RATE: Overcoming Noise and Sparsity of Textual Features in Real-Time Location Estimation

Yu Zhang[1], Wei Wei[2], Binxuan Huang[2], Kathleen M. Carley[2], Yan Zhang[1]
[1]Key Laboratory of Machine Perception (MOE), Peking University, Beijing, China
[2]School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
yuz9@illinois.edu, {weiwei, binxuanh, kathleen.carley}@cs.cmu.edu, zhy@cis.pku.edu.cn

ABSTRACT

Real-time location inference of social media users is the fundamental of some spatial applications such as localized search and event detection. While tweet text is the most commonly used feature in location estimation, most of the prior works suffer from either the noise or the sparsity of textual features. In this paper, we aim to tackle these two problems. We use topic modeling as a building block to characterize the geographic topic variation and lexical variation so that "one-hot" encoding vectors will no longer be directly used. We also incorporate other features which can be extracted through the Twitter streaming API to overcome the noise problem. Experimental results show that our RATE algorithm outperforms several benchmark methods, both in the precision of region classification and the mean distance error of latitude and longitude regression.

CCS CONCEPTS

• **Information systems** → *Blogs*; *Social networking sites*; *Spatial-temporal systems*;

KEYWORDS

microblog; location inference; real-time; text mining

1 INTRODUCTION

Micro-blogging services such as Twitter, Tumblr and Weibo are regarded as indispensable platforms for information sharing and social networking. In recent years, estimating the location information of social media users has become a popular topic with some important applications. For example, the ability to select a group of users in the specific spatial range can enable analysis on real-time disaster information, localized friendship recommendation or investigations on the geographic variation in habits.

For Twitter users, while tweet text is the most commonly used feature in location inference, most of the prior works suffer from the following two problems of textual features.

Noise. Due to the length limitation of a tweet, there are always lots of non-standard usages of the tweeting language including abbreviations, typos, and emoji.

Sparsity. In contrast with the entire corpus, there is a very small proportion of words appearing in each short tweet. Therefore, the "one-hot" encoding vectors of each tweet will be sparse and hard to deal with.

In this paper, we propose RATE to overcome noise and sparsity of textual features in ReAl-Time location Estimation.

To tackle the noise problem, we incorporate other information available from the retrieved tweet. Figure 1 shows an example of a tweet with 8 metadata extracted through the Twitter streaming API [13]. Note that the exact latitude and longitude can be extracted only if users open their GPS service. But few users choose to do so for the concern of privacy. The absence of GPS signals forces us to rely on other information.[1]

In Figure 1, user's residence, name and description are free-text fields completed during the registration. We can combine tweet text and some free-text fields into a single "document". We name it as *textual features*. Besides, we also have several *categorical features* including time zone, UTC offset, tweet language and user language.

To tackle the sparsity problem, we use Latent Dirichlet Allocation [2] as a building block to deal with textual features so that "one-hot" encoding vectors will no longer be directly used as input.

Our generative model assumes that each location has different distributions over topics, words and categorical features. Therefore, we can infer the location through the observable features. Our model can be applied in either region classification or latitude and longitude regression, and experimental results show that RATE outperforms several benchmark methods in both tasks. As a byproduct, it can also be used to discover the real-time hot topics and find lexical variation of different regions.

Related Work. There is a long sequence of studies in location estimation of social media users. For more details, please refer to [7].

Among various approaches used in location inference, finding location indication words is the most common one. For example, Cheng et al. [4] propose a local word filtering method. Eisenstein et al. [6] incorporate Correlated Topic Models (CTM) [1] to describe the relationships among different local topics. Chen et al. [3] further add user's interests into their topic model.

Another widely used technique relies on social network relationships, so as to infer a user's location from that of its followers and followees [5, 8, 10]. But in our context, a classifier needs to deal with the additional challenge of having to rely only on the information in a single tweet.

[1]In this paper, we adopt the same scenario as [13]. We want to solve the location estimation problem for real-time Twitter streams. In this scenario, it becomes infeasible to retrieve follower-followee relationships or to make plenty of queries to an access-limited database. Therefore, we cannot rely on social connections or some third-party information although it is easy to put them in the model from the technical perspective.

CIKM'17, November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133067

```
{
    [text] I'm at London @HeathrowAirport (LHR) in Hounslow, London https://t.co/XXXX
                                        /*Tweet Text*/
    [lang] en                           /*Tweet Language (Categorical)*/
    [user]{
        [timezone] Atlantic Time (Canada) /*Time Zone (Categorical)*/
        [utc_offset] -10800             /*Offset (Categorical)*/
        [uloc] Quebec                   /*User Location (Free-text)*/
        [lang] en                       /*User Language (Categorical)*/
        [name] John Smith               /*Name (Free-text)*/
        [description] #FightForBigMike  /*Description (Free-text)*/
    }
    [place]{                            /*Location*/
        [country] United Kingdom
        [bounding_box][coordinates] [[-0.508684, 51.455116], [-0.508684, 51.619006],
                                    [-0.376038, 51.619006], [-0.376038, 51.455116]]
}
```

Figure 1: Example of a geotagged tweet with 8 metadata [13]

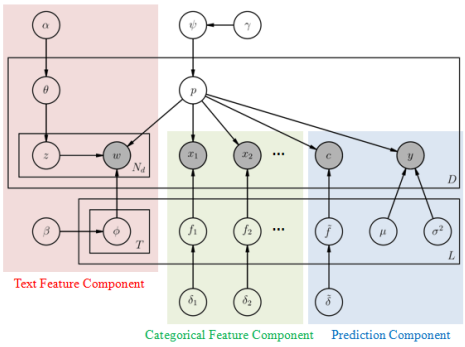

Figure 2: Plate diagram of RATE

Our method is inspired by Zubiaga et al.'s work [13]. They use tweet metadata to train a Maximum Entropy classifier (i.e., logistic regression). However, they adopt "one-hot" representation and ignore the sparsity of textual features.

2 MODEL

We use a Bayesian graphical model RATE to characterize the relationship between tweets, topics and regions. Using plate notation, Figure 2 illustrates the structure of our model.

There are L latent areas. Each area i has T topics $\phi_{i,j}$ (represented by multinomial distributions over the words), F categorical features $f_{u,i}$ (represented by multinomial distributions over the categories) and a region distribution $\tilde{f}_i$. As normal practice, we suppose that all multinomial distributions have a Dirichlet prior. We also assume that the latitude and longitude are extracted from a two-dimensional Gaussian distribution governed by the area's geographical center μ_i and variance σ_i^2. Given the area indicator p, we use the distributions of p to generate words, categorical features, the region indicator and coordinates of the tweet.

RATE has three major components: the textual feature component, the categorical feature component and the prediction component, which have been marked in Figure 2.

The textual feature component has a similar structure with Latent Dirichlet Allocation (LDA) [2]. The only difference is that in LDA, the word is selected by a per-token hidden variable z, while in RATE, the word is selected jointly by a topic index z and a per-tweet area index p.

In the categorical feature component, features such as user language, tweet language and time zone are generated by the multinomial distribution $f_{u,i}$.

The prediction component includes the region indicator c_i and the coordinates y_i of the tweet. In the training set, they are the features which help us cluster the tweets and infer the latent parameters. And in the testing set, they are no longer observable and are the variables we want to predict.

We conclude our generative story as follows:

As common sense, different regions may have different popular topics. Even for the same topic, there exists a geographic lexical variation [6]. Different regions may also have distinct distributions of user's features. For example, French will be the dominating element of the user language distribution in France, while it will not cover a considerable proportion in Germany. Therefore, the observable text and user features of the tweet are strong spatial indicators.

Algorithm 1 Generative Process for RATE

1. Sample the distribution over areas $\psi \sim \mathrm{Dir}(\gamma)$
2. for each area $i = 1, 2, ..., L$
   (1) for each topic $j = 1, 2, ..., T$
      Sample the distribution over words $\phi_{i,j} \sim \mathrm{Dir}(\beta)$
   (2) for each categorical feature $u = 1, 2, ..., F$
      Sample the distribution over categories $f_{u,i} \sim \mathrm{Dir}(\delta_u)$
   (3) Sample the distribution over regions $\tilde{f}_i \sim \mathrm{Dir}(\tilde{\delta}_i)$
   (4) Sample area center and variance $\mu_i \sim N(a, b^2 I)$, $\sigma_i^2 \sim \Gamma(c, d)$
3. for each document $k = 1, 2, ..., D$
   (1) Sample area indicator $p_k \sim \mathrm{Mul}(\psi)$
   (2) Sample the distribution over topics $\theta_k \sim \mathrm{Dir}(\alpha)$
   (3) for each word position $l = 1, 2, ..., N_k$
      Sample topic indicator $z_{k,l} \sim \mathrm{Mul}(\theta_k)$
      Sample word $w_{k,l} \sim \mathrm{Mul}(\phi_{z_{k,l},p_k})$
   (4) for each categorical feature $u = 1, 2, ..., F$
      Sample category indicator $x_{u,k} \sim \mathrm{Mul}(f_{u,p_k})$
   (5) Sample region indicator $c_k \sim \mathrm{Mul}(\tilde{f}_{p_k})$
   (6) Sample coordinates $y_k \sim N(\mu_{p_k}, \sigma_{p_k}^2 I)$

Inference. We use a Gibbs-EM algorithm [11] to infer the model parameters. During the E step, we assume that μ and σ^2 are already known as the result of a previous M step. We then use Collapsed Gibbs Sampling to generate samples for z and p and use the average of these samples to approximate the expectation:

$$\Pr(z_{kl} = z|\neg z_{kl}) \propto (n_{k,*}^{z,*-} + \alpha_z) \cdot \frac{n_{*,r}^{z,p_k-} + \beta_r}{\sum_{r=1}^{V}(n_{*,r}^{z,p_k-} + \beta_r)},$$

and

$$\Pr(p_k = p|\neg p_k) \propto \frac{1}{\sigma_p^2} \exp(-\frac{||y_k - \mu_p||_2^2}{2\sigma_p^2}) \cdot \prod_{l=0}^{N_k-1} (n_{*,*}^{*,p-} + \gamma_p + l) \cdot$$

$$\prod_{j,r:n_{k,r}^{j,*}>0} \frac{\prod_{l=0}^{n_{k,r}^{j,*}-1}(n_{*,r}^{j,p-} + \beta_r + l)}{\prod_{l=0}^{n_{k,*}^{j,*}-1}(\sum_{r=1}^{V}(n_{*,r}^{j,p-} + \beta_r) + l)} \cdot \prod_{u=1}^{F+1} \frac{m_{*,u}^{p,x_{ku}-} + \delta_{u,x_{ku}}}{\sum_{v=1}^{C_u}(m_{*,u}^{p,v-} + \delta_{u,v})}.$$

Table 1: Location prediction results.

Dataset	Europe			UK		France	
Method	Precision	MDE(km)	Time(ms)	Precision	MDE(km)	Precision	MDE(km)
NB	0.8770	427.0	0.20	0.3842	215.0	0.5218	284.6
SVM	0.8796	426.3	3.56	0.4188	205.8	0.5395	**275.8**
GeoTM	0.7973	529.8	16.5	0.4260	195.6	0.5414	277.9
LR-Text	0.7229	726.9	2.48	0.3983	224.6	0.5316	310.9
LR-Full	0.8890	429.2	2.46	0.4207	214.6	0.5413	285.4
RATE	**0.8922**	**372.8**	17.5	**0.4325**	**194.8**	**0.5586**	284.9

Here $n_{k,r}^{j,i}$ denotes the number of times that a document k has a word r that falls into topic j in area i, and $m_{k,u}^{i,v}$ is the number of times that a feature u of document k falls into category v of area i. Note that c_i and y_i are observable in the training set. Therefore, we regard c_i as the $(F+1)$-th categorical feature of tweet i.

In the M step, we estimate μ and σ^2 by maximizing the likelihood function, which is defined as the average over all samples drawn from the E step:

$$Q(\mu, \sigma^2) = \frac{1}{S}\sum_{s=1}^{S} \log(\Pr(O, \Omega^{(s)}|\mu, \sigma^2)) - \frac{1}{2}\lambda||\sigma||_2^2.$$

By solving the equations $\frac{\partial Q}{\partial \mu_p} = 0$ and $\frac{\partial Q}{\partial \sigma_p} = 0$, we acquire an MLE estimation for the center and variance of each region:

$$\mu_p = \frac{\sum_{s=1}^{S}\sum_{k:p_k^{(s)}=p} y_k}{\sum_{s=1}^{S}\sum_{k:p_k^{(s)}=p} 1},$$

and σ_p^2 is the positive root of the following biquadratic equation

$$\sum_{s=1}^{S}\sum_{k:p_k^{(s)}=p} (\lambda\sigma_p^4 + \sigma_p^2 - \frac{1}{3}||y_k - \mu_p||_2^2) = 0.$$

For a tweet in the test set, we can either make a point estimation on the latitude and longitude or make a region classification.

For latitude and longitude regression, we have

$$\hat{y}_k = \frac{\sum_{s=1}^{S} \mu_{p_k^{(s)}}/\sigma^2_{p_k^{(s)}}}{\sum_{s=1}^{S} 1/\sigma^2_{p_k^{(s)}}}.$$

For the classification, we have

$$\hat{c}_k = \arg\max_C \prod_{s=1}^{S} \tilde{f}_{p_k^{(s)},C}.$$

3 EXPERIMENTS

Dataset. We extract a Twitter dataset within the geographical boundary of Europe from October 2015 to December 2015. The boundary is defined by the (latitude, longitude) point (-13.97, 33.81) in the lower-left corner and (41.40, 58.73) in the upper right corner. We remove the users who posted less than 10 tweets during these 3 months and get 376,356 users left. To avoid bias in the dataset, we randomly select one tweet for each user. We name this dataset as Europe.

In Europe, we select all the tweets from UK/France to form another 2 datasets. After the filtering, we have 77,852 and 36,451 tweets in UK and France respectively.

In all of the 3 datasets, we use 60% of tweets for training, 20% for tuning the parameters, and the remaining 20% for final testing. For the tweet text, we remove URLs and the words that occur less than 10 times in the whole corpus. But we retain mentions ("@username"), hashtags ("#topic") and stop words. After the preprocessing, we have a vocabulary of approximately 30K words.

For each tweet, we combine tweet text and the user's profile location into a single "document". It is our textual feature. Besides, we use user language, time zone and tweet language as categorical features.

Evaluation Metrics. For the region classification task, we use *precision* to evaluate the performance, which is defined as the percent of tweets which are predicted in the same region where they are published. Note that in Europe, we directly conduct country-level classification. In UK and France, we divide each country into 4 regions according to the result of K-means.

For the coordinates regression task, we use *mean distance error* (MDE) as our metric. It is the average error distance (on the sphere) between predicted location and actual location.

Effectiveness. We select the following benchmark methods, which are also applicable in the real-time scenario, to compare with our approach.

(1) Naive Bayes (NB) is a basic classification method using only categorical features.

(2) SVM trains a linear Support Vector Machine with both categorical features and textual features.

(3) LR-Text [13] trains a Logistic Regression classifier using only textual features with "one-hot" representation.

(4) LR-Full [13] is similar to LR-Text, but it incorporates both categorical features and textual features. According to the original paper, this combination performs the best.

(5) RATE is the method proposed in this paper.[2]

(6) GeoTM is a simplification of RATE, using only textual features. It is also similar with the method in [6]. The only difference is that Eisenstein et al. use CTM, while we use LDA.

Table 1 shows the location prediction results of the methods mentioned above. As expected, RATE significantly outperforms all the benchmark methods, both in region classification and coordinates regression.

[2]The code as well as the dataset is available at https://github.com/yuzhimanhua/Location-Inference/.

Methods only using textual features, such as GeoTM and LR-Text, perform not so well in country-level classification because categorical features do help a lot in dealing with the noise problem in coarse-grained tasks. However, textual features show their power in fine-grained tasks. We can see that GeoTM performs the second best in UK and France. where categorical features may have less contribution in location estimation. Therefore, if we want to balance the performances of our algorithm in both coarse-grained tasks and fine-grained ones, it will be effective to incorporate both textual features and categorical features into our model. Moreover, we should note that RATE adopts a better way than SVM and LR-Full in dealing with the sparsity of textual features.

Efficiency. Table 1 also shows the running time each algorithm spends on each tweet in Europe. Note that we only calculate testing time and do not take the training phase into account. We can observe that RATE, SVM and LR-Full are almost at the same order of magnitude in efficiency. Since we only adopt the original Collapsed Gibbs Sampling method in RATE, we believe that RATE can be even faster with the help of some acceleration strategies of sampling [9].

Parameter Study. As common practice, we set α to be $50/(LT)$ and other Dirichlet priors to be 0.01.

Figure 3 shows the MDE of RATE in Europe with different numbers of regions (L) and topics (T). We can observe that for each fixed L, the model always performs the best in the case $T = 1$. Therefore we no longer need to sample θ and z, and the structure of β, ϕ, w, p, ψ and γ will be identical to the DMM model [12], which has proved to be effective in dealing with short text like tweets [12].

The "best" L is 30 in Europe, which approximately equals to the number of countries.

Words and Topics. As a byproduct in the training process of RATE, we show the top 8 words in the top 5 regions in Europe in Table 2. The top words can be divided into four categories: temporal words (e.g., "today" and "october"), location names (e.g., "paris" and "spain"), local characteristic words (e.g., "rain" and "wind" in Britain and "love" in France) and hashtags. These four kinds of words correspond to time, locations, topics and events respectively.

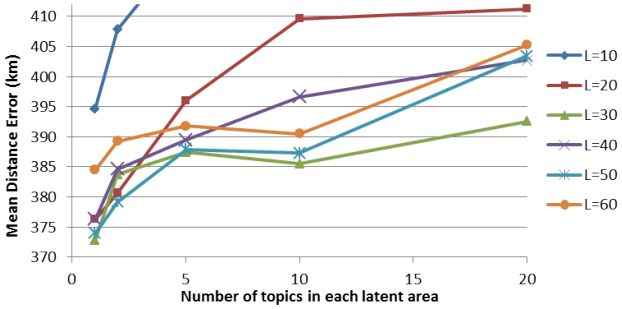

Figure 3: Mean Distance Errors of RATE in Europe with different L and T.

Table 2: Top 8 words of the top 5 region in Europe. Stop words have been removed. The italic words in the brackets are the English translation.

Location	Word
(36.48,31,36) Turkey	türkiye / istanbul / üniversitesi (*university*) izmir / ankara / lisesi (*high school*) / antalya fakültesi (*faculty*)
(56.19,-1.73) UK	london / night / wind / rain tonight / october / year / life
(39.82,-3.49) Spain	hoy (*today*) / madrid / mañana (*morning*) vida (*lifetime*) / spain / octubre (*october*) #gala4gh16 / jueves (*thursday*)
(47.69,2.76) France	paris / france / demain (*tomorrow*) mdr (*lol*) / soir (*evening*) / vie (*life*) journée (*day*) / aime (*love*)
(43.66,11.03) Italy	milano / italy / #xf9 / italia #gf14 / oggi (*today*) / roma / milan

4 CONCLUSION

In this paper, we propose a Bayesian graphical model to overcome the noise and sparsity problems in real-time location estimation on Twitter. The key ideas of our model are that: (1) we use the combination of text information and user profile information to tackle the noise problem. (2) we use topic modeling characterizing the geographic lexical variation to tackle the sparsity problem. Quantitative analysis justifies our model on several Twitter datasets by showing that our approach outperforms several benchmark methods. Qualitative analysis shows that our model is also useful in extracting location-relevant topics.

Acknowledgements. We would like to thank Yujie Qian and Matthew Benigni for valuable discussions, and anonymous reviewers for useful feedback. This work is supported by NSFC under Grant No.61532001 and No.61370054.

REFERENCES

[1] D. M. Blei and J. D. Lafferty. 2006. Correlated topic models. In *NIPS'06*. MIT Press, 147–154.
[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2002. Latent dirichlet allocation. In *NIPS'02*. MIT Press, 601–608.
[3] Y. Chen, J. Zhao, X. Hu, X. Zhang, Z. Li, and T.-S. Chua. 2013. From Interest to Function: Location Estimation in Social Media.. In *AAAI'13*. AAAI, 180–186.
[4] Z. Cheng, J. Caverlee, and K. Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM'10*. ACM, 759–768.
[5] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *KDD'14*. ACM, 15–24.
[6] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. 2010. A latent variable model for geographic lexical variation. In *EMNLP'10*. ACL, 1277–1287.
[7] B. Han, P. Cook, and T. Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49 (2014), 451–500.
[8] D. Jurgens. 2013. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships.. In *ICWSM'13*. AAAI, 273–282.
[9] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola. 2014. Reducing the sampling complexity of topic models. In *KDD'14*. ACM, 891–900.
[10] A. Sadilek, H. Kautz, and J. P. Bigham. 2012. Finding your friends and following them to where you are. In *WSDM'12*. ACM, 723–732.
[11] H. M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *ICML'06*. ACM, 977–984.
[12] J. Yin and J. Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *KDD'14*. ACM, 233–242.
[13] A. Zubiaga, A. Voss, R. Procter, M. Liakata, B. Wang, and A. Tsakalidis. 2017. Towards real-time, country-level location classification of worldwide tweets. *IEEE TKDE* 29 (2017), 2053–2066.