Information Systems 57 (2016) 77–87

Contents lists available at ScienceDirect

Information Systems

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

Home is where your friends are: Utilizing the social graph to locate twitter users in a city

Dimitrios Kotzias [a], Theodoros Lappas [b], Dimitrios Gunopulos [c]

[a] University of California, Irvine, United States
[b] Stevens Institute of Technology, United States
[c] University of Athens, Greece

ARTICLE INFO

Article history:
Received 13 December 2014
Received in revised form
9 July 2015
Accepted 27 October 2015
Available online 5 November 2015

Keywords:
Social networks
Data sparsity
Location profiling

ABSTRACT

Micro-blogging services such as Twitter have gained enormous popularity over the last few years leading to massive volumes of user generated content. A portion of this content is shared via geo-aware mobile devices, such as smartphones. Pieces of information shared on such a device can be tagged with the user's location, conditional on the user's settings. These geostamps enable a number of mainstream applications, such as emergency response, disease tracking, news reporting, and advertising. Unfortunately, informative geostamps are typically sparse, since content is often shared via devices that do not support geo-tagging, such as desktop or laptop computers. In addition, even if a mobile device is used, a flawed geo-location service can lead to missing geostamps, or geostamps that are too general to be informative. In this work, we address this sparsity issue via a new approach that identifies users attached to a given location of interest, such as a city. We then focus on retrieving specific tweets at a finer granularity within the given location, such as specific blocks within a city. Our approach leverages the correlation between strong connectivity in the social graph and proximity in the real world, while utilizing both textual tweet content and Twitter's underlying social graph. Previous relevant work assumes that all required Twitter data is available without access restrictions. This is an unrealistic assumption, since Twitter limits the number of data requests per user and charges a subscription fee for unrestricted access. Therefore, in order to increase the number of practitioners and applications that can benefit from our work, we optimize our method to work with the minimum amount of queries to the Twitter API. Finally, our experiments demonstrate the efficacy of our work via both a quantitative and qualitative evaluation.

1. Introduction

Over the last decade, online platforms where individuals generate and contribute content have gained massive popularity. In most of these platforms, individuals connect with each other and establish social networks. A typical example is Twitter, which currently hosts more than 200 million users contributing more than 400 million tweets per day [15]. Twitter is unique among such social networks, due to its ability to propagate diverse information to an immense number of people at an even faster pace than conventional news networks [30]. The use of mobile access points, such as smartphones, further enriches the Twitter network with a spatial dimension by enabling users to generate and share content from a variety of locations in *real time*. The unique nature of the Twitter network has motivated a number of research efforts and applications focusing on spatio-temporal data

E-mail addresses: dkotzias@ics.uci.edu (D. Kotzias), tlappas@stevens.edu (T. Lappas), dg@di.uoa.gr (D. Gunopulos).

http://dx.doi.org/10.1016/j.is.2015.10.011
0306-4379/

[23,16,7] that include the time and location of a tweet's creation. Location information is crucial, since it serves as a direct way to transfer and apply knowledge from an online setting back to the physical world. This connection enables a plethora of mainstream applications, such as the tracking of diseases [5,33], the detection and management of emergency situations [34], the prompt delivery and propagation of news updates on local events [23], the analysis of the behavioral and mobility patterns of people within a city [7,11], and targeted advertising [1]. Location information is also necessary for the design, improvement and optimal resource allocation of modern cities. In recent relevant work, Gionis et al. [13] use geo-tagged information from social networks to recommend customized tours in urban settings. Cranshaw et al. [9] identify spatial clusters in urban areas and motivate potential applications in urban planning. Van Gennip et al. [38] use geo-tagged police records to determine gang membership and model gang violence in the city of Los Angeles. Galbrun et al. [12] combine crime data from the cities of Chicago and Philadelphia to model crime in these areas provide optimal route in terms of safety and distance. The importance of location information has also been recognized by major online platforms: Twitter recently started reporting local trends.[1] while Google's search engine considers the user's location when returning relevant results.[2]

The underlying assumption for applications based on location information is that the available data will be sufficiently dense to support their functionality. In practice, however, location information is very sparse. Research on Twitter data suggests that a big percentage of users either do not provide their location information in their profiles or submit noisy data [6], with only 48% of users providing an actual location at the city level or better [14].

Moreover, the number of actual tweets with geographical coordinates is much lower, in the order of 1% [35]. Reasonably, this sparsity of information constitutes a major issue for all the aforementioned applications. The problem caused by the sparsity of geo-tagged tweets is exacerbated by the fact that practitioners and applications typically have access to just a subset of the Twitter data. Twitter only allows for a limited number of requests for data through its free API, through rate limiting or sampling of tweets.[3] Even though recent updates to the Twitter API allow querying for tweets from a specific location, even if a geostamp is not present,[4] the sample of tweets returned is not adequate for any of the above applications. This is demonstrated in Fig. 1, which shows the number of new tweets per minute that we were able to crawl for the city of Dublin using the API. Gaining unlimited access is a possible but costly option[5] that is simply not available to most users and researchers.

Taking into consideration the limitations of the free API, we design our method to achieve high quality results while respecting an upper bound on the number of data requests. Given a specific area of interest, such as a city, our approach operates in two steps: first, we identify users attached to the area based on their connections within the social graph. We then use the tweets of these geo-located users to retrieve more tweets and densify the available data for the given location of interest.

Our contribution is threefold:

- We study and verify the strong connection between the geographical proximity of users and their distance in the social graph
- We provide a framework for (1) identifying more users at a location of a city-level granularity and (2) attaching geographical coordinates to individual tweets within that city
- We introduce the first approach for user geo-location that takes into consideration the limitations imposed by social platforms on data access, such as those imposed by the Twitter API.

Fig. 2 illustrates our methodology. We start with a set of users from a specific location and consider the social graph formed by their connections. The set is then filtered to retrieve a seed of nodes which is provided as input to the `MaxEdge` algorithm, which allows us to discover more users in the given location of interest.

2. Related work

This paper builds upon our previous work on the sparsity of location information on Twitter [20]. In this new extended version of our work, we provide (i) a much more detailed discussion of the experimental results, which takes into consideration the demographics of the cities included in our datasets (Section 5), (ii) a quantitative analysis that motivates the need for methods that respect Twitter's API limitations (Section 1), (iii) a theoretical analysis that verifies the hardness of the problem of retrieving users from a given region of interest (Section 2), (iv) a more complete discussion of related work, including more recent work (Section 2), (v) a discussion on open problems and future work (Section 6).

Related work by other researchers has focused on: (1) identifying the location of a given user, (2) identifying the location of an individual tweet, and (3) attempting to model the spatial distribution of individuals. Next, we discuss each of these three categories in more detail.

The fundamental difference between our own work and previous papers from the first category is that they focus on the geo-location of a user that is provided as input, while our goal is to retrieve *new* users and tweets that are associated with a given region. Eisensten at al. [10] attempt to solve the user geo-location problem through geographical topic models . They capture the difference in the use of language for a specific topic among people from distant areas. They are able to predict the location of a user with an error mean distance of 900 km, and achieve a 27% accuracy when predicting the state of a user. Their approach assumes that there is a significant distance and language difference between the different areas and can

[1] https://blog.twitter.com/2010/now-trending-local-trends
[2] http://www.google.com/landing/now/
[3] https://dev.twitter.com/rest/public/rate-limiting
[4] https://dev.twitter.com/rest/reference/get/search/tweets
[5] http://gnip.com/sources/twitter/realtime/