Where are You Tweeting? A Context and User Movement Based Approach

Zhi Liu
Computer Science and Engineering
University of North Texas
Denton, TX 76207
zhiliu@my.unt.edu

Yan Huang
Computer Science and Engineering
University of North Texas
Denton, TX 76207
Yan.Huang@unt.edu

ABSTRACT

Geotagged tweets allow one to extract geo-information-trend, search local events, and identify natural disasters. In this paper, we propose a Hidden-Markov-based model to integrate tweet contents and user movements for geotagging. A language model is obtained for different locations from training datasets and movements of users among cities are analyzed. Home cities of users are considered in modeling the patterns of user movements. Evaluation on a large Twitter dataset shows that our method can significantly improve geotagging accuracy by 55% for home cities and 2% for other non-home cities as well as reduce error distances by orders of magnitude compared with pure text-based methods.

Keywords

Geotagging, Tweet, Hidden Markov Model

1. INTRODUCTION

Social media and mobile devices provide users platforms to share their thoughts, opinions, interests, and whereabouts. With 310 millions monthly active users¹, Twitter is one of the busiest microblogs worldwide. Geotagged tweets allow one to extract geo-information-trend [3], search local events [16], and identify natural disasters [13]. However, geotags are not consistently available due to privacy concerns and other reasons. In fact, only less than 1% of tweets contain a geotag [1] and nearly 90% of Twitter users do not have geo-location enabled². Geotagging is the process of assigning a location to a tweet indicating where it is published. Accurate geotagging, the focus of this work, will provide added values to social media contents for applications such as emergency response, event search, trend detection, and smart transportation.

The availability of user geotagged tweets allows one to profile the geographic distributions of terms in tweets and

these distributions have been used in several existing work for geotagging. Bayesian based method generates a language model for a location and calculates the probability of a tweet belonging to different locations based on the Bayesian model [7]. Given the local language model θ_L , the probability of a tweet T belonging to a location L can be calculated as: $P(L|T) = \frac{P(T|\theta_L)P(L)}{P(T)}$ and $P(T|\theta_L) = \prod_i P(t_i|\theta_L)$.

The location with the maximal probability is assigned to a tweet as its geotag. In [4], the authors also propose a similar method to estimate the locations of tweets based on the distributions of words. Kullback-Leibler (KL) divergence based method measures the difference between a tweet and the local language model [7, 12]. The location with minimal KL value is assigned to the tweet. Gaussian model or Gaussian Mixture Model (GMM) fits geotags of n -grams in tweets into Gaussian distributions. The center of the Gaussian distribution [6] or a weighted sum of GMMs [11] is assigned as the geotag of tweets.

Another line of work uses location relevant words from contents as an important clue. In [2], local geo-parsing is considered to suggest locations. Geotagging is done through classification. Location indicative words (LIWs) in the content are detected mainly by three measurements: term frequency, inverse city frequency, and information gain. These LIWs are used as features to classify tweets into cities. In [15], the authors combine different indicators including content information, user profile, and time zone in estimating the locations of tweets.

All of the above approaches rely solely on social media content for geotagging. However, pure content-based methods leave out the movement patterns of a user. The sequence of geotags of tweets published by a user is aligned with the location sequence of the same user. The movement of a user is affected by various factors such as home location, travel distance, recent activities, and the characteristics of cities. Using user movement patterns can potentially help improve the geotagging accuracy.

Thus in this work, we propose a framework that integrates content and user movement to achieve better geotagging performance. Specifically, this paper makes the following contributions: 1) We introduce movement patterns of users into geotagging; 2) We propose a Hidden Markov Model to integrate language model and user movements for geotagging; 3) The evaluation on a large Twitter dataset shows that our model increases geotagging accuracy by 2% for non-home cities and 55% for home cities as well as reduce

¹<https://investor.twitterinc.com/results.cfm>

²<http://www.beevolve.com/twitter-statistics/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'16, October 24-November 28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2983323.2983881>

error distances by orders of magnitude compared with pure text-based methods discussed above.

2. INTEGRATING MOVEMENT PATTERN IN GEOTAGGING

The problem of geotagging tweets is equivalent to estimating the locations of users where they publish the tweets. Locations of users are not only revealed by tweet content but also by a user's previous and next locations because traveling preferences have a great influence on the next locations of users. In this section, we propose a Hidden-Markov-based model to integrate user movement and tweet content in geotagging. The model also considers the home location of a user to better represent the transition probability for users of different home cities.

In our model, the states of the Hidden Markov Model are the city level locations of users and the state observations are tweets. The state (city) is not directly visible but the observation (tweet) is visible. Therefore, the sequence of tweets generated gives some information about the sequence of cities. Each tweet corresponds to a state which represents the city where the user publishes the tweet. The transition probabilities of the HMM are observed from large number of user movement records.

Figure 1: Estimating geotags of tweets by Hidden Markov Model

More specifically, as it is shown in Figure 1, for each user with N published tweets, he/she will have N corresponding states. Each state represents the location of the user where he/she publishes the tweet. The discrete states of the HMM are the set of all the cities. The goal is to predict a series of states (cities) from a series of observations (tweets) of the user.

Transition Probabilities: Transition probabilities are observed from a large number of user traveling records. For each user u with a list of published tweets $T^u = \langle T_1^u, T_2^u, \dots \rangle$, we obtain the list of cities where he has visited through the geotags of his tweets: $C^u = \langle C_1^u, C_2^u, C_3^u, \dots \rangle$. Given a city C_i , the transition probability from C_i to city C_j , $P(C_j|C_i)$, is defined as:

$$P(C_j|C_i) = \frac{\sum_u C_i^u \rightarrow C_j^u}{\sum_u C_i^u \rightarrow C^u} \quad (1)$$

where $C_i^u \rightarrow C_j^u$ denotes a pair of consecutive cities in the sequence C^u . So $\sum_u C_i^u \rightarrow C_j^u$ is the total frequency of users traveling from city C_i to C_j , and $\sum_u C_i^u \rightarrow C^u$ is the frequency of users traveling from city C_i to any other city. So this probability can be explained as: for the users who publish a tweet in city C_i , they have a chance of $P(C_j|C_i)$ to publish next tweet in city C_j , which is also the transition probability in our model.

Home City Influence on Transition Probabilities: The home location of a user has a great influence on the transition probability. For two users u_1 and u_2 in the same city C where u_1 's home location is in C and u_2 is a visitor, the probability of u_1 to stay in C is much higher than u_2 and the probability of u_2 to return to his home city is very high as well. So, home city greatly influences the transition probability between two cities.

To address this issue, we train transition probabilities separately based on home cities of users. We use U to denote all users in the dataset and U_{C_i} is the users with home city C_i . For a user u with home city C_i : 1) When u moves between city C_i and C_j (includes $j = i$), $P(C_j|C_i)$ are observed from all the users in U_{C_i} ; 2) When u moves from city C_j to city C_k , $i \neq j$ & $i \neq k$, the training data are from users in $\{U - U_{C_j}\}$. For the second case, we combine all the users whose home city is not C_j together due to the lack of enough training data for users in every city.

Observation Probabilities: The observation probability gives the likelihood of an observation from a given state. For geotagging, given a tweet T , there is an observation probability $P(T|C)$ for each city C . From the geotagged tweets, we build the language model θ_C for each city. The probability of term t from a tweet which published in city C can be estimated as: $p(t|\theta_C) = tf_{(t,C)}/dl_C$, where $tf_{(t,C)}$ is the frequency of term t in city C and dl_C is the total frequencies of all terms in city C [10]. Then given the state of city C , the observation probability $P(T|C)$ is defined as:

$$P(T|C) = \prod_{t_i \in T} P(t_i|\theta_C) \quad (2)$$

In this work, we smooth the probability of $P(t_i|\theta_C)$ by using the Dirichlet smoothing [17]. The stop words³ are removed before calculating the observation probabilities.

Inference: The probability of the initial state, $P(S_0 = C_j)$, is given by the transition probability of $P(C_j|C_i)$, where C_i is the home city of the user. This is based on the observation that a user is likely to travel to any city from his home city and the first tweet observation is not necessarily from his home city. Then based on the Viterbi algorithm, we find the city sequence which can maximize the product of the observation probabilities and transition probabilities as the geotags of tweets. Assuming the total number of tweets of a user is N , the number of cities is $|C|$, and each tweet contains $|T|$ terms, the time complexity of estimating the location of these N tweets is $O(N \times |C|^2 + N \times |C| \times |T|)$.

3. GEOTAGGED TWITTER DATASET

We collected tweets published from May 21 to July 6, 2015 using Twitter Streaming API⁴ with the spatial bounding box $[24^\circ 3'N, 127^\circ 7'W, 48^\circ 8'N, 65^\circ 5'W]$, which covers U.S. mainland. The dataset consists of over 60 million geotagged tweets from 2,200,402 Twitter users.

Knowing a user's home helps to understand user movement patterns. Users are bound by travel distance from his home and the types of outbound cities from home affect visiting possibility. However, only 16% users register their city

³<http://xpo6.com/list-of-english-stop-words/>

⁴<https://dev.twitter.com/streaming/overview>

(a) The distribution of number of users in 40 selected U.S. cities. (b) The distribution of number of tweets in 40 selected U.S. cities.

Figure 2: The distributions of users and tweets in the dataset.

	Overall			Non-Home City			Home City		
Methods	Mean	Med.	Acc.	Mean	Med.	Acc.	Mean	Med.	Acc.
Random	2847	2450	3.86	2252	2334	4.9	2845	2439	3.8
Bayesian	1402	1022	36.7	1245	765	31.8	1375	994	36.8
KL divergence	1318	954	38.8	1191	741	32.8	1301	944	38.9
HMM-based	52.6	1.82	95.1	997	181	34.6	52.4	1.82	96.3

Table 1: Mean Error Distance (km), Median Error Distance (km), and Accuracy(%) of the geotag estimation results.

level home locations [8] and collecting the home location from user profile is strictly limited by the rules of Twitter API.

Previous work has provided different methods for estimating the home locations of users [9] based on analyzing content information or geographic distribution of friends. In this work, we assign home location following the method used in [14, 5]. A set of tweets of each Twitter user is selected to estimate the home location (this set will not be used for building and testing geotagging models). These tweets are assigned into different cities according to their geotags. We use city boundaries from the United States Census Bureau⁵ to decide which city a geotagged tweet belongs to. Then we assign the city with the most number of tweets as a user’s home. In [14], the authors show that the home location obtained in this way has an 85 percent accuracy.

We then select all cities with more than 5,000 users. There are 40 cities C in our dataset. We only keep the tweets: (1) tweeted by users in C ; (2) published in C . Figure 2 shows the number of users and the number of geotagged tweets in these 40 cities. The top three cities with the most number of users and tweets are Los Angeles (46,508 users; 1,183,634 tweets), Chicago (30,689 users; 634,665 tweets), and New York (30,689 users; 419,130 tweets).

4. EXPERIMENT

We implement four algorithms and compare their performances under different cases. To reduce the impact of spammers and robots, we only keep users with number of geo-tagged tweets between 5 and 500 in our dataset [12]. The first quarter of tweets (sorted by published time) of users are used to estimate home cities. We then run a 10 fold cross validation on the rest of the dataset. Here it

should be noted that all the tweets from the same user will be assigned to the same fold.

Methods: We use the random selection method as the baseline method and select two methods proposed by the related works to compare with our algorithm.

- **Random selection (Random):** Select a city as the geotag location of tweet randomly.
- **Bayesian:** The Bayesian-based method described in introduction [7].
- **Kullback-Leibler divergence (KL divergence):** The KL divergence between tweet T and city C is calculated as equation 3 and the geotag is based on the ranking of the KL divergence [7, 12].

$$KL(\theta_T|\theta_C) = \sum_t p(t|\theta_T) \log \frac{p(t|\theta_T)}{p(t|\theta_C)} \quad (3)$$

- **HMM-based:** The method proposed in this paper.

Measurements: A tweet is assigned to a city as its geotag. We then assign the coordinate of the city center as the predicated location of the tweet. The city center we used is listed in Wikipedia⁶. We use three measures:

- **Accuracy (Acc).** The percent of tweets which are predicted in the same city where they are published.
- **Mean Error Distance (km):** The average error distance between predicted location and actual location.
- **Median Error Distance (km):** The median error distance between predicted location and actual location.

⁵<http://www.census.gov/en.html>

⁶https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

In this paper, we define the distance as the geographic distance.

Table 1 shows that our HMM-based methods can significantly improve the performance of geotagging compared with related works. The overall accuracy is improved by more than 55%; and the mean error distance and median error distance of our method are orders of magnitude better.

We are interested in learning if the improvement is mainly due to the reason that more than 90% of tweets are published in home cities which tend to be easier to estimate. We separate the performance for home cities and non-home cities. Since more than 90% tweets are published in home cities of users, all of these algorithms achieve similar performances in the home cities with the overall case. For the difficult case of non-home cities, the two algorithms from related work, the Bayesian-based method and the KL divergence method, perform better than they do in home cities. This illustrates that when users travel to non-home cities, they may tend to use terms with local characteristics. By contrast, when users stay in home cities, their words are more commonly used with less regional information. However, for the tweets published in these non-home cities, our HMM-based model still outperforms other two methods by nearly 2% in accuracy and the error distances are significantly smaller at the same time. Compared with related works, the improvement on error distances demonstrates that even for the incorrectly estimated cases, our model can locate the tweet with a closer city from the real location of the tweet due to the benefits from the usage of patterns of user movements. In general, the results demonstrate that integrating home city and movement will benefit geotagging in both home and non-home cities.

5. CONCLUSION

In this paper, we propose an HMM-based algorithm for geotagging by integrating a local language model and user movements. We evaluate our model on a large Twitter dataset under different cases. The results show that our method can improve the accuracy by more than 55% percent compared with the baseline algorithms. Even for the tweets published in non-home cities, our method still achieves better results on accuracy and error distances.

6. REFERENCES

- [1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventtweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.
- [2] H. Bo and P. C. T. BALDWIN. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062, 2012.
- [3] C. Budak, T. Georgiou, D. Agrawal, and A. El Abbadi. Geoscope: Online detection of geo-correlated information trends in social networks. *Proceedings of the VLDB Endowment*, 7(4):229–240, 2013.
- [4] 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*, pages 759–768. ACM, 2010.
- [5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD*, pages 1082–1090. ACM, 2011.
- [6] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 127–136. ACM, 2015.
- [7] S. Kinsella, V. Murdock, and N. O’Hare. I’m eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.
- [8] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD*, pages 1023–1031. ACM, 2012.
- [9] Z. Liu and Y. Huang. Closeness and structure of friends help to estimate user locations. In *International Conference on Database Systems for Advanced Applications*, pages 33–48. Springer, 2016.
- [10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.
- [11] 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 and social computing*, pages 1523–1536. ACM, 2014.
- [12] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.
- [13] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- [14] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.
- [15] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *ICWSM*, 2013.
- [16] J. Weng and B.-S. Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.
- [17] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.