An Unsupervised Approach to Identify Location based on the Content of User's Tweet History

Satya Katragadda, Miao Jin, and Vijay Raghavan

Center for Advanced Computer Studies, University of Louisiana at Lafayette, Lafayette, USA
satya@louisiana.edu, mjin@cacs.louisiana.edu, raghavan@louisiana.edu

Abstract. We propose and evaluate an unsupervised approach to identify the location of a user purely based on tweet history of that user. We combine the location references from tweets of a user with gazetteers like DBPedia to identify the geolocation of that user at a city level. This can be used for location based personalization services like targeted advertisements, recommendations and services on a finer level. In this paper, we use convex hull and k-center clustering, to identify the location of a user at a city level. The main contributions of this paper are: (i) reliability on just the contents of a tweet, without the need for manual intervention or training data; (ii) a novel approach to handle ambiguous location entries; and (iii) a computational geometric solution to narrow down the location of the user from a set of points corresponding to location references. Experimental results show that the system is able to identify a location for each user with high accuracy within a tolerance range. We also study the effect of tolerance on accuracy and average error distance.

Keywords: Social Media, Location Analysis, Computational Geometry, Convex Hull, K-Center Clustering

1 Introduction

Social Media usage has increased exponentially in the recent years. Increase in the usage of Twitter, a micro-blogging service, is a good example. These microblogs enable people to post smaller, frequent updates on events happening around them, as well as opinions and details of their personal life. This popularity has led to massive amounts of user-generated data, which provides variety of opportunities and challenges for research. They range from election prediction [32] to event detection [27] and stock market prediction [4]. Mining user centric data like location would help improve the accuracy of the above applications. Location also plays an important role in identifying information appropriate to the user, or events near him.

Traditionally, most of the applications identify the locations of their users by their IP addresses. The reliability of these techniques varies depending on

the country and the continent[1] . Yuval et al. estimates that the accuracy of IP prediction in United States is 44% within 40km radius and 80% within 100km radius [29]. Same methods perform better across in European countries ranging from 78% over 40km and 90% within 100km radius. However, the use of VPN networks and dynamic allocation of IP addresses by ISP mask the original location of the user. Moreover, this information is not available to the research community from either Twitter or ISP's due to privacy concerns. Thus, we require other methods to identify location of the user.

There are two ways that Twitter shares the location of the user: geotagging the tweet with location of the user when it is tweeted for those users who have their location services enabled on mobile devices and the location field of the profile. However, less than 1% of the publicly available tweets are geo-tagged due to privacy concerns [14] and about 34% of the users provide incomplete or inaccurate information in their location fields [17]. This number is further reduced when we are looking at the city level location information and the number of people using street level data is extremely small. Another important factor is that people do not update their information as they move from place to place.

We try to solve this location identification problem by identifying the location of a user using references to locations present in his/her tweets. The location here refers to the area a user resides in or a place the user is associated with or interested in at a city level. By city level we refer to n miles around a geolocation (latitude and longitude), the value of n is 30, 60 and 90 miles for this paper. We use the content of the tweets to identify location information associated from the tweets. We try to extract these location cues to predict the location of user.

One of the main challenge with this approach is the nature of social media. Tweets are usually noisy, containing a variety of information. We need to separate location information from this noise. The second problem is the ambiguity of location names. For example a location name can refer to multiple locations e.g., *Lafayette* can refer to 20 towns around the world, not to mention *Lafayette Park* or *Mount Lafayette*. This is one of the major problems with using content information from tweets, as the 140 characters for a tweet does not provide the context to identify the location. The third problem we encounter is that the interests of a user are fleeting; he may refer to multiple events around the world. A user might be talking about *St. Patrick's Day parade* in *New York* and then refer to *Crimea* in the next Tweet. We need to narrow down the location of the user. Our intuition is that the user mostly refers to things happening around him or those he is interested in, instead of the events happening across the world. Finally, the user may have multiple locations e.g., during a vacation or travel. It might be difficult to narrow down the location of the user.

Most of the previous works in this area as introduced in Section 2 concentrate on building language models or complex graph based analysis techniques to identify location information. These techniques require a lot of training data and time to process the input before identifying the location. In this paper, we solve the problem of location identification by using traditional clustering

[1] `http://www.maxmind.com/en/city_accuracy`

and computational geometric approaches to identify the location of a user at a city level with high accuracy without any training data. We also handle the ambiguous location references in a Tweet by considering all the relevant locations rather than just one location suggester by a gazetteer.

The rest of the paper is organized as follows: An overview of previous work is given in Section 2. Section 3 contains the methodology of our approach. Section 4 explains the data that we used in our experiments. In Section 5, we evaluate our techniques and compare them to baseline and alternative techniques. Section 6 presents some extensions and future work to our approach along with conclusions and future work.

2 Previous Work

The task of identifying geolocation from IP addresses webpages has been studied in detail in the past [5] along with location information from the content of webpages [1, 30, 34] and search query logs [2, 33]. Location was also predicted from various user-generated content like blog posts [11], Wikipedia edit logs [20], Flickr tags [8,16,24] and Facebook locations based on graph framework [3] based on the assumption that users talk more about locations they are familiar to them.

Location prediction on Twitter user based on models built using content of tweets was first studied by Cheng et al. [7]. They built a probabilistic framework to identify the location of user at a city level by identifying words specific to a particular location. This approach requires manual identification of words local to a particular location to build a statistical predictive model. They used about 4,124,660 tweets from 130,689 users to train their model and predict the list of cities in decreasing of probability of the city being the user's home city. They report an accuracy of 51% within 100 miles for 5,190 users with about 5 million tweets with about 1000 tweets per user.

Kinsella et al. [18] created language models of locations using the coordinates extracted from geotagged tweets. They predict the location of an individual tweet using the language models. The location of the user is then arrived at by aggregating all the locations of his/her tweets. They used 7.3 million tweets from Twitter Firehose that were reverse geo-coded using Yahoo Placemaker service. They used 80% of data for training, 10% for tuning and 10% for testing their model. They report a prediction accuracy of 53% for country level, 31% for state level, 29% for city level and 13% for zip code level for tweet locations. For location prediction on users, they report a accuracy of 76%, 34%, 28% and 14% for country, state, city and zip code level predictions respectively.

Hetch et al. [17] built a Multinomial Naïve Bayes model to identify words with a local focus and their respective locations. They use locations from user profile in their model, resolved using Wikipedia-based geocoder and further refined using data from ESRI and US census. They used about 99,296 users from 4 countries who had more than 10 tweets for training the model. Their model identifies the country for 2500 users with 73% accuracy and 30% accuracy for 500 users on state level. Eisenstein et al. [10] built topic models for linguistically

consistent regions and states to predict location of a user. They used 9,500 users and 380,000 tweets to predict the location of users with an accuracy of 58% over a region and 24% to predict on a state level.

Mahmud et al. [21] designed an ensemble of location classifiers using tweet content, tweeting behaviours and timezone to identify location of a user at a city level. They used USGS gazetteer to reverse geocode the locations. They report an accuracy of 67% at a city level for 9500 users. Schulz et al. [28] designed a multi indicator model incorporating multiple spatial indicators, like locations from tweets, location from profile information, time zone etc. to predict the location of the tweet. They represent all these indicators in the form of a polygon to identify location of the tweet. They report that they were able to geotag 92% of tweets with a median error of 30km and predict 79% of users location in a 100 mile radius.

Unlike previous work on this topic, we use just the names of locations and landmarks mentioned in the tweet to predict the location of the user. The approaches mentioned above [7, 10, 17, 18, 21], rely on supervised and semi-supervised techniques to identify words related to a location, and then predict the location of the user. In addition, our approach also handles the disambiguation of location, unlike [21, 28] which relies on a gazetteer to identify the best match for a location reference in the tweet.

There has been some work done on identifying ambiguous location references in tweets. Gelernter et al. [12] used named-entity recognition software to extract location names from tweets and found that most of them do not perform location identification from tweets out of the box. Sultanik et al. [31] use named-entity recognizer on n-grams extracted from tweets. Their method then matches these location entries to RapidGeo gazetteer. Paradesi [23] designed a model to predict location of tweet using a POS tagger to identify noun phrases and then compared them to USGS location database to identify all the location names in a tweet. They identify the location of the tweet by calculating the distance between the location references from USGS database to the one in user profile to pick the nearest location. The author reports an accuracy of 15% from 2000 tweets in his/her dataset.

There has also been research done in identifying location of a user based on relationship and location of his friends. Nowell et al. [19] studied the influence of distance in social ties. Backstrom et al. [3] used social ties between friends on Facebook to identify location of a user. Rout et al. [25] and Sadilek [26] built a classifier to identify the location of a user based on social ties of their friends and their location. All these approaches use location fields of users to identify the location of their friends.

3 Methodology

The location identification system is made up of three main components: Extracting location-based entries from tweets, resolving ambiguity of location names and finally predicting the location of the user.

3.1 Extracting Location References

In order to identify location-based entries from tweets, we use gazetteers to assign a unique identifier to each location-based entity. DBPedia[2], Geonames[3] and US census[4] are some of the large-scale gazetteers with millions of entries for states, cities, streets and landmarks. We use DBPedia as the main resource for identifying location-based entries because it helps us in ambiguity resolution which we will explain in the next step. Each record in DBPedia is assigned a Unique Reference Identifier (URI). The problem of location identification is to match the reference to location from the tweet to a DBPedia URI.

Although DBPedia based entity recognition algorithms like DBPedia spotlight [9] are available, their performance is limited by the size of the tweet [22] and their disambiguation resolution in absence of context is error prone.

We use ark-nlp [13] a POS tagger specially designed for Twitter, to identify parts of speech in tweets with the aim of identifying noun phrases. We eliminate all adjectives, verbs and prepositions from a tweet and extract n-grams from it. The maximum number of n-grams are limited to 4 to reduce the computation and look up time. These n-grams are compared to DBPedia entries with geographic coordinates, if a match is found it is identified as a location reference. If there are multiple locations are identified for subset of an n-gram, for example *District of Columbia* vs *Columbia*, the country, we retain the largest n-gram in this case *District of Columbia*. DBPedia disambiguation pages and page redirects are also used to identify as many location-based entries as possible and retain information from the largest n-gram. When we identify the URI, we extract the geolocation (latitude and longitude) for that URI along with other information like region, state and country information associated with that URI.

3.2 Ambiguity Resolution

There are various difficulties in identifying geographic coordinates for text. Different granularity levels for a location are considered namely street - neighborhood - city - state - country. All geolocations are represented on a city level as much as possible to retain the accuracy of the model, even though we still retain geolocation from finer levels of granularity like street or neighborhood coordinates. We now divide these entries into ambiguous entries and unambiguous entries based on the following criteria.

– If we encounter multiple locations entries in the same tweet we retain all of those entries. However, if they follow the hierarchy, we consider the lowest granularity possible given they are in an order like *Seattle* and *WA* or *London, UK*. If those entries don't follow an order e.g. *Chile* and *Brazil*, we retain all those entries. These entries are categorized as unambiguous entries.

[2] `http://dbpedia.org/About`
[3] `http://www.geonames.org`
[4] `https://www.census.gov/geo/maps-data/data/gazetteer.html`

– In case of multiple entries for same text on different levels of granularity like *Washington D.C.* vs *Washington* state, we look at descriptive words of corresponding granularity levels like "city", "state" etc. and try to identify their granularity level and get their location. They are categorized as unambiguous entries.
– In case of ambiguous location reference like *Lafayette*, we extract all the locations with that name. All these locations are categorized as ambiguous.
– All the location entries under disambiguation section of DBPedia are categorized as ambiguous.

After all location-based entries are categorized, the granularity of the location reference is drilled down to the city level by converting state and country level entries to the city level. This drilling down is done to reduce the possibility of error in predicting the location of the user as much as possible. An extra location entry is added to all cities that have the state or country entry that match the current state or country entry. For example, for a state level entry *Texas*, all the city level location entries with state information as *Texas* for the current user will be incremented by one.

3.3 Location Prediction

Once all the location references of the user are extracted, the home location of the user needs to be identified. Geolocations for all these location references are extracted from DBPedia. We plot all these geolocation coordinates in euclidean space and try to identify the location of the user from the coordinates. The area encompassing these coordinates is considered to be the active area of the user, which he/she referred to in the recent past. The task of location identification is to predict the home location of user from the active area. To achieve this task computational geometric and clustering approaches are used to predict the location of the user.

Convex Hull Approach with Onion Peeling. The convex hull of a finite set of points Q in a plane is the smallest convex polygon P that encloses Q. convex hull is widely used in various fields like pattern recognition, GIS and image processing. Convex hull is generated for all the geolocations extracted from DBPedia for the user. The home location of the user is within the convex hull, and the most obvious solution is the centroid of the polygon. This would be the most optimal solution if area of the polygon is small. However, if the polygon is really large, as it is in the case of location references in tweets where the user talks about events half way across the world, we need to find a solution where these outliers does not affect our location identification. The number of outliers may vary for different users, so we apply onion peeling algorithm [6] to remove those outliers and predict the location of a user.

Let Q be a finite set of points in a plane. The set of convex layers of Q, denoted by $C(Q)$ is the set of convex polygons defined iteratively as follows:

1. Compute the convex hull of Q.
2. remove the geolocations on convex hull from Q if there are still geolocations in Q.
3. repeat until no more convex hulls can be formed.

The three possibilities of the remainder of points are either a polygon or a line or a single geolocation. The location of the user would be the center of polygon or the center of the line or the last remaining geolocation.

K-Center Clustering. We model the location identification as a clustering problem, the main aim is to cluster all near by geolocations together so that the largest cluster with most location references is considered to have the home location of the user, at the same time, we also need to control the size of the largest cluster as small as possible in order to reduce the average error distance. The clustering algorithm should not be sensitive to outliers, while detecting locations that are near each other. Considering all the requirements, we need a clustering algorithm that clusters all the geolocations in a dataset with a large number of outliers into a minimum number of clusters. We use k-center clustering, which is used to solve resource allocation problems.

In k-center clustering problem of a group of points P, we are required to find K points from P and the smallest radius R such that if disks with radius R are placed on those centers then every point in P is covered [15]. We calculate the location of a user from geolocation in the following steps

1. Randomly pick a geolocation x and assign it as the center of a cluster.
2. Assign all geolocations within d distance of x to the cluster.
3. repeat steps 1 and 2 with unclustered geolocation farthest from centers of all the clusters, until there are no unclustered geolocations left.
4. Predict the center of the cluster with the maximum number of geolocations as the home location of the user.

4 Data

Twitter REST API is used to collect the most recent tweets of a user (less if the user has none within the past year). We started with three profiles UL Lafayette[5], LSU[6] and Huffington Post[7]. We extract 200 most recent tweets from all public profiles of the followers of these three profiles in a depth first format. Our final data set had 2578567 tweets generated from 12,893 users. Of this 2,528 users had their location fields empty or had incomplete or nonsensical data that cannot be geotagged e.g. solar system. Table 1 shows the granularity of location information from user profiles on different levels. This location field is reverse geotagged using DBPedia to extract their geolocation (latitude and longitude

[5] `https://twitter.com/ULLafayette`
[6] `https://twitter.com/lsu`
[7] `https://twitter.com/HuffingtonPost`

Table 1. Granularity of location field in user profile

Granularity Level	Number of Users	Percentage of Users
Street Level	698	5.43%
City Level	8998	69.95%
State Level	213	1.66%
Country Level	456	3.55%

Table 2. Granularity of location references in tweets

Granularity Level	Number of Tweets	Percentage of Tweets
Street Level	96538	10.29%
City Level	439554	46.85%
State Level	125938	13.42%
Country Level	276294	29.45%

coordinates). In case of an ambiguity, we manually identified the location of the user based on the profile text of the user. All users whose information cannot reliably be geotagged are ignored. We finally ended up with 8740 users and 1,734,937 tweets. This geolocation for each user is considered the ground truth data and compared against the predicted location of the user. We had 516,335 (29.8%) of tweets, which had 938,314 reference to locations in them. Table 2 shows the number of location references we encountered for each granularity level.

5 Discussion of Results

5.1 Evaluation Methods

For the evaluation, we measure the accuracy of an identified location as follows
Average Error Distance (AED). The average distance between the location of the user's profile to the predicted geolocation of the user.
Accuracy (ACC). The percentage of correctly predicted locations over all the users at a city level. The tolerance value in this case is 0.
Accuracy within N miles (ACC@N). The percentage of predicted locations that are within N miles of the location in user's profile. For example, ACC@30 measures the percentage of predicted locations that are within 30 miles of location from the user profile. We use ACC@30, ACC@60, and ACC@90 to calculate accuracy within 30, 60 and 90 miles respectively. N is the tolerance value.

5.2 Prediction Models

We analyze the results for the following methods
Random Method (RDM). For each user, we randomly select a geolocation from all the location references in his history based on the probability of the locations.

Table 3. Average Distance Error

Method	Average Distance Error(Miles)
RDM	765.6
TM	632.3
COP	124.6
KCC@30	115.4
KCC@60	123.2
KCC@90	138.7

Trivial Method (TM). For each user, we simply select the geolocation that appears most frequently in user's history.
Convex Hull with Onion Peeling (COP). For each user, we select the geolocation predicted by convex hull onion peeling algorithm.
K-Center Clustering (KCC). For each user, we select the center of circle with maximum location references within the circle. Radius of the circle is the tolerance value and is same as N in Accuracy@N.

We use random and trivial methods as a baseline against convex hull and k-center clustering methods. We report the Accuracy and Accuracy@N for all the above models for all the location references in user history along with just the unambiguous entries. We also study the effect of the number of tweets in user history and their effect on accuracy of location prediction. Accuracy for precious approaches were presented in Section 2, we compare our results with their accuracy.

5.3 Results

In this section we present a detailed analysis of location identification of a user from his/her tweet history. The goal of these experiments is to understand: (i) if the accuracy of location identifier improves from usage of convex hull and k-center clustering approaches; (ii) the effect of distance on accuracy of the models; (iii) the number of tweets in the user's history affects the accuracy of the model.

Location Identification. Table 3 shows the average error distance for all the approaches with the average error decreasing considerably for convex hull and k-center approaches. This is due to the way convex hull and k-center clustering smoothes disambiguates in location references, compared to trivial method which predicts the location as the most frequent location used by the user or random method which randomly picks a location from the user's location references. For example, when a user refers to *London*, it can be the *city of London, UK* or *London, California*, a mistake in identifying correct location can offset an error of 5300 miles. It should be noted that as radius of the circle in k-center approach increases, the average error distance also increases. The average distance error for KCC@30 is 115 miles compared to 139 miles for KCC@90. This is due to the fact that as radius of cluster increases, the centre moves further away from original home location of the user.

Table 4. Accuracy of different models for various amounts of tweets

Method	Unambiguous Location References				All Location References			
	ACC	**ACC@30**	**ACC@60**	**ACC@90**	**ACC**	**ACC@30**	**ACC@60**	**ACC@90**
100 Tweets								
RDM	0.492	0.507	0.519	0.543	0.407	0.448	0.473	0.491
TM	**0.521**	0.528	0.537	0.557	**0.623**	0.613	0.628	0.632
COP	0.459	0.535	0.543	0.584	0.614	0.708	**0.726**	**0.728**
KCC	**0.521**	**0.559**	**0.613**	**0.626**	**0.623**	**0.713**	0.721	0.726
150 Tweets								
RDM	0.514	0.521	0.546	0.57	0.427	0.48	0.517	0.533
TM	**0.537**	0.541	0.572	0.577	**0.671**	0.684	0.694	0.712
COP	0.54	0.584	0.595	0.608	0.652	0.78	0.785	0.798
KCC	**0.537**	**0.632**	**0.647**	**0.673**	**0.671**	**0.793**	**0.805**	**0.813**
175 Tweets								
RDM	0.522	0.538	0.564	0.583	0.482	0.517	0.539	0.547
TM	0.563	0.596	0.614	0.628	**0.685**	0.702	0.711	0.729
COP	**0.571**	0.592	0.616	0.631	0.671	0.80	0.807	0.812
KCC	0.563	**0.656**	**0.668**	**0.68**	**0.685**	**0.803**	**0.814**	**0.828**
200 Tweets								
RDM	0.532	0.548	0.572	0.594	0.488	0.529	0.543	0.556
TM	**0.578**	0.617	0.631	0.639	**0.691**	0.715	0.724	0.734
COP	0.575	0.623	0.641	0.649	0.69	**0.817**	0.824	0.827
KCC	**0.578**	**0.661**	**0.673**	**0.685**	**0.691**	0.815	**0.829**	**0.831**

Table 4 shows the accuracy(ACC) and accuracy@N($ACC@N$) for all prediction models on both unambiguous location references and all the location references for all the users in the dataset. It should be noted that there is not a huge increase in accuracy of COP and KCC over baseline for unambiguous location references. This seems to be inline with the results from Mahmud et al. [21] where a pure location based classifier gave an accuracy of 64% compared to 69% for our models. The COP and KCC gives a borderline better results by smoothing the latitude and longitude coordinates over a radius of N for unambiguous location entries. However, when considering all the location references including the ambiguous references, all the methods except the random method show an increase in accuracy. This can be explained that the increased number of locations increases the factor of randomness. The convex hull and k-center algorithms out perform the baseline methods in these cases with a maximum accuracy of 83% within a 90-mile radius, compared to 79% accuray by Schulz et al. [28]. K-center approach performs better than the convex hull, but there is not a huge difference between the methods. Even though, the $ACC@N$ of method increases as the value of N increases , the average distance error also increases. Depending on the application, we have to choose optimal value of N that gives good accuracy and average error distance.

Effect of Number of Tweets. We also looked into the effect of the number of tweets on the accuracy of the methods. The accuracy of location prediction increases with the number of tweets, but does seem to taper off as the number of tweets increases beyond 175. Unlike Cheng et al. [7], we do not require 1000 tweets to predict the location of the user. There is no huge difference in execution time for generation of coordinates for both methods. The main problem is the availability of data and the freshness of the tweets. There is a huge increase jump in accuracy for *ACC@90* from 72.6% to 83.1% for k-center and 72.8% to 82.7% for convex hull method.

6 Conclusion and Future Work

In this paper, we provide a computational geometric approach to identify the location of a user based on his tweet history at a city level. Our method performs better than earlier work in this area, by retaining all ambiguous locations and predicting home location of the user, instead of identifying best possible match to the location from gazetteer. It is also important to note that our method does not require any training or manual intervention to build the model. We are able to predict the location of the user with 71% accuracy within 30 miles of home location of the user using 100 tweets and increases accuracy to 83% within 90 miles of home location using 200 tweets. We also studied the effect of increase in average error distance for k-center clustering as the value of N increases. The average distance error increases from 114 miles for N=30 to 139 miles for N=90.

In future we would like to identify the location references that are being talked about by public at large, compared to location references unique to the user. We are also interested in improving convex hull and k-center approaches by assigning weights to important location references. Instead of predicting a single location for a user, we can also include a temporal factor in identifying vacation or a visiting a new place for a user. We also hope to identify the location of the user on a finer level of granularity i.e. zip code and neighborhood level of granularity.

Acknowledgments

This work is supported in part by the awards NSF/IIP - 1160958 from the Computer and Information Science and Engineering (CISE) Directorate of the National Science Foundation and NSF CCF-1054996.

References

1. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 273–280. ACM (2004)

2. Backstrom, L., Kleinberg, J., Kumar, R., Novak, J.: Spatial variation in search engine queries. In: Proceedings of the 17th international conference on World Wide Web. pp. 357–366. ACM (2008)
3. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World wide web. pp. 61–70. ACM (2010)
4. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science (2011)
5. Buyukkokten, O., Cho, J., Garcia-molina, H., Gravano, L., SHivakumar, N.: Exploiting geographical location information of web pages. In: In Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB99. pp. 91–96 (1999)
6. Chazelle, B.: On the convex layers of a planar set. Information Theory, IEEE Transactions on 31(4), 509–517 (1985)
7. Cheng, Z., Caverlee, J., Lee, K.: 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. pp. 759–768. ACM (2010)
8. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th international conference on World wide web. pp. 761–770. ACM (2009)
9. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
10. Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 1277–1287. Association for Computational Linguistics (2010)
11. Fink, C., Piatko, C.D., Mayfield, J., Finin, T., Martineau, J.: Geolocating blogs from their textual content. In: AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0. pp. 25–26 (2009)
12. Gelernter, J., Mushegian, N.: Geo-parsing messages from microtext. Transactions in GIS 15(6), 753–773 (2011)
13. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 42–47. Association for Computational Linguistics (2011)
14. Graham, M., Hale, S.A., Gaffney, D.: Where in the world are you? geolocation and language identification in twitter. CoRR abs/1308.0683 (2013)
15. Guha, S.: Tight results for clustering and summarizing data streams. In: Proceedings of the 12th International Conference on Database Theory. pp. 268–275. ACM (2009)
16. Hauff, C., Houben, G.J.: Geo-location estimation of flickr images: social web based enrichment. In: Advances in Information Retrieval, pp. 85–96. Springer (2012)
17. Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In: Tan, D.S., Amershi, S., Begole, B., Kellogg, W.A., Tungare, M. (eds.) CHI. pp. 237–246. ACM (2011)
18. Kinsella, S., Murdock, V., O'Hare, N.: 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. pp. 61–68. ACM (2011)

19. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences of the United States of America 102(33), 11623–11628 (2005)
20. Lieberman, M.D., Lin, J.: You are where you edit: Locating wikipedia contributors through edit histories. In: ICWSM (2009)
21. Mahmud, J., Nichols, J., Drews, C.: Where is this tweet from? inferring home locations of twitter users. In: ICWSM (2012)
22. Meij, E., Weerkamp, W., de Rijke, M.: Adding semantics to microblog posts. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 563–572. ACM (2012)
23. Paradesi, S.M.: Geotagging tweets using their content. In: FLAIRS Conference (2011)
24. Popescu, A., Grefenstette, G., et al.: Mining user home location and gender from flickr tags. In: ICWSM (2010)
25. Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., Cohn, T.: Where's@ wally?: a classification approach to geolocating users based on their social ties. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. pp. 11–20. ACM (2013)
26. Sadilek, A., Kautz, H., Bigham, J.P.: Finding your friends and following them to where you are. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 723–732. ACM (2012)
27. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web. pp. 851–860. ACM (2010)
28. Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., Mühlhäuser, M.: A multi-indicator approach for geolocalization of tweets. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
29. Shavitt, Y., Zilberman, N.: A study of geolocation databases. CoRR abs/1005.5674 (2010)
30. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N.: Adding geographic scopes to web resources. Computers, Environment and Urban Systems 30(4), 378–399 (2006)
31. Sultanik, E.A., Fink, C.: Rapid geotagging and disambiguation of social media text via an indexed gazetteer. Proceedings of ISCRAM12 pp. 1–10 (2012)
32. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 178–185 (2010)
33. Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.Y., Li, Y.: Detecting dominant locations from search queries. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 424–431. ACM (2005)
34. Zong, W., Wu, D., Sun, A., Lim, E.P., Goh, D.H.L.: On assigning place names to geography related web pages. In: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries. pp. 354–362. ACM (2005)