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[3]. To make these tweets informative and actionable, identifying the originating location of the tweet (and hence the user) is very important. [2] 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.

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. 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.

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. The main disadvantage of this method is that to identify tweets from a given location, it is required that we have a set of good quality of tweets from this location to train the model. 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 test dataset contained 5119 users with 1000+ tweets of each user. 51% of these users could be located down to their city-level. The average error distance was reported as 535.564 miles. A network based solution requires the network information of a given user.

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. 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.
3. 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.