Tweet Geolocation Error Estimation

E. Holbrook[1], G. Kaur[1], J. Bond[1], J. Imbriani[1], C. E. Grant[1], and E. O. Nsoesie[2]

[1]University of Oklahoma, School of Computer Science
Email: {erik; cgrant; gkaur; jared.t.bond-1; joshimbriani}@ou.edu

[2]University of Washington, Institute for Health Metrics and Evaluation
Email: en22@uw.edu

Abstract

Tweet location is important for researchers who study real-time human activity. However, few studies have examined the reliability of social media user-supplied location and description information, and most who do use highly disparate measurements of accuracy. We examined the accuracy of predicting Tweet origin locations based on these features, and found an average accuracy of 1941 km. We created a machine learning regressor to evaluate the predictive accuracy of the textual content of these fields, and obtained an average accuracy of 256 km. In a dataset of 325788 tweets over eight days, we obtained city-level accuracy for approximately 29% of users based only on their location field. We describe a new method of measuring location accuracy.

1. Introduction

With the rise of micro-blogging services and publicly available social media posts, the problem of location identification has become increasingly important. Accurately assigning a geolocation to a tweet is extremely useful to a broad range of applications, like tracking the spread of disease through social media (Paul and Dredze, 2011). However, few users report their true location online. Only around 1% of tweets contain accurate location information on where the tweet was created, as retrieved from the Twitter API. (Culotta, 2014).

Previous methods use machine learning models trained on the textual content of the tweet and the user's past tweets to identify city-level location information (see § 2). These methods are heavily dependent on heuristics and the granularity of the training data: any extension of the techniques would require intense restructuring of the classifiers. Investigation into a different geographic region would require new data, retraining the classifiers, and a complete rework of the location types to be identified. Furthermore, similar methods require large amounts of data, e.g. all available tweets from each user or a large social network graph. These methods are often time- and cost-prohibitive: access to a large portion of the Twitter stream is expensive.[1] In contrast, we attempt to predict geolocations given only a single tweet.

In this paper, we examine the reliability and accuracy of predicting tweet origin locations based on meta-content of the of the tweet itself, specifically, the user-supplied 'location' and 'description' fields, as well as propose a lightweight system for identifying the location of a user.

[1]http://readwrite.com/2010/11/17/twitter_to_sell_50_of_all_tweets_for_360kyear_thro/

2. Related Work

The Twitter public API allows for programmatic access to a subset of all public tweets, and with open-source libraries like Tweepy.org, developers can investigate social trends. Identifying the locations of these trends has been examined from several different angles (Compton *et al.*, 2014; Zhang and Gelernter, 2014; Priedhorsky *et al.*, 2014). The common focus is predicting the home location of a Twitter user based on their past tweets.[2]

Compton *et al.* (2014) sought to identify the home location of users through a large social network consisting of approximately 110 million users and reciprocated mentions between users. Home locations were estimated by the strength of connections between users with known locations. They reported city-level accuracy for around 90% of the users.

Priedhorsky *et al.* (2014) sought to combine social features with tweet content features, and to produce a prediction with a quantitative error estimation in terms of distance. To this end, they proposed a location-estimation method based on Gaussian mixture models, and presented several variations of their algorithm for this estimation, which ranged in accuracy from approximately 1700 to 8000 kilometers.

Each of these methods requires a large amount of operational overhead, i.e. a very large number of tweets from each user and other data obtained from the Twitter API, or a complicated prediction mechanism. There is also a lack of a uniform, straight-forward error measurement: each defines it in a different way, making direct method comparisons between methods difficult.

Our approach addresses these issues in several ways. First, we attempt to identify the origin location of a specific tweet, that is, the location of the user when they composed and delivered the tweet as opposed to the main home location of the user. This definition simplifies API requests and data scale requirements compared to other researchers. Secondly, we focus specifically on the meta-features of the tweet, namely the location and self-description, as opposed to the textual content of the tweet. Finally, we build a fast estimator. Our priority is to quickly and simply locate the origin of a tweet with an accuracy in term of distance to the true location.

3. Methods

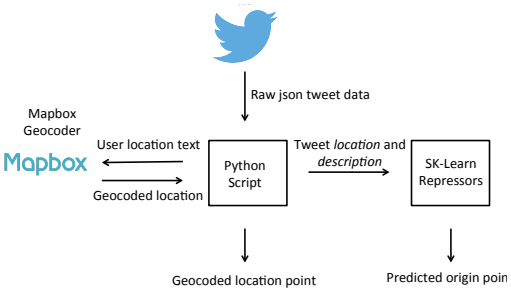
Our dataset consists of tweets collected from April 11 to April 19, 2016. The Twitter API was used to identify users in Oklahoma who supplied location information with their tweets. 325788 tweets were collected during this period. A python program sent the unaltered text in the user 'location' fields to the Mapbox geocoder[3]. In parallel, two Scikit-learn (Pedregosa *et al.*, 2011) SVR regressors were trained on the 'location' and 'user description' fields of the tweets.

4. Experiments

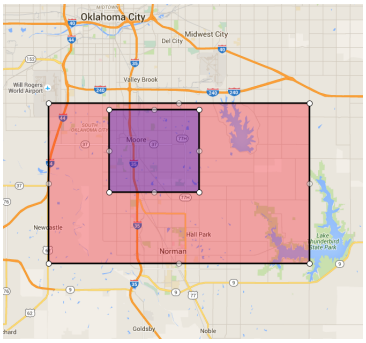
First, we determined what percentage of users in our data set reported a location which could be resolved (not necessarily correctly) with the Mapbox API. We found that approximately 56% of users' locations could be resolved. Next, we examine the accuracy of these locations as compared to the actual origin. The location field is geocoded with the Mapbox geocoder API. If a location is identified, its coordinates are determined via the API. Finally, we built two estimators with the Scikit-learn library to estimate the origin of the tweet based on the meta data to predict the latitude and longitude, respectively. The user's 'location' and 'user description' fields were tokenized into uni- and bi-grams, and each regressor was trained independently on the first 90% of the tweet

[2]Twitter Geolocation `https://dev.twitter.com/overview/terms/geo-developer-guidelines`.

[3]`https://www.mapbox.com/api-documentation/#geocoding`



(a) High level system architecture.



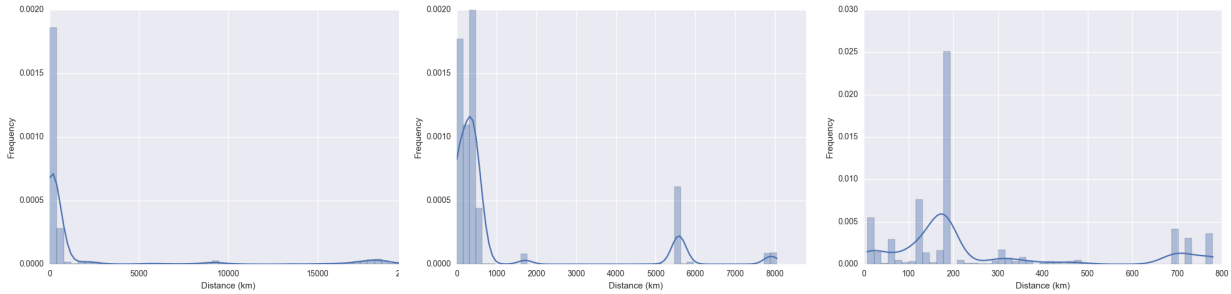
(b) Example tweet Mapbox geocoder and Twitter API bounding box.

Figure 1: Figure 1a is the high level architecture. In Figure 1b the outer, red box represents Mapbox geocoder results. The inner purple box corresponds to the Twitter API location data.

dataset, leaving the remaining 10% for testing.

Since both the geocoder and the Twitter API describe users' locations as bounding boxes, we require an error calculation method that can describe an average distance, and also the potential range of that distance to deal with the problem of overlap. We model the distance between the actual and predicted regions as the average distance between any two random points selected respectively from the two region. We assume that any point within the regions could be the origin point with an equal probability. Therefore, the two bounding boxes are simply two uniform probability distributions, whose mean distance is simply the distance between the two centroids: $\mu_{AB} = \mu_{center_A} - \mu_{center_B}$.

To address the problem of overlap and containment, we propose examining the standard deviation of the mean of μ_{AB} : $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2}$, where σ^2 is in the latitudinal or longitudinal direction. If two regions are centered exactly on the same point, the mean distance will be zero, but this calculation reveals how much the two regions differ in size.



(a) Total error. **(b)** Standard deviations. **(c)** Regressor prediction error.

Figure 2: Distribution of error from geocoding.

5. Discussion and Summary

This method of error calculation relies on two assumptions. First, the tweet could have originated from any point with uniform probability. Second, both regions are rectangles, whose respective

sides are parallel. The first holds for relatively small regions (i.e. U.S. state-sized areas) which are far from the poles. The second assumption is satisfied by the fact that the Twitter API returns bounding boxes a few kilometers in size.

We found almost 30% of the tweets with *location* field data can be resolved to city-level accuracy by simply geocoding the text with no alterations. Our mean accuracy was approximately 1941 km, though the long tail distribution in Figure 2a suggests that future work on filtering could be applied to improve this result greatly.

Hecht *et al.* (2011) reported that approximately 34% of users had either non-geographic information or simply nothing entered in the 'location' field. Our results suggest that automated techniques could eliminate these from the dataset and thus improve overall prediction accuracy. We found that locations with a mean distance greater than 500 km had sarcastic or nonsensical locations, for instance: `Im out here` and `most likely school`. These locations were still resolved by Mapbox to physical locations.

The results from the two regressors are intended to demonstrate the usefulness and validity of the error estimation, and to give context to the geocoding results. Two regressors yield a prediction accuracy of 256 km, which is expressed in terms of a mean distance from the actual origin location. Our method yields a reliable prediction error estimation for all points. Classification here, while useful for the correctly classified tweets, is useless for the incorrectly classified data points.

In this work, we made several steps toward location prediction of social media users. We have examined the accuracy and reliability of geocoding users' location information as a way of predicting tweet origin locations. We have also demonstrated the potential utility of this information through machine learning on the textual content. Finally, we have established a robust and straightforward method for accuracy measurement in terms of the distance between the predicted location and the ground-truth location of tweets. In the future, we will engineer several techniques to improve accuracy and evaluate the results on a larger data set.

6. Acknowledgements

Thanks to Le Gruenwald for her helpful comments and the Robert Wood Johnson Foundation for their generous support.

References

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE.

Culotta, A. (2014). Estimating county health statistics with twitter. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1335–1344, New York, NY, USA. ACM.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Priedhorsky, R., Culotta, A., and Del Valle, S. Y. (2014). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536. ACM.

Zhang, W. and Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70.