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. As more and more people are taking to microblogging networks, like Twitter, as a primary source of their communication with the rest of the world, the analysis of user generated content has become increasingly significant for crisis management. To make the contents of a tweet actionable, we need to be able to determine the location of the user. A recent study has shown that approximately only 3.17% of tweets are tagged with the user location. 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. But the biggest challenge in this approach is the requirement of a training dataset and creation of the statistical model, which can be a time consuming process. To this end, we propose an approach that uses Wikipedia as a background knowledge to analyse tweets in order to predict the location of the users. The main advantage of the proposed approach is that the use of Wikipedia 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

People turned to Twitter during "Hurricane Sandy" to communicate, express, share and receive information about the disaster. Twitter reported that its users had generated more than 20 Million tweets regarding the hurricane in a span of three days. It is also reported that the volume of tweets spiked compared to other days and most of these tweets (close to 35%) contained either news or information regarding the disaster[1]. These enourmous utilization of social networking platforms during disasters has opened up a new research focus to utilize the platforms to manage and assist people during disaster [7].

On the other hand, only 3.17% tweets generated are taged with geographical information [6]. In order to assist and manage disasters by relying on Twitter information, location of users plays a prominent role. For instance, a user with the following tweet "Need based tweet for food?" who is in need of information related to RedCross camps near by will have higher chances of getting required information if his/her location is known.

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

Existing approaches to location prediction of users are either network basedcite or supervised content-based. Both approaches require prior data either about the users (user's network) or training data for each location (supervised content-based). During real-world events, this requirement turns to a challenge, specifically for content-based approaches. In order to overcome this issue, in this work we present an approach that utilizes Wikipedia as a source of knowledge to predict users' location based on the content generated. Briefly, our approach uses the graph structure of Wikipedia to find local entities for each location. The presence of these entities in users' tweets help determine/predict their locations. Our intuition is that, more the users talk about local entities of a particular location, most likely they are from the corresponding location. Preliminary evaluation of our approach with a random sample from the dataset shared by prev approach[cite] has shown promise and performs better than the baseline.

In the rest of this paper, we will first discuss the related work on location prediction on Twitter in Section **??**, 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 **??**.

2 Related Work

Predicting location of Twitter users has take two directions (1) Network based: based on the asumption that users are connected to most people from his/her location. (2) Content-based: based on the premise that the online content of a user is influenced by the geographical location of the user

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 then used to narrow down the location of any user. Cheng et al. [2] 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 in probabilistic model was determined by its spatial variation (cite spatial variation) across united states. Their approach on a dataset of X users with X+ tweets performed with an accuracy of X within Y miles. One disadvantage of this approach was the assumption that a "term" is spatially significant to a particular location/city. This challenge was addressed by Phillies work [cite] by modeling the variations as Gausian mixture model, which in turn provided with better results. The results on the same dataset used by Chen et al showed an improvement of X within Y miles. [3] created language models at different granularity levels from zip code to country level using a training dataset of 5.8 million geotagged tweets.

Network based solutions requires the network information of a given user. McGee et al. [4] train 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 these network-based

approaches is the availability of location information of people in the appropriate user's network.

The above mentioned approaches require prior training dataset (either content or netowrk), which forms a bottle neck during disaster management. The approach presented in the paper intends to overcome this requirement of training data for each city by leveraging Wikipedia as the knowledge source.

3 Approach

Previous research [2] [1] have established that the content of a user's posts reflects his/her geographical location. We propose to use the information available in Wikipedia to establish words that are most representative of a given location.

3.1 Dataset

We selected 1670 cities in the United States of America having population greater than 20000. We randomly selected 600 users containing 1000+ tweets each, from the dataset made publicly available by Cheng et al[2].

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* [2] 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*.

The entire collection of Wikipedia is available for download[3]. 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.

3.3 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

[2] http://en.wikipedia.org/wiki/Boston
[3] http://en.wikipedia.org/wiki/Wikipedia:Database_download

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. [5] 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:

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

3.4 Location Estimation

We used Zemanta[4] 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.

4 Evaluation

4.1 Evaluation Metrics

We use the two metrics defined in [2] 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.

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

Location	**Wikipedia Links from User Tweets**
Portland, Maine	Portland Museum of Art; Old Port; Peaks Island; Portland Head Light
Las Vegas, Nevada	Bellagio (resort and casino); Hard Rock Cafe; Mandalay Bay;Las Vegas Hotel and Casino; Showgirls; McCarran International Airport
Detroit	Hockeytown; Detroit Red Wings; General Motors; Detroit Yacht Club; Belle Isle Park
Phoenix, Arizona	Chase Field; Downtown Phoenix; Cave Creek, Arizona;Maricopa County, Arizona
Pittsburgh	Forbes; Fort Pitt Tunnel; Pittsburgh Steelers; Luke Ravenstahl; University of Pittsburgh; Station Square; PNC Park; Pittsburgh Penguins;

Table 1. Wikipedia Links Annotated in Tweets

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. 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.
3. Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *arXiv preprint arXiv:1207.0246*, 2012.
4. 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.
5. 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.
6. 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.
7. 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.