Discover Overlapping Topical Regions by Geo-semantic Clustering of Tweets

Yuta Taniguchi
Tokushima University
Tokushima 770-8506, Japan
Email: yuta.taniguchi@tokushima-u.ac.jp

Daiki Monzen
Kyushu University
Fukuoka 819-0395, Japan

Lutfiana Sari Ariestien
Kyushu University
Fukuoka 819-0395, Japan

Daisuke Ikeda
Kyushu University
Fukuoka 819-0395, Japan

Abstract

Geotagging is an interesting feature of social media services which adds metadata of geographical locations to photos, web sites or messages. From a different perspective, geotagging can be seen as annotating geographical locations conversely by images or texts. It is a challenging task to summarize such annotations and uncover topical geographical regions characterized by specific topics locally since such knowledge is useful for location-based advertising and so on. Determining topical regions is not trivial since topical region's topic and geographical area are dependent on each other. In this paper, we aim to discover overlapping topical regions from geotagged text messages (tweets) collected from Twitter. To this end, we employ Mean Shift clustering algorithm and an integrated vector space of a geographic and semantic vector spaces. Running Mean Shift algorithm on the vector space, we can evaluate both geographical density and semantic density of tweets simultaneously. Subsequently, our method determines regions of clusters detected by Mean Shift algorithm applying the kernel density estimation on clustered tweets in the geographical space. Our experiments show clusters get broken into several sub-clusters that overlap each other when we increase the weight of semantic density over that of geographical density.

1. Introduction

With the growth of mobile devices, *geotagging* is getting more and more popular among social media services such as Flickr, del.icio.us and Twitter, which add metadata of geographical locations to photos, web sites or messages. A number of studies worked on semantic annotations of web contents employing geographical annotations and textual annotations.

From a different perspective, geotagging can be seen as annotating geographical locations conversely by images or texts. We can consider geotagged social media as a socially-made map with rich annotations. It is a challenging task to summarize such annotations and uncover geographical regions characterized by specific topics locally since such knowledge is useful for location-based advertising and so on.

Determining a topical region is not trivial since its topic and geographical region are dependent on each other. Unlike non-semantic spatial clustering which only consider spatial distribution of data, topical regions can overlap on a geographic space since two regions at the same location could be distinguished by their semantics. Thus we have to consider the distribution of topics and locations simultaneously. Furthermore, regions could take a variety of forms and we should not assume only elliptical regions. The representation of topics is also important. For the affinity for methods in information retrieval or machine learning, it would be better to define a topic as a vector of a vector space model.

In this paper, we aim to discover overlapping topical regions from geotagged text messages (tweets) collected from Twitter. To this end, we employ Mean Shift clustering algorithm [1] and an integrated vector space of a geographic and semantic vector spaces. Running Mean Shift algorithm on the vector space, we

can evaluate both geographical density and semantic density of tweets simultaneously. Subsequently, our method determines regions of clusters detected by Mean Shift algorithm applying the kernel density estimation on clustered tweets in the geographical space. Our experiments show clusters get broken into several sub-clusters that overlap each other when we increase the weight of semantic density over geographical density.

2. Related Work

There are many researches analyzing the relationship between location and semantics of photographs [2]–[4]. In these works, usually tags of photographs are analyzed. Thomee and Rae [5] proposed an algorithm that uncover regions of terms. Those studies mainly target a single term, not a set of words that represents a more complex topic.

Other studies [6], [7] models the relationship among topics, locations, users etc. based on probabilistic models. However, such models needs to be trained, and usually training data cost very much.

3. Method

We consider a topical region is a set of geographical points which are closely related in terms of locations and topics. However, in our method, we aim to find sets of geotagged tweets which are mutually related and then to determine geographical areas from those sets of tweets. We formulate the former as a problem of clustering tweets which are represented as ponits in a combined feature space of spatial and topical spaces.

3.1.Tweet Representation

In our approach, every tweet is represented as a point $(x, y, w_1, \ldots, w_N)$ of $(2+N)$-dimensional vector space. The first two components x and y are geographical coordinates, i.e. longitude and latitude respectively, of tweets. The remaining components are term weights of a bag-of-words representation of tweets.

There are many term weighting schemes such as TF-IDF [8] and BM25 [9]. Given a set of documents D, those scheme weights a term $t \in d$ in a document $d \in D$ considering *inverse document frequency* (IDF), the number of documents in D that contains t, in addition to *term frequency* (TF), a frequency of t in d. It is considered that terms with high IDF values don't specify a topic of a document since such terms are too common or general. Hence, IDF values are successfully used in those schemes to suppress the weights of stop words.

However, we don't use IDF values for weighting terms in our method. Instead, we just use normalized values of TF as term weights. This is because we think generality and spatial distribution of a term are dependent on each other. Therefore, to avoid taking generality of a term into account doubly, we don't consider IDF values.

Let f_{ij} be the frequency of term t_i in a document d_j, and $w_i^{(j)}$ be the term weights of the document d_j. We define the weights as follows:

$$w_i^{(j)} = \frac{s f_{ij}}{\sqrt{\sum_i f_{ij}^2}},$$

where s is a positive real number. Since geographic coordinates and term weights have very different scale, we introduce a scalingfactor s to control the balance between geographic characteristic and semantic characteristic of a tweet.

3.2.Clustering

Mean Shift [1] is a non-parametric algorithm that finds local maxima of density of data points based on kernel density estimation. It is proposed and heavily used in the field of computer vision, and it has been used for tasks such as image segmentation and image smoothing to find regions in which pixels have similar colors.

It has several good properties for region discovery:

1) it doesn't assume hyperspherical clusters,
2) it determines the number of clusters automatically, and
3) it has only a single parameter to control.

These properties are also useful for our task, and hence we employ Mean Shift for discovering regions.

In Mean Shift algorithm, a kernel is used for density estimation. We employ the flat kernel (uniform kernel) for efficiency, since it only needs nearest neighbors within a finite distance. The bandwidth parameter r of the kernel can be manually adjusted to control the geographical size of clusters independently of s.

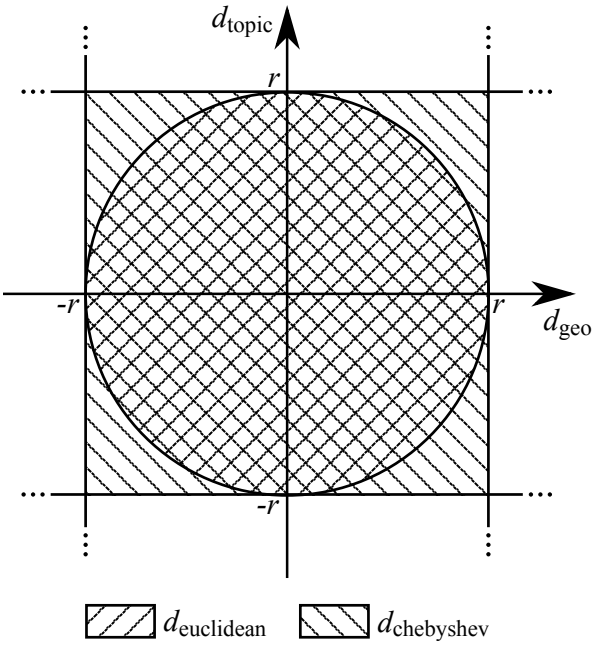


Figure 1. Area within the distance r from the origin in the cases of Euclidean distance and Chebyshev distance.

3.3.Distance Function

A distance between tweets is considered in computation of Mean Shift. We use Euclidean distance of the feature space described above. The distance between tweets $d_1 = (x^{(1)}, y^{(1)}, w_1^{(1)}, \ldots, w_N^{(1)})$ and $d_2 = (x^{(2)}, y^{(2)}, w_1^{(2)}, \ldots, w_N^{(2)})$ is computed as follows:

$$d_{\text{euclidean}}(d_1, d_2) = \sqrt{d_{\text{geo}}^2(d_1, d_2) + d_{\text{topic}}^2(d_1, d_2)},$$

where

$$d_{\text{geo}}(d_1, d_2) = \sqrt{(x^{(1)} - x^{(2)})^2 + (y^{(1)} - y^{(2)})^2}$$
$$d_{\text{topic}}(d_1, d_2) = \sqrt{\sum_i (w_i^{(1)} - w_i^{(2)})^2}.$$

We also employs another distance function, Chebyshev distance:

$$\begin{aligned} &d_{\text{chebyshev}}(d_1, d_2) \\ &= \lim_{k \to \infty} \left(|d_{\text{geo}}(d_1, d_2)|^k + |d_{\text{topic}}(d_1, d_2)|^k \right)^{1/k} \\ &= \max \left(|d_{\text{geo}}(d_1, d_2)|, |d_{\text{topic}}(d_1, d_2)| \right). \end{aligned}$$

Figure 1 shows the difference between areas within a distance r in the case of Euclidean distance and in the case of Chebyshev distance. With Chebyshev distance, we can control the scaling parameter more easily since $d_{\text{euclidean}}$ and $d_{\text{chebyshev}}$ are independently restricted with this distance.

4. Experiment

4.1.Data

We collected tweets using REST API[1] provided by Twitter. Our collection consists of only geotagged tweets that are located within the circle at N33°35′25" E130°24′6" with radius 20 kilometer, around Fukuoka city, Japan. There are 20,117 tweets in the collection.

4.2.Terms

Most of the tweets in the collection are written in Japanese. We used a Japanese morphological analyzer MeCab [10] to extract terms from Japanese sentences. Furthermore, we filtered out terms other than adjectives, verbs, adverbs and nouns using part-of-speech labels given by MeCab.

For the rest of tweets, we simply extracted sequences of alphabetic characters as terms. We didn't filter obtained terms by part-of-speech unlike Japanese terms.

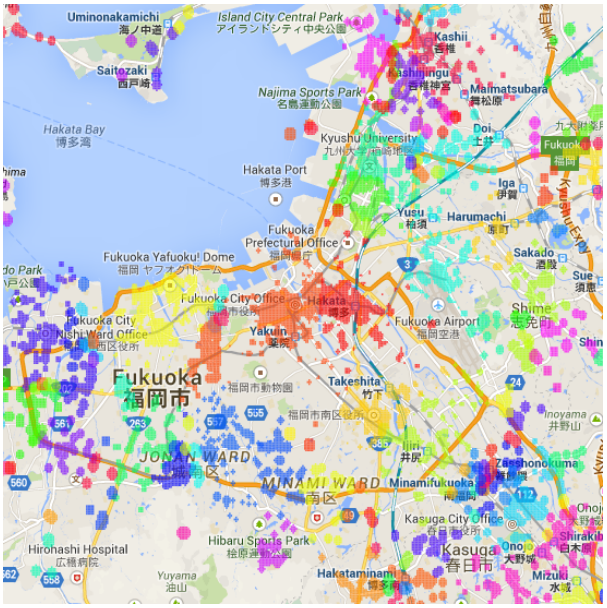
4.3.Region Determination

We determined an actual area of clusters obtained by our method to visualize topical regions. First, we computed the density of tweets for each cluster by kernel density estimation. We used Gaussian kernel with bandwidth 0.0005 for the computation. Then we consider regions with density larger than 1 belongs to a cluster.

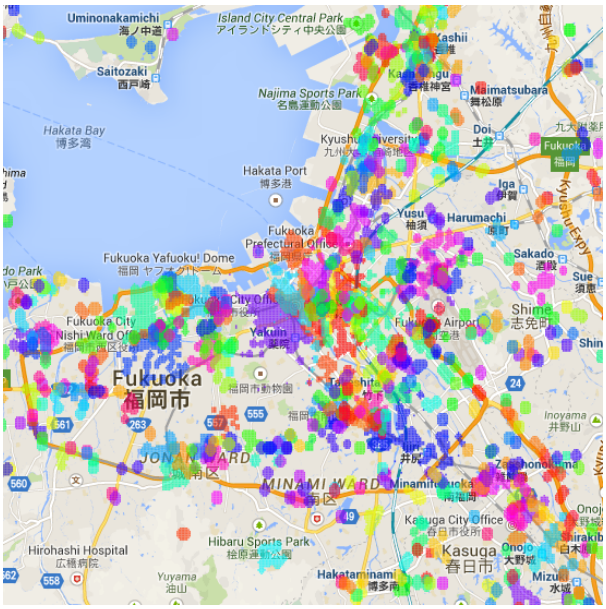
4.4.Results

Figure 2 shows the obtained topical regions on maps: Fig. 2(a) shows the result based on only a geographical distribution of tweets, and Fig. 2(b) and

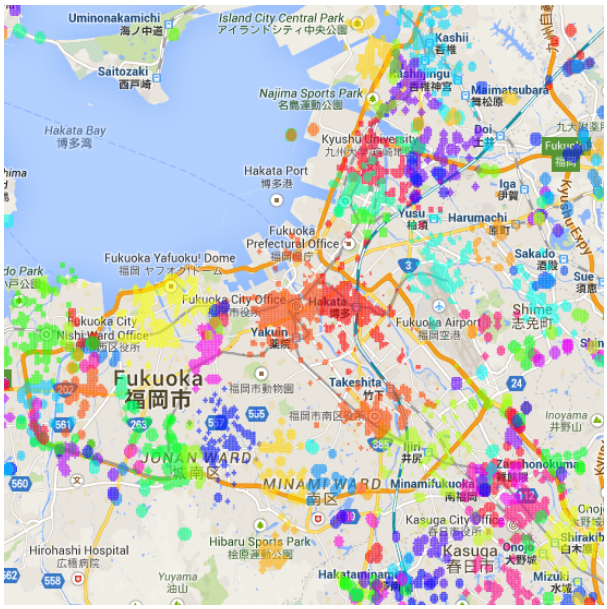
1. https://dev.twitter.com/rest/public



(a) Result of clustering based on only a geographical distribution. ($r = 0.01$, $s = 0$)



(b) Result of geo-semantic clustering with Euclidean distance. ($r = 0.01$, $s = 0.009$)



(c) Result of geo-semantic clustering with Chebyshev distance. ($r = 0.01$, $s = 0.009$)

Figure 2. Obtained topical regions shown on maps.

Fig. 2(c) show the results of geo-semantic clustering with Euclidean and Chebyshev distances respectively.

From the maps we can see that the cluster size vary according to parameter values. Increasing the scaling

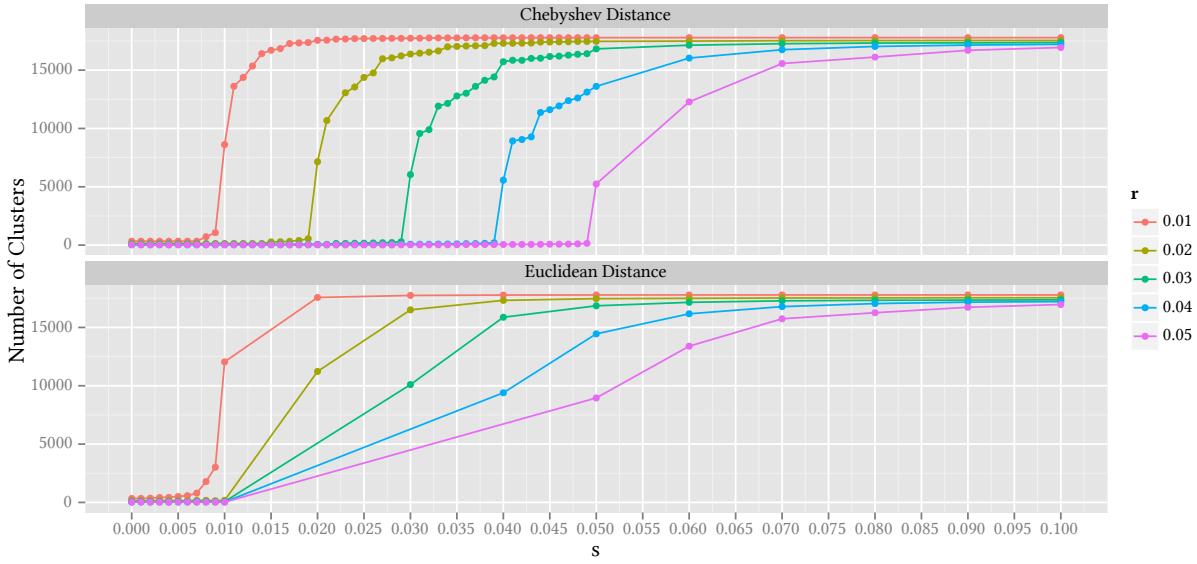


Figure 3. Simulation results

factor s seems to divides clusters into sub-clusters. Comparing distance functions, we can say the resulting clusters are very sensitive to the value of s especially in the case of Euclidean distance.

This observation is shown more clearly in Fig. 3, which shows how the number of clusters changes when we vary the scaling parameter s for each distance. In both cases, we can see that the larger r is the more slowly the number of clusters increase. Furthermore, it is shown that there are obvious threshold for s where the numbers of clusters increases suddenly.

5. Conclusion

We studied on topical region discovery problem. We formulates the problem as a clustering problem in a combined feature space of geographic and semantic spaces. Our method is based on Mean Shift algorithm and Euclidean or Chebyshev distance functions. We performed experiments to show the impact of parameters on resulting topical regions.

Acknowledgment

This work was supported by KAKENHI Grant 24300059.

References

[1] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[2] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 103–110.

[3] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries for large collections of geo-referenced photographs," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 853–854.

[4] L. Cao, J. Yu, J. Luo, and T. S. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 125–134.

[5] B. Thomee and A. Rae, "Uncovering locally characterizing regions within geotagged data," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 1285–1296.

[6] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 769–778.

[7] A. Ahmed, L. Hong, and A. J. Smola, "Hierarchical geographical modeling of user locations from social media posts," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 25–36.

[8] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[9] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994, pp. 232–241.

[10] T. Kudo and Y. Matsumoto, "Japanese dependency structure analysis based on support vector machines," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics, 2000, pp. 18–25.