Where's @wally? A Classification Approach to Geolocating Users Based on their Social Ties

Dominic Rout*
Dept. of Computer Science
University of Sheffield
d.rout@shef.ac.uk

Daniel Preoţiuc-Pietro*
Dept. of Computer Science
University of Sheffield
daniel@dcs.shef.ac.uk

Kalina Bontcheva
Dept. of Computer Science
University of Sheffield
k.bontcheva@dcs.shef.ac.uk

Trevor Cohn
Dept. of Computer Science
University of Sheffield
tcohn@dcs.shef.ac.uk

* These authors have contributed equally

ABSTRACT

This paper presents an approach to geolocating users of online social networks, based solely on their 'friendship' connections. We observe that users interact more regularly with those closer to themselves and hypothesise that, in many cases, a person's social network is sufficient to reveal their location.

The geolocation problem is formulated as a classification task, where the most likely city for a user without an explicit location is chosen amongst the known locations of their social ties. Our method uses an SVM classifier and a number of features that reflect different aspects and characteristics of Twitter user networks.

The SVM classifier is trained and evaluated on a dataset of Twitter users with known locations. Our method outperforms a state-of-the-art method for geolocating users based on their social ties.

Categories and Subject Descriptors

[**Information systems**]: [World Wide Web, Web applications, Social networks]

Keywords

Social networks, Geolocation, Twitter, classification, Support Vector Machines

1. INTRODUCTION

Proponents of the internet, global connectedness and ubiquitous social networking might argue that these technologies render geographical distance irrelevant - it is just as easy, one can reasonably claim, to maintain a relationship with someone a ten hour flight away as with someone a ten minute walk away. While relationships between people undoubtedly do form spontaneously on the Internet, most online friendships follow real world interactions. Users might meet at school, a party or a social event and connect online afterwards.

Geography is still a major factor in the shaping of online social networks. In this study we explore the effect of geography on the Twitter social network. While our analysis and method are generalizable to any network, we wish to explore and highlight some peculiarities of this particular platform and gain more insight into these Internet relationships. We propose a method of automatically discovering home locations of social network users, with potential applications in the domains of advertising, privacy, political opinion analysis and content targeting.

Twitter allows users to share their location in two ways: attaching geographical coordinates to tweets and/or setting a home town in their public profile. The former requires deliberate effort and is consequently hardly used. In contrast, profile-level location sharing is far more common. Even though it is optional, as many as 26% of Twitter users enter correct information in this field and supply an actual city or town [2]. We make use of this profile data and acquire a corpus of over 200,000 Twitter users with known locations.

In this paper, we consider the problem of geolocating users on the Internet. Geolocation here means the process of assigning to each user a 'home location' or the place where they spends most of their time. Home locations will be described as a pair of geo-coordinates and as the name of a city or region. While there are diverse methods for geolocating users automatically (see Section 2), our focus is on using only information about social relationships (e.g. followers and friends) for this task. Our experiments confirm that knowledge of a small part of the social network is sufficient for obtaining potentially useful results.

The contributions of this paper are:

- In contrast to previous work, we formulate geolocation as a classification task, where each user is assigned a class, corresponding to their true location at the finest possible level of granularity (e.g. a city, region or a country).
- We examine how geography influences relationships in a social network and whether it is possible to infer a user's location, based on their social ties with good

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France
Copyright 2013 ACM 978-1-4503-1967-6/13/05 ...$15.00

accuracy. The underlying assumption is that Twitter users are more likely to follow others that are geographically closer to them.

- We show how variations in population density across regions and cities can be taken into account, in order to improve geolocation accuracy.
- We investigate how best to account for the noisy nature of online networks, where social ties do not necessarily reflect real world connections and friendships do not need be reciprocated. As a result, two other features indicative of real-world friendships (reciprocation and triangle ties) are integrated in our method.
- We conduct quantitative evaluation experiments on a dataset of around 200,000 users, comparing our method against several baselines and previous work. The overall accuracy is close to that of IP-location, without the associated costs [12] or the need to have the user's IP address.

2. RELATED WORK

The most popular existing method of geolocating users of online services is that of IP tracking, where a user can be located within the U.K. with an accuracy of 72% within 25 miles based on their IP address[1]. Unfortunately, most of the databases containing this information are not freely available to end-users and can also get out-of-date without frequent maintenance, as this information evolves over time. Secondly, in many applications information such as the IP address of a user is not available at all. Doubts have also been cast upon the accuracy of these methods[12].

The problems with IP-geolocation have given impetus to research on automatic methods for geolocating users of social networks based on publicly available data about a user, such as their profile, posts and social network. Broadly speaking, these methods fall into two different categories: content-based (i.e. using the textual posts of a given user) and network-based (i.e. using the social network).

Content-based methods ('you are where you write about') typically gather the textual content produced by the user and infer their location based on features, such as mentions of local place names [5] and use of local dialect. In the work of [4, 2], region-specific terms and language that might be relevant to the geolocation of users were discovered automatically. A classification approach is devised in [9] that also incorporates specific mentions of places near to the user. One disadvantage to this method is the fact that someone might be writing about a popular global event which is of no relevance to his actual location. Another is that users might take deliberate steps to hide their true location by altering the style of their posts or not referencing local landmarks.

In contrast, network-based geolocation methods ('you are where your friends are') aim to use the user's social network to infer their location. To the best of our knowledge, the only existing method of this kind (i.e., relying on the user's social network alone) is the work of [1], who first create a model for the distribution of distance between pairs of friends, before using this to find the most likely location for a given user. The influence of distance in social network ties is demonstrated by the earlier work of [8]. The main disadvantage of their approach is that it assumes that globally all users have the same distribution of friends in terms of distance. Also, they do not account explicitly for the density of people in an area, which is a key factor in our approach.

Our method for collecting a large dataset of users with known, ground truth locations involves parsing the user defined 'location' field in their Twitter profile. A thorough analysis of how users use this field is presented in [6]. Other work has relied instead on small amounts of geotagged data (e.g Foursquare checkins) as an extra feature for user location or as the only way of locating users. One method has used checkins as part of location discovery[9], leading to high accuracy results but somewhat limited by the very small amount of geo-tagged data available. It is possible that users are reluctant to geo-tag data due to practical and privacy constraints, such as battery drain on mobile devices.

3. DATASET

The data used in this study comes from users of Twitter within the United Kingdom. While other studies have concentrated on the U.S. [1], we chose the U.K. for a number of reasons. Firstly, it has fewer ambiguous place names, which makes it easy to derive a high-quality gold standard corpus by parsing the location field of the Twitter user profiles. The U.K. also encompasses only one timezone and is much smaller than the U.S., reducing the potential effects of regionalisation. Cities in the U.K. are generally closer together than their U.S. counterparts, with fewer regional hubs. We hypothesise that this makes user location classification harder for U.K. users. A distribution of population density in the U.K. on Twitter according to our data can be seen in Figure 1.

We conducted these experiments on Twitter, because it is one of the most popular social networks, with around 140 million active users[2] and a publicly accessible API. Additionally, Twitter is an interest-graph network[3], which encourages users to form connections with others based on shared interests, regardless of whether they know the other person in real life, i.e. online relationships are not always dependent on geography. Twitter allows for unidirectional social connections, where a user can be friends with another user without the need for reciprocation. These types of relations create ties that introduce noise in the social network graph and make network-based geolocation of Twitter users particularly challenging. A deeper analysis of Twitter and its social network structure is presented in [7].

3.1 Collection

In order to train and evaluate our method, we sampled a subset of the Twitter social graph concentrated on a specific country. The aim was to collect a number of users that could be located with a high degree of confidence. The Gardenhose Twitter stream (~10% of the entire Twitter stream) was monitored during the month of November 2011, in order to gather a list of Twitter users that had posted publicly. For this study, the list was restricted automatically to users located in the U.K.. We discarded users who were not in Greenwich Mean Time (GMT) timezone and did not post in English, as determined automatically by using the system introduced in [14].

We did not discard any user on the basis of amount of

[1] http://www.maxmind.com/app/city_accuracy

[2] http://blog.twitter.com/2012/03/twitter-turns-six.html

[3] http://techcrunch.com/2010/10/16/why-twitter-is-massively-undervalued-compared-to-facebook/

Number of users	206,200
Mean in/out degree	48.75
Median indegree	11
Median outdegree	27
Reciprocity	37%

Table 1: Dataset statistics

activity or number of friends, although we needed to observe at least one tweet in a month to be aware of the user, thus eliminating completely inactive users. With around 600,000 users remaining after the elimination of time zones other than GMT and non-English posts, we used the Twitter API to download each user's public profile. A high precision method described in Section 3.2 was then applied to resolve users to actual locations in the U.K.

After discarding users that were not placed U.K. by our method, 206,200 twitterers remained. Some statistics of the dataset are presented in Table 1. The median count of incoming connections or followers (indegree) was much lower than the mean, perhaps due to outliers such as celebrity accounts which act as hubs and have very high indegree. The existence of celebrity accounts has an impact on location resolution, as a connection to one is perhaps unlikely to be conditioned on geographical closeness. Reciprocity in the graph is relatively low at 37%, meaning that many Twitter relationships are uni-directional and those that are followed do not always follow back.

Figure 3 shows the distribution of twitter 'friendships' with their geographical distance for users within the U.K., demonstrating that the number of friendships clearly diminishes as distance increases. This relationship between friendship and distance roughly follows a power law distribution of the form x^b with $b = -0.59$.

Since all user locations in our set are known reliably, testing sets can be generated easily by removing the user-location assignments where necessary. We were also able to control the amount of data made available for supervised training of classifiers and to selectively redact locations (e.g. large cities), in order to perform error analysis.

The anonymised dataset of users and their social ties, together with the compiled list of locations in the U.K. is freely available for download[4].

3.2 Location resolution

In order to train and evaluate our method we had to gather ground-truth locations for real users. We focus on obtaining a mapping between a user and a city within the U.K. corresponding to their 'home' location. Since this dataset is used for evaluation, it is important that this mapping be as precise as possible.

Twitter users can supply as part of their profile a short, non structured string of text representing their location. In practice, not all users provide a valid string for this field. We parse these locations, identify the place names mentioned, and disambiguate them to an actual inhabited place (e.g. city, village), region, and country. This is a non-trivial problem, because place names, even when restricted to a given country, are still ambiguous (e.g. Richmond in London vs. Yorkshire, Cobham in Kent vs. Surrey).

Recently, Linked Open Data (LOD) resources, in particular DBpedia[5] and Geonames[6], have emerged as large-scale resources, where knowledge about millions of locations is encoded using Unique Reference Identifiers (URIs). For these experiments, we have adopted DBPedia as the dataset providing the URIs of inhabited places in the U.K. The problem of disambiguating the Twitter user's location field is then cast as the problem of assigning the URI of the corresponding DBpedia entry (e.g. 'London' is assigned 'http://dbpedia.org/page/London').

There are a number of general purpose, DBpedia-based entity disambiguation algorithms (e.g. DBpedia Spotlight [11] and Zemanta[7]), however, they were not best suited to disambiguating the user location field, due to their dependence on wider linguistic context. For instance, an evaluation of DBpedia Spotlight [11] on a tweet dataset has shown significantly poorer performance [10].

Rather than rely on an existing entity disambiguation approach, we developed a domain-specific, high precision method, using regular expressions to look up the content of the location field (ignoring region, county and country names and casing) against a gazetteer list of U.K. place names and their corresponding URIs, extracted from DBpedia. Where there are several locations with the same name and in the absence of other context, the city with the largest population (according to DBpedia) is selected.

Since our gazetteer was compiled from DBPedia, the matched locations are also assigned additional metadata encoded in DBPedia, via the assigned URI, e.g. geographical coordinates (latitude and longitude) and local region (London, South West, South East etc). The resulting plot of the distribution of U.K. Twitter users in our sample is shown in Figure 1, which is created using the automatically identified coordinates.

The accuracy of this location resolution method was evaluated on a random sample of 1,000 users. Using two independent human annotators, we have found that we are correctly assigning an average of 97.2% users. The method, as currently implemented, is conservative and focused on precision. Future improvements will address the annotation of informally described locations and non-canonical names (Thorn for Thorngumbald, Laandan for London etc), which are currently not annotated. These can be handled by enhancing the place name lexicalisations encoded in DBPedia with name variants collected from Wikipedia redirect pages (contain synonyms and abbreviations), disambiguation pages (for multiple entities with the same name), and anchor texts used when linking to the place name's Wikipedia page (which has a direct, 1-to-1 mapping to a DBPedia URI).

In total, the DBPedia-based gazetteer list comprises of 17,521 different settlements, from which 4,295 locations were mentioned at least once in our dataset of U.K. Twitter users. The ranked distribution of the number of users per city in our dataset is shown in Figure 2. Population is extremely skewed, with the majority of Twitter users located within a small number of cities. The most populous five cities in our dataset are shown in Table 2. It should be noted that the method described here does not rely on geotagged tweets, but instead locations are considered at the settle-

[4] http://dcs.shef.ac.uk/~dominic/

[5] http://wiki.dbpedia.org/Datasets

[6] http://www.geonames.org/about.html

[7] http://www.zemanta.com/

City	Users
London	57054
Manchester	10699
Liverpool	5684
Birmingham	4870
Glasgow	4650

Table 2: Most represented cities in our dataset

ment level. A preliminary investigation showed that a minority of tweets actually contain GPS coordinates, whereas users with a specified location in their profile are much more common. This distinguishes our approach from related methods which focus only on geotagged tweets, specifically containing GPS tags [3, 13, 4].

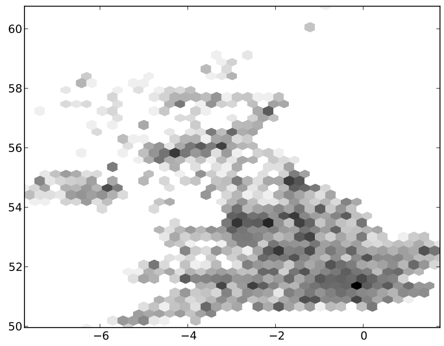

Figure 1: Distribution of the 'Twitter population' in our U.K. sample

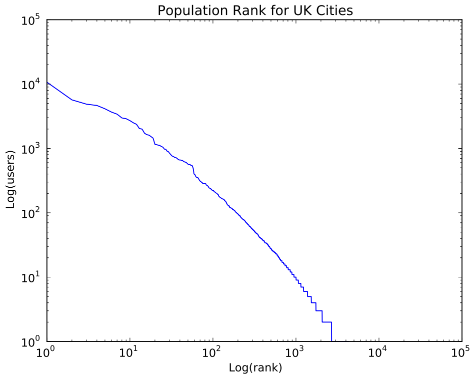

Figure 2: Distribution of the 'Twitter population' in cities in the U.K.

4. METHODOLOGY

Given a set of users and the social graph which binds and connects them, the task that we address in this paper is to accurately locate a subset of these users, for whom the true location is not observed. This reflects the situation that exists on several online social networks, including Twitter and Facebook, in which some, but not all, users can be traced to their real location by some deterministic method (e.g. the one presented in Section 3.2).

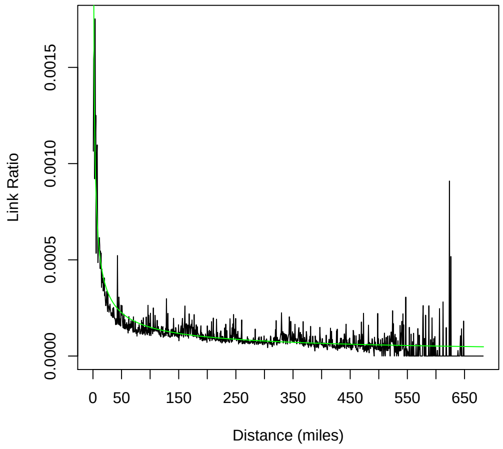

Figure 3: Distribution of distance ratio in social ties

Central to our approach is the reformulation of the problem as a classification task, coupled with a number of lightweight features, derived from the social network alone. These features are easy to gather, and could be used to produce results quickly in an 'online' setting. They attempt to model different effects, like the influence of more densely populated areas or closeness on the social graph.

One limitation introduced by the decision to infer a user's location from social ties alone is that a sudden change of address might be very slow to materialize in their social network. For example, a student that moves from their hometown to a university in another city will need time to gain friends there in order to counterweight the ties they have from their home. This problem could be tackled by also looking at the chronology of the friendship links, but these are generally hard to gather without monitoring the social network over time.

The granularity of location we use is the settlement of origin for any user, which could be a city, town, village or hamlet. This granularity was chosen, since it is sufficiently fine-grained in terms of geo-coordinates, while also being discrete (i.e. the cities can be numbered) and thus amenable to traditional classification. Moreover, many of the real-world applications considered for the geolocation problem are targeting location at a city level. Thus, given a user and the entire social network, the task is to select the city or town where they are based, or the one closest to their actual location.

It should also be noted that the same method could be applied at different granularities. Thus, we could use our algorithm where cities are replaced with other meaningful geographical units (e.g. counties, regions or countries) and the task would be to locate a user into one of these units.

As discussed in Section 3.1, this method is evaluated on a large set of Twitter users. Relationships on Twitter can form for a number of reasons, for example, bona fide friendship, usually appearing as mutual relationships; following of celebrities, which are usually unidirectional; and following of businesses, which may or may not be reciprocated. These different kinds of relationships carry differing probabilities that the users involved will be collocated, introducing noise to the classification task. In addition, our method also attempts to model the size of regions/cities and their population and how this affects the locations of a user's online friends.

4.1 Task Definition

As discussed above, we formulate the user geolocation task as follows: given a user and their entire social network, the algorithm needs to assign to the user their true location or one closest to this, at a chosen geographical unit of granularity (typically a city). The entire social network is restricted to the subset that contains only users with known locations, discarding the rest for feature selection. A method that also considers users with unknown location in a global inference algorithm is the subject of future work.

Classification tasks targeting a large number of possible classes (locations such as cities in our case) can make classification difficult. Therefore, we constrain further the location classification problem as follows.

Given a list of the user's friends (Twitter 'friends' are considered outgoing edges in the social graph), a restricted list of candidate locations is assembled. For the user u, let E_u be the set of all of their immediate neighbours, defined as outgoing edges on their social graph and let L_u be the known location for user u. The set of candidate locations C_u is given by:

$$C_u = \{L_f : f \in E_u\} \tag{1}$$

If we let $d(a, b)$ indicate the Haversine distance (accounting for the curvature of the Earth) between locations a and b, the objective, then, is to choose the closest candidate location to u, L'_u such that:

$$L'_u = \underset{c \in C_u}{\operatorname{argmin}} \left(d(L_u, c) \right) \tag{2}$$

That is, the closest location c in the candidate set C_u to the user's true location L_u from the gold-standard.

4.2 Classification

The simplest feature we employ is the number of friends a user has in a candidate location c. That is:

$$|x \in E_u : L_x = c| \tag{3}$$

Our initial hypothesis is that users with more friends in a given place are more likely to reside in that location. A setup in which this hypothesis is evaluated alone is described in Section 4.4.

However, given the noise present in online social networks additional features need to be integrated for better accuracy. In order to select from the candidate locations C_u using arbitrary features, we define our method as follows:

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(p(L_x = c) \right) \tag{4}$$

Where $p(L_x = c)$ is some estimator of the likelihood that the user's location L_x is the candidate location c.

Candidate locations for each user are used to create training examples, with features initially for the number of friends in that location, number of friends not in that location. As a source of supervision, we use the binary value for whether the user in the example is actually collocated with the subset of friends in the candidate. When assigning a location to a user we take for them the location which yields the greatest certainty of a collocation, or positive judgement.

For our more complex feature sets, we learn estimators for the likelihood of collocation using an SVM (support vector machine) classifier with a Radial Basis Function kernel, because it can be used probabilistically and because some of the features such as number of triads are noisy and fairly non-linear. We use the SVM implementation provided by LIBSVM[8].

4.3 Features

For a given location, the basic features are the count friends that are in that location and the count of friends not in that location. The following more complex features have been proposed to model more complex aspects of the user's social graph.

4.3.1 City size

As a large settlement is more likely to appear in a person's candidate location set by coincidence (e.g. because celebrities live there or because friends move there), we hypothesise that friends from smaller cities give more information about a user's possible location than connections to people from larger ones.

This is analogous to the idea of the entropy - that is the quantity of information that is gained by observing a social tie in a densely populous area is smaller than the quantity of information given by a social tie in a sparsely populated area.

This entropy problem is similar to the one found in information retrieval, wherein certain terms appear in almost any context (auxiliaries, closed-class terms), regardless of content. These words are consequently less indicative. One approach to dealing with these low-information, very common terms is to multiply their frequency by Inverse Document Frequency - that is, the inverse of the number of documents d that a term t appears in, normalised by the size of the document set D.

$$IDF_t = log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{5}$$

Inspired by this solution, we implemented a ratio we call Inverse City Frequency (ICF), which allows the method to take into account, in our city selection, that some cities have a much higher population. As a proxy for population, we use the number of Twitter users located in that city within our dataset, and apply the following formula for a location l:

$$ICF_l = \frac{|U|}{|\{u \in U : L_u = l\}|} \tag{6}$$

We select the city with the highest count of friends, multiplied by ICF. That is:

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(|\{x \in E_u : L_x = c\}| \cdot ICF_c \right) \tag{7}$$

[8] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

In Figure 2 we present as a justification of this method a ranked list of settlements in the U.K. by population. The long tail shows that many locations are somewhat under-represented, though the distribution does not appear to be power law. Locations such as London remain well represented.

4.3.2 City population bins

As an alternative to the ICF metric, cities can be divided into groups by population. However, in practice this measure is over-simplistic, penalising some locations too strongly. We use Support Vector Machines (SVM) to combine features more effectively.

As a new feature, we create population 'bins' which are ranges of population size, into which each location can be placed. The bins were selected according to population rank, in such a way as to keep them roughly of an equal size. For our dataset, the bins used are $[1, 2, 4, 12, 57054]$. For example, a city with 50 Twitter users is placed in bin 5. The likelihood that a user is in a location can then be modelled both on the number of followers and the number of the bin in which the location is placed, using the SVM classifier, as described in Section 4.2.

Bins are used instead of population numbers directly due to the somewhat eccentric distribution of population size - they increase dramatically and unpredictably. Using bins instead allows our model to capture more nuance about the exact kinds of relationships over which we aggregate.

4.3.3 Triads

In many cases, users will have relationships that form triads with other users. That is, they will have friends in common. These form local networks which are likely to engender social closeness (because they will both belong to some community). The addition of this feature to our method is motivated by the hypothesis that relationships with a greater number of triads (more friends in common) will also indicate geographical closeness.

To define these features more precisely, let $\gamma_{a,b}$ be the set of common neighbours for users a and b. For each candidate location l and user u, given a set of buckets $B = [0, 5, 10, 20, 40]$ we generate features of the form:

$$|\{x \in E_u : L_x = l,\ b_{n-1} < \gamma_{u,x} \leq b_n\}| : b_{n-1}, b_n \in B \quad (8)$$

For each bucket size, this is the number of friends the user has in the location with which they share as many common neighbours as that bucket allows. These counts are normalised so that they sum to 1 by dividing each by the total number of friends.

We arranged these features in such a way as to reduce the dimensionality of the number of common neighbours for each friend in the candidate location as little as possible (less so than, for example, taking the maximum or the mean), whilst still creating an arrangement where a fixed number of features would be produced.

4.3.4 Reciprocated friendships

One peculiarity of the Twitter social network is that the friendship relationship is unidirectional. A user can choose who to be 'friends' without of the need for their approval (with a few exceptions) or the need of the relationship to be reciprocated.

We hypothesise that reciprocated connections on Twitter are more indicative of real-world friendships. Many users on Twitter are followers of different celebrities and institutional accounts (e.g. companies, news agencies) and use this service for information purposes instead of or in addition to social purposes, however, these celebrity accounts rarely follow back.

We therefore formulate a feature for each candidate location corresponding to the portion of the user's friends in that location that have reciprocated their friendship.

4.4 Baselines and Upper Bound

One limitation created by approaching user geolocation by restricting a priori the candidate locations is that a theoretical upper bound is placed on performance. If the list of candidate cities does not include a user's own (i.e. they do not live in the same city as any of their friends), any method that formulates the problem in this way cannot possibly select the correct one. In this case, the upper bound is shown by choosing the closest city to the distance. We will refer to this upper bound as the oracle performance for our class of methods in the remainder of the paper. Section 5.1 shows the actual upper bound figures on the experimental dataset.

Even though our assumption that at least one friend from the user's immediate graph is collocated with them may seem restrictive, this represents a trade-off which makes our classification task much more straightforward by limiting the number of candidate classes. The same assumption was made to restrict the search space in comparable work by [1]. As demonstrated by the evaluation results (see Section 5), a user will most often have at least one friend very close to their actual location.

In order to establish the relative performance of our method, a number of baselines were implemented for the task of assigning a user to a home location.

The first baseline is random choice, i.e. we calculate the accuracy when locations are selected completely at random from a user's candidate set. The rationale behind the random baseline is to demonstrate that our proposed features are informative.

The second baseline for user geolocation is to assign to each user the most common location from amongst all of their friends. That is, an estimate for a user's location L'_u is derived as follows:

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(|x \in E_u : L_x = c| \right) \quad (9)$$

4.5 Method Comparison

In addition to the two baselines, we have reimplemented the geolocation method proposed in [1], since this is the only other approach directly comparable to ours, as it uses only information from user's the social network to perform geolocation.

The evaluation of the method from [1] was performed on a dataset of U.S. users of the Facebook social network, which has properties very different to those of Twitter networks. Firstly, Facebook relationships are bidirectional and generally arise following real-world encounters. Additionally, U.S. geography is very different to that of the U.K., with more regional hubs and very big distances between them. In the U.K. however, the population is more evenly spread across the territory, making the possibility of confusion higher.

Taken together, these properties make geolocation of Twitter users in the U.K. a significantly harder task.

The method takes into account the friendship links of a user and their length and shows that the distribution of friendship distances roughly follows a power law. Then, [1] hypothesise that by considering this distribution it is possible to find the most probable home location by maximum likelihood.

As our Twitter dataset is very different, we have reimplemented the method and recomputed the power law fit to the distance distribution set. The power law estimate is a function of the form $x^{-0.59}$ and is shown in Figure 3. As can be seen, the power law fit is fairly poor on our Twitter dataset, most likely due to the unidirectional nature of the friend relationship on Twitter.

Because of this, and in order not to understate the performance of the method, we have performed a line search over the parameter space of the power law function in order to find the parameter that gives the best predictive performance under 50 miles. The power law is in this case $x^{-1.6}$. These methods are denoted in the experiments results reported below as Backstrom and Backstrom-Best, respectively.

5. EVALUATION RESULTS

Given that user geolocation is performed at city level, we hold the mappings between users and their city's geographical coordinates. This gives rise to two modes of evaluation; city level accuracy, which is equal to the percentage of users placed in the correct city, and the error distribution, showing the distribution of the distance between the 'home' city and the predicted one. While city level accuracy is easier to interpret, the error curves provide a more complete picture of the actual performance, while performance at k miles also allows us to compare our results to other approaches.

Consequently, this section reports both accuracy figures and error curves for our methods. Accuracy at increasing distances is shown in Table 3. Additionally, we use error curves to demonstrate this increasing accuracy visually.

Our training sets were chosen at random, with testing sets of 20,000 users held out. 10-fold cross-validation was performed; intervals reported are 99% confidence according to student's T-test. We do not report confidence for the two Backstrom methods as the power-law must be parameterised using a time-consuming and inexact empirical search.

As described in Section 4.2, we chose a SVM classifier to combine our features. This was after considering other learning algorithms and finding them to perform poorly or to learn very slowly. For example, Figure 4 shows that when linear regression is used for learning, performance does not increase as more training examples are supplied, beyond a certain point.

5.1 Upper bound

The oracle performance on the testing data is calculated by assigning to each user the candidate location closest to their own real location. It is impossible for our system to produce better results than this because it does not consider locations that are not linked at all to the target user. The upper limit on overall performance is shown in Figure 5.

This oracle performance is around 78%. In practice, it is possible that some of the users that should be impossible to locate could end up being placed in a city very near to their homes.

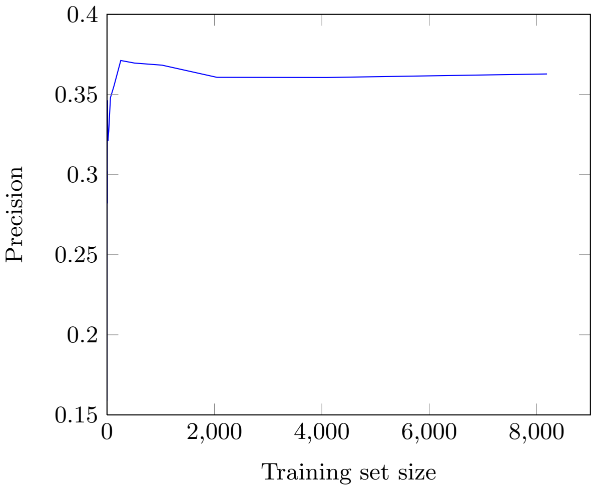

Figure 4: Performance using linear regression for learning

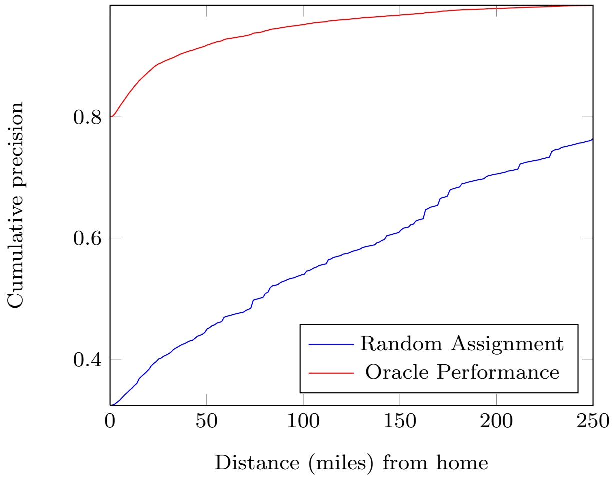

Figure 5: Oracle, random and simple edge counting performance

5.2 Baselines

The graph in Figure 5 shows the lower bound on performance, wherein cities are chosen from the candidate set completely at random. Note, however, that a more practical comparison might be made against the proportion of users that actually report their true location in the relevant field (26% according to [2]). We also calculated that an accuracy of 27.31% is attainable simply by assigning London as the location of every user.

The most frequent location baseline method, as defined in Section 4.4, works surprisingly well, with an accuracy of 39.18% to the nearest town/city (Figure 5).

We believe that this relatively high accuracy is attained by such a simple baseline, because online relationships are more probable when preceded by real world relationships (which are typically governed by geographical constraints).

Method	≤ 0 m	≤ 50 m	≤ 100 m	≤ 150 m	≤ 200 m
Random assignment	31.61% ± 0.44%	43.61% ± 0.50%	52.71% ± 0.44%	59.56% ± 0.38%	68.89% ± 0.23%
Oracle performance	78.15% ± 0.25%	89.59% ± 0.27%	92.97% ± 0.27%	94.50% ± 0.21%	95.59% ± 0.19%
Backstrom [1]	9.17%	47.36%	59.39%	67.23%	75.71%
Backstrom Best [1]	25.48%	52.02%	59.75%	65.78%	74.04%
Simple Friendship Count	39.49% ± 0.39%	47.79% ± 0.40%	55.18% ± 0.37%	60.66% ± 0.34%	69.69% ± 0.29%
Inverse City Frequency	10.66% ± 0.19%	40.59% ± 0.41%	51.34% ± 0.38%	59.99% ± 0.33%	69.27% ± 0.24%
Population	45.94% ± 0.47%	58.20% ± 0.46%	65.44% ± 0.40%	70.46% ± 0.27%	77.20% ± 0.29%
Population & Neighbours	47.13% ± 0.42%	59.24% ± 0.42%	66.39% ± 0.27%	71.43% ± 0.29%	77.97% ± 0.27%
Pop. & Neighb. & Recip.	**50.08% ± 0.54%**	**62.08% ± 0.53%**	**69.03% ± 0.43%**	**73.83% ± 0.47%**	**79.99% ± 0.34%**

Table 3: Comparison of performance of all methods, with 99% confidence intervals

The Twitter graph in our dataset shows a great deal of collocation between connected vertices, where users are far more likely to be friends if they live close to one another. This is in line with our model of relationships by distance, argued by [1] to fit a power law distribution. The distribution of distance ratio in social ties is illustrated in Figure 3.

5.3 Previous method

We have evaluated the algorithm presented in [1] in 2 variants, as described in Section 4.5. The results are lower on our Twitter datasets, than those presented in their original analysis of Facebook networks. This confirms our hypothesis that our dataset is more challenging, due to the unidirectional nature of Twitter relationships and the different, more geographically dense region under study. We also note that the power law inferred from the distance distribution did not give the best results.

In our view, the weakness of this algorithm is that it does not explicitly take in account that some areas are more densely populated than others and, due to the nature of our data, is very sensitive to populous cities and the noise created by the unidirectional friendship relationship on Twitter.

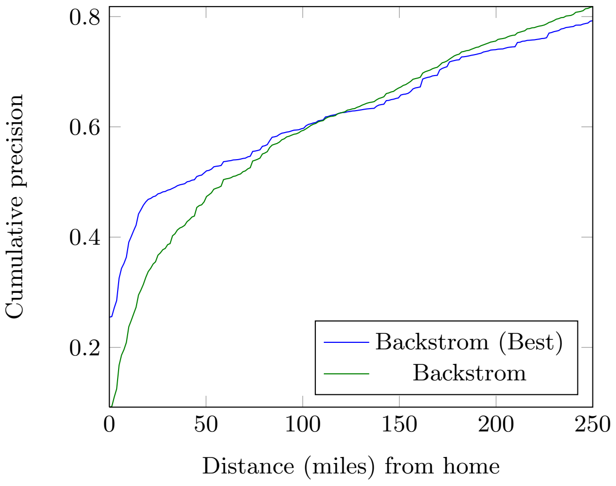

Figure 6: Performance using previous method [1]

5.4 City size

Incorporating the city size by simply multiplying our ratio by the number of users in that location did not perform as well as expected. The performance dropped to 10.85%, which is considerably worse than choosing at random, because large cities like London (which account for much of the data) were being penalised too harshly and never assigned.

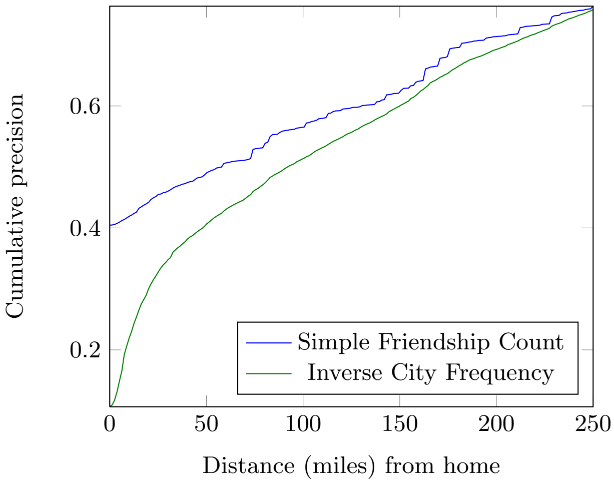

Figure 7: Performance using simple counts and ICF weighting

5.5 Classification

As discussed in Section 4.2, we use an SVM classifier with a Gaussian radial basis function (RBF) kernel in order to learn these probabilities.

Incorporating the population of a city created a considerable increase in performance, as shown in Figure 8. Adding information about the number of common neighbours yielded only a slight improvement, although the result of adding a feature for reciprocated relationships was more noticeable (Figure 8).

5.6 Training data size

The SVM classifier takes substantially longer to train as it is given more examples. Eventually, the performance gains from increasing the dataset size are somewhat outweighed by the cost in terms of extra training time. Figure 9 demonstrates the relationship between increased dataset size and increased performance, using all available features. It shows the diminishing returns as the training set grows very large.

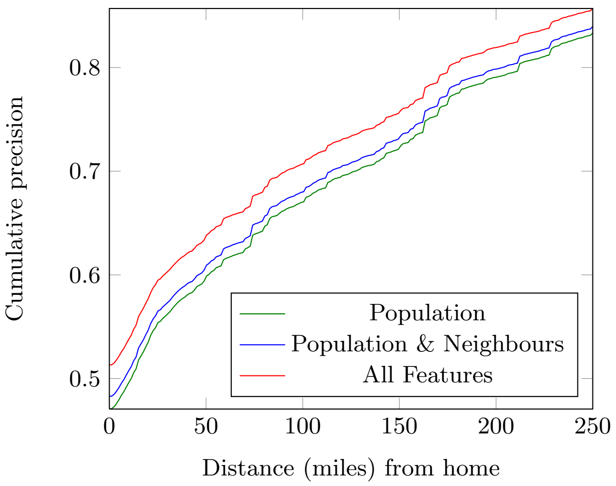

Figure 8: Performance using all features

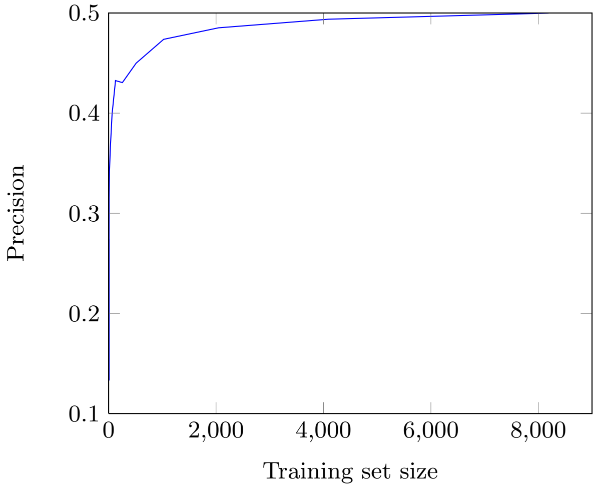

Figure 9: Performance for increasing size of training set

5.7 Iterating

In early experimentation, we also allowed users to have their location updated once one had been assigned. In subsequent passes of the algorithm, the current state of friendships would be taken into account and location assignments updated. This is a form of percolation - allowing the method to overcome the limitation caused by only observing the graph at depth 1 and allowing more users to be allocated to at least some location.

However, we found that making our method iterative consistently reduced location accuracy. An error analysis demonstrated that the population distribution tended to become 'flatter', with fewer people assigned to major cities and overall they were spread more evenly. Since most people do actually live in densely populated areas, this was disadvantageous. Iterating like this would perhaps be more effective when geolocating within large, rural areas.

6. CONCLUSIONS AND FUTURE WORK

This paper investigated factors that influence social 'closeness' in an online setting, modelled them as features within a classification task and demonstrated that they can be used to better geolocate users automatically.

Our experiments demonstrated that social ties alone can be enough to locate a user of a social network. Of the methods evaluated here, the best performing approach was to select amongst location candidates using collocation, population density and reciprocation features, wherein nearly half of users could be geolocated correctly to their town or city. This is in addition to the 26% that can already be unambiguously geolocated using their profile fields.

In contrast to previous methods, in this paper geolocation is formulated as a classification problem. Although the set of possible target locations is large, computation is made feasible through a simplifying assumption that users share their location with at least one of their friends. We have shown exactly how limiting this assumption is for our Twitter dataset.

This paper has extended previous understanding of the geolocation task by demonstrating that the population of a settlement affects the likelihood that a user with friends there, also lives in that location. Additionally, we have shown that properties of a user's social network, including the distribution of triads and reciprocated connections, can help predict the friends to whom they are geographically closer.

It might be argued that the finding of this work shown in Figure 5 is relevant to online privacy concerns. It demonstrates that regardless of textual content or the content of a user's profile field, 50% of Twitter users could be located from their social connections alone, by identifying which social links are geographically local. Our method is a very shallow one, relying only on features of the graph itself. Another method that utilises greater knowledge or even a manual approach could yield much higher results, demonstrating that even if one is very careful, it is not necessarily possible to remain truly anonymous online.

This paper has conducted a review of different geolocation methods, based on user's social ties, as well as other features. Even though not directly comparable due to the use of different datasets, our results are better than those reported by [9] and [2]. Moreover, both methods rely on features which are more difficult to collect. One key advantage of our method is that it is very lightweight and we believe it would be practical to apply in an online setting.

Our experiments have concentrated on a single geographical area. The evaluation methodology necessitated a high precision parser of location names for Twitter profile fields, which was achieved by using a gazetteer list of unambiguous place names in the U.K.. We feel that the U.K. is a reasonable region with which to evaluate our system, because it features many clusters of population in a relatively small area. In future work, we will develop a more robust location recogniser for Twitter profile fields and evaluate on a geographically wider dataset.

The method described here does not use any textual content from tweets whatsoever. This allowed us to explore what is possible using straightforward analysis of the social graph alone. In future work we plan to replicate some of the content features from [9], as well as measures of social closeness between users, which can be derived from actual tweets (e.g. counts of user mentions and re-tweets). Textual content may be the key to disambiguating the kind of relationship a particular link represents, thus allowing a geolocation system to deal better with noise.

It is also possible to predict first a region for a user, then their specific settlement within that region. This is an area for future investigation which we believe may enable our algorithm to scale to a wider geographical area whilst maintaining similar performance.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1) and the Trendminer project, EU FP7-ICT Programme, grant agreement no.287863.

7. REFERENCES

[1] L. Backstrom, E. Sun, and 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*, pages 61–70. ACM, 2010.

[2] Z. Cheng, J. Caverlee, and L. Kyumin. 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, 2010.

[3] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.

[4] J. Eisenstein, B. O'Connor, N. Smith, and E. Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[5] C. Fink, C. Piatko, J. Mayfield, T. Finin, and J. Martineau. Geolocating Blogs from their Textual Content. In *Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 1–2. AAAI Press, 2008.

[6] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, New York, NY, USA, 2011.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[8] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences*, 102(33):11623–11628, August 2005.

[9] J. Mahmud, J. Nichols, and C. Drews. Where is this Tweet From? Inferring Home Locations of Twitter Users. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.

[10] E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In *Proceedings of the Fifth International Conference on Web Search and Data Mining*, 2012.

[11] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBPedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.

[12] J. A. Muir and P. C. Oorschot. Internet Geolocation and Evasion. Technical report, School of Computer Science, Carleton University, 2006.

[13] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic Analysis of Citizen Sensor Data: Challenges and Experiences. *Web Information Systems Engineering-WISE 2009*, pages 539–553, 2009.

[14] D. Preoţiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Workshop on Real-Time Analysis and Mining of Social Streams*, 2012.