A Novel Method for Geographical Social Event Detection in Social Media

Xingyu Gao[1,2,3], Juan Cao[1,2], Qin He[1,2,3], Jintao Li[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{gaoxingyu, caojuan, heqin, jtli}@ict.ac.cn

ABSTRACT

Popular microblogging service has attracted much attention around the world recently. With tremendous amount of tweets published each day, social event detection is becoming one of the most challenging research topics, especially for geographical social event. This paper proposes a novel geographical social event detection approach by mining geographical temporal pattern and analyzing the content of tweets. For the tweets published by users in the geographical area at each time unit, we first estimate its geographical temporal pattern based on the alternation regularity of tweets. Furthermore, we discovery the unusual geographical area by more frequent alternation of tweet count, and adopt adaptive *K-means* clustering algorithm for the tweets published in the geographical area. Finally, the geographical social event is detected by the number of the tweets in the cluster. We implement and validate our approach on realistic data collected from real-world social media websites. Experimental results show that our method can detect geographical social event with better performance than traditional methods. In addition, vivid demonstration of geographical social event can be effectively performed by our method.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial Databases and GIS; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithm, Design, Experimentation, Performance

Keywords

Social Event Detection, Geographical Temporal Pattern, Adaptive *K-means* Clustering

1. INTRODUCTION

Nowadays, popular social network websites such as Twitter[1], Sina Weibo[2], and Tencent Weibo[3] have already published colossal size of highly-dynamic real-time data in the form of messages and status updates. Subsequently, increasing microblogging service has attracted much attention around the world, especially for location-sharing social media services like Foursquare, Gowalla and so on. Millions of thousands of people utilize the online social network to express their opinions or share their interesting activities socially connected to their relatives, friends and colleagues through their computers and mobile phones.

With tremendous amount of tweets published in social network each day, a challenging task is the grouping of tweets according to the underlying events for the researchers. Many research work has been published on these microblog systems, especially during recent years. Sasaki et al. [6] proposed a message analysis approach for using Twitter as a cue to detect geographical social events, and developed a natural disaster alarming system by using Twitter occurrences from the event locations for some special types of cases, such as earthquakes in Japan by using Twitter user as social sensor. However, they must manually design the input query. So their approach could not be applied to conduct a various types of socio-geographic analysis including unknown and unexpected events. Sankaranarayanan et al. [7] presented a system so-called TwitterStand which captured tweets that corresponded to breaking news. The method combining Part-Of-Speech (POS) tagging and Named-Entity Recognition (NER) was adopted to extract geographic nouns from contextual posts so as to cooperate geographic information into events. Pozdnoukhov et al. [5] adopted a streaming Latent Dirichlet Allocation topic model trained with an incremental variational Bayes method to explore space-time dynamic structures of the topical content of short textual messages. Hong et al. [1] utilized both statistical topic models and sparse coding techniques to provide a principled method for uncovering different language patterns and common interests in the twitter stream.

Most researches have focused the textual analysis for tweet, and some researches [4, 8, 9] on social event detection, visualization or search for photos posted on social media sites. However, relatively few researches address geographical pattern of social event. Lee et al. [2, 3] presented measurement for geographical regularities deduced from the usual behavior patterns of crowds. They detected any unusual event

ICIMCS '13, Aug. 17-19, 2013, Huangshan, Anhui, China.
Copyright 2013 ACM 978-1-4503-2252-2/13/08 ...$15.00.

[1] `http://twitter.com/`
[2] `http://weibo.com/`
[3] `http://t.qq.com/`

happening in the monitored geographical area by comparing the regularity. Nevertheless, this work doesn't further analyze the content of posted tweets, which cannot vividly demonstrate diverse geographical social events taken place in the geographical areas respectively.

In this paper, we propose a novel geographical social event detection method by geographical temporal pattern mining and content analysis. We first mine the geographical temporal pattern of the tweets in social activities, and discover the unusual geographical region by this pattern. Then we adopt the adaptive *K-means* clustering algorithm for the content of tweets, which involves in the area where unusual geographical area found. Next, the geographical social event is detected by the number of the tweets in the cluster. Also, the detected geographical social events can be intuitively displayed by the highly-frequent keywords involved in the cluster. Furthermore, we evaluate the performance of our proposed approach. Our experiments show that, using geographical temporal pattern mining and adaptive *K-means* clustering, our method can better detect geographical social event, also vivid demonstration of geographical social event can be effectively performed.

The rest of this paper is organized as follows. In Section 2, we introduce unusual geographical area discovery by geographical temporal pattern mining. Section 3 describes content analysis for tweet by adaptive *K-means* clustering method in detail. In Section 4, we show the experimental results of geographical social event detection, and we give the conclusion in Section 5.

2. UNUSUAL GEOGRAPHICAL AREA DISCOVERY

In this section, we firstly give our description for social event. Then, geographical temporal pattern mining method for the event is introduced in Section 2.1. Event is a group of participated behaviors aggregating activities and discovered at a time unit. We describe the social event by the behaviors: number of tweets posted within a specific time period in the geographical district. The behaviors are regarded as posting tweets in this paper. We count the total occurrence of tweets during a specific period of time.

In general, social event has the geographical attribute. So we adopt geographical temporal pattern mining approach to discover the unusual geographical area. The details are described in the following sections.

2.1 Geographical Temporal Pattern Definition

For the purpose of convenient analysis, one day is equally divided into 4 non-overlapping partitions (early morning: 0:00-6:00, morning: 6:00-12:00, afternoon: 12:00-18:00 and night: 18:00-24:00), each time unit with 6 hours. It is reasonable that we set the administrative district as the geographical area. In order to better reflect change regularity of behaviors in the geographical area, we define the geographical temporal pattern (GTP) of behaviors as

$$\delta_i^t(tu) = |C_i^t(tu) - C_{i-1}^t(tu)| \tag{1}$$

$$GTP_i^t(tu) = \frac{\delta_i^t(tu) - min\{\delta_i^t(tu)\}}{max\{\delta_i^t(tu)\} - min\{\delta_i^t(tu)\}} \tag{2}$$

where $C_i^t(tu)$ is number of tweets posted in a certain geographical area at the time unit of the i^{th} day (the current day), and $C_{i-1}^t(tu)$ denotes number of tweets posted in the same geographical area at the same time unit of the $(i-1)^{th}$ day (the day before the current day).

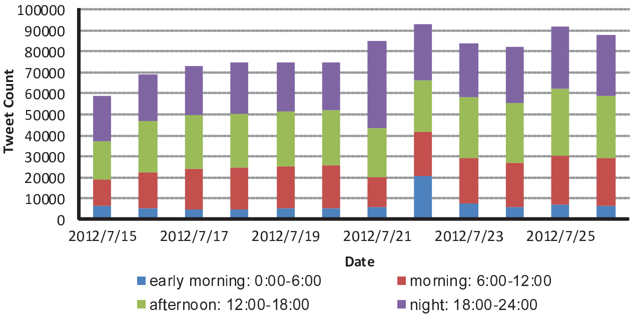


Figure 1: Tweet count in Beijing area.

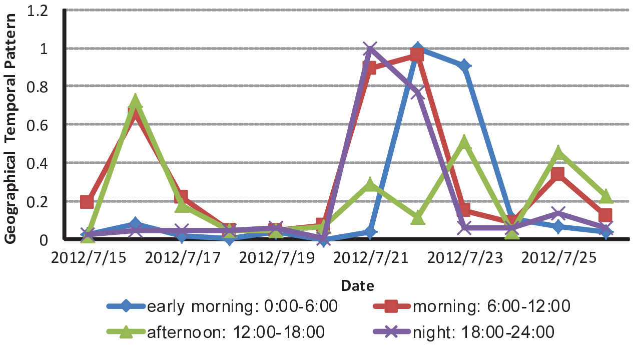


Figure 2: Geographical temporal pattern in Beijing area at the time units.

Then, we can mine the geographical temporal pattern of social event by Equation (2). The tweet count of Beijing area is shown in Figure 1, and the estimated geographical temporal pattern of behaviors is shown in Figure 2.

2.2 Unusual Geographical Area Discovery by Pattern

From the analysis by Section 2.1, we can estimate the geographical pattern of regional social activities. If the aggregated tweet in the regional social activities is more frequent, we discover that this geographical region is unusual. Assuming that geographical temporal pattern (GTP) is calculated by Equation (2), consequently, the unusual geographical area (GA) at time unit is explored by

$$GA_{unusual}(tu) = \begin{cases} true, & if\ GTP_i^t(tu) > \theta \\ false, & otherwise \end{cases} \tag{3}$$

where θ is the threshold. Equation (3) denotes that if the geographical temporal pattern value exceeds the threshold, then the unusual geographical region is discovered. For example, the geographical temporal pattern values in the night of 2012/7/21 and the early morning of 2012/07/22 are exceptionally high, so this area at these two time units is regarded as an unusual region.

3. CONTENT ANALYSIS

Besides the geographical attribute, from the perspective of tweet content, social event is comprised of a series of highly-frequent keywords in the event happened over time. It is crucial that we should vividly display diverse events found in the area to the user, while Lee's method [2, 3] only can detect the unusual Region-of-Interest (ROI) where the event happened at time unit. As discovered the unusual geographical area at the time units, we further analyze the content of the tweets posted in the unusual district during this period.

3.1 Adaptive K-means Clustering

We adopt clustering method to analyze the tweet content in order to detect all sorts of geographical social events by clusters, and the event can be intuitively displayed by the keywords involved in the cluster. We firstly use Natural Language Processing (NLP) tool to divide each tweet to the words, and Vector Space Model (VSM) is employed to represent the words. Due to the short length of tweet content, we use Term Frequency (TF) as element of the vector though TF-IDF is widely applied in textual clustering, the distance is measured by the *Cosine* similarity between the vectors. Then, we adopt the adaptive *K-means* clustering for tweet clustering so as to automatically find the number of clusters K without user input which vary across every time unit. The details of the adaptive *K-means* clustering process is presented as follows:

Let samples $\{S_i|i = 1, 2, \cdots, N\}$ be the tweets published in the discovered unusual geographical area. Through the clustering process, the tweets will be clustered with content similarity.

Step 1 Compute Cosine similarity between each two sample-pairs and set their average as $basicStep$;

Step 2 Compute Cosine similarity between i^{th} and other samples, then the density value of this samples is set as the number of samples, whose distances are smaller than $basicStep$;

Step 3 Sort all density values by descending order;

Step 4 *If* the ratio of the biggest density and N is greater than 0.5, the step threshold is set as $0.8 \times basicStep$ and the density threshold is set as 100; *Else If* the ratio of the biggest density and N is smaller than 0.01, the step threshold is set as $1.5 \times basicStep$ and the density threshold is set as 10;

Step 5 Set K' as the number of pseudo clustering, where K' denotes the number of samples whose density values are greater than the density threshold, and the number of final clustering K is set as 0;

Step 6 For $i = \{1, 2, \cdots, K'\}$, compute the Euclidean distance between i^{th} sample and other $K' - i$ samples. If there is no sample whose distance is smaller than the step threshold, then execute $K++$; otherwise, K holds no change;

Step 7 As K is the number of final clustering, its corresponding sample is set as the initial clustering center, which is input as Step 1 of *K-means* clustering;

Step 8 Repeat from Step 2 to Step 4 of *K-means* clustering to complete the whole clustering process.

3.2 Geographical Social Event Detection by Clustering

From Section 3.1, we can obtain the clustering of the tweets published in the unusual geographical area at the time units. If the cluster contains more tweets, we regard the cluster as an event. Assuming that the tweets published in the unusual geographical area is clustered by adaptive *K-means* clustering, then the occurred geographical social event is detected by

$$E_{occur} = \begin{cases} true, & if\ N_t(C_i) > \eta \\ false, & otherwise \end{cases} \tag{4}$$

where η is the threshold. Equation (4) denotes that if the number of tweets in the Cluster C_i exceeds the threshold, then geographical social event is detected.

4. EXPERIMENTS

At first, we introduce the experimental dataset from our data collection, then show the performance evaluation through comparison both our approach and geographical social event detection method introduced in [2, 3].

4.1 Dataset

Sina Weibo is a very popular microblogging website in China, and it allows people to express their opinions by posting short messages with limit of 140 Chinese characters. Every user in this social media website can fill in his profile and messages to the public. Because of the streaming API provided by Sina Weibo with the quantity limits of crawled users and tweets, we developed our data crawler by template-based method to obtain: 1) user with his profile, including user ID, user name, his/her geographical area and tags, etc.; 2) tweets posted by users, containing author of the tweet, content, distribution time and so on.

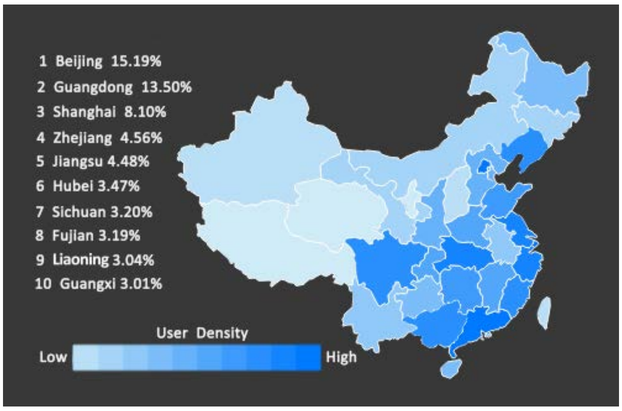


Figure 3: The geographical distribution of users in our dataset.

We conduct the experiment on our real world tweet dataset named MCG-Weibo to evaluate our proposed method. The dataset consists of 170 thousand users, and their published tweets with 3 billion tweets totally. The geographical distribution of users in the dataset is shown in Figure 3, which lists the top 10 geographical area from ranking the user density. We can see the top three geographical areas are Beijing, Guangdong and Shanghai, which matches network development degree of these areas, besides a faithful representation of local economic development level.

4.2 Evaluation

To evaluate the performance of our proposed method, we choose the tweets posted during the period of two months as training set, from 2012/06/26 to 2012/08/26, with 32,830,277 tweets, and testing period of 2012/07/21 to 2012/07/26 with 3,368,803 tweets. We set the threshold for geographical temporal pattern θ as 0.75 and the threshold for number of tweets in the cluster η as 56. We provide 20 interesting events for the test dataset to detect in Beijing area. As a result, we can detect 13 events among the prepared 20 events in the geographical area. Thus, our method has better performance where the accuracy can reach 65%, which outperforms Lee's method with 60% accuracy.

4.3 Detection of Geographical Social Event

In order to validate the advantage of our method of geographical social detection, which not only detects geographical social event well but also can display the diverse events to users, we conduct the experiment on our dataset MCG-Weibo. Based on our approach, two examples of the detected geographical social events are shown in Figure 4. The upper parts of Figure 4 (a) and (b) are the detected events at the same geographical area, including the posted tweets and photos about the events. The wordle is demonstrated at the bottom parts of Figure 4 (a) and (b). It is very convenient to let users understand the content of event, people can also vividly know the viewpoint of netizens.

5. CONCLUSIONS

In this paper, we have presented our novel approach to detect geographical social events for tweets in Sina Weibo. Firstly, our detection method by geographical temporal pattern mining and content analysis provides an effective way to detect events. Secondly, realistic dataset is conducted to evaluate the methods, and the results show that our method works well. Finally, the experiments indicate that our method can provide vivid demonstration for the detected geographical social events. We plan to extend our work to track and visualize the events.

6. ACKNOWLEDGMENTS

This work is supported by National Nature Science Foundation of China (61172153), National Key Technology Research and Development Program of China (2012BAH39B02).

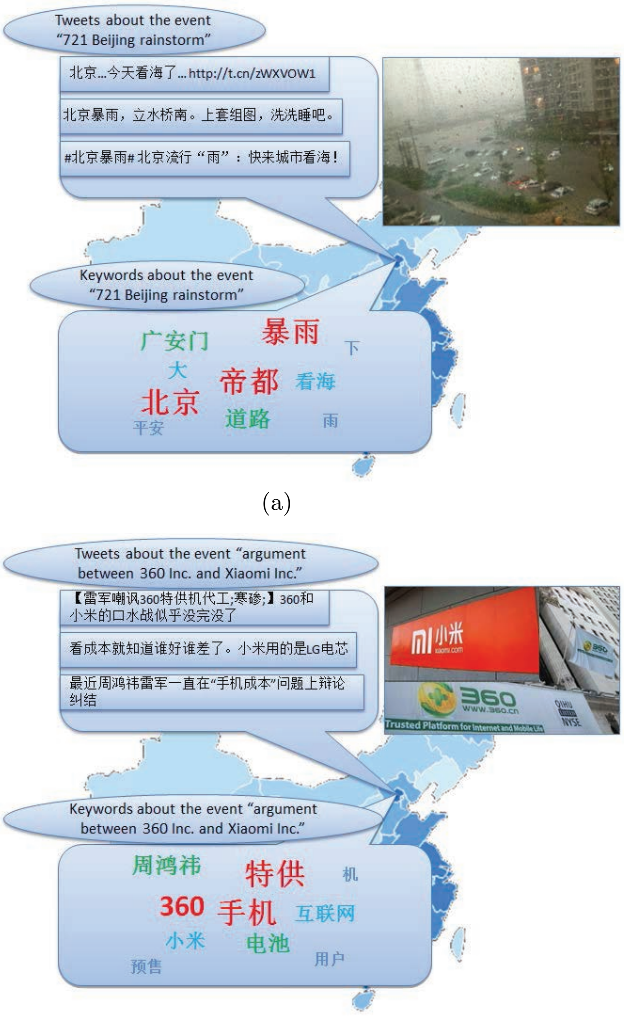
7. REFERENCES

[1] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[2] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *GIS-LBSN*, pages 1–10, 2010.

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

[4] Y. Nakaji and K. Yanai. Visualization of real-world events with geotagged tweet photos. In *ICME Workshops*, pages 272–277, 2012.



北京...今天看海了...http://t.cn/zWXVOW1

北京暴雨，立水桥南。上套组图，洗洗睡吧。

#北京暴雨#北京流行“雨”：快来城市看海！

(a)

Tweets about the event "argument
between 360 Inc. and Xiaomi Inc."

【雷军嘲讽360特供机代工;寒碜;】360和小米的口水战似乎没完没了

看成本就知道谁好谁差了。小米用的是LG电芯

最近周鸿祎雷军一直在“手机成本”问题上辩论纠结

mi 小米
xiaomi.com
360
WWW.360.cn
Trusted Platform for Internet and Mobile Life
QIHU
NYSE
Keywords about the event "argument
between 360 Inc. and Xiaomi Inc."
周鸿祎
特供
机
360
手机
互联网
小米
电池
预售
用户

(b)

Figure 4: Two examples of the detected geographical social events. (a) is the tweets, photo and keywords related to the event "721 Beijing rainstorm", and (b) is about the event "argument between 360 Inc. and Xiaomi Inc.".

[5] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *GIS-LBSN*, pages 1–8, 2011.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.*, 25(4):919–931, 2013.

[7] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51, 2009.

[8] M. Wang, K. Yang, X.-S. Hua, and H. Zhang. Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 12(8):829–842, 2010.

[9] Y. Wang, H. Sundaram, and L. Xie. Social event detection with interaction graph modeling. In *ACM Multimedia*, pages 865–868, 2012.