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

The power of social media was demonstrated during Hurricane Sandy when more that 20 million tweets related to the hurricane were posted in a span of three days. Individuals and organizations have both turned to Twitter to coordinate relief efforts. Mining these tweets can provide valuable information to assist people in a timely manner[6]. To make these tweets informative and actionable, identifying the originating location of the tweet (and hence the user) is very important. [5] showed that only 3.17% of tweets are tagged with geographical information. Thus, geo tagging tweets is an important problem to solve.

Current approaches to detect location based on the content of tweets focus on building statistical models using a training dataset of tweets. In this paper we propose an approach that uses Wikipedia as a knowledge base to identify words with a local geographic scope. Our objective is to show that words having a strong association with a particular location, can be determined using the information available in Wikipedia.

Wikipedia is a large 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 teams, cityscape, transportation etc.The information is very detailed for a bigger city like Paris[1] (population of 12 M) and significantly lower for a smaller county like Monroe County, Wisconsin[2] (population of 44 K).

Our hypothesis is that by spotting entity in tweets and correlating them with the occurrence of Wikipedia entities for a given city, we can identify the location of a user. As in statistical methods, we do not need to manually collect seed words for each city of interest nor do we need a large training dataset to select the local words.

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, (2) Using the network information of the user.

The first approach is 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. [1] 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 estimate the spatial dispersion of a word and apply Laplace smoothing to overcome the sparsity of words across locations in their dataset. Their training dataset consisted of 130,689 users with 4,124,960 status updates. Their test dataset contained 5119 users with 1000+ tweets of each user. 51% of these users could be located down to their city-level within 100 miles. The average error distance was reported as 535.564 miles. Kinsella et al. [2] created language models at different granularity levels from zip code to country level using a training dataset of 5.8 million geotagged tweets. They reported their results on two datasets - SPRITZER containing 5% of the public twitter stream of 4 weeks and FIREHOSE containing 700,000 tweets from the Twitter Firehose. At the city-level, they reported an accuracy of 65.7% and 29.8% on the SPRITZER and the FIREHOSE dataset respectively.

A network based solution requires the network information of a given user. McGee et al. [3] used the relationship between users on Twitter to determine their location. Their training dataset consisted of 1.6 million twitter users and their network information. On a test dataset of 249,584 users, they reported 63.9% accuracy in determining the location within 25 miles. Rout et al. [7] formulated this task as a classification task and trained an SVM classifier on twitter users with known location, to use a person's social network to locate them. Their training dataset contained 200,000 twitter users. They tested their approach on a random sample of 1000 users and reported 50.08%accuracy at the city level.

[1] http://en.wikipedia.org/wiki/Paris

[2] http://en.wikipedia.org/wiki/Monroe_County,_Wisconsin

3 Implementation

3.1 Dataset

We selected 1670 cities in the United States of America having population greater than 20000.

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* [3] 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 association with 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[4]. We use this dump to extract the the internal Wikipedia links of all the cities in our dataset.

3.3 Scoring City-specific Entities

Mendes et al. [4] proposed *Inverse Candidate Frequency*

References

1. 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.
2. Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *arXiv preprint arXiv:1207.0246*, 2012.
3. 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.
4. 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.

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

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

5. 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.
6. 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.
7. 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.