ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Fast track article

Improving geolocation of social media posts

Elizabeth Williams *, Jeff Gray, Brandon Dixon

University of Alabama, Box 870290 Tuscaloosa, AL 35487, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Social media
Geolocation
Pervasive computing

ABSTRACT

Pervasive social systems often take advantage of geographical information to provide real-time information to users based on their location. However, due to privacy concerns, many social media users do not share their exact geographical coordinates. In this paper, we describe our technique that predicts locations of posts that are not associated with explicit coordinates, a process called geolocation. Existing research has utilized the content of a post as well as the post author's social media relationships with other users to estimate location. Our research provides a novel approach to geolocation by combining multiple techniques, as well as adding a new technique: estimating location by clustering similar social media posts that are centered in a geographical area.

1. Introduction

Pervasive computing systems are often adaptive and react to changes in a user's environment. A common source for personalization within websites and mobile applications is a user's location. For example, retailers with mobile apps such as Target often send notifications when a user enters their store. Websites frequently show advertisements based on a visitor's location.

Social media platforms have also provided ways to incorporate their user's location. On many social media platforms, users have the ability to tag their posts with their current location information. Because of the fast dissemination of information on social media, it has been used as a method for detecting natural disasters and gathering information about major events such as the 2013 Boston Marathon explosions [1]. Knowing the location information for these types of events can allow first responders to pinpoint the exact location, as well as the reach, of the event. For example, using a form of geolocation, Sakaki et al. [2] discovered the epicenter of earthquakes as well as the location of the aftershocks. In order to perform a full analysis on social media posts to identify major events or discover people's opinions in different geographical areas, location information for the posts is required. As discussed in [3], social media may not always provide locations that are as accurate as GPS, but social media provides public access to location data as well as a social aspect. Also, discovering locations from social media can be useful when traditional location services such as IP-address geolocation give inconsistent results [4].

City-wide geolocation can be useful in certain instances. For example, geolocation can be used to discover protests and organized movements occurring in various cities [1]. However, because the ultimate goal of geolocation is predicting all possible locations with 100% accuracy, existing geolocation methods have room for improvement.

However, on the social networking site Twitter, as few as 0.87% of tweets are geotagged, or associated explicitly with geographical coordinates [5]. In our previous research, we found that only 2.63% of tweets we analyzed were tagged with explicit location information [6]. Without many tweets being associated with a specific location, adaptive systems are unable

* Corresponding author. Fax: +1 205 348 6959.
E-mail addresses: eawilliams2@crimson.ua.edu (E. Williams), gray@cs.ua.edu (J. Gray), dixon@cs.ua.edu (B. Dixon).

http://dx.doi.org/10.1016/j.pmcj.2016.09.015
1574-1192/

to take advantage of the location of the tweets to provide location-specific information. This paper describes our system, GeoContext Locator (GCL), which provides a method for predicting the location of, or geolocating, social media posts.

Like most related research (e.g., [7,8,4]), as described in Section 2, we chose Twitter[1] to implement GCL for several reasons. First, Twitter allows users to write short posts with freeform text that can be analyzed to discover locations. Also, Twitter allows users to connect locations to posts in two ways. Users can attach geographical coordinates directly to the post, called geotagging. Alternatively, Twitter users can tag tweets using Twitter Places,[2] which allows a user to tag a post with the name of a location. This tag also includes a bounding box of geographical coordinates around the location. In addition to location information, Twitter allows users to establish relationships with other users using the concept of *following*. If User A follows User B, User A can receive User B's tweets on their Twitter home page. Twitter describes users who User A follows as *friends* and users who follow User A as *followers*.

Although we chose to utilize Twitter for GCL, it can be adapted easily to any other social network. GCL simply processes JSON objects from a social media stream that have content and location information, so any other social network or information provider that attaches geographical coordinates to shared information could be used. In this paper, we describe our system, GCL, for geolocating a stream of tweets. Our system is unique in several ways:

1. GCL combines analysis of the content of the tweet, the location specified on the user's account, and locations of the user's friends and followers to perform geolocation. Previous research has not utilized all of these aspects together to discover a tweet's location.
2. We present a novel approach to geolocation by estimating a tweet's location by analyzing the locations of tweets with similar content in real-time. We evaluate the effectiveness of this new approach in Section 3.5.
3. GCL introduces new ways of extracting location information from tweets by analyzing various combinations of tokens within the text of the tweet. GCL utilizes cognitive computing resources AlchemyAPI[3] and Dbpedia,[4] as well as the Google Places API, to map text to geographical coordinates.

The structure of this paper is as follows: first, in Section 2, we discuss existing work in the area of geolocation of social media posts. In Section 3, we define our approach, GCL, for predicting tweet location. We then describe our experimental setup and results in Section 4. Finally, we conclude with a discussion of future work.

2. Related work

Existing research on geolocation has focused mainly on two areas: (1) geolocation based on the content of the social media post, and (2) geolocation based on the relationships of the user with other users on the social media network. We first discuss research based on the content of the post.

2.1. Content-based geolocation

Several approaches are based on comparing a tweet to previous tweets with known locations to discover similarities between the tweets. Tweets that are determined to be similar can be inferred to have similar locations. A significant number of geotagged tweets occur as a result of the user having other location-based social networks, such as Foursquare, that send automatic geotagged messages to their Twitter account. Watanabe et al. [9] created a database from tweets that were posted via Foursquare. They were then able to use the database to look up place names in non-geotagged tweets and predict the location of the non-geotagged tweets. Ikawa et al.'s [10] approach for predicting user locations involves extracting keywords from tweets in a training set. Keywords are then extracted from the test set tweets, and the keywords are compared to those in the training set. Cosine similarity is computed between the keywords, and the location associated with the keyword set in the training set is estimated as the location of the tweet in the test set.

Several approaches predict locations within a grid cell rather than as geographical coordinates. Wing and Baldridge [11] ran their geolocation algorithm on Wikipedia documents, rather than a more traditional type of social media platform such as Twitter. However, like the previously described related work, the authors also utilize the content of the document to predict a location of the text. Their approach divides the Earth into varying sized cells and predicts a cell for each document. Their model calculates the distribution of words over different locations and compares the word distribution of each document to the word distribution of each geographic cell, eventually choosing the cell with the highest similarity. Baldwin et al. [12] also predict the location of the author of each post within a grid cell on a map. Their approach utilized a naive Bayes classifier to approach the problem of geolocation. They split each Twitter post into tokens and consider each token as a feature in the classifier.

Some approaches utilized existing web services or other APIs in order to perform geolocation. Jaiswal et al. [5] utilized a named-entity extraction module, ANNIE, to extract possible locations from the content of Twitter posts. The locations were

1 http://dev.twitter.com.
2 http://dev.twitter.com/overview/api/places.
3 http://www.alchemyapi.com/.
4 http://dbpedia.org.

Please cite this article in press as: E. Williams, et al., Improving geolocation of social media posts, Pervasive and Mobile Computing (2016), http://dx.doi.org/10.1016/j.pmcj.2016.09.015