Knowledge Enabled Approach to Predict the Location of Twitter Users

Revathy Krishnamurthy, Pavan Kapanipathi(✉), Amit P. Sheth, and Krishnaprasad Thirunarayan

Kno.e.sis Center, Wright State University, Dayton, USA
{revathy,pavan,amit,tkprasad}@knoesis.org

Abstract. Knowledge bases have been used to improve performance in applications ranging from web search and event detection to entity recognition and disambiguation. More recently, knowledge bases have been used to analyze social data. A key challenge in social data analysis has been the identification of the geographic location of online users in a social network such as Twitter. Existing approaches to predict the location of users, based on their tweets, rely solely on social media features or probabilistic language models. These approaches are supervised and require large training dataset of geo-tagged tweets to build their models. As most Twitter users are reluctant to publish their location, the collection of geo-tagged tweets is a time intensive process. To address this issue, we present an alternative, knowledge-based approach to predict a Twitter user's location at the city level. Our approach utilizes Wikipedia as a source of knowledge base by exploiting its hyperlink structure. Our experiments, on a publicly available dataset demonstrate comparable performance to the state of the art techniques.

Keywords: Wikipedia · Twitter · Location prediction · Semantics · Social data · Knowledge graphs

1 Introduction

Location of Twitter users is a prominent attribute for many applications such as emergency management and disaster response [15], trend prediction [1], and event detection [24]. Twitter users can choose to publish their location information by way of (1) geo-tagging their tweets, or (2) specifying it in the location

This material is complemented in part based upon work supported by the National Institute of Health under Grant No. 1R01DA039454-01 and National Science Foundation under Grant No. IIS-1111182. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the employer or funding organization. We would like to thank: (1) Zemanta for their support; (2) Derek Doran, Lu Chen, and Wenbo Wang for their invaluable feedback.

R. Krishnamurthy and P. Kapanipathi—Joint first authors.

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 187–201, 2015.
DOI: 10.1007/978-3-319-18818-8_12

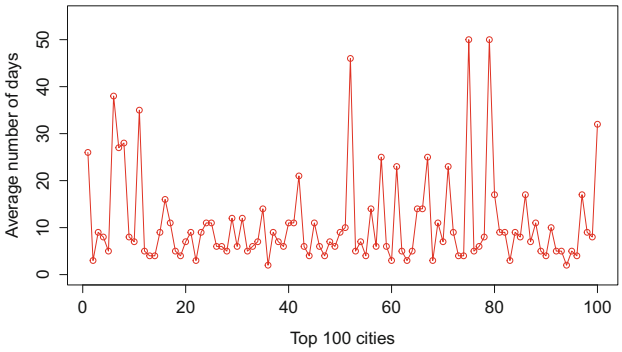

Fig. 1. Estimates of average number of days to collect geo-tagged tweets for top 100 cities in the training dataset of [5]

field of their Twitter profile. However, recent studies have shown that less than 4 % of tweets are geo-tagged [13,19]. Also, while many users choose to leave the location field of their profile empty or enter invalid information, others specify location at different granularity such as city, state, and country. Thus, most of the information entered in this field cannot be reverse geocoded to a city. For instance, Cheng et al. [5] found that, in their dataset comprising of 1 million Twitter users, only 26 % of the users shared their location at the city level.

Existing approaches to predict the location of Twitter users, based on their tweets, use supervised learning techniques [4,5,17]. They are built on the hypothesis that the geographic location of users influences the content of their tweets. These approaches are data-driven and require large training dataset of geo-tagged tweets to build statistical models that predict a user's location. Cheng et al. [5] created a training dataset comprising of 4,124,960 geo-tagged tweets from 130,689 users in continental United States. The collection of this dataset was time intensive and done over a period of 5 months from September 2009 to January 2010. However, in the recent times we have seen a rapid growth of Twitter. Hence, we examined the effort required to create a similar data set in the present day. We selected the top 100 cities with the maximum count of tweets in the dataset of [5] and collected geo-tagged tweets from these cities over a period of 5 days. Based on the tweets collected in this duration, Fig. 1 shows the average number of days required to collect tweets comparable in volume to [5]. We see that for some cities it would take up to 50 days for creating a high quality training data set. This makes it a time intensive process; consequently, making the approach challenging to adapt to newer cities. We address this weakness by proposing a knowledge based solution.

Knowledge bases have been used to either propose alternatives to learning approaches [11,12] or in combination with learning approaches to improve their performance [8,25]. This work falls in the former category. Our approach can be organized into three steps: (1) First, the *Creation of a Location Specific Knowledge base*, which exploits the hyperlink structure of Wikipedia to build a

knowledge base of location specific entities. Additionally, we weight each entity by its ability to discriminate between locations. (2) Second, *User Profile Generation*, which creates a semantic profile of a Twitter user whose location is to be determined. The user profile consists of wikipedia entities found in their tweets and are weighted to reflect their importance to the user. (3) Finally, we use the overlap between the entities in the tweets of a user and the location specific knowledge base to predict the user location in the *Location Prediction* step. Concretely, we make the following contributions:

- We propose a novel knowledge based approach to predict the location of a Twitter user at the city level.
- We introduce the concept of *local entities* which are entities that can discriminate between geographic locations.
- We evaluate our approach using a benchmark dataset published by Cheng et al. [5] and show that our approach, which does not rely on a training dataset, performs comparable to the state of the art approaches.

The rest of the paper is organized as follows. Section 2 describes the creation of a location specific knowledge base. Section 3 describes our approach to predict the location of a user using the location specific knowledge base. Section 4 describes the evaluation and results of our approach. In Sect. 5, we explain the related work on location prediction. Finally, Sect. 6 concludes with suggestions for future work.

2 Creation of Location Specific Knowledge Base

To create a location specific knowledge base, we (1) identify the local entities of a city, and (2) compute their localness measure with respect to the city.

2.1 Local Entities

Previous research that address the problem of location prediction of Twitter users, have established that the content of a user's posts reflects his/her location. Cheng et al. [5] introduced the idea of *local words* which are words that convey a strong sense of location. For example, they found that the word *rockets* is local to Houston whereas words such as *world* and *peace* are more generic and do not exhibit an association to any particular location. Using the same intuition, we introduce the concept of *local entities*. Local entities are wikipedia entities that can distinguish between locations.

We leverage Wikipedia to identify the local entities for each city. While there are many knowledge bases, such as Yago[1], DMOZ[2], and Geo Names[3], we choose Wikipedia because (1) it is comprehensive, (2) it contains dedicated pages for

[1] http://www.mpi-inf.mpg.de/yago-naga/yago.
[2] http://www.dmoz.org.
[3] http://download.geonames.org/export/dump/.

cities, and (3) it has a hyperlink structure that can be exploited for our purposes. A wikipedia page comprises of links to other wikipedia pages. These links are referred to as *internal links*[4] of the wikipedia page and are semantically related to the page (or a portion of it) [21]. Consequently, we consider the entities represented by internal links in the wikipedia page of a city as local entities of that city. For example, the wikipedia page of *San Francisco* contains a link to the wikipedia page of *Golden Gate Bridge*. Thus, we consider *Golden Gate Bridge* as a local entity with respect to *San Francisco*. Note that while a wikipedia page does not contain link to itself, we consider the city as a local entity to itself because location names in tweets provide important cues towards the actual location of the user.

2.2 Localness Measures

All the local entities of a city are not equally local with respect to the city. For example, consider *San Francisco Giants* and *Major League Baseball* that are local entities of the city *San Francisco*[5]. While the *San Francisco Giants* are a baseball team based out of San Francisco, *Major League Baseball* is a professional baseball organization in North America. Intuitively, the entity *San Francisco Giants* has a higher potential than *Major League Baseball* to distinguish *San Francisco* from other cities in United States. Therefore, we introduce the concept of localness measure for each local entity such that the localness score reflects the distinguishing ability of the local entity with respect to a city. We experiment with four measures to determine the localness of an entity. These measures can be classified into three categories: (1) association based measure, (2) graph based measure, and (3) semantic overlap based measures.

Association Based Measure. In information theory, pointwise mutual information is a standard measure of association. It is used to determine association between terms based on the probability of their co-occurrence. The intuitive basis for using an association measure to establish localness of an entity is that, higher the co-occurrence of a local entity with the city in wikipedia pages, higher is the localness of the entity with respect to the city. In order to determine the association between a local entity and a city, we utilize the whole Wikipedia corpus. We define the PMI of a city and its local entity as:

$$PMI(le, c) = \log_2 \frac{P(le, c)}{P(le)P(c)} \tag{1}$$

where c is the city and le is a local entity of the city.

We compute the joint probability of occurrence, $P(le, c)$ as the fraction of the wikipedia pages that contain links to the Wikipedia pages of both the city and the entity. Additionally, the individual probabilities of the city $P(c)$ and the local entity $P(le)$ are computed as the fraction of the wikipedia pages that contain links to the wikipedia page of the city and the local entity alone respectively.

[4] http://en.wikipedia.org/wiki/Help:Link#Wikilinks.
[5] http://en.wikipedia.org/wiki/San_Francisco.

Graph Based Measure. The Wikipedia hyperlink structure can also be represented as a directed graph whose vertices are the wikipedia pages. An edge in this graph represents a link from the wikipedia page of the source node to the wikipedia page of the target node. Since the hyperlink structure of Wikipedia allows us to represent a city and its local entities as a graph, we use a graph theoretic measure to compute the localness of the local entities.

To construct the graph of local entities for a city, we prune the Wikipedia hyperlink graph by selecting only those edges that connect the local entities of the city. For instance, *San Francisco Giants* and *Major League Baseball* are nodes in the graph of local entities of San Francisco. A directed edge from *San Francisco Giants* to *Major League Baseball* represents the link from the wikipedia page of *San Francisco Giants* to the wikipedia page of *Major League Baseball*.

Betweenness centrality has been used extensively to find influential nodes in a network. Our hypothesis is that, the relative importance of a node in the graph of local entities, reflects the localness of the local entity with respect to the city. It is defined as follows:

$$C_B(le, c) = \sum_{le_i \neq le \neq le_j} \frac{\sigma_{le_i le_j}(le)}{\sigma_{le_i le_j}} \tag{2}$$

where c is a city, le, le_i, le_j are local entities of c, $\sigma_{le_i le_j}$ represents the total number of shortest paths from le_i to le_j and $\sigma_{le_i le_j}(le)$ is the number of shortest paths from le_i to le_j through le. We normalize the measure by dividing C_B by $(n-1)(n-2)$ where n is the number of nodes in the directed graph.

Semantic Overlap Measure. Halaschek et al. [10] measure the relatedness between concepts using the idea that related concepts are connected to similar entities. Similarly, we measure the localness of an entity with respect to a city as the overlap between the internal links of the entity and the internal links of the city. To compute this semantic overlap, we use the following set based measures: (1) Jaccard Index, and (2) Tversky Index.

Jaccard Index is a symmetric measure of overlap between two sets and is normalized for their sizes. Jaccard Index for a city c and its local entity le is defined as follows:

$$jaccard(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cup IL(le)|} \tag{3}$$

where $IL(c)$ and $IL(le)$ are the internal links found in the wikipedia page of city c and local entity le respectively.

Tversky Index is an asymmetric measure of overlap of two sets [26]. While the Jaccard Index determines the overlap between a city and a local entity, a local entity generally represents a part of the city. For example, consider the local entity *Boston Red Sox*[6] of the city *Boston*. Its internal links may not symmetrically overlap with that of *Boston* because internal links of *Boston* are from different categories such as *Climate*, *Geography* and *History*. Hence, we

[6] Boston Red Sox is the baseball team of Boston.

adapt Tversky Index to measure unidirectional overlap of the local entity le to the city c as follows:

$$ti(le, c) = \frac{|IL(c) \cap IL(le)|}{|IL(c) \cap IL(le)| + \alpha|IL(c) - IL(le)| + \beta|IL(le) - IL(c)|} \quad (4)$$

where we choose $\alpha = 0$ and $\beta = 1$ to penalize the local entity, for every internal link in its page not found in the wikipedia page of the city.

3 Knowledge Enabled Location Prediction

In Sect. 2, we created a location specific knowledge base comprising of local entities and their localness measures. Now, we describe our algorithm to predict the location of a Twitter user using the location specific knowledge base.

3.1 User Profile Generation

Our approach is based exclusively on the content of a user's tweets. We create a semantic profile of the user whose location is to be predicted. It comprises of wikipedia entities mentioned in their tweets. From this profile, entities that are local entities of a city are used to predict the location of the user. The *User Profile Generation* can be explained in two steps: (1) Entity Recognition from user's tweets; (2) Entity Scoring to measure the extent of the usage of the entity by the Twitter user.

Entity Recognition. Entity recognition is the process of recognizing information like people, organization, location, and numeric expressions[7]. To perform this task on tweets, we utilize existing APIs since the focus of this paper is to predict a Twitter user's location. We opted for Zemanta because of the following reasons: (1) It has been shown to be superior to others as evaluated against other entity recognition and linking services, by Derczynski et al. [6]; (2) Zemanta's web service[8] also links entities from the tweets to their wikipedia pages. This allows an easy mapping between the Zemanta annotations and our knowledge base extracted from Wikipedia; and (3) It provides co-reference resolution for the entities.[9]

Entity Weighting. We weight each entity with the frequency of its occurrence in a user's tweets. Frequency of mentions of an entity indicates the significance of the entity to the user.

3.2 Location Prediction

To predict the location of a user, we compute a score for each city whose local entities are found in the profile of the user, defined as follows:

[7] More details on entity recognition can be found in [20].
[8] http://developer.zemanta.com/docs/suggest/.
[9] We thank Zemanta for their support.

$$locScore(u, c) = \sum_{e \in LE_{cu}} locl(e, c) \times s_e \quad (5)$$

where LE_{cu} is the set of local entities of c found in the profile of user u, $locl(e, c)$ is the localness measure of the entity e with respect to the city c and s_e is the weight of the local entity in the user profile. The location of the user is determined by ranking the cities in the descending order of $locScore(u, c)$.

4 Evaluation

First, we compare our approach with the four localness measures explained in Sect. 2.2. Then, we use the best performing measure to evaluate against the state of the art content based location prediction algorithms.

4.1 Dataset

For a fair comparison of our approach against the existing approaches, we use the dataset published by Cheng et al. [5]. The dataset contains 5119 users, from the continental United States, with approximately 1000 tweets of each user. These users have published their location in their profile in the form of latitude and longitude coordinates. These locations are considered to be the ground truth. Spammers and bots are filtered out from this dataset using Lee et al.'s [16] work.

To create the location specific knowledge base, we consider all the cities of United States with population greater than 5000, as published in the census estimates of 2012. Accordingly, our location specific knowledge base comprises of 4,661 cities with 500,714 local entities.

4.2 Evaluation-Metrics

We adopt the following four evaluation measures used by the existing location prediction approaches [5]:

- *Accuracy* (ACC): The percentage of users identified within 100 miles of their actual location.
- *Average Error Distance* (AED): The average of the error distance across all users. *Error distance* is the distance between the actual location of the user and the estimated location by our algorithm.
- *Accuracy@k* (ACC@k): The percentage of users whose actual locations are within the *top-k* predicted locations of the user, with an error distance of 100 miles.
- *Average Error Distance@k* (AED@k): The Average of error distance, between the closest predicted location at *top-k* to the actual location, across all the users in the dataset.

4.3 Baseline

We implement a baseline system which considers all the entities of a city to be equally local to the city. To predict the location of a user, we compute the score for each city by aggregating the count of local entities of the city found in the user's tweets and selecting the city with the maximum score. In other words, the localness score (*locl*) of each entity in Eq. 5 is 1.

4.4 Results

Table 1 reports the results for location prediction using the (1) Baseline, (2) Pointwise Mutual Information (PMI), (3) Betweenness Centrality (BC), (4) Semantic Overlap Measures - Jaccard Index (JC), and (5) Semantic Overlap Measures - Tversky Index (TI). We see that Tversky Index is the best performing localness measure with approximately 55 % ACC and 429 miles of AED. The ACC is doubled compared to the baseline. However, compared to Jaccard Index, there is only a slight improvement in ACC from 53.21 % to 54.48 % and decrease in AED from 433 to 429 miles.

Table 1. Location prediction using different localness measures

Method	ACC	AvgErrDist (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	38.48	599.408	49.85	56.06	64.15
BC	47.91	478.14	57.39	62.18	66.98
JC	53.21	433.62	67.41	73.56	78.84
TI	**54.48**	**429.00**	**68.72**	**74.68**	**79.99**

The top k cities for a user are determined by ordering the aggregate score for each city (defined in Eq. 5). As shown in Fig. 2, the ACC of our approach increases with k. At $k = 5$, using Tversky Index as the localness measure, we are able to predict the exact location of approximately 80 % of the users. Similarly, as shown in Fig. 3, the AED decreases as k increases. Figure 4 shows the accuracy of prediction within increasing radius (in miles). As seen in the graph, we can predict approximately 46 % of the users within 30 miles of the actual location of the user.

Performance of Localness Measures. Table 1, Figs. 2 and 3 have compared the results of our approach using the localness measures described in Sect. 2.2. In this section, we discuss our findings and analysis on why some localness measures performed better than others in the location prediction task.

Pointwise mutual information measure is sensitive to low frequency data [3]. This led to high absolute PMI scores for the local entities of a city like *Glen Rock, New Jersey* as compared to that of *San Francisco* due to the low occurrence of

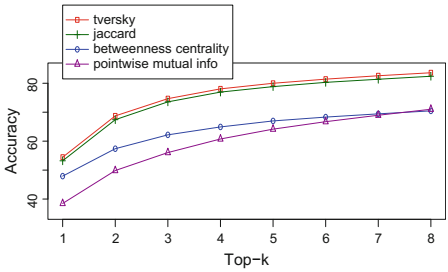

Fig. 2. Top-k accuracy

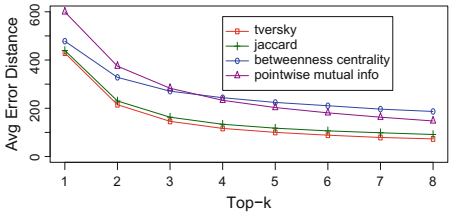

Fig. 3. Top-k average error distance

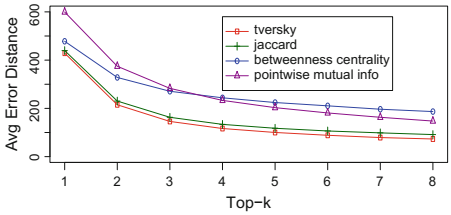

Fig. 4. Accuracy of prediction at increasing miles of radius

the former as compared to the latter in the wikipedia corpus. Nevertheless, the prediction results using PMI show a significant improvement over the baseline.

Betweenness centrality, as a localness measure, performs better than PMI. Although, betweenness centrality addresses the sensitivity to low frequency, it weights certain generic entities higher than more specific entities. For example, in our knowledge base, *United States* is a local entity with respect to the city *San Francisco*. We found that there are multiple shortest paths through *United States* in the graph of local entities of *San Francisco*, thus increasing its importance. However, the entity *United States* is fairly generic and cannot discriminate between the cities in our knowledge base.

The semantic overlap measures overcome the disadvantages of both betweenness centrality and PMI. The primary distinction between the two semantic overlap measures is that Jaccard Index is symmetric while Tversky Index is assymmetric. Jaccard Index is biased against local entities that have less internal links. For example, consider the two entities *Eureka Valley, San Francisco* and *California*. Both are local entities of the city *San Francisco*. Intuitively, we would expect *Eureka Valley, San Francisco* (a residential neighbourhood in San Francisco) to be more local than *California* with respect to the city *San Francisco* but with Jaccard Index the result is opposite. This problem motivated the use of an asymmetric measure. Using the Tversky Index, the localness measure of an entity is the highest when all its internal links are subsumed by those in the wikipedia page of the city. Furthermore, the local entity is penalized for only the internal links in its page not present in the city. Therefore, in the above

Table 2. Examples of local entities found in tweets

City	Entities
New York City, NY	New York City; Brooklyn; Harlem; Queens; New York Knicks; The Bronx; Manhattan; Train station; Metro-North Railroad; Rapping; Times Square; Broadway theatre; New York Yankees; Staten Island; Brooklyn Nets; Hudson River;
Houston, TX	Houston; Houston Texans; Houston Astros; Interstate 45; Houston Chronicle; Greater Houston; Harris County, Texas; Galveston, Texas; Downtown Houston; Houston Rockets;
Nashville, TN	Nashville, Tennessee;Belmont University; Frist Center for the Visual Arts; Southeastern Conference; Centennial Park (Nashville); Gaylord Opryland Resort & Convention Center; Nashville Symphony; Cheekwood Botanical Garden and Museum of Art;

example it is able to assign a higher degree of localness to *Eureka Valley, San Francisco* than *California* with respect to the city *San Francisco*. This approach to weighting the local entities performs better than Jaccard's index with improved accuracy and lower average error distance. Table 2 shows examples of local entities extracted from the tweets of users. These examples illustrate that local entities of various types such as sports teams, landmarks, organizations, local television networks and famous people are used to predict the location of a user.

Comparison with Existing Approaches. For the location prediction task based on user's tweets, the state of the art approaches require a training dataset of geo tagged tweets. Their models are trained using a dataset of 4.1 million tweets collected over 5 months between 2009 and 2010. From Fig. 1, we can see that the collection of geo-tagged tweets for the *top-100* cities (ranked based on the number of geo-tagged tweets from the cities used to train the models [4,5]) in 2015 can take up to 50 days[10]. On the other hand, our approach requires a pre-processing step of creating an index of the wikipedia links (or Dbpedia wikilinks that can be easily downloaded from DBpedia[11]).

We compare the existing approaches against our approach with the best performing localness measure, i.e. Tversky Index (see Table 3). For a fair comparison in the results, we have evaluated our approach on the same test dataset as Cheng et al. [5], Chang et al. [4] and Jalal et al. [17]. As reported in Table 3, our approach performs comparable to the state of the art approaches.

[10] This experiment was performed keeping in mind the extensive growth of Twitter from 2009 to 2015.

[11] http://wiki.dbpedia.org/Downloads2014.

Table 3. Location prediction results compared to existing approaches

Method	ACC
Cheng et al. 2010 [5]	51.00
Chang et al. 2012 [4]	49.9
Jalal et al. 2014 [17]	**55.00**
Our approach with TI	**54.48**

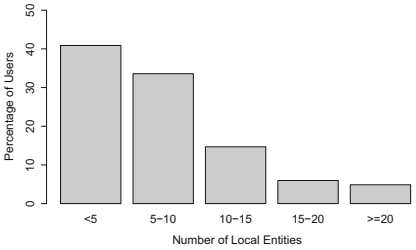

Fig. 5. Percentage of users with the count of distinct local entities from the predicted city

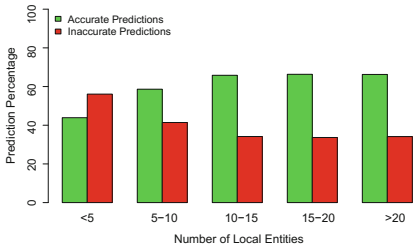

Fig. 6. Predictions corresponding to the count of distinct local entities in users' tweets (Tversky Index)

Impact of Local Entities on Prediction Accuracy. Twitter users have varying frequency of mentioning local entities of their city in tweets. From the test dataset of 5119 users (with approximately 1000 tweets per user), Fig. 5 shows the percentage of users against the frequency of *distinct local entity* mentions. While 40 % of users have mentioned less than 5 distinct local entities, 25 % of the users have more than 10 distinct local entities in their tweets. The impact in determining the location of users based on this varying frequency of local entity mentions is shown in Fig. 6. The accuracy of prediction increases with increase in the number of distinct local entities in the tweets of a user. The accuracy of prediction is 66 % for users who mention more than 10 local entities in their tweets.

Geographic Distribution of Predictions. The count of local entities for the cities in our knowledge base ranges from 11 (for *Island Lake, Illinois*) to 1095 (for *Chicago*). This reflects the information available on wikipedia about the city. The first thought is that, these variations can impact the performance of our approach. In order to analyze this issue, we performed experiments to check if any bias exists towards specific cities as a result of the number of local entities of the city. Hence, we plotted the distribution of our accurate predictions on a map of United States (Fig. 7) and the distribution of test users in the dataset as shown in Fig. 8. We can see that despite the variation in the amount of information available for each city, our algorithm was able to predict locations

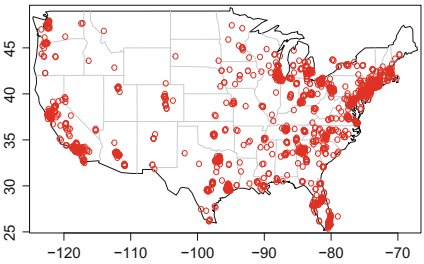

Fig. 7. Distribution of users predicted within 100 miles of their location

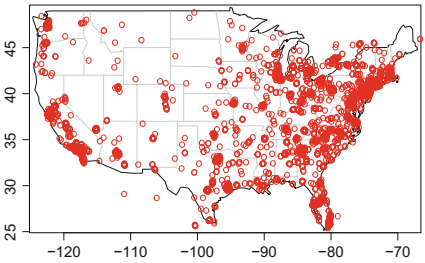

Fig. 8. Distribution of all users in the dataset

of users from all over United States. The knowledge base for these accurately predicted cities ranged between 40 to 1095 local entities.[12]

5 Related Work

Geo-locating twitter users has gained a lot of traction due to its potential applications. Existing approaches to solve this problem can be grouped in to classes: (1) content based location prediction, and (2) network based location prediction.

Content-based location prediction approaches are grounded on the premise that the online content of a user is influenced by their geographical location. It relies on a significantly large training dataset to build a statistical model that identifies words with a local geographic scope. Cheng et al. [5] proposed a probabilistic framework for estimating a Twitter user's city-level location based on the content of approximately 1000 tweets of each user. They formulated the task of identifying local words as a decision problem. They used the model of spatial variation proposed by [2] to train a decision tree classifier using a hand-curated list of 19,178 words. Their approach on a test dataset of 5119 users, could locate 51 % of the users within 100 miles with an average error distance of 535 miles. The disadvantage of this approach was the assumption that a "term" is spatially significant to or characteristic of only one location/city. This challenge was addressed by Chang et al. [4] by modeling the variations as a Gaussian mixture model. While this approach still required a training dataset of geo-tagged tweets, it did not need a labeled set of seed words. Their tests on the same dataset showed an accuracy (within 100 miles) of 49.9 % with 509.3 miles of average error distance. Eisenstein et al. [7] proposed cascading topic models to identify lexical variation across geographic locations. Using the regional distribution of words, determined from these models, they predicted the locations of twitter users. Their dataset comprised of users from United States. Their

[12] Further information on the evaluation, datasets and code can be found at the Wiki page of this project http://wiki.knoesis.org/index.php/Location_Prediction_of_Twitter_Users.

accuracy at the region and state level was 58 % and 27 % respectively. Kinsella et al. [14] addressed two problems, namely, (1) predicting the location of an individual tweet and (2) predicting the location of a user. They created language models for each location at different granularity levels of country, state, city and zipcode, by estimating a distribution of terms associated with the location. Jalal et al. [17] used an ensemble of statistical and heuristic classifiers. These classifiers used words, hashtags, and location names as features. A low level classifier, that predicts location at the city level, needs to discriminate among many locations. To alleviate that, they propose an ensemble of hierarchical classifiers that predict the location at time zone, state and city level. However, their approach is also supervised and relies on a training dataset.

Network based solutions are grounded in the assumption that the locations of the people in a user's network and their online interaction with the user can be used to predict his/her location. McGee et al. [18] used the interaction between users in a network to train a Decision Tree to distinguish between pairs of users likely to live close by. They reported an accuracy of 64 % (within 25 miles). Rout et al. [23] formulated this task as a classification task and trained an SVM classifier with features based on the information of users' followers-followees who have their location information available. They tested their approach on a random sample of 1000 users and reported 50.08 % accuracy at the city level. However, a network based approach can only be used to determine the location of users who have other users in their network whose location is already known.

In the Twitter domain, Wikipedia has been leveraged for many tasks. Osborne et al. [22] have shown that Wikipedia can enhance the performance of first story detection on Twitter. The graph structure of Wikipedia has been utilized by Genc et al. [9] to classify tweets. Also, the Wikipedia graph has been leveraged by Kapanipathi et al. [12], with an adaptation of spreading activation theory to determine the hierarchical interests of users based on their tweets.

6 Conclusion and Future Work

In this paper, we presented a novel knowledge based approach that uses Wikipedia to predict the location of Twitter users. We introduced the concept of local entities for each city and demonstrated the results of different measures to compute the localness of the entities with respect to a city. Without any training dataset, our approach performs comparable to the state of the art content based approaches. Furthermore, our approach can expand the knowledge base to include other cities which is remarkably less laborious than creating and modeling a training dataset.

In future, we will explore the use of semantic types of the Wikipedia entities to improve the accuracy of the location prediction and decrease the average error distance. We also plan to augment our knowledge base with location information from other knowledge bases such as Geo Names and Wikitravel. Additionally, we will examine how to adapt our approach to predict the location of a user at a finer granularity level like the neighborhoods in a city.

References

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707. IEEE (2011)
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. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the Biennial GSCL Conference (2009)
4. Chang, H.-W., Lee, D., Eltaher, M., Lee, J.: @ Phillies tweeting from philly? predicting twitter user locations with spatial word usage. In: ASONAM 2012 (2012)
5. 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)
6. Derczynski, L., Maynard, D., Aswani, N., Bontcheva, K.: Microblog-genre noise and impact on semantic annotation accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media, pp 21–30. ACM (2013)
7. 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)
8. Gabrilovich, E., Markovitch, S.: Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge. In: AAAI, vol. 6, pp. 1301–1306 (2006)
9. Genc, Y., Sakamoto, Y., Nickerson, J.V.: Discovering context: classifying tweets through a semantic transform based on wikipedia. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) FAC 2011. LNCS, vol. 6780, pp. 484–492. Springer, Heidelberg(2011)
10. Halaschek, C., Aleman-Meza, B., Arpinar, I.B., Sheth, A.P.: Discovering and ranking semantic associations over a large rdf metabase. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, pp. 1317–1320. VLDB Endowment (2004)
11. Hu, X., Zhang, X., Lu, C., Park, E.K, Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 389–396. ACM (2009)
12. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on twitter using a hierarchical knowledge base. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 99–113. Springer, Heidelberg (2014)
13. Khanwalkar, S., Seldin, M., Srivastava, A., Kumar, A., Colbath, S.: Content-based geo-location detection for placing tweets pertaining to trending news on map. In: The Fourth International Workshop on Mining Ubiquitous and Social Environments, p. 37 (2013)
14. 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)
15. Kireyev, K., Palen, L., Anderson, K.: Applications of topics models to analysis of disaster-related twitter data. In: NIPS Workshop on Applications for Topic Models: Text and Beyond, vol. 1 (2009)

16. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010)
17. Mahmud, J., Nichols, J., Drews, C.: Home location identification of twitter users. ACM Trans. Intell. Syst. Technol. (TIST) **5**, 47 (2014)
18. McGee, J., Caverlee, J., Cheng, Z.: Location prediction in social media based on tie strength. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 459–468. ACM (2013)
19. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitters streaming api with twitters firehose. In: Proceedings of ICWSM (2013)
20. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)
21. Nuzzolese, A.G., Gangemi, A., Presutti, V., Ciancarini, P.: Encyclopedic knowledge patterns from wikipedia links. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 520–536. Springer, Heidelberg (2011)
22. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: first story detection using twitter and wikipedia. In: Proceedings of the Workshop on Time-aware Information Access, TAIA, vol. 12 (2012)
23. Rout, D., Bontcheva, K., Preoţiuc-Pietro, D., Cohn, T.: Where's @wally?: a classification approach to geolocating users based on their social ties. In: HT (2013)
24. 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)
25. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Volume Three, pp. 2330–2336. AAAI Press (2011)
26. Tversky, A.: Features of similarity. Psychol. Rev. **84**(4), 327 (1977)