MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-408

Conditional Density Estimation of Tweet Location: A Feature-Dependent Approach

Hayate ISO, Shoko WAKAMIYA, Eiji Aramaki

Nara Institute of Science and Technology (NAIST), Japan

Abstract

Twitter-based public health surveillance systems have achieved many successes. Underlying this success, much useful information has been associated with tweets such as temporal and spatial information. For fine-grained investigation of disease propagation, this information is attributed a more important role. Unlike temporal information that is always available, spatial information is less available because of privacy concerns. To extend the availability of spatial information, many geographic identification systems have been developed. However, almost no origin of the user location can be identified, even if a human reads the tweet contents. This study estimates the geographic origin of tweets with reliability using a density estimation approach. Our method reveals how the model interprets the origin of user location according to the spread of estimated density.

Keywords:

Social Media; Geographic Mapping; Disease Outbreak

Introduction

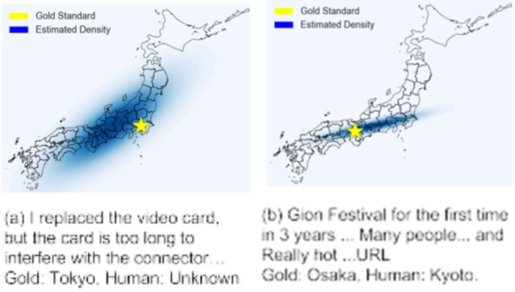
The recent rise in popularity and scale of social media has created a growing necessity for social-media-based public health surveillance. The feasibility of such approaches has been demonstrated using various associated information, including temporal information [1-2] and spatial information [3-5]. Temporal information is associated with all tweets, but spatial information is often unavailable for privacy reasons. One report described that fewer than 0.5% of tweets include GPS information [6]. This problem has become an important motivation for many studies of location estimation [7-11].

Recent studies have elucidated the characteristics of geo-tagged tweets using various approaches. These include location specificity of user attributes such as gender and age [12], linguistic variation [13], temporal effects on location classification accuracy [14], population biases [15], and content-based geographic density of tweets [16]. Our study is aimed at exploring the content-based characteristics of tweet location (System estimated location vs. Geo-tagged location) further. We also investigate differences between the estimated results and the interpretability of human estimation (System estimated location vs. human estimated location). The motivating examples of comparison to humans are shown in Figure 1. Based on these examples, even a human would have difficulty estimating the precise location. However, by some clues, we were able to infer the tweet origin weakly. As this example shows, we can have an idea of a tweet's general region of origin. Therefore, unlike previous studies aimed at estimating the concrete region, our task is to estimate the probability density of the origin location, which more naturally fits human understanding.

A recent study [16] was undertaken to estimate the location as a density estimation problem. Although their motivations resemble ours, this research represents the geo-location identifiability of a given tweet as a combination of word-specific or *n*-gram-specific Gaussian Mixture Model (GMM).

We summarize the contributions of this study as follows:

- We provide simpler and more reasonable approaches to estimate the geographic region of a tweet. Although an earlier study [16] estimated GMM in each word independently, our method handles tweet contents in a vector representation.
- We examine the relation between human inference and geographic biases of geo-tagged tweets.
- We objectively and quantitatively evaluate the differences between human and model inferences.



(a) I replaced the video card, but the card is too long to interfere with the connector…
Gold: Tokyo, Human: Unknown

(b) Gion Festival for the first time in 3 years ... Many people... and Really hot ...URL
Gold: Osaka, Human: Kyoto.

Figure 1 –Estimated geographic distribution of a tweet. "Gold" represents the true origin of the tweet and "Human" represents the human interpretation of the origin of user location from the tweet content.

Materials

Tweet Dataset

We collected 554,320 geo-tagged Japanese tweets for a week (July 15, 2012 – July 21, 2012). For our purposes, we extracted tweets that were posted by Official Twitter clients: *Twitter for iPhone* and *Twitter for Android*. Consequently, our corpus comprises 204,748 tweets.

To estimate the number of mixture components, we split tweets into a triplet of training, validation, and test data. We used 144,748 tweets for training data and 30,000 tweets each for validation and test data. This is the section where the authors describe the methods used at the level of detail necessary to convey the sample size, setting, procedure, datasets, analytic plan, and other relevant particulars to the reader.

Human Annotation Rule

To evaluate how our model interprets origin locations of tweets, humans annotated Japanese region (8-way) or prefecture (47-way) labels for 5,000 tweets that were sampled randomly from the test data. Tweet data with several granularity labels were annotated independently by two annotators. To avoid birthplace and residential place biases, we permitted rough searching for the word included in a tweet to identify its origin location. The human agreement rate was 93.6%, as measured by 100 tweets sampled randomly from the annotated ones. Table 1 presents the results, which indicate the difficulty of this task, in which most tweets were not identified by humans. Furthermore, goodness-of-fit test results showed that tweets with a position were identified by human beings. Results also showed that the multinomial distributions by prefecture differ for learning data tweets.

Table 1-Human annotated data summary: goodness of fit (GoF) test. We tested humans' correctly annotated data with overall geo-tagged tweets as a multinomial distribution at the prefecture level.

Level	Ratio (%)	Precision (%)
Annotated	10.4	87.8
Unknown	89.6	-
Prefecture	9.1	89.7
Region	1.3	74.6

Methods

Word-Specific Gaussian Mixture Model

In an existing approach [16], the GMM has been used for geographic density estimation of geo-tagged tweets. Each tweet is converted to *n*-gram features consisting of the number of *n*-gram occurrences in a corpus and geographic coordinates (longitude and latitude) of *n*-grams. GMM applied for each *n*-gram w_j is defined as word- specific GMM as

$$1)\ p(y|w_j) = \sum_{k=1}^{K} \pi_{kj} N(y|\mu_{kj}, \Sigma_{kj})$$

here $y \in \mathbb{R}^2$ represents geographic coordinates (latitude and longtude), w_j represents the word indexed in j, π_{kj} is the weight as the word w_j is assigned to the k-th mixture component, and $N(y|\mu_{kj}, \Sigma_{kj})$ is a multi variate Gaussian distribution with mean $\mu_{kj}$and covariance Σ_{kj}.

After estimating GMM for each *n*-gram, the weighted sum of word-specific GMM is combined as

$$2)\ p(y|x) = \sum_{j=1}^{J} \grave{\pi}_j\, p(y|w_j)$$

Where x = {w_1 … w_j} represents the words in a tweet, $\grave{\pi}_j$ is the weight of the GMM on the word w_j.

For the formula, it is important to ascertain the weight of each GMM $\pi_j(w_j)$. To date, most methods calibrate parameters for improving the prediction accuracy, but they merely consider the geo-location identifiability of *n*-grams to adjust weights. These approaches also merely consider whether a specific word is included or not. They do not consider the meanings of the respective tweets.

In contrast, this study applies *Gaussian Mixture Regression*, which allows expansion of word-specific GMM further by ascertaining weights of tweet GMM automatically from joint probability distributions of a tweet and geo-location. Consequently, we can impose any kind of feature to our model.

Feature-Dependent Density Estimation

To represent our location density model, we use Gaussian Mixture Regression (GMR) [18], which is formalized by a conditional distribution of a jointly estimated Gaussian Mixture Model (GMM).

Our model need not prepare a specific evaluation index for feature-dependent weight estimation. We can derive the weight of two-dimensional GMM from jointly estimated GMM. Therefore, we designate our model as *feature-dependent*. Depending on given feature vectors, the feature of GMR varies the mixture of weights and each component of Gaussian location and variances. Figure 1 presents examples of conditional geographic density of two tweets.

To obtain the GMR results, we first estimate the joint probability *p(y,x)* of *p*-dimensional tweet representation and two-dimensional Gaussian Mixture Model [17]. Then conditional distribution *p(y|x)* of geo-location *y* for a given tweet *x* can be derived analytically from the joint distribution *p(x,y)* as follows:

$$3)\ p(y|x) = \frac{p(x,y)}{\int_x p(x,y)dx} = \frac{p(x,y)}{p(x)} = \frac{\sum_{k=1}^{K} \pi_k N(x,y;\, \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x;\, \mu_{k,x}, \Sigma_{k,xx})} = \sum_{k=1}^{K} \grave{\pi}_k(x) N(y;\, \mu_{k,y|x}, \Sigma_{k,y|x})$$

where

$$4)\ \mu_k = \begin{pmatrix} \mu_{k,1} \\ \mu_{k,2} \end{pmatrix}, \Sigma\, k = \begin{pmatrix} \sigma_{k,1}^2 & \rho_k \sigma_{k,1} \sigma_{k,2} \\ \rho_k \sigma_{k,1} \sigma_{k,2} & \sigma_{k,1}^2 \end{pmatrix}$$

$$5)\ N(y, \mu_{k,y|x}, \Sigma_{k,y|x}) = \frac{N(x,y;\, \mu_k, \Sigma_k)}{N(x;\, \mu_{k,x}, \Sigma_{k,xx})}$$

$$6)\ \mu_{k,y|x} = \mu_{k,y} + \Sigma_{k,yx} \Sigma_{k,xx}^{-1} (x - \mu_x)$$

$$7)\ \Sigma_{k,y|x} = \Sigma_{k,yy} - \Sigma_{k,yx} \Sigma_{k,xx}^{-1} \Sigma_{k,xy}$$

$$8)\ \grave{\pi}_k = \frac{\pi_k N(x;\, \mu_{k,x}, \Sigma_{k,xx})}{\sum_{k=1}^{K} \pi_k N(x;\, \mu_{k,x}, \Sigma_{k,xx})}$$

The key point of GMR is that the weights of mixture parameters are changed flexibly depending on the feature vector *x*. Consequently, the tweet's vector representation defines the two-dimensional geographical probability density functions *p(y|x)*. GMR is useful for any feature as *p*-dimensional vectors.

As described in the paper, we use a continuous word vector learned by *fasttext* [19] to compress the dimensions of our tweet dataset. We compose a vector representation of tweet by averaging all the word vectors in tweets.

Results

To compare our model with human inference, we evaluated our model through several perspectives.

First, we calculated the distance between our model density and the origin of tweets in a different identifiability dataset such as the prefecture level or region level. We choose the mode value of the estimated distribution as the estimated location and we got the city name from estimated geographic coordinates using Google Map API.

As a baseline method, we used the regularized linear regression method, Elastic-Net [20]. We optimized the baseline model using the validation set.

Our first results are presented in Table 2 and 3. Although our model performs worse than the baseline model, our model monotonically improves prediction performance through human inference improvement.

Second, we comprehensively investigated cases in which our model revealed a result similar to a human's inference, when it failed to estimate origin locations, and when it outperforms human inference. Our characteristic examples are presented in

Figure 2. As the first example, Figure 2(a), shows the human label as coincident with the origin of tweets, in such cases, our model easily estimates the origin of locations. In the second example, our model has failed estimation with high accuracy. Our model density is widely spread. However, humans are also puzzled when specifying the location and labeled the region level. The last example reveals the superiority of density- based estimation. In the tweet, a user stated "Yokohama", which is a city of Kanagawa prefecture. Therefore, the human annotated Kanagawa prefecture for this tweet.

However, the speaker makes remarks at the boundary of the prefecture; the speech position is actually within Chiba prefecture. In a conventional classification approach, we misclassified this tweet, but it is apparent that our method estimates the distribution across both prefectures.

Our model incorporates the uncertainty of the user location estimated from tweet contents. We ascertained that a density-based approach is more reasonable to cover a wider range of class such as a prefecture or region than classification approaches for a difficult tweet to identify the geolocation. To improve the model validity further, feature vectors have an important responsibility. Although this research only employs the textual information for geolocation estimation, the many previous researches empower the geolocation performance via classification. We will further explore which feature has good effects for estimating the tweet's uncertainty.

In addition, we will apply our model for non-geotagged infectious diseases related tweets such as Influenza [2] to explore the regional trends of the infectious diseases.

Conclusion

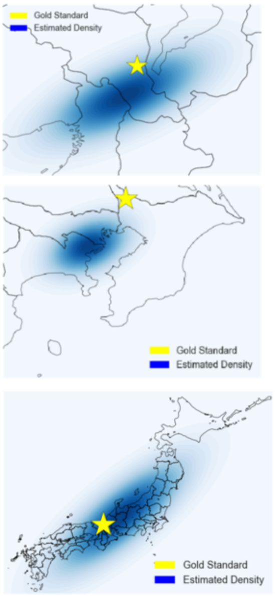
In this study, we demonstrated that GMR provides a new perspective for estimation of the tweet posting origin. We provide simpler and more reasonable approaches to estimate the geographic region of a tweet. Although an earlier study [16] estimated GMM in each word independently, our method handles tweet contents in a vector representation. We examine the relation between human interpretability and geographic biases of geo-tagged tweets.

Table 2-Mean distances: Region data include Prefecture data

Level	GMR (km)	Elastic-Net (km)
Prefecture	251	271
Region	242	268
Overall	278	272

Table 3-Median distances: Region data include Prefecture data

Level	GMR (km)	Elastic-Net (km)
Prefecture	154	181
Region	134	181
Overall	214	191



(a) Well then, from now on, we have lunch at Kanato Rakuhoku's "Kanato Himawari". Gold: Kyoto, Human: Kyoto.

(b) @reply Oh, I really care about sushi in Hokuriku. Gold: Fukui, Human: Hokuriku.

(c) The shirasu bowl served in Yokohama is super nice URL Gold: Chiba, Human: Kanagawa.

Figure 2-Characteristic estimated density examples.

Acknowledgements

Funding sources for the work and other relevant acknowledgements are noted here. Authors may also present disclosures or disclaimers to their work in this section.

References

[1] M.J. Paul, M. Dredze, and D. Broniatowski, Twitter improves influenza forecasting, *PLOS Currents Outbreaks* (2014).

[2] H. Iso, S. Wakamiya, and E. Aramaki, Forecasting word model: Twitter-based influenza surveillance and prediction, in: *Proceedings of the 26th International Conference on Computational Linguistics*, 2016, pp. 76-86.

[3] D.A. Broniatowski, M.J. Paul, and M. Dredze, National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic, *PLoS One* **8** (2013), e83672.

[4] A. Sadilek, H. Kautz, and V. Silenzio, Predicting disease transmission from geo-tagged micro-blog data, in: *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[5] S. Eubank, H. Guclu, V.A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, Modelling disease outbreaks in realistic urban social networks, *Nature* **429** (2004), 180-184.

[6] Z. Cheng, J. Caverlee, and K. Lee, You are where you tweet: a content-based approach to geo-locating twitter users, in: *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, 2010, pp. 759-768.

[7] O. Ajao, J. Hong, and W. Liu, A survey of location inference techniques on Twitter, *Journal of Information Science* **41** (2015), 855-864.

[8] D. Jurgens, T. Finethy, J. McCorriston, Y.T. Xu, and D. Ruths, Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice, in: *ICWSM*, 2015, pp. 188-197.

[9] B. Han, P. Cook, and T. Baldwin, Text-based twitter user geolocation prediction, *Journal of Artificial Intelligence Research* **49** (2014), 451-500.

[10] A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, Exploiting text and network context for geolocation of social media users, *arXiv preprint arXiv:1506.04803* (2015).

[11] Z. Liu and Y. Huang, Where are You Tweeting?: A Context and User Movement Based Approach, in: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ACM, 2016, pp. 1949-1952.

[12] U. Pavalanathan and J. Eisenstein, Confounds and consequences in geotagged Twitter data, *arXiv preprint arXiv:1506.02275* (2015).

[13] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing, A latent variable model for geographic lexical variation, in: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp. 1277-1287.

[14] M. Dredze, M. Osborne, and P. Kambadur, Geolocation for twitter: Timing matters, in: *Proceedings of NAACL-HLT*, 2016, pp. 1064-1069.

[15] M.M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, Population bias in geotagged tweets, in: *ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*, 2015, pp. 18-27.

[16] R. Priedhorsky, A. Culotta, and S.Y. Del Valle, Inferring the origin locations of tweets with quantitative confidence, in: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, 2014, pp. 1523-1536.

[17] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, *The annals of statistics* (1973), 209-230.

[18] H.G. Sung, *Gaussian mixture regression and classification*, Rice University, 2004.

[19] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, Bag of tricks for efficient text classification, *arXiv preprint arXiv:1607.01759* (2016).

[20] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2005), 301-320.

Address for Correspondence

Eiji ARAMAKI, Ph.D. <aramaki@is.naist.jp>