IBM researchers' algorithm explores tweets for home location cues

March 24 2014, by Nancy Owano

Credit: arXiv:1403.2345 [cs.SI]

(Phys.org) —By drawing on the content of users' tweets and their tweeting behavior, a team of three IBM researchers said they have a new algorithm to infer the home location of Twitter users at different granularities, including city, state, time zone or geographic region. The algorithm makes use of the person's last 200 tweets for tracking. The scientists described their approach as an "ensemble of statistical and heuristic classifiers" and with this approach they said they could predict locations and make use of a geographic gazetteer dictionary (USGS [United States Geological Survey] gazetteer) to identify place-name

entities. They analyzed movement variations of Twitter users, built a classifier to predict whether a user was travelling in a certain period of time and used that to further improve their detection accuracy.

The paper, "Home Location Identification of Twitter Users," submitted earlier this month on arXiv.org, is by Jalal Mahmud, Jeffrey Nichols and Clemens Drews of IBM Research. They said they had experimental evidence to suggest their algorithm works well in practice. In fact, they said it "outperforms the best existing algorithms for predicting the home location of Twitter users."

From July 2011 to Aug 2011, they collected tweets from the top 100 cities in US by population. They invoked the Twitter REST API to collect each user's 200 most recent tweets (less if that user had fewer than 200 total tweets). Some users discovered to have private profiles were eliminated. The final data set had 1.5 million tweets by 9551 users.

In listing their contributions, the IBM researchers said, when tested using the 1.52-million tweet dataset from 9551 users from 100 US cities, that their algorithm outperforms the best existing algorithms for home location prediction from tweets. "Our best method achieves accuracies of 64% for cities, 66% for states, 78% for time zones and 71% for regions."

Microblogging is of great interest to scientists seeking various research answers. As for Twitter, scientists regard this as an ideal laboratory for mining data. The paper's authors, however, noted that less than 1% of tweets are geotagged and they said that information available from the location fields in users' profiles is unreliable at best.

" In this paper, we aim to overcome this location sparseness problem by developing algorithms to predict the home, or primary, locations of Twitter users from the content of their tweets and their tweeting

behavior. Ultimately, we would like to be able to predict the location of each tweet and our work to predict a user's home location is a key step towards achieving that goal."

Among their future research goals are to incorporate more domain knowledge in their location prediction models, such as a landmark database. They said they hoped to integrate their algorithm "into various applications to explore its usefulness in real world deployments."

More information: Home Location Identification of Twitter Users, arXiv:1403.2345 [cs.SI] arxiv.org/abs/1403.2345

Abstract

We present a new algorithm for inferring the home location of Twitter users at different granularities, including city, state, time zone or geographic region, using the content of users tweets and their tweeting behavior. Unlike existing approaches, our algorithm uses an ensemble of statistical and heuristic classifiers to predict locations and makes use of a geographic gazetteer dictionary to identify place-name entities. We find that a hierarchical classification approach, where time zone, state or geographic region is predicted first and city is predicted next, can improve prediction accuracy. We have also analyzed movement variations of Twitter users, built a classifier to predict whether a user was travelling in a certain period of time and use that to further improve the location detection accuracy. Experimental evidence suggests that our algorithm works well in practice and outperforms the best existing algorithms for predicting the home location of Twitter users.

https://techxplore.com/news/2014-03-ibm-algorithm-explores-tweets-home.html