In addition, the system also shows the list of all tweets for the selected cell together with their attributes arranged in the form of a table (see Fig. 3 for example).

Note that besides the aggregate views based on the entire time period of data, it is also possible to compute the same views for finer time units such as weeks. When a specific week is selected on the time slider, the views are based on the values computed for a given week.

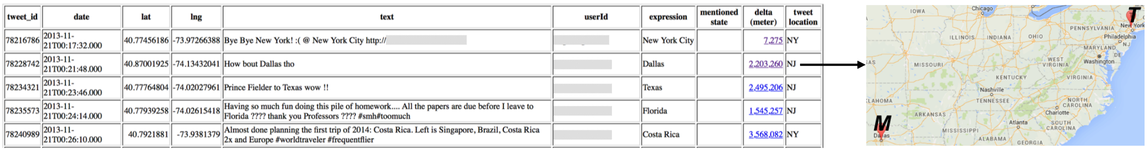
2.4 Pairwise Collective Spatial Attention

The collective spatial attention described above is either many-to-one or one-to-many type CSA (i.e., users from many areas collectively tweeting about the same spatial area or users from the same area collectively tweeting about different areas). In this section we focus on the pairwise relation between locations from where users tweet and the locations they tweet about. In other words, we extend the above types of spatial attention to one-to-one type spatial attention (users from a single area collectively tweeting about another area). To display such Pairwise CSA the system draws the top-k mention arrows on a map. A mention arrow, represented as (c_o, c_d), is defined as directed pair of an origin cell c_o and a destination cell c_d such that many users from c_o collectively refer to c_d. First, the frequency of tweets and the unique number of users at c_o referring to c_d at the same time w are calculated by the following functions, $Count_t((c_o, c_d), w)$ and $Count_u((c_o, c_d), w)$, respectively. Mention arrows are then drawn based on either of these values. When selecting views for a given week $w \in W$ the frequency of tweets of an arrow (c_o, c_d) can be normalized as follows.

$$Norm_t((c_o, c_d), w) = \frac{Count_t((c_o, c_d), w)}{Count_t((c_o, c_d), W)} \quad (1)$$

The frequency of users of the arrow is normalized in a similar way.

Each mention arrow gets highlighted in blue when selected. At the same time, all the arrows pointing to the same destination as the selected arrow (i.e., when multiple cells \link" to the same cell) become highlighted in gray. Note that we focus on the arrows having the same destina-



tweet_id	date	lat	lng	text	userId	expression	mentioned state	delta (meter)	tweet location
78216786	2013-11-21T00:17:32.000	40.77456186	-73.97266388	Bye Bye New York! :(@ New York City http://		New York City		7,275	NY
78228742	2013-11-21T00:21:48.000	40.87001925	-74.13432041	How bout Dallas tho		Dallas		2,203,260	NJ
78234321	2013-11-21T00:23:46.000	40.77764804	-74.02027961	Prince Fielder to Texas wow !!		Texas		2,495,206	NJ
78235573	2013-11-21T00:24:14.000	40.77939258	-74.02615418	Having so much fun doing this pile of homework.... All the papers are due before I leave to Florida ???? thank you Professors ???? #smh#toomuch		Florida		1,545,257	NJ
78240989	2013-11-21T00:26:10.000	40.7921881	-73.9381379	Almost done planning the first trip of 2014: Costa Rica. Left is Singapore, Brazil, Costa Rica 2x and Europe #worldtraveler #frequentflier		Costa Rica		3,568,082	NY

Figure 3: An example of the list of tweets of a selected cell or an arrow. When clicking the value of \delta" attribute of a tweet, a map is presented with two pins mapped showing the location stamp (T) and disambiguated location mention (M).

tion instead of the ones sharing the same origin, since the former are often the result of some events taking place in the destination area. Arrows to the centers of states are shown in red and sport-related arrows (see Sec 2.4.2) are in violet.

2.4.1 Explaining Pairwise CSA

Mention arrows are superimposed on a selected contextual view that shows global CSA (see Sec 2.3) in order to provide context for understanding Pairwise CSA according to the Focus+Context paradigm [5]. For example, when the intensity view based on location mentions is used as the context for arrows, it is possible to compare the amount of global CSA received by a given cell with the characteristics of mention arrows ending at the cell. Similarly, the distance views would allow comparing the average distance of CSA with the length of the arrow.

The system also provides information for explaining any selected mention arrow in a pop-up window when hovering a cursor over the arrow. This window consists of two panels (see Fig. 5). The upper part of the left-hand side panel shows the general information concerning the selected arrow. These are origin, destination, rank, user count, number of arrows pointing to the same destination cell and the probability of representing a sport-related event. The lower part displays the list of feature words related to the arrow. The feature words are extracted and ranked based on TF-iAF (Term Frequency - inverse Arrow Frequency) scores. The system regards all the arrows in a given view as the collection of virtual documents and displays the top-100 words for each arrow based on their TF-iAF scores. The feature terms are color-coded for facilitating understanding of differences between arrows directing the same destination. In particular, red-colored terms mean terms peculiar to the selected arrow, while white terms denote terms shared among arrows directing the same destination cell.

The right-hand side panel shows temporal information in the form of two graphs based on the Focus+Context visualization style. The upper graph is the temporal local view showing either the frequency of tweets (or, depending on selection, the number of unique users) during a week. On the other hand, the lower graph shows the same quantity over the entire time period with a yellow bar highlighting the position of the selected week. In both the graphs, red lines display the intensity of CSA between the origin cell and the destination cell. White lines represent the average frequency of tweets (or the number of users) aggregated over all the arrows pointing to the same destination.

Finally, when an arrow is clicked, a new window, similar to the one for explaining cells, is shown to display the attributes and content of tweets that underlie the arrow (see Fig. 3).

2.4.2 Detecting CSA of Sport Events

The system can filter mention arrows according to a specific topic. In the current implementation we divide arrows into ones related to sport events and others, due to large number of arrows that relate to sport events like football matches. To classify an arrow (c_o, c_d) to the sport category, we compare feature terms $T_{(c_o,c_d)}$ with sport-related terms.

We calculate the sport words' frequency of a pair of two cells, $Freq_{sp}((c_o, c_d))$, as follows.

$$Freq_{sp}((c_o, c_d)) = \frac{|splist_{(c_o,c_d)}|}{|T_{(c_o,c_d)}|} \tag{2}$$

$$splist_{(c_o,c_d)} = \{t \,|\, Match(t, splist) = 1, t \in T_{(c_o,c_d)}\}$$

$splist$ is a list of sport-related terms generated from several sport vocabulary lists [2,3]. A mention arrow whose sport words' frequency is over a threshold is deemed to represent a sport-related Pairwise CSA.

3. CASE STUDIES

3.1 Dataset

The data currently used by our system has been collected with few short breaks from Sept. 25, 2013 to Jan. 17, 2014 from USA. After removing tweets in other languages than English[4] we obtained 158M tweets. We then extracted location mentions using Stanford CoreNLP tagger [5] and disambiguated them using GeoNames service[6] as well as our own rule-based mechanisms. More details of the disambiguation process are provided in [1]. The final dataset contains 4.3M spatially annotated tweets issued in USA by 28% of the users of the original dataset.

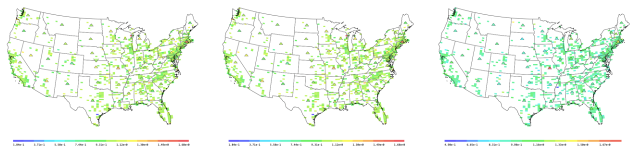
3.2 Examples

3.2.1 CSA

We make here several observations by comparing different contextual views. By contrasting the intensity views based on location stamps with the ones based on location mentions for the whole period as shown in Fig. 1(a) and Fig. 1(b), we observe that the most populous cities in USA such as New York, Los Angels, Chicago, Houston, and Philadelphia gather much CSA from both their citizens as well as users at different locations. We next look at the distance views based on location stamps and location mentions (Fig. 2(a) and Fig. 2(b)). Fig. 2(a) shows that users at locations in the east part of USA tend to be generally more referring to far away places than the users in the west part of USA.

We next compare CSA in different weeks. The heatmaps of Fig. 4 show the distance views based on location mentions for three consecutive weeks (Nov. 18-24, Nov. 25-Dec. 1, and Dec. 2-8). Interestingly, many cells of the heatmaps of the weeks from Nov. 18 and Dec. 2 are coloured in green,

[2] ht tp://www.enchantedlearning.com/wordlist/sports.html
[3] http://www.vegau.com/resources/
[4] http://code.google.com/p/language-detection
[5] http://stanfordnlp.github.io/CoreNLP/
[6] http://www.geonames.org/source-code/javadoc/



(a) Nov. 18-24 (b) Nov. 25-Dec. 1 (c) Dec. 2-8

Figure 4: Distance views of CSA based on location mentions during three di erent weeks.

while those of the heatmaps of the weeks from Nov. 18 and Nov. 25 are predominantly light green which indicates further distance. This suggests that many tweets tend to be issued towards distant places before and during the week of Thanksgiving Day (Nov. 28), as people plan to travel or contact relatives who may live far away.

3.2.2 Pairwise CSA

When analyzing Pairwise CSA we could observe quite many spatial relationships due to sport events. Fig. 5 portrays Pairwise CSA using the distance view based on location mentions as underlying context (CSA). Lets take as an example the top-scored mention arrow (rank 1) in the week from Dec. 23. Looking at the pop-up window for this arrow (see the bottom-right pop-up in Fig. 5) we can know that it represents Pairwise CSA that originates from Marlton, New Jersey (city very close to Philadelphia) and is directed to Dallas. The arrow is categorized as a sport-related one. Indeed, its representative terms (in red) suggest that it is related to the national football game: Philadelphia Eagles (Philadelphia) vs. Dallas Cowboys (Dallas) (e.g., words such as \cover,", \defender" and \#goeagles"). The game took place in Dallas on Dec. 29. The temporal graphs tell us that the event lasted only half a day and it was the rst time when the Pairwise CSA between the same pair of cells occurred within 4 months' long time period. This can be observed from the temporal graph of the entire time period.

We also notice that in the same week there were 17 arrows to the same destination when looking at the entire map. Thanks to ranking and color coding of words we can understand that the di erence between the arrows relates to which teams users from di erent locations support. For instance, we found ve words concerning \eagles" and only two words for \dallas" in the top-15 words of the 1st top-scored arrow. On the other hand, the common point of all the 17 arrows is represented by the word \dallas" (see the word lists in the pop-ups in Fig. 5). In fact, another arrow ranked as the 11th top-scored sport-related arrow in the same week originates from San Antonio, Texas. Unlike the arrow from Marlton, this one is characterized by the words supporting Dallas Cowboys (e.g., \#indallaswetrust"). Also, the red line in the lower temporal graph indicates that there were other times (e.g., in October) of high Pairwise CSA between the same origin and destination cells as the cells of this arrow. Generally, comparing red and white lines in both temporal graphs helps to detect geo-social events.

In another example, we could detect the impact of the 2013 United States general elections held on Nov. 5 by looking at the Pairwise CSA towards the Commonwealth of Virginia from cities in New Jersey, Virginia and San Francisco. The Pairwise CSA is represented by the words such as \virginia,"\richmond,"\win,"\elect," and \governor." Finally, we also notice the e ect of the government shutdown

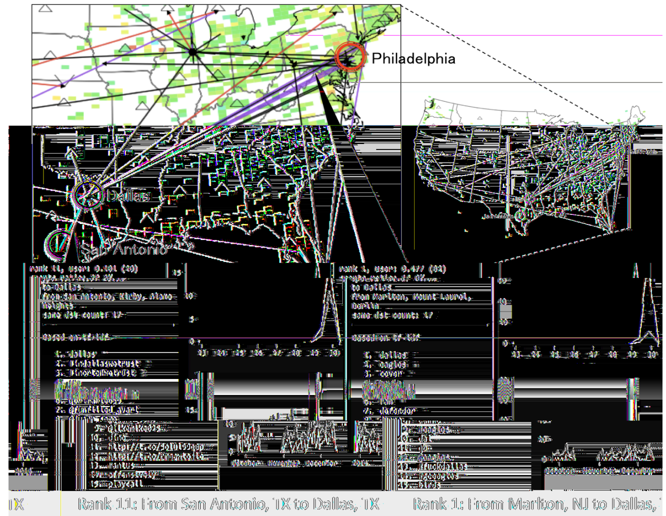


Figure 5: Pairwise CSA view overlaid on the intensity view of the location stamps in the week of Dec. 23. The top-75 mention arrows based on the number of users are displayed.

on Oct. 1 which gathered crowd attention due to lots of mention arrows (20 arrows in the top-100 arrows) appearing in that week which are towards USA (i.e., point to the central inverted triangle).

4. CONCLUSIONS

In this paper we demonstrate an interactive system for analyzing the collective interest of users directed towards or originating from given geographical areas. It shows the global and pairwise types of the collective spatial attention, their temporal uctuations as well as associated topics. The main application is fostering social studies that aim at using social media for inferring space-related knowledge.

5. ACKNOWLEDGMENTS

This research was supported in part by Strategic Information and Communications R&D Promotion Programme (SCOPE), the Ministry of Internal A airs and Communications of Japan, and JSPS KAKENHI Grant Numbers 26280042 and 15K00162.

6. REFERENCES

[1] E. Antoine, A. Jatowt, S. Wakamiya, Y. Kawai, and T. Akiyama. Portraying collective spatial attention in twitter. In *KDD '15*, pages 39–48, 2015.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW '10*, pages 61–70, 2010.

[3] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *CIKM '10*, pages 759–768, 2010.

[4] O. Goonetilleke, T. Sellis, X. Zhang, and S. Sathe. Twitter analytics: A big data management perspective. *SIGKDD Explor. Newsl.*, 16(1):11–20, Sept. 2014.

[5] J. Lamping, R. Rao, and P. Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *CHI '95*, pages 401–408, 1995.

[6] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *ICDMW '12*, pages 571–578, 2012.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW '10*, pages 851–860, 2010.