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. Prior knowledge of geographical locations of Twitter users will facilitate improved crisis management. 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 experiments beats the baselines of existing approaches with an accuracy of approximately 30%.

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]. This kind of extensive use of social networking platforms during time-sensitive situations such as disasters has paved the way for new areas of research which focus on the leveraging these platforms for disaster management [9].

Locations of Twitter users play a promient role for crisis management to inform or advice with necessary information. However, only 3.17% of tweets are tagged with geographic coordinates [8]. In improving emergency response, using information from Twitter, the location of an online user plays an important role. Current approaches to predict location of Twitter users based on their tweets, focus on building statistical models. In the event of a disaster, to identify user location 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

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

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 and performs better than their baseline.

In the rest of this paper, we will first discuss the related work on location prediction of Twitter users in Section 2, followed with a detailed explanation of our approach in Section 3. Section 4 discusses a preliminary evaluation of our approach and the paper concludes with discussion future directions we plan to take 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. Use of these words in tweets are used to narrow down the location of any user. 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 130689 users with 1000+ tweets each, 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 Gausian mixture model. Their tests on the same test 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 given 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.

The above mentioned approaches require prior training dataset (of either the content or network), which can be a bottleneck during disaster management. Our goal is to overcome this requirement of training data for each city by leveraging Wikipedia as the knowledge source.

3 Approach

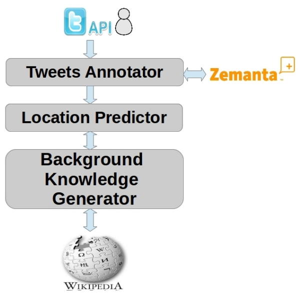

Fig. 1. Architecture

Previous research [1, 3] have established that the content of a user's posts reflects his/her geographical location. We follow the same line of content-based but with an unsupervised approach. An overview of the approach is shown in Figure 1. Tha approach comprises of three components (1) Tweets Annotator: Extracts Wikipedia entities from users tweets, (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 Annotation of Users' Tweets

Derczynski et.al, in their latest work [4] have compared three state of art *entity recognition and linking* systems for tweets. The systems compared with corresponding *Precision, Recall and F-Measures* are 1. Dbpedia Spotlight [7] (P=20.1, R=47.4, F=28.3), 2. Zemanta[2] (P=57.5, R=31.8, F=**41.0**) and 3. TextRazor[3] (P=**64.6**, R=26.9, F=38.0) . We used Zemanta[4] because of its relatively superior performance and also because of their rate limit extension (10,000 per day) provided for research purposes, on request[5].

3.2 Creation of Background Knowledge

Wikipedia is a crowd sourced encyclopedia. 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 understanding of a user about the given page. For instance, the Wikipedia page of *Boston, Massachusetts* [6] 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

[2] `http://developer.zemanta.com/`
[3] `http://www.textrazor.com/technology`
[4] `http://developer.zemanta.com/docs/suggest/`
[5] We thanks Zemanta for their support.
[6] http://en.wikipedia.org/wiki/Boston

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.*

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[7]. We use the dump dated 14-Feb-2014 to extract the internal links from the Wikipedia pages of all the cities in our dataset. Figure 1 shows the distribution of the count of internal links among all the city pages. From our dataset, *Pittsburgh* had 2684 as the largest count of internal links and *Round Lake Beach, Illinois* had 33 as the smallest count of internal links.

Scoring city-specific Entities Given a set of internal links for a city, we score each link to determine the degree of its relevance to the city. The more a given internal link is common to the cities in our dataset, the less it maybe relevant to one particular city. For example, in our dataset of 1650 cities, an internal link to the Wikipedia page of *Barack Obama* appears 105 times as opposed to *Southern California* and *Golden Gate Bridge* which appear 50 and 6 times respectively.

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 ability of an internal link with respect to a city. Let C be the set of cities in our dataset. Let I be the set of internal links for a city c ∈ C. The ICF of an internal link i ∈ I, that appears in n cities, is defined as:

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

3.3 Location Estimation

We used Zemanta[8] to annotate tweets. It maps entities in the input text to Wikipedia pages.

For a user U, let T_u be the set of their tweets, $Z_u = \{z_1,z_2,...,z_k\}$ be the set of entities annotated by Zemanta that map to a Wikipedia url. Let —z_k— represent the cardinality z_k in T_u. Let C be the set of cities in out dataset and $\forall$ $c_j \in$ C, let L be the set of its internal wiki links where $ICF(l_i)$ is the score $\forall$ $l_i \in$ L.

For the user U we compute the score of each city in our set as:

$$Score(c_j) = \sum_{i=1}^{I} |l_i| \times ICF(l_i) \qquad \forall l_i \in Z_u \tag{2}$$

We tag the city with the maximum score as the location of the user. (PAVAN: User argmax/argmin here – make it more a mathematical notation).

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

[8] http://www.zemanta.com

4 Evaluation

4.1 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.

4.2 Experimental Results

We evaluated our approach on 594 users with 1000+ tweets each. These users are distributed across United States. Figure X shows the distribution of the users. Using our approach we could locate 30% of the users within 100 miles of their actual location. Table 1 shows 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

In this paper, we have presented an approach that leverages Wikipedia to determine location of Twitter users. With a preliminary evaluation we have showed that the approach performs well with an accuracy of approximately 30% for 1000 random users selected from the datasets exposed by other existing approaches. This performance beats the baseline by over 10%. While the existing approaches (network-based and content-based) require training data for predicting users' locations, with this approach we have introduced an alternative that can perform the same task by leveraging crowd-sourced background knowledge.

In future we would like to explore other scoring techniques for entities for both 1. creating background knowledge for cities and 2. entities scoring from users' tweets . Specifically, the creating of background knowledge presently uses *ICF* which reflects the discriminative ability of the entity. However, we need to focus the usage of the entity for a particular city which is yet to be explored. We also acknowledge the limitation of this approach to be the coverage of Wikipedia., i.e. cities that are not present in Wikipedia will be ignored by our approach. We intend to explore other geo-datasets on LOD that can provide us with appropriate information to adapt to our approach.

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.