Information Systems ■ (■■■■) ■■■–■■■

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

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/

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

thus not be applied to a city-level granularity, in which people are more likely to have similar conversational habits. In recent work on the same dataset, Ahmed et al. [2] proposed a tree-like hierarchical structure of topics at which the lower levels of the tree represent more specific versions of the general topics at the parent nodes. In this way, they are able to extract location-specific topics and place users with an average error of 298 km. Further, Cheng et al. [6] utilized the locality of phrases rather than topics, and manage to pinpoint the city of 51% of users within 100 miles of their actual location. Mahmud et. al. in [29] improve this method by identifying *named local words*. These are terms with a strong geographical association, such as the name of a location, as retrieved from services such as foursquare. They use a Multinomial Naive Bayes classifier and test hierarchical algorithms that first predict the country, state, and city of a user. They achieve an accuracy of 58% for a 100 miles radius. The common denominator of these approaches is that they take advantage of the difference in the use of language for specific topics or words which identify the location of users. However, such reasoning is at a disadvantage in the case of smaller areas where there are not many language differences. Our work differentiates at two fundamental levels. First, we take advantage of user's writings as well as their relationships in the social graph, which enables us to predict location with greater accuracy and geographical details. Second, we propose a way to attach exact geographical coordinates at the tweet level, after we have identified the location of the user who created it.

Ren et al. [32] place each user at the location associated with the majority of his friends. While they achieve a precision of 59.3%, they only test their methodology on 704 users. Working on a much larger dataset of Facebook data, Backstrom et al. [4] suggest a correlation between friendship and distance on the map, and build a more elaborate model to find the probability of a users home location, given the location of his contacts. While they place 67.7% of users correctly, they also assume knowledge of all the friends of the user we are trying to locate, an assumption that is rarely valid for very large graphs, such as those created on social platforms. Our method aims to minimize the number of queries to the social graph and, more importantly, operates within realistic data restrictions, which translates to the absence of the knowledge of the location of all the friends of a user. More recently, Li et al. [25] create an enriched social graph, in which each user is connected with other users via weighted edges. The weight of an edge represents the probability that the two users share the same location. A given user is then geo-located based on the geo-stamps of her neighbors, as well as the weights of the edges that connect her to them.

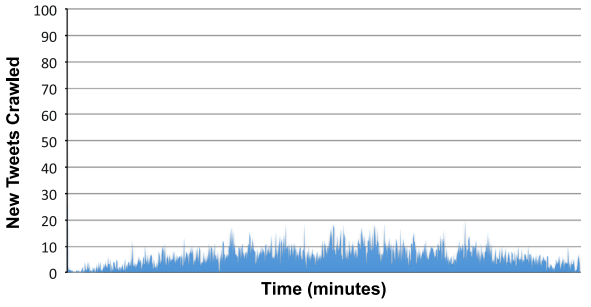

Fig. 1. Number of new tweets per minute for a 15 miles radius around the city of Dublin.

The second group of relevant research focuses on the geolocation of given tweets. Our work differs in that (1) our goal is retrieve new tweets from a given region, and (2) we focus on the retrieval of users, as well as tweets. Ikawa et al. [17] attempt to estimate the location of a given tweet by associating expressions with locations. For each tweet, they find the location with the closest word list and place it there. However, the underlying assumption that people will tweet about the place they are in does not always hold, leading to an accuracy of 14% for a radius of 5 kilometers. Li et al. [26] achieve a better precision for the same problem as they attempt to identify Places of Interest (POI). They first build a Language Model for each POI, based on tweets that occurred there and information crawled from the POI's website. They then rank the KL-divergences between the tweet and each language model. They test their method for the 10 top POIs at a city, and reach an accuracy of more than 60% for their best case. However, their accuracy fluctuates greatly based on the number of available tweets for each POI, and their premise is somehow unrealistic since, in a real-world scenario, there are much more than 10 possible locations within a city. Kinsella et al. [19] provide a method for pinpointing the location of both users and tweets in a variety of granularities. They build language models for each location and test a Query Likelihood model, in order to predict the location of a tweet. Their best result is an accuracy of 31.9% for users at the town granularity and 13.9% for tweet location at the zip-code granularity. While they use the same model for both problems, our methodology employs a customized approach for each task, which enables us to create a more robust system.

The third group of relevant work focuses on modeling the spatial density of users. Cho et al. [7], propose a

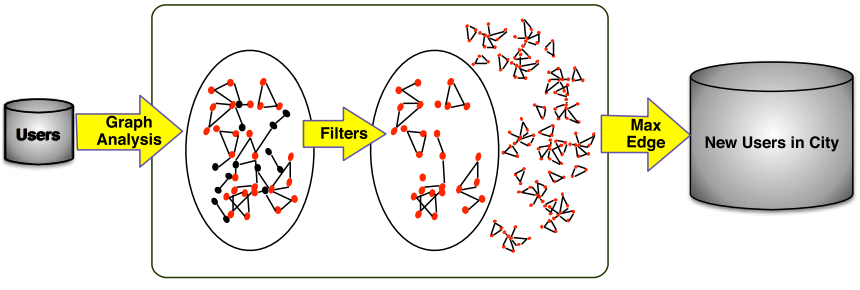

Fig. 2. An overview of our methodology.

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

mixture model of Gaussians to capture the diversity of human behavioral patterns. Lichman et al. [27] improve upon this method, by exploring Kernel Density Estimators (KDE), with some impressive results. They use a mixture of KDE's, which helps their model avoid overfitting and deal with the sparsity of data. Their work is a stellar example of applications that can benefit from methods such as our own, which aim to address the sparsity issue in location data.

To the best of our knowledge, despite the large number of papers that focus on estimating the location of a single user, the problem of identifying more users from a specific location has yet to be addressed. In contrast to most of the previous work done in this area we: (i) attempt to identify the location of a user at the *city level*, (ii) combine this information to attach geographical coordinates to a tweet, by taking under consideration *both* the relationships of users in the social graph as well as the content of their writings, and (iii) use a computational model based on Twitter, optimized for the number of queries.

3. Problem definition

Given a specific city, our goal is to (i) discover users that are geo-located in the city and (ii) associate tweets by the discovered users to specific blocks within the city. The two problems are formally defined below:

Problem 1. *Geographically-focused user retrieval* : Given a set of users U who tweeted from a specific city C, identify as many users from the same city as possible, by asking at most k-queries about the social graph.
Problem 2. *Tweet geo-location* : Given a set of tweets T_1 for which we know the exact geographical location within a city C and a set of tweets T_2 for which we do not have location information, attach coordinates within C to every tweet in T_2.

The primary focus of our work is placed on the first problem, since the availability of users from a specific location is essential for a plethora of mainstream applications in multiple domains [5,33,34,23,7,11,1]. We then address the second problem by building on the solution of the first one: as we demonstrate in our experiments, the location with the strongest association with a tweet's creator is a highly accurate predictor of the location of the tweet itself (even more so than the text of the tweet). Therefore, the second tweet geo-location problem provides a natural platform for us to verify the efficacy of our algorithms for geographically-focused user retrieval.

4. Methodology

In this section we describe our methodology for the two problems addressed in our work. We begin in Section 4.1 with an approach that retrieves users attached to a given city of interest. Section 4.2 then describes a method for the accurate geo-location of tweets within a city.

4.1. Problem 1: user discovery

Our methodology for user location utilizes the connections in Twitter's social graph to identify users from a given city. The underlying assumption is thus that the proximity in the social graph is correlated with actual geographical proximity: *users who live in the same region are also likely to be close to each other in the social graph*. Previous work has indeed verified this assumption [21,37,25,4], as well as a strong small world effect in local communities [3]. In order to further test the validity of this assumption, we present a relevant motivational study. We start by using Twitter's API to retrieve tweets from three different cities: Dublin, Manchester, and Boston. The wider metropolitan area of each city was considered. The data was collected by specifying the area's bounds as a pair of [*latitude*, *longitude*], corresponding to the area's south-west and north-east corners. We then retrieve the profiles of the creators of these tweets. In order to identify the location of each creator, we use a gazetteer provided by GeoNames.[6] This contains a list of all the areas and neighborhoods of a city. Thus, given the location listed on a user's profile, we can check whether it is indeed included in the city. Table 1 provides the following descriptive statistics for all three datasets, which were collected between April and August of 2013: the total number of users from the city, the number of tweets posted by all included users, and the density of the city's metropolitan area. The considerable difference in density between Boston and the other two cities allows us to evaluate our methodology in the context of different levels of data sparsity.

This first phase gives us a set of users for each city. Next, we explore the connectivity of these users within the Twitter graph. We begin by building the subgraph that includes all the social connections among the users from a given city. The prevalent connection on Twitter is the “follow” action: a user may choose to follow other users in order to receive updates on their tweets. A “friendship” connection is then created between two users that follow each other. Given the subgraph of these connections, we apply the following 3-step filtering process toward the evaluation of the relationship between geographical and online social proximity:

1. Discard users with more than 30,000 followers. This eliminates celebrities and bots that are likely to follow be followed by multiple people from around the globe.
2. Discard users that are not included in the largest connected component of the remaining graph. The goal is to capture the intuition that a person with multiple connections to members of a local community is more likely to actually reside there. This step only considers friendship connections (mutual follows), since these are more likely to occur between two people who live in the same location and actually know each other.
3. Discard users who are just visiting a city and are thus not geographically attached to it. This is implemented by only maintaining users who tweeted repeatedly and

[6] www.geonames.org

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

Table 1
Dataset details.

City\Statistics	Users	Tweets	Density
Dublin	43k	1.9 M	3943/km^2
Manchester	40k	1.3 M	4051/km^2
Boston	55k	1.5 M	909/km^2

for an extended period of time from the city. Specifically, we only keep users who tweeted more than 4 times in a period greater than 100 h.

The process of discarding users with more than 30,000 followers and less than 4 tweets in 100 h (less than 1 tweet per day) is meant to eliminate extreme outliers (celebrities and barely active users) that cannot be associated with a specific location. Specifically, only 0.08% of the users in our data have more than 30,000 followers. While tuning these numbers is beyond the scope of the motivational study described in this section, it is intuitive that decreasing the upper bound on followers or increasing the lower bound on the post-rate numbers will only result in stricter filtering and a smaller graph.

The results of this analysis for all three cities are shown in Fig. 3. The figure shows 3 columns for each city, with each column representing the percentage of remaining users after each of the three filtering steps described above. Each column reports, from bottom to top, the percentage of (1) users who fall within the city of interest, (2) users who fall within a broader district that includes the city: Ireland for Dublin, England for Manchester and the state of Massachusetts for Boston, (3) users who did not specify their location (i.e. had a Null value) in their profiles, and (4) users who have specified a location outside the city.

We observe that the first and second filtering steps eliminate a significant number of the users who fall outside the city or did not provide a location in their profiles. In contrast, the number of users within the city of interest is either unaffected or only drops slightly. These results verify our intuition and previous work that suggested correlation between geographical proximity and online friendship [4]. Specifically, our study provides evidence that the social graph can be used to (1) eliminate users that have no attachment to a city and (2) retrieve users from a specific city by identifying clusters of individuals who replicate their real-life proximity and friendship on Twitter.

The encouraging findings of our study motivate the design and use of our `MaxEdge` Algorithm for user geolocation. The algorithm utilizes the Twitter API, which can be used to retrieve users attached to a given geocode.[7] However, This functionality is rate-limited and can thus only be applied to retrieve a sample of such users. Taking this into consideration; `MaxEdge` accepts as input an upper bound k on the number of requests to the Twitter

[7] https://dev.twitter.com/rest/reference/get/search/tweets

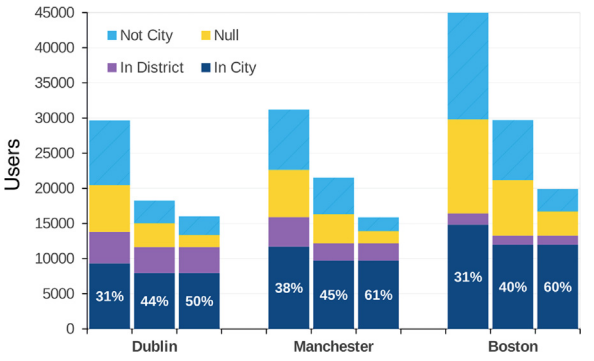

Fig. 3. Graph analytics for each city. Column 1: all users, Column 2: largest connected component, Column 3: time tweets filter.

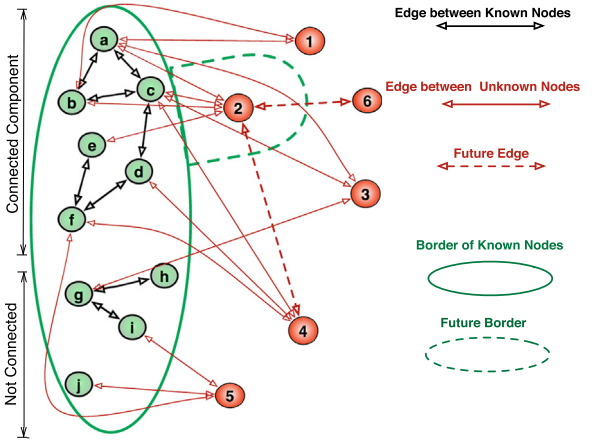

Fig. 4. Seed and frontier example.

API. The second input is a geocode R representing a region of interest, such as a specific city.

The algorithm begins by retrieving a sample set of users attached to the given region. The set is filtered using the 3-step filtering process described earlier in this Section. The resulting filtered set of users serves as the *seed*, which is used to further explore the social graph and retrieve additional users from the region as follows: First, we use the API to retrieve the id's of friends of all the nodes in the seed, which allows us to build a model graph. We refer to the set of these "friends" nodes that are not also in the seed as the *frontier*. We then repeatedly retrieve the user from the *frontier* with the largest number of connections to the seed and use the API to retrieve the location of this *winner* node. If the location falls within the region of interest R, the user is added to the seed and the *frontier* is extended accordingly. The process is repeated until we have exhausted all k allowed API requests.

Every iteration of the `MaxEdge` algorithm maximizes the conductance of the seed over the frontier. Higher conductance is desirable, since it captures the quality of the community structure of a subgraph [24,8,18]. Given a set of nodes S, conductance is formally defined as:

$$\phi(S) = \frac{\sum_{i \in S j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}$$

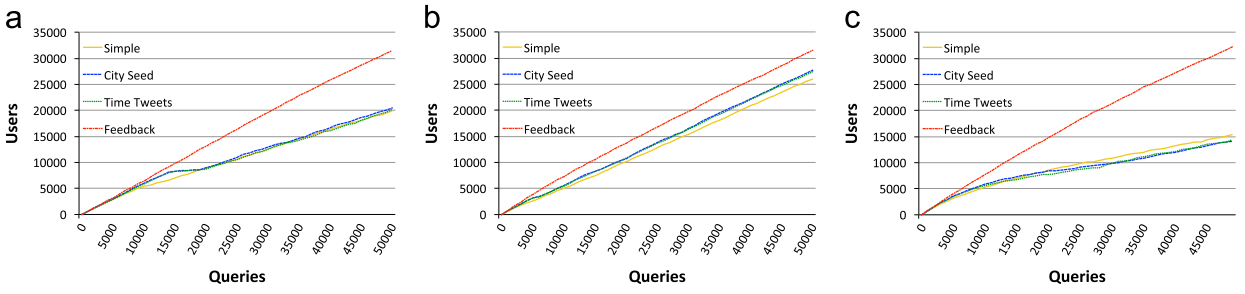

Fig. 5. Users in C vs number of queries in Twitter. (a) Dublin. (b) Manchester. (c) Boston.

where A is the adjacency matrix of the graph and $A(S) = \sum_{i \in S} \sum_{j \in S} A_{ij}$.

In our context, conductance is computed as the number of edges from the seed to the frontier, divided by the number of edges within the seed. At every step, we take the maximum known number of edges that cross this boundary and bring them in it, thus locally maximizing the community structure of our known users.

An illustration of `MaxEdge` is shown in Fig. 4: the nodes within the green bounded ellipse on the left represent the *seed* nodes. The red nodes numbered 1–5 on the right side represent the current *frontier*. These are nodes with a still unknown location that are connected with at least one node from the *seed*. In this example, node 2 has four edges in the *seed*, more than any other node. After this winner node has been moved from the frontier to the *seed*, its edges are going to become known, hence we will discover node 6, and update the count for node 4. The next winner be selected and handled in the same manner.

Algorithm 1 shows a weighted version of the `MaxEdge` algorithm, which we refer to as `Feedback`. This weighted version rewards nodes that pointed to a "correct" (i.e. a user that falls within the region of interest), and penalizes nodes that pointed to an incorrect one. Specifically, if the location of the *winner* node is within the region of interest, all the edges of the nodes that pointed to that winner, get an increase in weight. Similarly, if the node explored was not in the region, all the nodes that pointed to it will be penalized. This allows for users that live in a city, but have many outside connections, to weigh less in the decision of the next winner, while promoting the votes of those with strong in-city ties.

Algorithm 1. The `FeedBack` algorithm.

Input: Region of interest R, maximum number of API requests k
Output: Set of users from region R.
$\mathcal{U} = \{\text{creator}(T): T \in \mathcal{T}\}$
$seed \leftarrow \text{filter}(\mathcal{U})$ // 3-step filtering from Section 4.1
$frontier \leftarrow \bigcup_{u \in seed} friends(u) \setminus seed$ //Use Twitter API for $friends(\cdot)$
while $k \geq 0$ **do**
  $winner = \arg\max_{u \in frontier} |(u, v): v \in seed|$
  if $location(winner) \in R$ **then**
    $seed \leftarrow seed \cup \{winner\}$
    $frontier \leftarrow \cup\, friends(winner)$ // Use Twitter API for $friends(\cdot)$
  end if
  $updateWeights()$ // Feedback
  $frontier \leftarrow frontier \setminus \{winner\}$
  $k \leftarrow k - 1$
end while
return $seed$

Connection with existing algorithms: Next, we briefly discuss the connection of the `MaxEdge` and `Feedback` algorithms with well-known approaches. We begin by considering similarities with *Topological Sorting*. Similar to our two approaches, topological sorting provides an order of the nodes in a given directed graph, based on the direction of its edges. It assumes knowledge of the entire graph and a fixed direction for all edges. Instead, our algorithms makes no such assumptions. First, only parts of the graph can become gradually accessible, due to the limitation on the number of queries. Second, the direction of the edges is a result of the order in which they are processed, which is the exact opposite of topological sorting. The algorithm however bares similarity to the PageRank algorithm [31], in the sense that it ranks first, the node with the highest number of incoming edges. This is desirable, however our method differs as it only counts edges from a subset of all the existing nodes.

Furthermore, the voting-based selection employed by `MaxEdge` is completely absent in Topological sorting or PageRank as is the weighted version employed by our `Feedback` approach. In fact, the weighted voting mechanism employed by `Feedback` is inspired by similar aggregation techniques typically used in Ensemble Learning [28], in which the weight of vote of each classifier is determined by its previous performance.

4.2. Problem 2: tweet location discovery

In this section we describe our methodology for the geo-location of tweets. We evaluate two different approaches: one assuming that we only have knowledge on the text of a tweet, and the other assuming that we also have information about the user who created it. For both methods, we initially segment the plane into square cells of equal side length a.

For the first approach, which we refer to as **QL**, we employ a variation of the Query Likelihood Model [22,19]. Specifically, we consider all the tweets that are known to come from the same cell C as a single document. We then assign a new tweet to the cell C^* that has the maximum likelihood of having produced that tweet. Formally, let θ_C be the language model build on all the tweets associated with cell C. Then, given a tweet T, **QL** assigns the tweet to

the cell C^* such that:

$$C^* = \arg\max_C \prod_{w \in T} P(w|\theta_C)^{tf(w,T)} \quad (1)$$

where $w \in T$ iterates over all the distinct terms included in tweet T and $tf(w,T)$ returns the frequency of term w in tweet T.

The second method, which we refer to as **QLU**, utilizes information about the user and her tweeting habits. Formally, Given a cell C and a user u, let N_u be the total number of tweets posted by u, and let $N_{u,C}$ be the number of tweets posted by u from C. Then, given a new tweet T posted by u, **QLU** assigns the tweet to the cell C^* such that:

$$C^* = \arg\max_C \frac{N_{u,C}}{N_u} \quad (2)$$

In the case of ties, or if information regarding user u is not available, then **QLU** falls back to the content-based **QL** method.

We observe that both methods will always assign a tweet to the cell that maximizes their respective objective functions, even if the score of the winning cell is low. This could happen in ambiguous cases where neither the tweet's content nor the mobility patterns of its creator provide sufficient information. Such ambiguous tweets could be erroneously assigned to random cells or, in the case of **QLU**, trivially assigned to high-activity cells, leading to a rich-get-richer effect. Such unwanted occurrences can be addressed by setting a lower bound on the value of the objective function: if the winning cell scores lower than the bound, then no assignment is made. This would be a preferable option in applications where unassigned tweets are preferable to false positives (e.g. applications relevant to security or emergency response). Finally, while our own goal is to verify the utility of knowing a tweet's creator for the task of assigning the tweet to a location, one could envision alternative algorithms that leverage the tweet's content and creator information in different ways. We provide a relevant discussion in the final section of the paper.

5. Experiments

In this section we present the experiments that we conducted toward the qualitative and quantitative evaluation of the proposed methodology. The three datasets described in Section 4.1 are used throughout are evaluation.

5.1. User geo-location

In this experiment we evaluate our methodology on the discovery of users from a given city of interest. First, we use the Dublin dataset as a basis for a theoretical analysis that verifies the difficulty of the user geo-location task and the need for an intelligent way to select users. We then move on to show the results achieved by different approached on all three datasets.

Out of the 1.1 million people living in Dublin, only 11% have a twitter account.[8] Therefore, even in the best case scenario, we will be able to retrieve 110,000 possible actual users from the city. Given the hard limit on the number of graph queries allowed by the twitter API, a brute force approach that randomly picks candidates from Twitter's set of 300 million users[9] would clearly be ineffective. Instead, let us consider if random picking would be effective given the seed of 43,000 location-verified users in the Dublin dataset. Given this initial seed, the resulting *frontier* (i.e. the set of users with at least one connection to the seed) has about 2 million users. Considering that the maximum number of retrievable Dublin users is 110,000, and we already know 43,000 of them, a random-picking algorithm would have a probability equal to 67,000/2,000,000 = 0.0335 of actually identifying a user from Dublin. In fact, this is an optimistic estimate that does not account for the fact that, for every newly identified Dublin user, the frontier will grow substantially, while the set of retrievable users will become smaller by exactly 1. Therefore, the already trivial probability of success will only become smaller. Instead, as we demonstrate in this section, our approach maintained a precision of 62.9% after 50,000 queries.

We illustrate our results for Boston, Dublin, and Manchester, for a budget of k queries, for $k \in \{5k, 10k, 15k, 20k, 25k, 30k, 35k, 40k, 45k, 50k\}$. We define C_s to be the seed of all the users who tweeted from city C and had an exact geo-location in their tweets. We test the performance of our `Max Edge` algorithm with three different strategies for selecting the seed:

- **Simple** refers to the algorithm described in Section 4.1, which takes all available users as the seed.
- **City Seed** populates the seed with the largest connected component of users who *declare* in their profiles that they live in C
- **TimeTweets**(n, t) populates the seed with the users who tweeted at least $n = 4$ times from the city, with a difference of $t = 100$ h between the first and last tweet.

We also evaluate the `Feedback` method, described in Section 4.1, which rewards the nodes that point to a correct user and penalizes those that pointed to an incorrect one. For this experiment, the reward and penalty is obtained by multiplying the current weight with a coefficient $1 \pm c$. For our experiments, we set $c = 0.01$. This `City Seed` strategy is used to select the seed for the `Feedback` method.

The considered approaches are evaluated based on their achieved precision: given the budget of queries k, each method identifies and recommends a set of candidates as "associated with the region of interest". The geostamp of each recommended user is then retrieved via the Twitter API, to verify the user's association with the region. The set of verified users is thus a subset of the

[8] http://www.ipsosmrbi.com/twitter-users-on-the-increase.html

[9] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

recommended set. The method's precision is then defined as:

$$precision = \frac{\text{Number of verified users}}{\text{Total Number of all recommended candidates}} \quad (3)$$

Fig. 5 shows our results for every city. The y-axis shows the number of discovered users from the city of interest (y-axis), after a certain number of queries (x-axis). All methods perform similarly for the first 10–15 thousand queries. Beyond this point, however, the three variations of the `MaxEdge` algorithm are clearly outperformed by the `Feedback` method, which finishes with the highest precision in all three cities: 62.9% for Dublin, 63.0% for Manchester ,and 65.6% for Boston. The weighted scheme employed by `Feedback` allows it to expand the seed with users that are indeed located in the city of interest. On the other hand, the fact that the other approaches do not consider the previous success of the node leads to the introduction of an increasing number of non-local users. The problem is exacerbated as the graph becomes larger and the number of users with non-local connections in the seed increases.

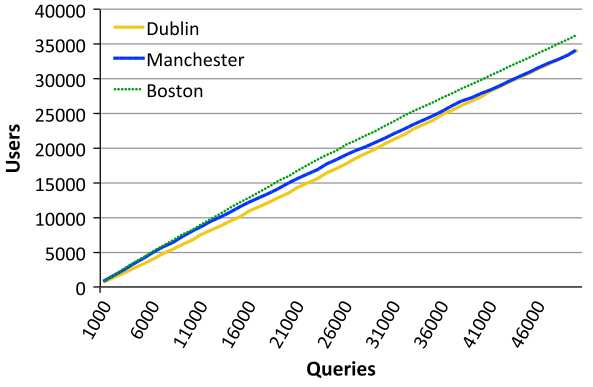

Fig. 6. Users in city C if prior knowledge vs number of queries.

5.1.1. User discovery with prior knowledge

This experiment aims to perform the task of user discovery in a city, assuming we have some prior knowledge about the search space. Specifically, we assume that all candidate users are attached to a region that includes the given city of interest, rather than to arbitrary locations. Fig. 6 illustrates the results of this experiment. The figure shows a large improvement for the `Simple` algorithm, but only a minor improvement for the `Feedback` approach. As the search space is considerably smaller, we expect this precision to increase, however, the fact that this model behaves marginally better than our `Feedback` algorithm, pinpoints to the success of the user voting scheme, as a method to counterbalance the loss of prior knowledge.

5.2. Tweet geo-location

In this section we evaluate the **QL** and **QLU** methods for the geo-location tweets. As described in Section 4.2, the first method focuses exclusively on the textual content of the tweet, while the second method also takes into consideration the geo-stamp attached to the user who creates the tweet. First, we segment the area of the city into square blocks with a fixed side length equal to a, for $a \in \{500\text{ m}, 1\text{ km}, 2\text{ km}, 4\text{ km}\}$. We then report the aver-

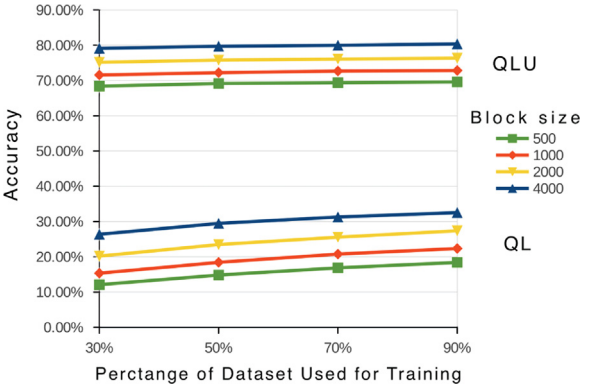

Fig. 8. Sensitivity of the **QLU** and **QL** methods to the size of the training set.

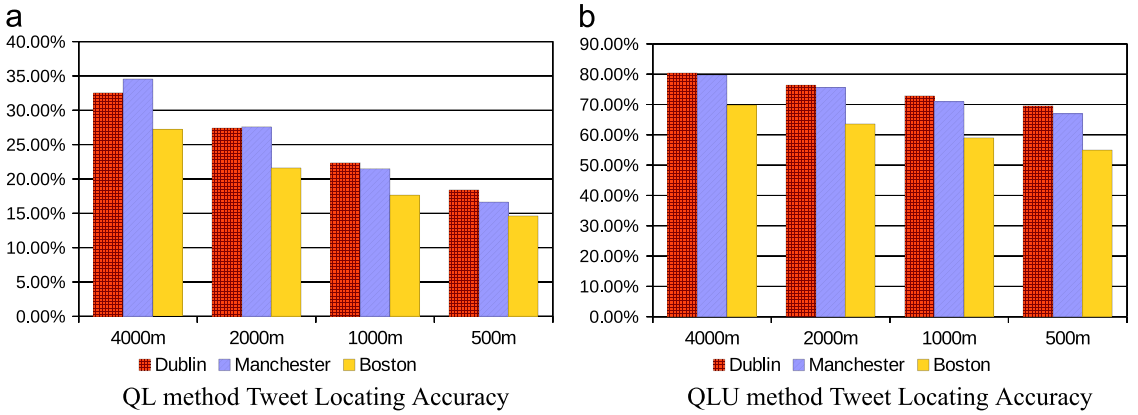

Fig. 7. Size of area – Boston: 39 × 39 km, Manchester: 21 × 22 km, Dublin: 34 × 36 km. (a) QL method Tweet locating accuracy. (b) QLU method Tweet locating accuracy.

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

age precision achieved by each method over a 10-fold cross validation. The results for each city are shown in Fig. 7(a).

As can be seen from the figure, the **QL** is clearly outperformed by **QLU** across cities and block sizes. This experiment verifies the intuition behind **QLU** and demonstrates the utility of considering the geostamp of the tweet's creator. The experiment also implies the consistent attachment of users to specific locations. In a different setting, with highly mobile users typically attached to a multitude of locations, the geo-stamp chosen to represent the tweet's creator would be less a credible source of information for our effort to geo-tag the tweet. Instead, the clear and persistent advantage of the QLU approach verifies the consistency of the geo-stamp associated with Twitter users.

Fig. 8 shows the sensitivity of the **QLU** and **QL** methods to the set of the training set. The upper lines represent the accuracy of the **QLU** method while the bottom ones represent that of **QL**. The figure demonstrates the robustness of the **QLU** method. The **QL** method is shown to be more sensitive to the size of the training corpus, with a clear upward trend observed in the achieved accuracy, as the size of the corpus increases.

5.3. Evaluation through topic models

Given the set of tweets retrieved for a given city, we can use topic modeling to capture the topics that are being discussed locally. This approach has numerous applications, such as identifying emergencies, events, and capturing urban trends. For this experiment, we use the established Latent Dirichlet Allocation (LDA) algorithm to independently compute topics on two different sets of tweets: (i) the verified, geo-located tweets that we collected for our experiments (see Table 1), and (ii) the extended set of tweets that includes both the tweets from first set and the new tweets retrieved by our `Feedback` algorithm. Next, we discuss the two different methods that we used to compare the two corresponding sets of computed topics. The Dublin dataset was used for this experiment.

First, we use the standard perplexity measure [36], which evaluates the ability of the learned model to accurately represent the given data sample. For both sets of tweets, we compute the perplexity values for a varying number of topics (which is required as input by the LDA algorithm). We also repeat the computation for different percentages of each set, in order to further explore the sensitivity of the results to the size of the provided corpus. Specifically, we compute perplexity for K topics, for $K \in 30, 40, 50, 60$ and for a percentage p of the entire set of tweets, for $p \in \{0.2, 0.4, 0.6, 0.8, 1\}$. We also normalize the reported values with the lowest perplexity achieved over both sets of tweets, and over all assignments of p and K, since the raw value of the perplexity is meaningless. The results for the original and extended sets of Dublin tweets are shown in Fig. 9(a) and (b), respectively. The figures verify the advantage of having the additional data retrieved by our approach: as more data becomes available, the observed perplexity consistently decreases. This trend, which is consistent for all values of the parameter K, is an encouraging finding that supports the utility of our approach for mainstream applications, such as topic modeling.

Motivated by the encouraging results of the perplexity study, we perform a qualitative test, for which we manually identify LDA-produced topics that can be associated with a coherent real-life topic of discussion. For this second study, we report the results for $K=60$ topics (the number for which the lowest perplexity values were reported for both the original and extended set of tweets) on a p% of each dataset, for $p \in \{0.2, 1\}$ (the smallest and largest values from the perplexity study). The results for other value assignments were similar and are omitted for

Table 2
Number of Coherent Topics (out of 60).

Dataset Used	Coherent topics
20% of the original Dublin Data	32
100% of the original Dublin Data	37
20% of the extended Dublin Data	42
100% of the extended Dublin Data	45

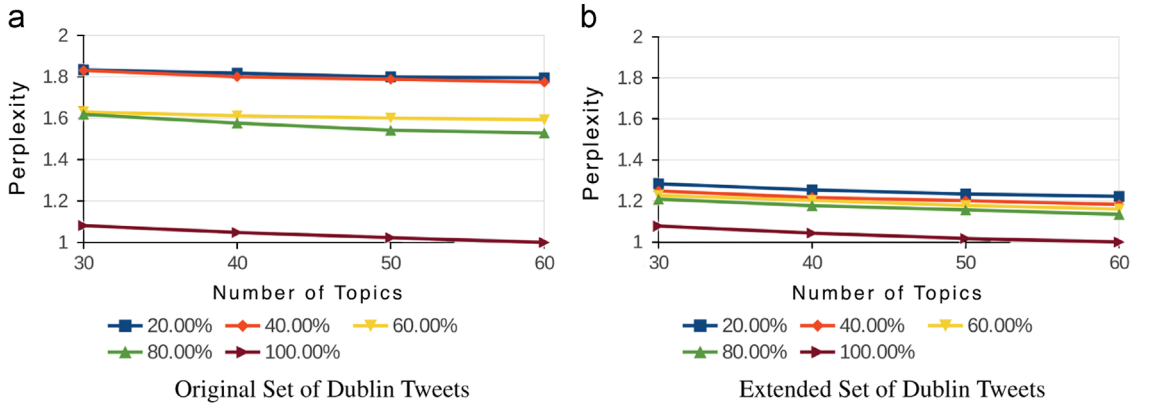

Fig. 9. Perplexity vs the number of topics, for various percentages of each set (lower values are preferable). (a) Original set of Dublin Tweets. (b) Extended set of Dublin Tweets.

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

Table 3
Annotated topics produced by LDA on both the Original (ORG) and extended (EXT) set of tweets.

20% ORG	Abortion	Ireland women abortion vote france vinb law issue politics country bill pro europe reading point government tax people bank
20% ORG	Tourism	Dublin ireland bar park photo guin dub street post tha cliath howth aerfort st bhaile square hotel temple green
20% ORG	Football	United mufc great castle de enjoy day pal weekend love mate massive yea matchday stanakatic caskett hun
20% ORG	Rugby	Lion leinster rugby ihl fitness lions coybib gatland trx bod fitfam pm ym seaofre final match anglotape fella heinekencup
20% ORG	*Incoherent*	Donegal christina tom bloomsday roscommon fe casa kenny patrick theapprentice hill floor princess rackard ding alley min
100% ORG	Places	Dublin ireland bar guin dub tha cliath aerfort bhaile pic hotel airport temple storehouse pub st pint trinity college
100% ORG	Football: UK	Player united season game play arsenal football sign team haha league fan goal win suarez man mate mufc chelsea
100% ORG	Music show	Check play rt music gig album video film listen show eventsindublin live song awesome watch band festival game cat
100% ORG	Rugby	Lion game great win match play final gaa leinster team rugby golf ireland today wimbledon player murray congrats dub
100% ORG	*Incoherent*	Watch man life show call god car made hit kid years laugh mr tom men face jesus dream break
20% EXT	Irish jobs	Business ireland job dublin social jobfairy marketing digital media hire great tip irishjob startup network tech sales online company
20% EXT	Politics: Economics	Ireland bank tax eu people europe pay report government uk protest minister year state home court news service job
20% EXT	Food	Food restaurant wine lunch taste special beer coffee free cocktail dinner menu delicious chef recipe bar eat lovely yum
20% EXT	Abortion	Abortion vinb life women ireland bill vote seanad pro people baby politics law party dail fg debate prolife labour
20% EXT	*Incoherent*	ur ya ye im goin yea dont wat tho pal ha gettin il wit ill ah nite jus bout
100% EXT	Politics: Economics	Ireland bank tax news eu britain report uk government police protest europe belfast human gold syria attack year minister
100% EXT	Research	Ireland great today health support student research eu event conference school day people children education week work launch europe
100% EXT	Abortion	Abortion ireland vinb women vote life people bill seanad law party support politics pro dail debate woman gay marriage
100% EXT	Irish jobs	Job dublin ireland business jobfairy irishbizparty manager hire sales recruit cork company bizhub engineer developer senior service client
100% EXT	*Incoherent*	ur ya ye ha goin haha wat tho gettin il wit im dat de jus day yea nite

lack of space. The results, which are shown in Table 2, indicate that the extended set of tweets retrieved by our approach delivered a larger number of cohesive topics, for all considered values of p. Table 3 provides additional qualitative evidence for multiple annotated topics produced by applying LDA on the original (coded ORG) and extended (EXT) sets of tweets.

6. Conclusions and future work

In this work we motivated and addressed the problem of retrieving tweets and users attached to a specific location of interest. Our work on user location verifies a correlation between strong connectivity in the social graph and proximity in the real world. In addition, we design our method to respect a maximum number of information requests that can be issued to the Twitter API. This widens the applicability of our approach by making it available to the average practitioner, who only has access to the free but rate-limited version of the API. The second part of our work delivers a method for the accurate geo-location of tweets within a city. Our full methodology was evaluated via both quantitative and qualitative experiments that demonstrate its effectiveness on real Twitter data.

Our work reveals a number of relevant problems that we intend to explore. First, the demonstrated utility of using the social graph to address the sparsity of location information on Twitter, as well the superiority of the `Feedback` algorithm, motivate us to consider an influence-based scheme in which every edge on the social graph is weighted by the strength of the connection between the two users. This would allow us to verify if users that interact more frequently or indeed more likely to be geographically closer. Another potentially valuable source of information for the task of retrieving users from a given region, that we did not consider in our work, is the nature and content of the tweets exchanged between users. Specifically, we intend to test the hypothesis that the tweets exchanged between users from the same geographical region have a significantly different structure or topic composition than tweets exchanged between users who are friends on Twitter but live in different regions. Finally, the encouraging results of the **QLU** method for the geo-location of tweets motivate us to consider methods that utilize different textual features mined from the tweet's content, as well as graph-based features associated with the tweet's creator.

Acknowledgments

This work has been co-financed by EU and Greek National funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Programs: THALIS - GeomComp, THALIS - DISFER, ARISTEIA - MMD” and the EU FP7 funded project INSIGHT (www.insight-ict.eu). This material is based upon work partially supported by the Office of Naval Research under MURI grant N00014-08-1-1015.

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011

The authors would like to thank Myrto Vlazaki and Ioannis Katakis for their valuable insights and recommendations.

References

[1] A. Agarwal, K. Hosanagar, M.D. Smith, Location, location, location: an analysis of profitability of position in online advertising markets, J. Mark. Res. (JMR) 48 (6) (2011) 1057–1073.
[2] A. Ahmed, L. Hong, A.J. Smola, Hierarchical geographical modeling of user locations from social media posts, in: Proceedings of the 22nd International Conference on World Wide Web. WWW '13, 2013, pp. 25–36.
[3] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, Four degrees of separation, in: WebSci, 2012, pp. 33–42.
[4] L. Backstrom, E. Sun, C. Marlow, Find me if you can: improving geographical prediction with social and spatial proximity, in: Proceedings of the 19th International Conference on World wide web, WWW '10, 2010, pp. 61–70.
[5] H.S. Burton, W.K. Tanner, G.C. Giraud-Carrier, H.J. West, D.M. Barnes, Right time, right place health communication on twitter: value and accuracy of location information, J. Med. Internet Res. 14 (November (6)) (2012) 156.
[6] Z. Cheng, J. Caverlee, K. 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, CIKM '10, 2010, pp. 759–768.
[7] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, 2011, pp. 1082–1090.
[8] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society, 1997.
[9] J. Cranshaw, R. Schwartz, J.I. Hong, N.M. Sadeh, The livehoods project: utilizing social media to understand the dynamics of a city, in: J.G. Breslin, N.B. Ellison, J.G. Shanahan, Z. Tufekci (Eds.), ICWSM, The AAAI Press, Dublin, Ireland, 2012.
[10] J. Eisenstein, B. O'Connor, N.A. Smith, E.P. Xing, A latent variable model for geographic lexical variation, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, 2010, pp. 1277–1287.
[11] G. Fuchs, G. Andrienko, N. Andrienko, P. Jankowski, Extracting personal behavioral patterns from geo-referenced tweets, 2013.
[12] E. Galbrun, K. Pelechrinis, E. Terzi, Safe navigation in urban environments, 2014.
[13] A. Gionis, T. Lappas, K. Pelechrinis, E. Terzi, Customized tour recommendations in urban areas, in: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14, ACM, New York, NY, USA, 2014, pp. 313–322.
[14] B. Hecht, L. Hong, B. Suh, E.H. Chi, Tweets from justin bieber's heart: the dynamics of the location field in user profiles, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, 2011, pp. 237–246.
[15] R. Holt, Twitter in numbers, March 2013, ⟨http://www.telegraph.co.uk URL http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html⟩.
[16] L. Hong, A. Ahmed, S. Gurumurthy, A.J. Smola, K. Tsioutsiouliklis, Discovering geographical topics in the twitter stream, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12, ACM, New York, NY, USA, 2012, pp. 769–778.
[17] Y. Ikawa, M. Enoki, M. Tatsubori, Location inference using microblog messages, in: Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, 2012, pp. 687–690.
[18] R. Kannan, S. Vempala, A. Vetta, On clusterings: good, bad and spectral, J. ACM 51 (May (3)) (2004) 497–515. URL ⟨http://doi.acm.org/10.1145/990308.990313⟩.
[19] S. Kinsella, V. Murdock, N. O'Hare, "I'm eating a sandwich in Glasgow": modeling locations with tweets, in: Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, SMUC '11, 2011, pp. 61–68.
[20] D. Kotzias, T. Lappas, D. Gunopulos, Addressing the sparsity of location information on twitter, in: EDBT/ICDT Workshops, 2014, pp. 339–346.
[21] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media? in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 591–600.
[22] J. Lafferty, C. Zhai, Document language models, query models, and risk minimization for information retrieval, in: Proceedings of the 24th Annual International ACM SIGIR, SIGIR '01, 2001, pp. 111–119.
[23] T. Lappas, M.R. Vieira, D. Gunopulos, V.J. Tsotras, On the spatio-temporal burstiness of terms, Proc. VLDB Endow. 5 (May (9)) (2012) 836–847.
[24] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Statistical properties of community structure in large social and information networks, in: WWW, 2008, pp. 695–704.
[25] R. Li, S. Wang, H. Deng, R. Wang, K.C.-C. Chang, Towards social user profiling: unified and discriminative influence model for inferring home locations, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, 2012, pp. 1023–1031.
[26] W. Li, P. Serdyukov, A. de Vries, C. Eickhoff, M., Larson, The where in the tweet, in: Proceedings of the 20th ACM International Conference on Information and knowledge management, CIKM '11, 2011, pp. 2473–2476.
[27] M. Lichman, P. Smyth, Modeling human location data with mixtures of kernel densities, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, ACM, New York, NY, USA, 2014, pp. 35–44. URL ⟨http://doi.acm.org/10.1145/2623330.2623681⟩.
[28] R. Maclin, D. Opitz, Popular ensemble methods: an empirical study, J. Artif. Intell. Res., 1999.
[29] J. Mahmud, J. Nichols, C. Drews, Where is this tweet from? Inferring home locations of twitter users, in: J.G. Breslin, N.B. Ellison, J. G. Shanahan, Z. Tufekci (Eds.), ICWSM, The AAAI Press, Dublin, Ireland, 2012.
[30] S. Murray, Twitter breaks news of whitney houston death 27 minutes before press, February 2012, ⟨http://mashable.com. URL http://mashable.com/2012/02/12/whitney-houston-twitter/⟩.
[31] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web, Technical Report 1999-66, Stanford InfoLab, previous number=SIDL-WP-1999-0120. November 1999. URL ⟨http://ilpubs.stanford.edu:8090/422/⟩.
[32] K. Ren, S. Zhang, H. Lin, Where are you settling down: geo-locating twitter users based on tweets and social networks, in: Information Retrieval Technology, Lecture Notes in Computer Science, vol. 7675, 2012, pp. 150–161.
[33] A. Sadilek, H. Kautz, V. Silenzio, Modeling spread of disease from social interactions, in: Sixth AAAI International Conference on Weblogs and Social Media (ICWSM), 2012.
[34] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, 2010, pp. 851–860.
[35] G. Valkanas, D. Gunopulos, Location extraction from social networks with commodity software and online data, in: ICDM Workshops, 2012, pp. 827–834.
[36] H.M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, 2009, pp. 1105–1112.
[37] S. Yardi, D. Boyd, Tweeting from the town square: measuring geographic local networks, in: ICWSM, 2010.
[38] Yves van Gennip, Blake Hunter, Raymond Ahn, Peter Elliott, Kyle Luh, Megan Halvorson, Shannon Reid, Matthew Valasik, James Wo, George E. Tita, Andrea L. Bertozzi, P. Jeffrey Brantingham, Community detection using spectral clustering on sparse geosocial data, SIAM J. Appl. Math. 73(1) (2013) 67–83.

Please cite this article as: D. Kotzias, et al., Home is where your friends are: Utilizing the social graph to locate twitter users in a city, Information Systems (2015), http://dx.doi.org/10.1016/j.is.2015.10.011