Geo Tagging Twitter Users using Wikipedia

Revathy Krishnamurthy, Pavan Kapanipathi, and Amit Sheth

Kno.e.sis Center, CSE Department
Wright State University, Dayton, OH - USA

Abstract. In this age of social media and citizen journalism, people are turning to Twitter to relay information in an emergency situation. To make the contents of their tweets actionable, we need to determine their location. However, a recent study has shown that only 3.17% of tweets are geotagged. Existing approaches are supervised and require training datasets to predict locations. Since crisis management is time-sensitive, the requirement of training data forms a bottleneck due to the time consuming process of creation of statistical models by crawling for tweets for each location needed. To this end, we propose an unsupervised approach that uses Wikipedia as a background knowledge to analyse tweets in order to predict the location of the users. This eliminates the need for a training dataset. We show that initial tests with this approach allows us to locate 30% of users within 100 miles of their actual location.

1 Introduction

The power of social media was demonstrated during Hurricane Sandy when more than 20 million tweets related to the hurricane were posted in a span of three days. It was reported that the volume of tweets doubled in these days as compared to the previous two days. 35% of these tweets contained news from media channels, information from government sources and eyewitness accounts[1]. The extensive use of social networking platforms during emergency situations has paved the way for new areas of research which focus on leveraging these platforms for disaster management [9].

Twitter has been used during emergencies to seek time sensitive information such as nearest shelters, food supplies, red cross locations etc. To provide personalized information, rescuers need the location of the victims. On the other hand, filtering tweets from the area of the disaster can help government agencies to mobilize rescue efforts. Thus, in improving emergency response using information from Twitter, the location of an online user plays an important role. However, only 3.17% of tweets are tagged with geographic coordinates [8].

As tweets are generally informal in nature and contain many acronyms and slang words, researchers have focussed on using statistical methods for identification of words with a strong geographic scope and then use these words to identify the location of a user. In the event of a disaster, to identify the location

[1] `http://www.journalism.org/2012/11/06/hurricane-sandy-and-twitter/`

of twitter users in real time, we need an approach that can be easily adapted to any geographic location. In order to overcome this challenge, we present an approach that utilizes Wikipedia as a source of background knowledge to predict users' location based on their online content. Briefly, our approach uses the graph structure of Wikipedia to find entities with a local geographic scope. The presence of these entities in users' tweets help estimate their location. Our intuition is that, more the users talk about entities with a local geographic scope, more are their chances of belonging to that location. Preliminary evaluation of our approach with a random sample from the dataset shared by Cheng et al[3] has shown promise.

In the rest of this paper, we will first discuss the related work on location prediction of Twitter users in Section 2, followed by a detailed explanation of our approach in Section 3. Section 4 discusses a preliminary evaluation of our approach and the paper concludes with a discussion on future work in Section 5.

2 Related Work

There have been two main approaches in addressing the problem of location identification of a twitter user: (1) Using the content of the tweets: based on the premise that the online content of a user is influenced by the geographical location of the user (2) Using the network information of the user: based on the assumption that the locations of the people in a user's network and their online interaction with the user can be used to determine the user's location.

Content-based location detection relies on a significantly large training dataset to build a statistical model that identifies words with a local scope. Cheng et al. [3] proposed a probabilistic framework for estimating a Twitter user's city-level location based on the content of approximately 1000+ tweets of each user. The locality of terms was determined by its spatial variation across the United States. Their approach on a test dataset of 5119 users, could locate 51% of the users within 100 miles and the average error distance was reported as 535.564 miles. The disadvantage of this approach was the assumption that a "term" is spatially significant to only one location/city. This challenge was addressed by Chang et al. [2] by modeling the variations as a Gaussian mixture model. Their tests on the same dataset showed an accuracy (within 100 miles) of 0.499 with 509.3 miles of average error distance. [5] created language models at different granularity levels from zip code to country level using a training dataset of 5.8 million geotagged tweets. At the city-level, they reported an accuracy of 65.7% and 29.8% on two different datasets.

Network based solutions requires the network information of a user. McGee et al. [6] 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 average error distance of 21 miles for 80% of their users. [10] 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, the limitation of a network-based approaches is the availability of location information of people in the given user's network.

3 Approach

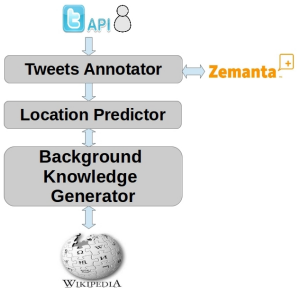

Fig. 1. Architecture

Previous research [1, 3] have established that the content of a user's posts reflects his/her geographical location. Using the same intuition we propose an approach that uses Wikipedia as a background knowledge to identify entities with a local geographic scope. An overview of the approach is shown in Figure 1. The approach comprises of three components (1) User Profile Generator: Extracts Wikipedia entities from users' tweets and scores them based on frequency, (2) Background Knowledge Generator: Generates background knowledge for each city using Wikipedia (3) Location Predictor: Utilizes the output of *Tweets Annotator* with *Background Knowledge* to predict the location of the user.

3.1 User Profile Generator

In this work, a user profile consists of the entities mentioned by a user in his/her tweets. Our hypothesis is that by correlating the occurrence of city specific entities from Wikipedia in a users tweets, we can estimate the location of the user. In order to extract entities from tweets we utilized Zemanta web service[2]. We opted for Zemanta because of its relatively superior performance [4] and also because of their rate limit extension (10,000 per day) provided for research purposes[3]. The extracted entities are scored based on their frequency of occurrence in the collection of a user's tweets.

Formally we define the profile of a user u as $P_u = \{(e, w_e)|e \in E, w_e \in R\}$ where E denotes the set of all Wikipedia entities spotted in their tweets and w_e is the frequency of mentions of entity e by user u.

3.2 Creation of Background Knowledge

Wikipedia is a crowd sourced encyclopedia containing dedicated pages for cities. Proportional to the size of the city, its Wikipedia page generally contains a variety of information about the city like its geography, culture, sports team, cityscape etc. Links to internal Wikipedia pages from a given page are an important feature of all Wikipedia pages. The aim of these links is to increase the

[2] `http://developer.zemanta.com/docs/suggest/`
[3] We thanks Zemanta for their support.

understanding of a user about the given page. For instance, the Wikipedia page of *Boston, Massachusetts* [4] mentions the *Boston Red Sox*, in the Sports section. It also provides a hyperlink to Boston Red Sox, that allows the user to navigate to the Wikipedia page of *Boston Red Sox*. We base our approach on the assumption that these internal links share varying degrees of relevance to the Wikipedia page of the city. As in the previous example, the Wikipedia page of Boston also contains an internal link to *Major League Baseball* which would be less representative of Boston than the *Boston Red Sox*.

Background knowledge for each city is represented by a weighted set of its internal links. We need to note that the internal links are also Wikipedia entities. Formally, we define a background knowledge for city c as $K_c = \{(l, s_l)|l \in I_c, s_l \in R\}$ where $I_c \in E$ denotes the set of Wikipedia entities that are the internal links in the Wikipedia page of city c and s_l is the score of internal link l that is calculated using ICF [7]. Mendes et al. [7] proposed *Inverse Candidate Frequency* for the task of entity disambiguation in DBPedia Spotlight. The idea behind ICF is that "a word commonly co-occuring with many resources is less discriminative overall". We use this intuition to identify the discriminative power of an internal link. If C is the set of all cities on Wikipedia, then ICF of an internal link l_i, that appears in n cities, is defined as:

$$ICF(l_i) = \log |C| - \log n \tag{1}$$

3.3 Location Estimation

For a user u with profile $P(u)$, in order to estimate the location, for each location c_i with knowledgebase K_{c_i} we find the overlapping Wikipedia entities set $O_{c_i u}$ between the user profile and location knowledgebase. Next, we use the following equation to estimate the most likely city of the user.

$$\underset{c_i}{\text{argmax}} \quad Score(c_i u) = \sum_{j=1}^{|O_{c_i u}|} w_{e_j} \times s_{e_j} \qquad \forall e_j \in O_{c_i u} \tag{2}$$

4 Evaluation

4.1 Dataset

From the test dataset published by [3], we randomly selected 935 users from United States. These users were distributed across 48 states. For each user, the dataset contains approximately 1000+ tweets. To create our knowledgebase, we selected 1670 cities in the United States of America having population greater than 20000. The entire collection of Wikipedia is available for download[5]. We use the dump dated 14-Feb-2014 to extract the internal links from the Wikipedia pages of all the cities in our dataset.

[4] http://en.wikipedia.org/wiki/Boston

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

4.2 Evaluation Metrics

We use the two metrics defined in [3] to evaluate our system (1) Accuracy (2) Average Error Distance . Accuracy is defined as the number 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 is the average of the error distance across all users.

4.3 Experimental Results

Our approach was able to locate 30.16% of the users within 100 miles of their actual location and the Average Error Distance across the 935 users was 886.25 miles. These users were distributed across 46 states. Table 1 shows a sample of the local words identified using Wikipedia, in the tweets of these users.

Location	**Wikipedia Links from User Tweets**
Chicago, Illinois	Chicago Cubs; North Center, Chicago;The Oprah Winfrey Show;Chicago White Sox
Las Vegas, Nevada	University of Nevada,Las Vegas; Las Vegas Boulevard; McCarran International Airport
Atlanta, Georgia	Atlanta Braves; Young Jeezy; Georgia Institute of Technology; Philips Arena; Buckhead (Atlanta)
Philadelphia, Pennsylvania	National Football League; Philadelphia Phillies; Philadelphia Eagles; Philadelphia Flyers
Detroit, Michigan	Eminem; General Motors; Detroit Red Wings; Greektown Casino Hotel;

Table 1. Wikipedia Links Annotated in Tweets

5 Conclusion and Future Work

As the role of social media continues to expand in emergency situations, the location of online users will play a crucial role in organizing relief efforts and disseminating information. In this paper, we have presented an approach that leverages Wikipedia to estimate the location of Twitter users. With a preliminary evaluation we have showed that the approach gives accuracy of 30% for 935 users selected randomly from existing datasets. While the current approaches (network-based and content-based) require a significant amount of training data for predicting users' locations, we have introduced an alternative that can perform the same task by using crowd-sourced background knowledge.

In this work, we used *ICF* to identify the discriminating power of an internal link. In future, we would like to use other scoring techniques to filter out seemingly irrelevant internal links from our consideration. In particular we would like to use the graph structure of Wikipedia and semantic types of Wikipedia pages,

to identify groups of internal links that display a stronger relationship to the city.

References

1. HAN Bo and Paul COOK1 Timothy BALDWIN. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062, 2012.
2. Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @ phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society, 2012.
3. 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.
4. 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*, HT '13, pages 21–30, New York, NY, USA, 2013. ACM.
5. Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *arXiv preprint arXiv:1207.0246*, 2012.
6. 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.
7. 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.
8. Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM*, 2013.
9. Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1), 2013.
10. 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 *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.