Location Prediction of Twitter Users using Wikipedia

Revathy Krishnamurthy, Pavan Kapanipathi, Amit Sheth, Krishnaprasad Thirunarayan
Kno.e.sis Center, Wright State University, Dayton, Ohio
{revathy, pavan, amit, t.k.prasad}@knoesis.org

ABSTRACT

The mining of user generated content in social media has proven very effective in domains ranging from personalization and recommendation systems to crisis management. The knowledge of online users' locations makes their tweets more informative and adds another dimension to their analysis. Existing approaches to predict the location of Twitter users are purely data-driven and require large training data sets of geo-tagged tweets. The collection and modelling process of tweets can be time intensive. To overcome this drawback, we propose a novel knowledge based approach that does not require any training data. Our approach uses information in Wikipedia, about cities in the geographical area of our interest, to score *entities* most relevant to a city. By semantically matching the scored entities of a city and the entities mentioned by the user in his/her tweets, we predict the most likely location of the user. Using a publicly available benchmark dataset, we achieve 3% increase in accuracy and 80 miles drop in the average error distance with respect to the state-of-the-art approaches.

Categories and Subject Descriptors

H.4 [**Information System Applications**]: Collaborative and social computing systems and tools

General Terms

Text Mining, Social Data Analysis

Keywords

Wikipedia, Twitter, Location Prediction, Semantics

1. INTRODUCTION

People are increasingly using microblogging platforms such as Twitter for a wide range of reasons, from sharing personal experiences to reaching out for help in emergency situations. Consequently, many applications such as brand tracking and recommendation engines have adopted Twitter as a prominent medium to gain insights. Furthermore, location based services such as emergency management and disaster response [18], local event detection system [29], and predicting trends [1] need the geographic location of Twitter users. On the other hand, Twitter users are often reluctant to share their location information and hence either, (1) choose to leave the location information in their profile empty; (2) enter invalid information; (3) or specify location at various levels of granularity like *city, state* and *country.* This has motivated research in automatic inferencing of geographic location of Twitter users.

Existing approaches to predict the location of Twitter users based on their tweets, [6, 7] are built around the intuition that the geographic location of users influences the content of their tweets. For instance, users are most likely to tweet about shops, restaurants, sports teams of their location and use location indicative slang words like *howdy* (Texas). These approaches are purely data-driven and need large training data sets of geo-tagged tweets to build statistical models that predict a user's location. The creation of a training dataset with representative tweets from all the cities of our interest and modelling the data is a tedious process.

To overcome these disadvantages, we propose a knowledge base enabled approach. Our intuition stems from the idea of *local words* proposed by Cheng et al. [7]. *Local Words* 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. We extend this idea to define *local entities* as entities that are able to discriminate between geographic locations. Our hypothesis is that the local entities appearing in a collection of a user's tweets can be used to predict his/her location. We leverage Wikipedia to determine local entities and in turn alleviate the difficulties in creating a training data set for location prediction. Note that our approach relies exclusively on the tweets of users and does not require other metadata such as user's profile or network information. We evaluate our approach against the state of the art content-based approaches using a benchmark dataset published by Cheng et al. [7]

Our contributions in this paper are as follows:

- We propose a novel knowledge base enabled approach as an alternative to existing supervised approaches to predict the location of Twitter users.
- We demonstrate the use of similarity measures to determine and score *Local Entities* with respect to a city.

- We show that without using any labelled training data, our approach achieves a comparable accuracy to the content-based state of the art approach and reduces the average error distance by 80 miles.

The rest of the paper is organized as follows: In Section 1.1, we provide relevant background on Wikipedia. In Section 2, we explain the related work on location prediction. Section 3 details our approach, while Section 4 describes the evaluation and results of our approach. Section 5 concludes with suggestions for future work.

1.1 Wikipedia

As humans, we use our general knowledge and experience in the interpretation of any text or discourse. For instance, when we read *"Buckeye State"* in a piece of text, we recognize it as the state of Ohio. Similarly, a knowledge base can improve a machine's ability to understand and interpret text. Knowledge bases have been successfully used in many domains such as clustering and classifications [5, 28], semantic relatedness [12, 15].

Wikipedia, a publicly available, online collaborative encyclopedia has been a prominent source of knowledge for humans as well as machines. It comprises of approximately 4.6 million articles that are comprehensive, well-formed with each article describing a single topic or entity [15]. Each page in Wikipedia contains links to other Wikipedia pages referred to as *wikilinks* or *internal links*[1]. The aim of these links is to enhance the user's understanding about the entity by providing pointers to related entities on the Wikipedia. For example, the Wikpedia page of *Boston*[2], has hyperlinks to *Boston Red Sox*, *American League* and *Major League Baseball*. Apart from allowing the user to navigate to the pages of related entities for better understanding, this feature also creates a hyperlink structure, that allows machines to use Wikipedia as a knowledge base of semantically linked entities. The Wikipedia hyperlink structure has been leveraged to accomplish tasks such as finding conceptual semantic relatedness [12], named entity disambiguation [14]. Our approach exploits this hyperlink structure of Wikipedia to determine entities related to a city and uses them to predict the locations of Twitter users.

2. RELATED WORK

Geo-locating content on the web has been studied in various contexts. IP addresses were discovered to be inadequate to determine the location of online users. Hence researchers focused on location entity recognition and disambiguation in textual content. Amitay et al. [2] proposed a geotagger based on a gazetteer to determine the geographical focus of web pages. Tietler et al. [26] extracted geographical information from newspaper articles and grouped them based on their location to display them in a map interface. Backstorm et al. [4] proposed a probabilistic model to discover the spatial distribution of search engine queries using query log data. More recently, geo-locating twitter users has gained a lot of traction. There have been two main approaches in predicting the location of a twitter user: (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. [7] 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 [4] 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. [6] by modelling the variations as a Gaussian mixture model. Furthermore, their approach to identify local words did not need a labelled 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. Our approach falls into this category of content-based location prediction. However, we use a knowledge base (Wikipedia) as an alternative to training data used by the other machine learning and statistical modelling approaches.

Eisenstein et al. [10] 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. Kinsella et al. [17] 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.

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. [19] 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. [24] 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 tasks such as first story detection [22], tweets classification [13], and identifying hierarchical interests of Twitter users [16]. Osborne et al. [22] in their work, they have shown that Wikipedia can enhance the performance of first story detection on Twitter. The graph structure of Wikipedia has

[1] http://en.wikipedia.org/wiki/Help:Link
[2] http://en.wikipedia.org/wiki/Boston

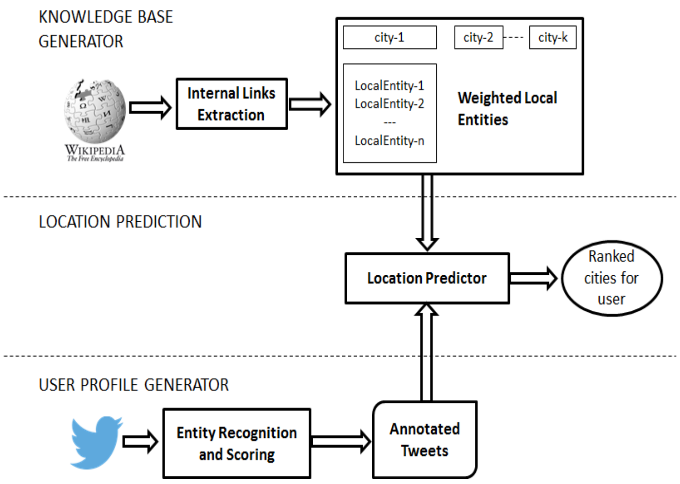

Figure 1: Framework of Location Prediction using Wikipedia

been utilized by Genc et al. [13] to classify tweets. Their approach first maps each tweet to the most relevant Wikipedia concept and further leverages the category structure to find the semantic distance between the mapped concepts for classification. The Wikipedia graph has also been utilized by Kapanipathi et al. [16], with an adaptation of spreading activation theory to determine the hierarchical interests of users based on their tweets.

3. KNOWLEDGE BASE ENABLED LOCATION PREDICTION

Previous research that address the problem of location prediction, have established that the content of a user's posts reflects his/her location. With the same intuition we propose an approach that uses Wikipedia to identify *local entities* and aggregates the occurrences of local entities in users' tweets to predict the location of users. As shown in Figure 1, our approach comprises of three primary components: (1) *Knowledge Base Generator* extracts local entities for each city from Wikipedia and scores them based on their relevance to the city (Section 3.1); (2) *User Profile Generator* extracts Wikipedia entities from the tweets of a user (Section 3.2); (3) *Location Predictor* uses the output of User Profile Generator and Knowledge Base Generator to predict the location of the user (Section 3.3).

3.1 KNOWLEDGE BASE GENERATOR

Wikipedia contains dedicated pages for geographical locations at various levels of granularity like village, town, city, downtown city, county, state and country. There are many knowledge bases which are publicly available and contain information about locations. For example, Yago [25], DMOZ[3], and Geo Names[4]. However, Wikipedia is comprehensive, dynamically updated by its users and has a hyperlink structure that can be exploited for our purposes. Due to these advantages over the other knowledge bases, we opted for Wikipedia. The Wikipedia page of each city links to other Wikipedia entities related to subtopics such as city's politics, culture, landmarks and sports teams. We base our approach on the assumption that these entities, are *local* to the city and vary in the degree of their *localness* to the city. For example, the Wikipedia page of *San Francisco* contains links to *San Francisco Bay Area* and *United States*. We consider them as local entities with respect to *San Francisco* where *San Francisco Bay Area* has more localness than *United States* with respect to *San Francisco*. In the rest of the paper, we refer to the internal links in the Wikipedia page of a city as the entities of that city.

3.1.1 Definitions

- **Wikipedia Hyperlink Structure** can be represented as a directed graph $G = (V_w, E_w)$ with a set of vertices $V_w \subseteq W$, where W is the set of all the Wikipedia pages and a set of edges E_w, where $E_w \subseteq V_w \times V_w$. There is a directed edge (v_1, v_2), if there is a link from Wikipedia page v_1 to v_2. For a given vertex v_i, $O(v_i)$ is the set of entities mentioned in the Wikipedia page v_i, i.e, $O(v_i)$ are the vertices that have an edge from v_i.

- **Local Entities.** The entities mentioned in a Wikipedia page of a city c are termed as *Local Entities* of the city c. From the hyperlink structure, these are the outgoing links $O(c)$ from each Wikipedia page of city c.

- **Knowledge base** for each city is represented by a weighted set of its *Local Entities*. Formally, we define a knowledge base K_c for city c as:

  $$K_c = \{(e, locl(c,e)) | e \in O(c), locl(c,e) \in R\} \qquad (1)$$

  where $e \in O(c)$, $O(c)$ is the set of *Local Entitities* of the city c, and $locl(c,e)$ is the localness score of entity e with respect to city c.

Given the set of *Local Entities* $O(c)$ of a city c, we want to determine the *localness* ($locl(c,e)$) of each entity with respect to the city. In the following sections, we describe four measures to compute the localness of an entity which is later used to predict the location of a user.

3.1.2 Pointwise Mutual Information

In information theory, pointwise mutual information [8] is a standard measure of association. It is used to determine association of terms based on their probability of co-occurrence. Similarly, we determine the association between a city and its local entities using their co-occurrences in the entire dump of Wikipedia. We define the PMI of a city and its local entity as shown in the Equation 2.

$$PMI(c,e) = \log_2 \frac{P(c,e)}{P(c)P(e)} \qquad (2)$$

where c is the city and e is a local entity of the city, $e \in c_O$.

We compute the joint probability of occurrence, $P(c,e)$ as the count of occurrences of the city and the local entity together in the other pages of the entire Wikipedia dump. Additionally, the individual probabilities of the city $P(c)$ and the local entity $P(e)$ are computed as the ratio of the count of their individual occurrences in the Wikipedia dump to the count of all entities in the Wikipedia dump.

[3] http://www.dmoz.org/

[4] http://download.geonames.org/export/dump/

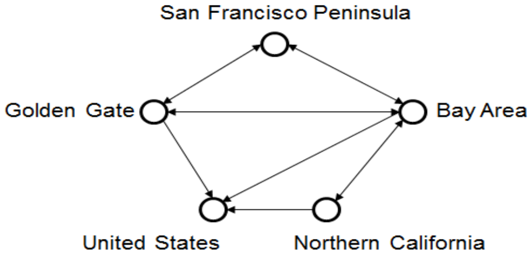

Figure 2: A pruned subgraph of San Francisco

3.1.3 Betweenness Centrality

Graph theoretic concepts have been used in social network analysis to understand and explain social phenomenon. Betweenness Centrality [11] is one such measure that has been extensively used to find influential people in a network. In this work, we use Betweenness Centrality to find the prominent entities in the *graph of Local Entities* for each city. The graph of Local Entities for each city is pruned from the Wikipedia hyperlink graph (Section 3.1.1) and consists of only those entities present in the corresponding city's Wikipedia page ($O(c)$). Formally, the graph for a city c is represented as $G_c = (V_c, E_c)$ where vertices $V_c \in (c \cup O(c))$ and edges $E_c \in V_c \times V_c$. There is an edge from v_{c_i} to v_{c_j} if the Wikipedia page of v_{c_i} has a link to entity v_{c_j}. An example of a subgraph is as shown in Figure 2. The nodes in this graph are a subset of entities mentioned in the Wikipedia page of *San Francisco*. We draw edges between entities based on the entity occurrences in their respective pages. For instance, an edge between *Golden Gate Bridge* and *San Francisco Bay Area* is indicative of the presence of the latter in the former's Wikipedia page.

Betweenness Centrality is defined as follows:

$$C_B(c, e) = \sum_{e_i \neq e \neq e_j} \frac{\sigma_{e_i e_j}(e)}{\sigma_{e_i e_j}} \tag{3}$$

where $e_i, e, e_j \in O(c)$, $\sigma_{e_i e_j}$ represents the total number of shortest paths from e_i to e_j and $\sigma_{e_i e_j}(e)$ is the number of shortest paths from e_i to e_j through e. Furthermore, we normalize the measure by dividing C_B by $(n-1)(n-2)$ where n is the number of nodes in the directed graph.

3.1.4 Semantic Overlap Measures

SemRank [3], a search results ranking system, measures the relatedness between concepts with the intuition that related concepts are connected to similar entities. Similarly, we use the Wikipedia hyperlink graph to determine the extent of relatedness between a city and an entity. We term this as *Semantic Overlap*. We use the following two standard set based measures to compute the semantic overlap between a city and an entity (1) *Jaccard Index*, and (2) *Tversky Index*.

Jaccard Index measures the overlap between two sets and is normalized for their sizes. We use this measure to find the similarity between a city and its entities. For example, to compute the localness of *Golden Gate Bridge* to *San Francisco*, we compute the Jaccard Index of the two sets containing the entities from the Wikipedia page of *Golden Gate Bridge* and *San Francisco* respectively. Jaccard Index for a city c and entity e ($e \in O(c)$) is defined as shown in Equation 4.

$$jaccard(c, e) = \frac{|O(c) \cap O(e)|}{|O(c) \cup O(e)|} \tag{4}$$

The idea behind using Jaccard Index is that larger the overlap between the entities of a local entity and a city, higher is the localness of the local entity with respect to the city.

Tversky Index is an asymmetric similarity measure of two sets [27]. While the Jaccard Index determines the similarity 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* of the city *Boston*. *Boston Red Sox* is the baseball team of Boston and will not completely overlap with all the entities of *Boston* which are from different categories like *Climate*, *Geography* and *History*. Thus we use Tversky Index which is a unidirectional measure of similarity of the local entity to the city. The Tversky Index is defined as shown in Equation 5.

$$ti(c, e) = \frac{|O(c) \cap O(e)|}{|O(c) \cap O(e)| + \alpha|O(c) - O(e)| + \beta|O(e) - O(c)|} \tag{5}$$

where we choose $\alpha = 0$ and $\beta = 1$, with no weight given to the entities of the city. Thus we only penalize the local entity e, for every entity in its page not found in the Wikipedia page of the city c.

3.2 User Profile Generator

Our approach is based exclusively on the contents of a user's tweets. We create a profile for each user that consists of Wikipedia *entities* spotted in their tweets. The *User Profile Generator* 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 is the process of recognizing information like people, organization, location, and numeric expressions[5] from twitter messages. As explained in [23], the concise nature of tweets (<= 140 characters) and the informal nature of their content have challenged the traditional entity recognition techniques. In this paper, our main focus is on the location prediction of Twitter users. Hence, we use the APIs available for Entity Recognition. In [9] authors have compared three different state of the art systems namely Dbpedia Spotlight [20], Zemanta[6] and TextRazor[7] for entity recognition in tweets. These results are summarized in Table 1. We opted for Zemanta because: (1) It has been shown to be superior to others; (2) Zemanta's web service[8] also links entities from the tweets to their Wikipedia articles. This allows an easy mapping between the Zemanta annotations and our knowledge base extracted from Wikipedia; (3) The web service provides co-reference resolution for the entities. I.e., if Barack Obama and Obama are mentioned

[5] More details on entity recognition can be found in [21]
[6] `http://developer.zemanta.com/`
[7] `http://www.textrazor.com/technology`
[8] `http://developer.zemanta.com/docs/suggest/`

Extractors	**Precision**	**Recall**	**F-Measure**	**Rate Limit**
Spotlight	20.1	47.5	28.3	N/A
TextRazor	**64.6**	26.9	38.0	500/day
Zemanta	57.7	31.8	**41.0**	**10,000/day**

Table 1: Evaluation of Web Services for Entity Resolution and Linking

in the twitter message, they are both linked to Wikipedia page for Barack Obama[9]; (4) Zemanta provides a higher rate limit of their API to 10,000 per day for research purposes.[10]

Entity Scoring. We consider the frequency of occurrence of a local entity in a user's tweets to predict his/her location. Formally, we define the profile of a user u as $P_u = \{(e, s) | m \in W, w \in R\}$ where W denotes the set of all Wikipedia entities and s is the frequency of mentions of entity e by user u.

3.3 Location Prediction

We compute an aggregate score based on all the local entities found in the user profile. In other words, to estimate the location for a user u with profile P_u, for each location c with knowledge base K_c, we find the intersection of the set of entities I_{cu} associated with the user profile and the Local Entities of the city. Next, we use the following equation to estimate the score for each city for a user.

$$locScore(c, u) = \sum_{j=1}^{|I_{cu}|} locl(c, e_j) \times s_{e_j} \tag{6}$$

where $e_j \in I_{cu}$, $locl(c, e_j)$ is the localness score of the entity e_j with respect to the city c, determined by one of the localness measure explained above. s_{e_j} is the score of the entity in the user profile P_u. The city for the user is determined by ranking the cities based on the $locScore(c, u)$ in descending order.

4. EVALUATION

First, we compare the four localness measures explained in Section 3.1 and then use the best performing measure to evaluate against the state of the art content based approach for location prediction.

4.1 Dataset

For a fair comparison of our approach against the state of art approaches, we use the dataset published by Cheng et al [7]. The dataset was collected from September 2009 to January 2010 by crawling through Twitter's public timeline APIs[11]. The dataset contains 5119 active users, from the continental United States, with approximately 1000 tweets of each user. The users' locations are listed in the form of latitude and longitude coordinates which is generally more reliable than the profile information. Spammers and bots are filtered to ensure a clean dataset. Additionally, we remove the word "RT" (referring to a re-tweet) from the tweets. We do this because Zemanta annotated "RT" in re-tweets incorrectly as *RT (TV Network)*[12] which affected the results as it is one of the local entities in our knowledge base.

To create our knowledge base, we consider all the cities of United States with population greater than 5000, as published in the census estimates of 2012. From the census estimates, we only include the locations listed as *city* and ignore the locations labelled as *village*, *town*, *county* or *CDP(Census Designated Place)*. The entire collection of Wikipedia articles is available as an XML dump[13]. The Wikipedia pages of two cities *Irondale, Alabama* and *Mills River, North Carolina* are marked as stubs[14] and hence are not included in our knowledge base. Although a Wikipedia page does not link to itself, we include the name of each city in its knowledge base.Finally, we have a knowledge base with 4,661 cities and 500,714 entities. To compute the distance between the actual and the predicted location we extract the latitude and longitude information of each city in our knowledge base from the infobox[15] of their corresponding Wikipedia page.

4.2 Evaluation-Metrics

We use Accuracy and Average Error Distance as the two metrics to evaluate our approach. *Accuracy* (ACC) is defined as the percentage of users identified within 100 miles of their actual location. Error distance is the distance between the actual location of the user and the estimated location by our algorithm. *Average Error Distance* (AED) is the average of the error distance across all users.

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.

4.4 Results

4.4.1 Location Prediction using Local Entities

Table 2 reports the Accuracy and the Average Error Distance 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% accuracy and 429 miles of AED. The accuracy is doubled compared to the baseline approach. However, compared to Jaccard Index, there is only

[9] `http://en.wikipedia.org/wiki/Obama`

[10] We thank Zemanta for their support

[11] `http://search.twitter.com/`

[12] `http://en.wikipedia.org/wiki/RT_(TV_network)`

[13] `http://en.wikipedia.org/wiki/Wikipedia:Database_download`

[14] `http://en.wikipedia.org/wiki/Wikipedia:Stub`

[15] `http://en.wikipedia.org/wiki/Help:Infobox`

Method	ACC	AvgErrDist (in Miles)	ACC@2	ACC@3	ACC@5
Baseline	25.21	632.56	38.01	42.78	47.95
PMI	32.46	792.32	44.48	52.43	61.55
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

Table 2: Location Prediction using Local Entities

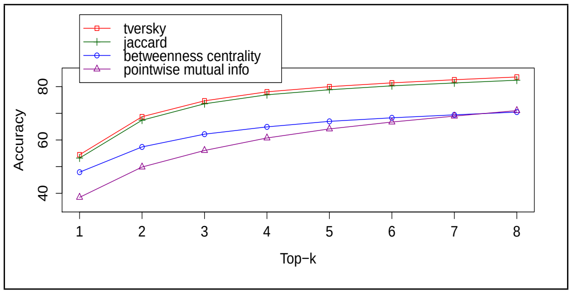

Figure 3: Top-k Accuracy

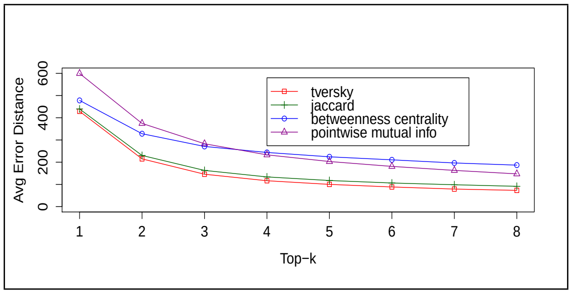

Figure 4: Top-k Average Error Distance

a slight improvement in accuracy from 53.21% to 54.48% and decrease in AED from 433 to 429 miles.

By ranking the cities for each user, based on the descending order of localness scores, we have also evaluated the accuracy of the approach at *top-k* ranks. Similar to accuracy, *accuracy@top-k* is calculated by the number of users whose home locations are determined correctly within the *top-k* locations in the generated ranked list of locations for the user, within an error distance of 100 miles. The AED @top-k is computed using distance between the closest predicted location@*top-k* to the actual location of the user. Figure 3 shows the change in accuracy across *top-8* locations determined using Tversky Index.

In order to calculate the error distance for a particular user for *top-k*, we picked the closest possible location predicted by our approach to the original location of the user within the *top-k* results. The error distance to this closest location is calculated and averaged across all the users to result in AED@top-k. Figure 4 shows that the AED decreases with inclusion of more top locations and similar to *accuracy@top-k*, Tversky Index performs the best.

4.4.2 Comparison with Existing Approaches

Method	ACC	AvgErrDist (in Miles)
Cheng et al.[7]	51.00	535.564
Chang et al.[6]	49.9	509.3
TI	**54.48**	**429.00**

Table 3: Location prediction results compared to existing approaches

For the location prediction task based on user's tweets, the state of the art approaches are purely data-driven. We have evaluated our approach on the same dataset as Cheng et al. [7] and Chang et al. [6]. As reported in Table 3, our approach performs better in terms of both the accuracy and the average error distance. Also, note that the other approaches are based on a training dataset of 4.1 million tweets while our approach is based exclusively on Wikipedia.

4.4.3 Impact of annotated entities

Figure 5 shows the count of all entities in the dataset annotated by Zemanta and Figure 6 shows the count of distinct local entities found in the tweets of users to predict their location. Note that these figures represent the predictions made using Tversky Index. From Figure 6, we see that when the number of local entities mentioned in the tweets are less than 5, the prediction drops by more than 12% (48% accuracy) compared to the overall accuracy of predictions. On the other hand, a prediction made on the basis of higher number of local entities is more reliable. The predictions made on the basis of 10 or more local entities were able to locate 66% of the users within 100 miles and 51% of the users within 20 miles.

4.5 Discussion

4.5.1 Performance of Localness Measures

We predict the location of a user based on the count of occurrences of local entities in their tweets and the localness measure of the entities with respect to a city. The pointwise mutual information measure of a city and its local entity is not normalized, making it sensitive to the count of their occurrences in the Wikipedia corpus. Consequently the absolute PMI scores of the local entities of a city like *Glen Rock, New Jersey* is higher than those of *San Francisco* because of the low occurrence of former as compared to the latter, in the Wikipedia corpus. This results in the location prediction to be skewed towards the cities that occur less frequently in the Wikipedia articles. Nevertheless, the prediction results using PMI show a significant improvement over the baseline. The localness of entities computed using betweenness centrality and the semantic overlap measures are normalized and yield better results than PMI.

The betweenness centrality of a node is based on the num-

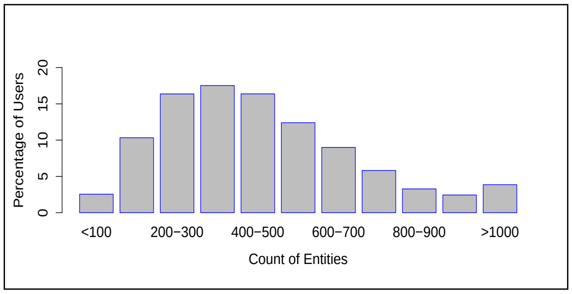

Figure 5: Percentage of users with the count of Wikipedia Entities extracted from their tweets

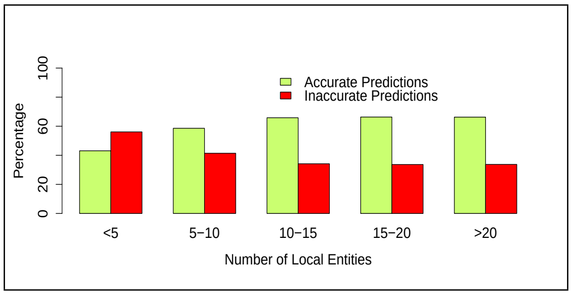

Figure 6: Predictions based on the number of Local Entities in users' tweets

ber of times a node occurs in the shortest path between two other nodes. We find that some entities which may not be local, get ranked higher because there are multiple shortest paths through them. Consider the snippet from the Wikipedia page of *Livingston, New York*, shown in Table 4. The underlined entities contain the shortest path to the rest of the entities of the city through *United States* thus increasing the importance of *United States* in the graph. Consider another example of the city *Endicott, New York*[16]. A section of the Wikepedia page of this city describes *IBM* and related entities like *Punched card* and *Circuit Board*. When we build a graph of the city, the shortest path between the *IBM* related entities and the rest of the entities of the city, is through *IBM*. This increases its betweenness centrality measure compared to the rest of the other local entities. As a result, when entities like *United States* or *IBM* occur frequently in a user's tweets, they lead to incorrect location prediction.

The idea behind using Jaccard Index is that larger the semantic overlap between the Wikipedia page of a city and an entity, higher is the localness of that entity with respect to the city. Thus it overcomes the disadvantage of Betweenness Centrality and is successful in assigning less localness to the more general entities like *IBM* and *United States*. However, we observe that it under-performs in measuring the localness of entities with fewer number of entities in comparison to the city. 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. Note that *San Francisco* has 717 entities, *Eureka Valley, San Francisco* has 36 entities and *California* has 940 entities. This is countered using the Tversky Index where the localness measure of an entity is highest when all of its entities are also present in the city. Furthermore, the localness of a local entity only diminishes for entities in its page not present in the city. Therefore, in the above 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 ranking the entities performs better than Jaccard's index with improved accuracy and lower average error distance. Table 5 shows examples of local entities from the tweets of users in the dataset used to predict their location.

Heritage
The residents of Livingston are descended from people of many nations, including:

- People from Oklahoma and other parts of the United States of America.
- Hindus, Sikhs and Muslims from India and Pakistan. Livingston has one of the largest communities of Sikhs in the United States.
- Mennonites from Germany and Russia.
- Armenians from Middle East

Table 4: A snippet from the Wikipedia page of Livingston, New York

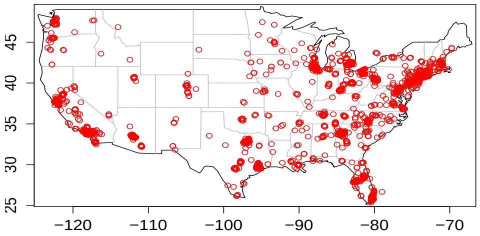

Figure 7: Distribution of users predicted within 100 miles of their location

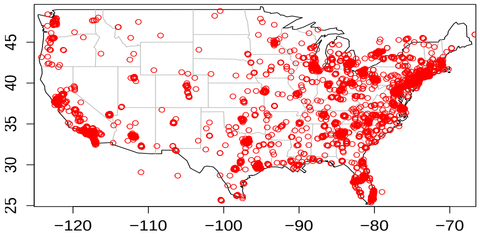

Figure 8: Distribution of all users in the dataset

4.5.2 *Size of Local Entities*

We analyzed the results to understand if the size of the knowledge base, i.e., the number of local entities per city, affect the accuracy of the prediction. The count of local entities in our knowledge base ranges from 11 (for *Island Lake, Illinois*) to 1095 (for *Chicago*). This reflects the information

[16] `http://en.wikipedia.org/wiki/Endicott,_New_York`

City	Entities
New York City	New York City, Brooklyn, Harlem, Queens, New York Knicks, The Bronx, Manhattan, National Football League, American Broadcasting Company, Train station, Rapping, Times Square, Fox Broadcasting Company, Broadway theatre, New York Yankees, Staten Island, Brooklyn Nets, Amtrak, Hudson River, Macy's Thanksgiving Day Parade
Houston	Houston; Houston Texans; NASA; Houston Astros; Interstate 45; Houston Chronicle; Greater Houston; Harris County, Texas; Galveston, Texas; Downtown Houston; Houston Rockets; Texas
Seattle	Seattle; Seattle Seahawks; Seattle metropolitan area; Kobe; Microsoft; Downtown Seattle; Light rail; Alki Point
Nashville, Tennessee	Nashville, Tennessee;Belmont University; Frist Center for the Visual Arts; Southeastern Conference; Centennial Park (Nashville); Gaylord Opryland Resort & Convention Center
Pittsburgh, Pennsylvania	Pittsburgh; Midwestern United States; PNC Park; Station Square; Squirrel Hill (Pittsburgh); Giant Eagle; Fort Pitt Tunnel;Pittsburgh Steelers; Luke Ravenstahl; University of Pittsburgh;

Table 5: Examples of Local Entities found in tweets

available in Wikipedia about the city. Despite the variation in the amount of information available for each city, we find that our algorithm was able to predict locations of users from 356 distinct cities from our knowledge base having local entities in the range of 40 to 1095. Figure 7 shows the distribution of the users, whose location were predicted accurately, across continental United States compared to the distribution of all users in the dataset as shown in Figure 8. We see that the accurate prediction (within 100 miles) is not restricted to few cities.

5. 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 better than 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 modelling a training data set.

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 neighbourhoods in a city.

6. REFERENCES

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting Flu Trends using Twitter Data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE, 2011.

[2] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging Web Content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280. ACM, 2004.

[3] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. Semrank: Ranking Complex Relationship Search Results on the Semantic Web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 117–127, New York, NY, USA, 2005. ACM.

[4] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial Variation in Search Engine Queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.

[5] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering Short Texts using Wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.

[6] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @ phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. In *ASONAM 2012*, 2012.

[7] Zhiyuan Cheng, James Caverlee, and Kyumin 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.

[8] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.*, 16(1):22–29, March 1990.

[9] Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

[10] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[11] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[12] Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[13] Yegin Genc, Yasuaki Sakamoto, and JeffreyV. Nickerson. Discovering context: Classifying tweets through a semantic transform based on wikipedia. In DylanD. Schmorrow and CaliM. Fidopiastis, editors, *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*,

volume 6780 of *Lecture Notes in Computer Science*, pages 484–492. Springer Berlin Heidelberg, 2011.

[14] Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 545–554, New York, NY, USA, 2012. ACM.

[15] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. Exploiting Wikipedia as External Knowledge for Document Clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. ACM, 2009.

[16] Pavan Kapanipathi, Prateek Jain, Chitra Venkataramani, and Amit Sheth. User Interests Identification on Twitter Using a Hierarchical Knowledge Base. In *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 99–113. Springer International Publishing, 2014.

[17] Sheila Kinsella, Vanessa Murdock, and Neil 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.

[18] Kirill Kireyev, Leysia Palen, and K Anderson. Applications of Topics Models to Analysis of Disaster-related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, volume 1, 2009.

[19] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location Prediction in Social Media Based on Tie Strength. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 459–468. ACM, 2013.

[20] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[21] David Nadeau and Satoshi Sekine. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[22] Miles Osborne, Saša Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. In *Proceedings of the Workshop on Time-aware Information Access. TAIA*, volume 12, 2012.

[23] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named Entity Recognition in Tweets: An Experimental Study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[24] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's @wally?: A Classification Approach to Geolocating Users Based on their Social Ties. In *HT*, 2013.

[25] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[26] Benjamin E Teitler, Michael D Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: A New View on News. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 18. ACM, 2008.

[27] Amos Tversky. Features of Similarity. *Psychological review*, 84(4):327, 1977.

[28] Pu Wang and Carlotta Domeniconi. Towards a Universal Text Classifier: Transfer Learning using Encyclopedic Knowledge. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 435–440. IEEE, 2009.

[29] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2541–2544. ACM, 2011.