Location Estimation of Event Based on Twitter

Twitter によるイベント位置の推定

学籍番号　47-166817
氏　名　楊　珂為（Yang, Kewei）
指導教員　瀬崎　薫 教授

Introduction

As we known the popular micro-blogging tool Twitter has been used in multiple fields due to the high value of its data. However, it is difficult to get the user's location directly now, people tend to shut off their Geotag function for protecting privacy. The goal of this research is to achieve the location estimation of events without Geotag.

To resolve this problem of non- Geotag I studied on the researches of detecting user's location. In the observation of the association of users and event I found that people would tweet event-related information in their tweets, and when the people is near to the event place there might post more tweets. The location estimation system is based on this thought.

Related work

In the research of detecting events, [1] proposed a method to discover temporal or geographical burst by using photos with location tag posted on Flickr. The method of [1] focus on analyzing the temporal and geographical distribution of these tags (New York, World Cup, dog, etc.) and assumed the tags with signify bias as the events. Walther [2] constructed a system to detect events on the geospatial space. They discussed what kind of feature can be useful in accurately detecting events, and reported that it can obtain good results by analyzing the number of users and topics of the post at a certain place. At the same time this research pointed that if people tweeting from the same place use the same words, it is likely that they talk about the same thing, which probably is some noteworthy event. In the user location estimation the key word also be important. Most studies are based on the research [3], it used a probabilistic framework to estimate city-level location based on the contents of tweets without considering other geospatial clues. Their approach achieved the accuracy on estimating user locations within 100 miles of error margin (at best) varying from 0.101 (baseline) to 0.498 (with local word filtering).

I concluded the relationship between keyword and user's location from these studies in the Fig.1. This relationship helped me in finding the idea of estimation.

Fig. 1 the relationship between keywords and users

Method

Above the researches, choosing a keyword is important in the location estimation. In the data crawling part the researchers always choose a selected set of words that show strong locality (termed as local words) instead of using entire corpus to improve the accuracy of predicting. And in this research I mainly discussed about how to estimate the events which has geographical locality. I designed a system to crawling the data from Twitter and analyzed these data by K-means clustering. Fig.2 shows the steps of system.

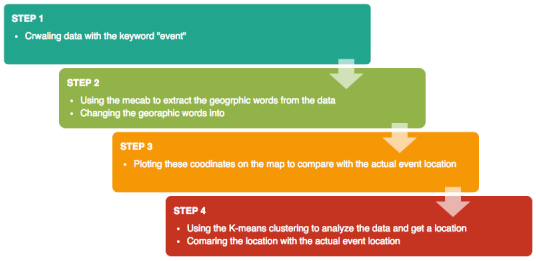

Fig. 2 the steps of location estimation

Considering the define of the event in the chapter 3.2, I choose “地震” as the “event”. Earthquake is easy to observe and has detailed data in the web, and in almost case it happened only at one place what help us to verify the results. I collected the data between 2018.6.21 to 2018.6.26 when the earthquakes frequently happened, and divided it into three datasets with the automatic stop of the API. All of these were crawled by the keyword “地震”.

For increasing the accuracy of result, I removed the spam tweets from the data. As we known these geographical words extracted from the data should contain the location of the event, I conserved the distribution of these words firstly. As the result I found that not all the locations would close to the earthquake location. The map did not show the actual frequency of the tweets posted. For solving this problem I choose the K-clustering to analyze these data.

K-clustering The K-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters K and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

Data assignment:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if c_i is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\underset{c_i \in c}{\operatorname{argmin}}\, dist(c_i, x)^2$$

where $dist(.)$ is the standard (L_2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.

Centroid update:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|}\sum_{x_{i\in S_I}} x_i$$

The algorithm iterates between steps one and two until a stopping criterion is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

For judging the appropriate value of k, in this research I used the elbow method - a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. I used this approach to define the value of k.

In the research I used the K-means clustering to analyze the coordinates changed from the location data to obtain centroids, the estimate location in this research.

Result

Table 1 shows the data I collected. It is obviously to say only few tweets contained Geotag.

Stats	Set A	Set B	Set C
Tweets	43600	248599	137600
Geo tag	11	91	88
The tweet contained word of "地震"	2066	4786	1893

Table 1 the result of crawling

In this paper I found the common feature of these sets is that only one major earthquake occurred at the same time. After observation, I found every clustering result always has a centroid just in Japan area, for example the table 3 shows the centroids of set 2 when k equals 2, 3 and 4. We can found only this centroid is just in Japan area. So in this research I mainly discuss the centroid in Japan.

K	Centroids
2	34.781,135.266
	34.854,135.695
3	34.854,135.695
	38.292,21.892
	29.910,-98.739
4	34.855,135.603
	36.659,-105.839
	38.570,15.589
	-11.548,-55.121

Table 3 The centroids of set 2 when k = 2,3 and 4

Table 4 shows the coordinates of centroids (it is be red in the Fig.3), and then I calculated the distance (error distance -- EDistance) between the centroid and actual location (the data is from [5]).

Set	K	Centroid	Actual	Error Distance(km)
1	3	34.854,135.695	35.131,140.248	432.689
2	2	34.727,135.279	34.832,135.622	33.469
3	4	35.676,140.040	35.348,140.345	45.793

Table 4 the result of error distance

As the result we can see there could be small error distance in the set 2 and set 3.

Discussion

To compare with related works, the research in this paper detected event location without using Geotag. But not all results have high precision. According to the data and results, I summarized the probably conditions what made this method succeed:

(1) It based on using the keyword “event”, the event that has a geographical limitation.

(2) The event was only happened at one place.

But there still has some problem in this work. When there was no earthquake happened, the tweets contained earthquake would be still posted by the people or bot, which will interfere the collection of data. After that, although I used the Mecab to extract the location successfully, the dictionary of Mecab is limited, what means this method is suitable in city level estimation. Besides that the population density may bother the result too, especially the high density of Tokyo and Osaka.

In the future work the major task is resolving this problem, and use other clustering algorithms for comparing the efficiency and quality. As the final goal I want to use this contents-based method in the user location estimation.

Conclusion

In this research I designed a location estimation system without using Geotag. It proved the availability of my method under limited conditions. With the experiment of using K-means clustering to analyze the data collected from the twitter, the result of experiment showed two of data sets get accurate event locations with small error distance -- the distance between the estimate location and actual location. At last, I discussed the possible methods such as adding a classification part in the data crawling, what could increase the accuracy of estimating the location of events based on Twitter in the future work.

References

[1] Rattenbury, T., Good, N. and Naaman, M.: Towards Automatic Extraction of Event and Place Semantics from Flickr Tags, SIGIR, pp.103–110 (2007).

[2] Walther, M. and Kaisser, M.: Geo-spatial Event Detection in the Twitter Stream, ECIR, pp.356–367 (2013).

[3] Z.Cheng, J.Caverlee,and K.Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” in ACM CIKM, Toronto, ON, Canada, 2010, pp. 759–768.

[4] Introduction to K-means Clustering, https://www.datascience.com/blog/k-means-clustering

[5] 気象庁震度データベース検索, https://www.data.jma.go.jp/svd/eqdb/data/shindo/index.php