Transactions in GIS

Transactions in GIS, 2011, 15(6): 735–751

Research Article

Inferring the Location of Twitter Messages Based on User Relationships

Clodoveu A. Davis Jr.
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Gisele L. Pappa
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Diogo Rennó Rocha de Oliveira
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Filipe de L. Arcanjo
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

Abstract

User interaction in social networks, such as Twitter and Facebook, is increasingly becoming a source of useful information on daily events. The online monitoring of short messages posted in such networks often provides insight on the repercussions of events of several different natures, such as (in the recent past) the earthquake and tsunami in Japan, the royal wedding in Britain and the death of Osama bin Laden. Studying the origins and the propagation of messages regarding such topics helps social scientists in their quest for improving the current understanding of human relationships and interactions. However, the actual location associated to a tweet or to a Facebook message can be rather uncertain. Some tweets are posted with an automatically determined location (from an IP address), or with a user-informed location, both in text form, usually the name of a city. We observe that most Twitter users opt not to publish their location, and many do so in a cryptic way, mentioning non-existing places or providing less specific place names (such as "Brazil"). In this article, we focus on the problem of enriching the location of tweets using alternative data, particularly the social relationships between Twitter users. Our strategy involves recursively expanding the network of locatable users using following-follower relationships. Verification is achieved using cross-validation techniques, in which the

Address for correspondence: Clodoveu A. Davis Jr., Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. E-mail: clodoveu@dcc.ufmg.br

doi: 10.1111/j.1467-9671.2011.01297.x

location of a fraction of the users with known locations is used to determine the location of the others, thus allowing us to compare the actual location to the inferred one and verify the quality of the estimation. With an estimate of the precision of the method, it can then be applied to locationless tweets. Our intention is to infer the location of as many users as possible, in order to increase the number of tweets that can be used in spatial analyses of social phenomena. The article demonstrates the feasibility of our approach using a dataset comprising tweets that mention keywords related to dengue fever, increasing by 45% the number of locatable tweets.

1 Introduction

With the expansion of the sources for geographic data on the Web, geocoding tools and techniques are becoming increasingly important. In the past, geocoding was usually restricted to street address data, usually obtained from applications on public health, education, public safety, transportation and others. Most research in this field worked with parsing algorithms, matching techniques, disambiguation, along with improvements to the addressing reference databases. The objective was to increase the match rate, i.e. the percentage of input addresses that could be associated with a pair of coordinates, and the positional accuracy, i.e. how close the output coordinates are to the desired location. This emphasis on geocoding addresses reflected the ongoing effort to transform geographic references, such as the ubiquitous postal address attribute in conventional databases, into actual geographic locations that could be visualized and analyzed using tools such as geographic information systems (GIS).

More recently, people have become aware of the increasing volume of online data and the richness of its content, especially considering the possibilities offered by current search engines. The fast evolution of free online mapping and location services also shows that people are ready to use and to contribute to geographic applications (Sui 2008). As a natural consequence, there is now much interest in associating Web documents to locations (Silva et al. 2006, Zong et al. 2005), but since this association is not always obvious, document contents or structure must be analyzed in the search for evidence of associated geographic locations (Borges et al. 2011) for natural language expressions that indicate a location (Delboni et al. 2005, Hall and Jones 2008), or by resolving toponyms from text (Alencar et al. 2010, Twaroch, Smart and Jones 2008). Place name catalogs such as gazetteers (Goldberg et al. 2007, Janowicz and Kessler 2008, Goodchild and Hill 2008), have gained a broader scope, so that they can serve as repositories of geographic knowledge, and include details on anything that people can associate to a place, including intra-urban place names, monuments, landmarks, and other distinctive features, along with semantic relationships between places (Machado et al. 2010).

Considering this broad range of interests, geocoding is increasingly understood and researched considering a wider scope. We want to be able to recognize related places from various sources, such as Web pages, keyword sets, mobile device trajectories, news sources, IP addresses, and others. Users of online social networks are an important item on this list. The intensive growth of online social networks throughout the last few years requires new ways to gather and analyze locations associated to events and phenomena discussed by users across the planet. Online social networks are quickly becoming a reflection of what calls the attention of the general public regarding daily events, to the point that the repercussions of some events can be more quickly felt as a "trending topic"

than from regular news sources. The focus of an individual's attention or concern can now quickly evolve from a short message to close friends that reaches the entire globe, in a propagation that is popularly called *viral*, for its speed and reach. We see examples of this phenomenon every day, ranging from trivial gossip about celebrities to political events of worldwide significance.

However, associating social network messages and users to locations is not an easy task. Services such as Twitter and Facebook leave it to the user to decide whether to associate herself with a location. Most people do not go through the trouble of activating this option. For the users that decide to do so, one of the options for the location is a free-text field, which allows for all kinds of responses, with various spatial granularity levels (city, state, country) or even invalid, inaccurate or humorous contents. Another option is to have the location automatically determined from the user's IP address as a city name. For mobile users, the GPS coordinates of the mobile device at the time a message is posted can be used for a location. Sifting through this kind of input is a challenge, which we tackle based on the assumption that most people are both cooperative and truthful regarding their social network profiles.

We have been involved in initiatives to monitor the online repercussions of events in real time, using sources that include, along with the aforementioned social networks, blogs and news. The overall strategy is called the *Observatory of the Web* (see http://www.observatorio.inweb.org.br for additional details), implemented as a pipeline with stages for data collection, extraction, analysis, and visualization (Gomide et al. 2011). Currently two very different contexts are being monitored: the Brazilian national soccer championship, and the dengue fever epidemic. Data from these contexts reinforced the need to include the spatial dimension in the analyses. In the case of the soccer championship, it is reasonable to assume that fan bases for most teams are concentrated in specific cities and regions, while for dengue fever we assume that more messages will spring up in regions that are more heavily affected by the epidemic.

This article presents our initiatives to enrich the geographic content of messages collected from Twitter. The number of messages that can effectively be associated to a location is relatively low. The idea is to increase that number by analyzing the following-follower relationships between Twitter users, thereby increasing the quantity of locatable dengue-related messages. The article is organized as follows: Section 2 presents some related work, both on geocoding and on social network analysis techniques; Section 3 introduces our methodology, which is then put to practice in Section 4, in which we present and discuss our results; Section 5 presents our conclusions and gives indications for future work.

2 Related Work

We agree with Goldberg et al. (2007) when they point out that geocoding is gaining a much wider range of applications. Research in this field is not limited to addresses, and seeks to recognize and understand indications of location associated with natural language terms and expressions, toponyms, telephone area codes, or even indirect evidence, such as references to landmarks, cultural events, typical cuisine, sports teams and many others. So far, however, there is no single method that is adequate for all of such a diverse set of sources.

The direct interpretation of the textual contents of Web pages can produce place-related terms and expressions, which in turn can be located in space. There are some

proposals in that direction. Delboni et al. (2005) propose recognizing relevant places in text by looking at the vicinity of positioning expressions such as "close to" or "at walking distance from". The authors were able to determine the most common positioning expressions in Portuguese from an analysis of Web pages using a set of regular expressions (Delboni et al. 2007). Zong et al. (2005) associate Web pages to places using a place name extraction algorithm based on a gazetteer, and then perform disambiguation. Twaroch et al. (2008) detect place names in Web queries using locative phrases and a reference gazetteer.

The interpretation of the contents of pages can focus on finding well-structured bits of information, such as postal codes, telephone area codes and postal addresses. Borges et al. (2007, 2011) present such a method, in which regular expressions for extraction are put together with the help of an ontology. Ahlers and Boll (2008) present address extraction from German Web pages with subsequent validation by a parser. Semantics often plays an important role for disambiguation and for distinguishing the expected information from the unstructured background. Cardoso et al. (2008) show how to do so for query expansion, also resorting to positioning expressions and their semantic interpretation.

Recognizing place names embedded in text can also be a challenge, considering that ambiguity with non-geographic entities is common. Leidner (2004) calls this activity *toponym resolution* and proposes the creation of shareable evaluation resources for recognition and disambiguation. Machado et al. (2010) implement a resource such as this, in the form of an ontologically-enhanced gazetteer. The proposed gazetteer keeps a record of spatial and semantic relationships between places, and also hosts lists of place-related terms and known ambiguities. Such place-related terms can be obtained from online sources such as Wikipedia, as demonstrated by Alencar and Davis (2011) using techniques initially proposed by Buscaldi and Rosso (2007).

Concerning location inference in social networks, Crandall et al. (2010) used an interesting approach, which estimates the probability of two users knowing each other from information indicating that they have been in approximately the same geographic location at approximately the same time, on multiple occasions.

Backstrom et al. (2010), in turn, used addresses provided by Facebook users and geocoded using TIGER data. Most users have at least one friend who has provided an address. Ambiguous or imprecise locations were left out of the study. Geocoded Facebook addresses were compared with GeoIP data; 57.2% were found within 25 miles of the geocoded Facebook address. The prediction method found 69.1% of the users within 25 miles using friendship data (16 or more locatable friends). Results indicate that people with five or more locatable friends are more precisely located than by using GeoIP. Notice that IP location techniques tend to be less reliable in developing countries, such as Brazil.

Finally, Gonzales et al. (2011) investigate the effects of locality in Twitter, focusing specially in following/follower location relations. One of their results shows that, in countries such as Brazil, where English is not the first language, there is a high intra-country locality among users and their followers, while English-speaking countries, such Australia, suffer from what they call external locality effect, having many of their followers in the US.

Apart from the works cited above, many other methods have been previously proposed to propagate and infer information from social networks, which can be easily represented as graphs (Lu and Getoor 2003, He et al. 2006, Agarwal et al. 2009). Many of these methods are adaptations of approaches initially created by the text-mining and information retrieval communities, where learning the relationship between words and

documents was the focus of research for a long time. Others are based on the ideas of homophyly and influence, which are not valid in the context of this article, as the similarity of two individuals gives no hints about their location.

In the case of homophyly and influence, methods based on link prediction are among the most popular (Lu and Zhou 2011). They consider the edges of the social network graph as specific attributes of a user, such as political view; these edges can be removed, added or altered by the algorithms. For other types of inference, such as the one addressed in this article, methods based on probability and causality, like Bayesian networks, have showed that many user attributes can be inferred based on the relationships of people in networks. Techniques such as spreading activation (Ziegler and Lausen 2004) could also be applied to take into account more than one level of propagation in the network for location inference.

3 Methodology

Knowing the location of users can lead to improvements in Web applications. The most basic example is to use location to improve search results. For example, if the user searches for coffee shops, he or she is probably more interested in the ones close to where he or she is. Here we focus on a different type of service: disease monitoring, although the proposed methodology can be applied to any context.

Our main objective is to observe, in real time, the intensity and distribution of discussions about the disease, thus attempting to identify regions where outbreaks may be taking place. The early identification of regions where epidemic diseases are starting to appear allows authorities to take preventive measures, such as intensifying public health campaigns, warning health centers and getting personnel and supplies ready to meet upcoming patient demands.

Having a set of relevant tweets, the architecture behind the Observatory runs a series of preprocessing steps, and at the end plots a map indicating where the tweets are coming from. Preliminary studies have already shown that there is a strong correlation between the source locations of users and recorded cases of the disease (Gomide et al. 2011). However, the problem is that most tweets are not associated to a geographic location. In Twitter, users can associate their profile with a location, but most people do not go through the trouble of activating this option. In our current assessments, less than 40% of the users have associated themselves with a location, and part of the declared locations is useless for automated analysis. Hence, obtaining more locations is still a big problem, therefore developing methods to infer user location from other evidence available in Twitter is the main objective of this article.

The idea of the proposed method is that the users whom people usually follow online are, the majority of the time, the same ones with whom they keep in contact in real life. This is specially true in Facebook, where relationships are bilateral. Twitter relationships, however, have a different nature, as Twitter mixes friendship relationships with other kinds of unbalanced, asymmetrical relationships. As investigated by Kwak et al. (2010), in Twitter users follow others not only for social networking, but also for information, which is not true for other types of networks with reciprocal relationships. Nevertheless, when we take an intersection between the followers of a user and the people he or she follows, we obtain reciprocal relationships, which constitute a more balanced subset of the entire set of Twitter relationships. Kwak et al. (2010) showed that only 22% of Twitter relationships

are mutual and, for the purposes of this article, are considered to be friendships. In accordance with the ideas of this article, Kwak et al. (2010) also concluded that Twitter users who have reciprocal relations with less than 2,000 friends are likely to be geographically close to most of them.

Hence, having the set of balanced reciprocal relationships, we can then infer a user's location if we know enough about his friends' locations, thereby achieving our goal to expand the number of locatable tweets for monitoring purposes. The following sections show how the data was collected, and then describe a simple method to infer user's locations considering their friends' geographical positions.

3.1 Database Collection

In order to provide a dataset for testing the proposed method, we first collected a set of tweets relevant to dengue monitoring. The tweets were collected using the Twitter API (see http://dev.twitter.com/ for additional details) and the words "Dengue" and "Aedes aegypti", the latter being the scientific name of the mosquito that transmits dengue fever. For each tweet collected, a preprocessing step checks whether the tweet is in Portuguese, since we are only interested in dengue fever outbreaks in Brazil.

Since we want to infer location based on friendship (recall that here this term refers to the set *followers* ∩ *followed by*), we need to identify the user behind the tweet, as well as his or her friendship network and his or her friends' locations. However, taking initially the set of users talking about dengue fever we were only able to generate a very sparse friendship network. This is not ideal because we also want to test the previously mentioned hypothesis, which stated that people in Twitter also keep in touch with those users next to them in real life. Hence, we decided to change the collection to obtain a network with more connections. We used the initial set of dengue users as seeds to a simple depth-first search that rebuilds part of the Twitter network. This procedure is described in Algorithm 1.

As described in Algorithm 1, for a set of seed users, we find the intersection among their followers and the users they follow. During the depth-first search we check, for all

Algorithm 1 Collector(seeds)

```
while (seeds <> ø) do
  remove (seed_i, seeds)
  friends ← getFollowers (seed_i) ∩ getFollowedBy (seed_i)
  createEdges (seed_i, friends)
  for all friend_j in friends do
    if user_k not collected yet and location available then
      if (location in gazetteer) then
        insert (friend_j, seeds)
      end if
    end if
  end for
end while
```

users, whether they had an associated location or not. As previously mentioned, Twitter users that enable the location field have three options: (1) to fill in a free-text field, which allows for all kinds of responses, ranging from the precise (e.g. a neighborhood name and a city name) to the casual (a city name), cautious (a country name) or tentatively humorous ("planet Earth") – and also for deliberately inaccurate ones; (2) a location field, automatically filled out as a city name, obtained from a service such as GeoIP using the user's IP address; and (3) in the case of mobile users, the GPS coordinates can be set automatically as their location. For monitoring purposes, knowing the city associated to each user or message would be enough.

Declared locations usually do not change over time. However, GPS and GeoIP locations may vary a lot, although they are present in the minority of the tweets, as discussed in Section 4. Another feature is that a tweet may be associated with more than one type of location simultaneously, such as declared location and GPS position. In this case, we choose as the user location the GPS position over the GeoIP location over the declared location. In order to guarantee that the location set to the user has a high confidence of being his or her most frequent location, we collected the set of n last tweets posted, and resolved his or her location to the most frequent city within that set. Different values of n were tested, as shown in Section 4. The rationale is that the user's most frequent location is most probably the one at which he or she lives.

For each user with a declared location, we checked in a gazetteer if the place name corresponds to a city name in Brazil. Invalid or ambiguous locations were disregarded. Note that the choice of using cities was arbitrary, and as next steps we intend to go one level down in the hierarchy, dealing with neighborhoods and perhaps with more specific locations, such as those provided by Foursquare (see https://foursquare.com/ for additional details).

3.2 Location Inference Method

Having a sample of the Twitter network, we want to infer the location of a user based on the location of his or her friends. The simplest way of doing that is to count the most popular locations among the friends of a user, using a simple majority voting scheme. In this case, the most popular location among friends is set as the location of a user. We first test this very naive approach in order to better understand the nature of the problem. Much more intelligent methods can be used for location inference, such as Bayesian networks and spreading activation, as listed in our future work.

However, this approach presents two main problems. First, the user can have too few locatable friends, and therefore the voting scheme must decide his or her location based on weak evidence. Second, the voting scheme can lead to tied results, and therefore a winner must be picked later.

These problems were solved by setting up three different parameters for the method. Two of them define the minimum and maximum number of friends a user should have in order to have his or her location correctly inferred. Preliminary experiments showed that having one friend or one million are both bad for the inference method. In the first case, there is no confidence in the data provided. In the second, the user is too popular to be trusted, being "friends" with everyone.

The third parameter refers to the minimum number of votes a location needs in order to be considered as the correct one. Locations with one vote are as reliable as users with one friend, and ties between many locations with one vote are neglected. The value of

these three parameters depends on the data being handled. As shown in the next section, they do have a large impact on the results obtained.

4 Experimental Evaluation

In order to test the efficacy of the proposed location inference method, we first colhyqjlected data from Twitter. As explained before, instead of considering all users talking about dengue fever to build the Twitter sample network, we selected a subset of them as seeds to build a friendship (*following* ∩ *follower*) network using a depth-first search.

The algorithm was executed using five random users talking about dengue fever as seeds, ending up with 61,400 users. From these 61,400 users, 24,767 had location information in their profiles (around 40%).

For each of these users, we collected their last 100 tweets (this is the limitation for each requisition submitted to the Twitter API). The time over which these 100 tweets were posted varies substantially according to the profile of the user. On average, 2.24 days passed between each of the 100 tweets posted, but the standard deviation was close to 8.7 days. The median, in contrast, was 0.55 days between each tweet, leading to a median total time of collection of 55 days.

For the 24,767 users with location information, 809 had at least one tweet posted via GPS, 1,801 had at least one tweet using GeoIP, and 22,874 had at least one tweet with a declared location. Hence, 171 users had more than one value for location. In this case, as explained before, the priority given for each value was GPS > GeoIP > Declared location.

Further experimentation showed that using only the last 10 tweets is enough to closely match the results with 100 tweets with a much lower computational time. This is due to the high number of declared locations in the tweets, which by nature should remain constant over time. The next section gives details about the collected dataset, followed by computational results.

4.1 Dataset

Figure 1 shows four histograms detailing the distribution of users according to the number of friends they have in different network configurations. Note that all histograms present up to 50 friends. Users with more than 50 friends were removed to make the histograms easier to read, since their frequency decays significantly. The histogram in Figure 1a shows the distribution of the originally collected data, considering people with and without location information, while the one in Figure 1b shows, for all users, how many of their friends have a valid location. Following this rationale, Figure 1c presents, for all users with location, their total number of friends, and Figure 1d shows the distribution of the data used in our experiments: users and friends with valid locations. Comparing the histograms, we note that the number of friends decreases abruptly as the location starts to be taken into account. Another metric that does not appear in the histograms but is extremely relevant is the number of people with no friends. We have 1,373 users with no friends at all (because the result of the intersection between following and followers was an empty set), and 763 with no friends with valid location information available. In total, 3.5% of the users in the dataset were not considered when inferring location.

The graphs show that the distribution follows a power-law, with many users having only one friend and only a few users having more than 50. Hence, we can say that there are many users that will not be helpful for location inference at all, because there is not enough information (friend's locations) to perform a confident inference. On the other hand, some users are too popular, and establish follower connections to a large number of people in the network. These people are also not regarded as important in the inference process, since their relationship with other users, although bilateral, cannot be viewed as actual friendship.

4.2 Experimental Results

Considering the dataset with the characteristics described in Figure 1d, the inference method was applied using a 10-fold cross-validation process (Witten and Frank 2005), i.e. the dataset was divided into 10 partitions, and the user locations from one of them were hidden. For the hidden partition, we used the proposed method to infer the users' locations. The results of these experiments are reported using two well-known metrics: precision and recall. Precision calculates the percentage of correctly inferred locations.

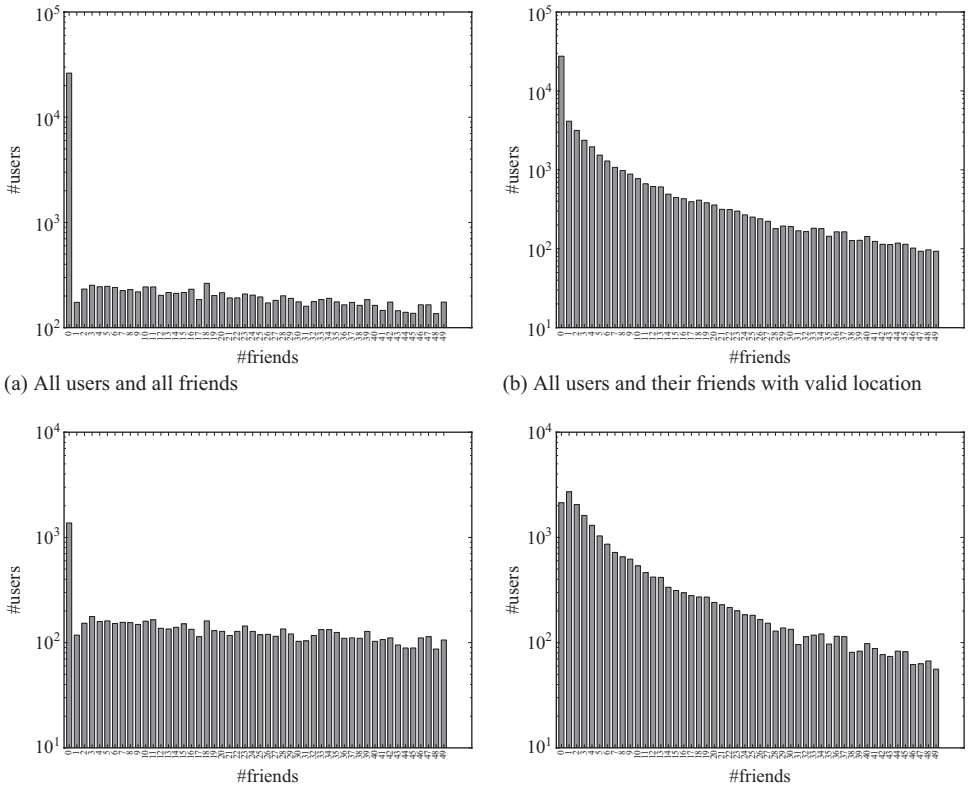


Figure 1 Distribution of the number of friends per user in different configurations of the network

Table 1 Precision and recall using different types of Twitter location

Data	No. of users	Precision	Recall
All	24,767	0.4013 ± 0.0042	0.9137 ± 0.0046
GPS	566	0.3262 ± 0.0776	0.3440 ± 0.0341
GeoIP	1,606	0.2178 ± 0.0289	0.4387 ± 0.0266
Declared	22,595	0.4047 ± 0.0050	0.9103 ± 0.0043

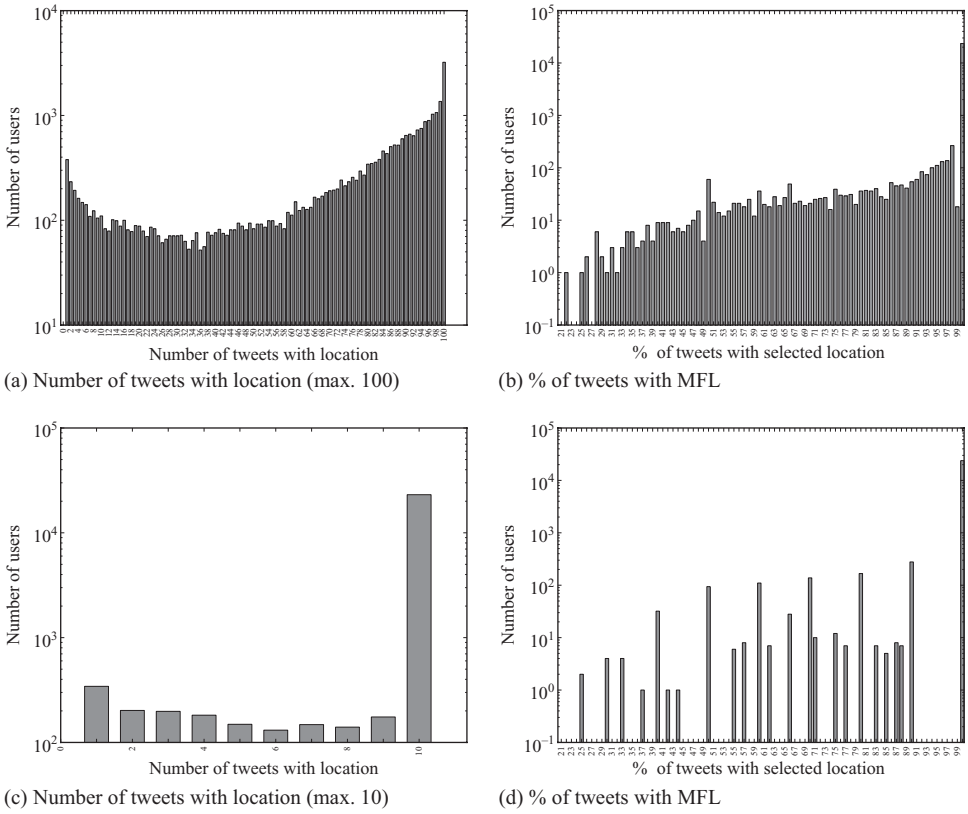
Recall represents the percentage of users for which we were able to infer a location. The location inference might not be possible in cases where the user has no friends or a tie occurs.

First we present the results using the complete dataset as well as different subsets of data according to how location was set in Twitter. Table 1 shows the type of data considered, the number of users in the network taken from the data and the av- erage precision and recall over a 10-fold cross-validation, followed by their standard deviations. As observed in the first line, using all data, we can only predict correctly around 40% of the locations for 91.4% of the users. The reason for such low accuracy is that, as explained before, many users have no friends or only one friend in their net- work. Performing a majority voting with one vote does not make sense. Because of the lack of information on friendship, the method proposed here is not applicable for these cases. On the other hand, users with a very large number of friends can also disrupt the inference method, as their intention is only to be popular and, in most cases, they do not bring useful location information. Examples of these user profiles are music bands, companies trying to promote a product, and institutional profiles, among others.

Table 1 also shows the results for networks built over tweets using GPS, GeoIP and declared location only. The numbers also give an idea of distribution of data in terms of GPS, GeoIP and declared locations. The latter method is the most used one and, at the same time, the one with the least reliable data source. Data from GPS and GeoIP are still a minority, and this explains the very low accuracies we got when considering only data coming from these sources.

Figure 2 shows the frequency of users according to the number of tweets collected that had an identified location, followed by the percentage of these tweets where the most frequent location appeared considering subsets of the 10 and 100 last posted tweets. These histograms show that there are around 3,500 users with 100% of locatable tweets (Figure 2a), and summing up the number of users with 50 to 100 locatable tweets we find a great part of the users in the dataset.

Figure 2b calculated, from the number of locatable tweets, the percentage which had the selected location. For instance, if 50 tweets had Rio as location and 20 had Campinas, we divided 50 by 70, generating the numbers shown in the graph. Note that, among more than 22,000 users with 100% of tweets with the most frequent location (MFL), are those with a single tweet (they are around 300, according to Figure 2a). However, this number corresponds to less than 1% of tweets with 100% of MFL. When we reduce the number of tweets from 100 to 10, the number of users with 100% of locatable tweets is higher than 20,000, as expected (see Figure 2c), and the MFL is higher than 50% for the majority of users.



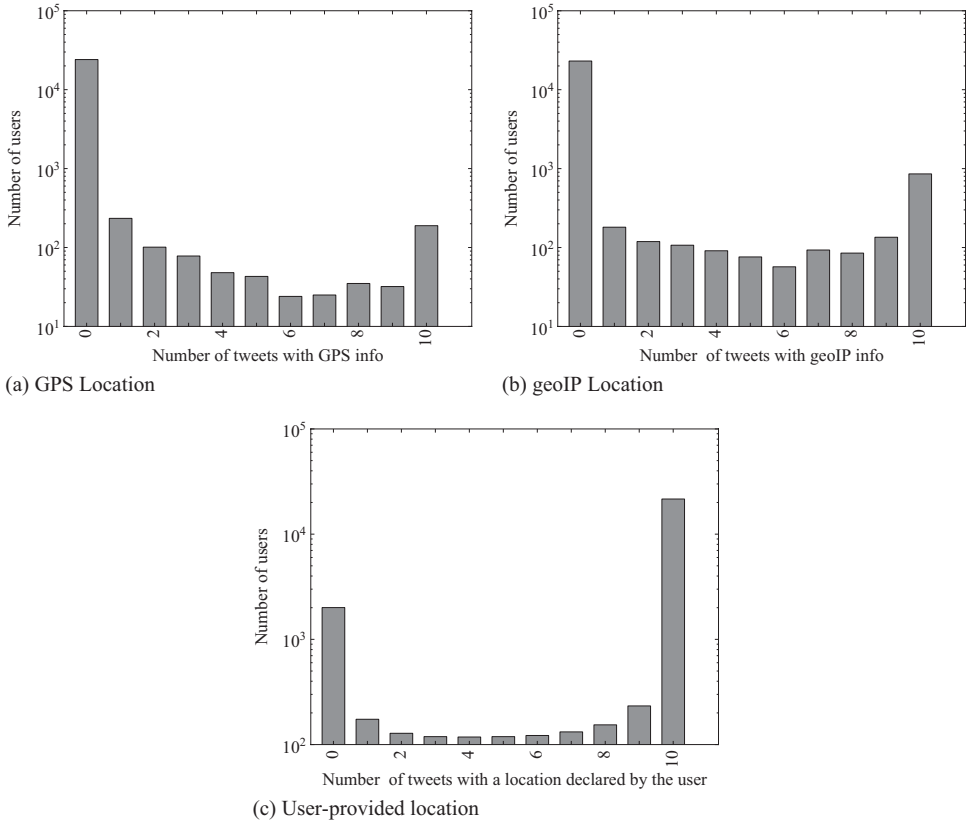
(a) Number of tweets with location (max. 100)
(b) % of tweets with MFL
(c) Number of tweets with location (max. 10)
(d) % of tweets with MFL

Figure 2 Total number of tweets with location and the percentage of these tweets where the most frequent location (MFL) occurred

It is important to mention that when users tweet using platforms that allow messages to be posted in many social networks simultaneously, such as TweetDeck (see http://www.tweetdeck.com/ for additional details), the location changes according to the one set in that particular application. Hence, in many cases you might have 100 tweets with identified location, with 60 posted from the Web and 40 identified by GeoIP (for instance, posted with Tweetdeck with no available declared location by available GeoIP location), but all of them with the same location.

Figure 3 shows the number of users with at least one GPS-location, GeoIP location and declared location. Recall that individual tweets can have more than one location source. Note that around 20,000 users have all their tweets with declared locations, as expected. The user will not change his or her location every time he or she tweets. At the same time, at least one of their tweets is also normally tagged with the GeoIP location or GPS location. That is the reason for the shapes in Figure 3.

This first set of experiments showed that there must be a minimum and maximum number of friends available to avoid the problems of lack of information and users too popular in Twitter, who are not necessarily real friends of a user. The simplest way to do that was to find an interval where, given a number of friends in that interval, it would be reasonable for the method to infer a location.



(a) GPS Location (b) geoIP Location

(c) User-provided location

Figure 3 Source of information from where the most frequent location was extracted

Using the minimum and maximum number of friends as a parameter, we tested some values according to the distributions given in the histograms presented in Figure 1. The minimum values were set to 1, 5, 10, 15 and 20, while maximum values were tested at 50, 100 and 200 friends. Figures 4 and 5 show the results of precision and recall and their respective 95% confidence intervals obtained with different minimum and maximum values in three different experiment configurations. In the first configuration, only the users we were inferring location to needed to have their number of friends within the interval defined by *min, max* (Figure 4a). In the second configuration, both the user and his friends need to have a number of friends in the defined interval (Figure 4b). The third configuration is explained further ahead.

Comparing Figures 4a and b we can observe that the limits of the interval do not influence greatly the results, and that the maximum value of precision in Figure 4a is 0.44, while in Figure 4b it reaches 0.65. This happens because, when ensuring that both the user and his/her friends are inside the defined interval, the confidence on the data is higher, and we know that no popular users or users with only one friend are influencing the results.

However, looking at Figures 5a and b, notice that the value of recall increases as the maximum limit reaches higher values, but is not influenced by the lower limit. At the same time, the gain in precision obtained by the restriction of having both users and their

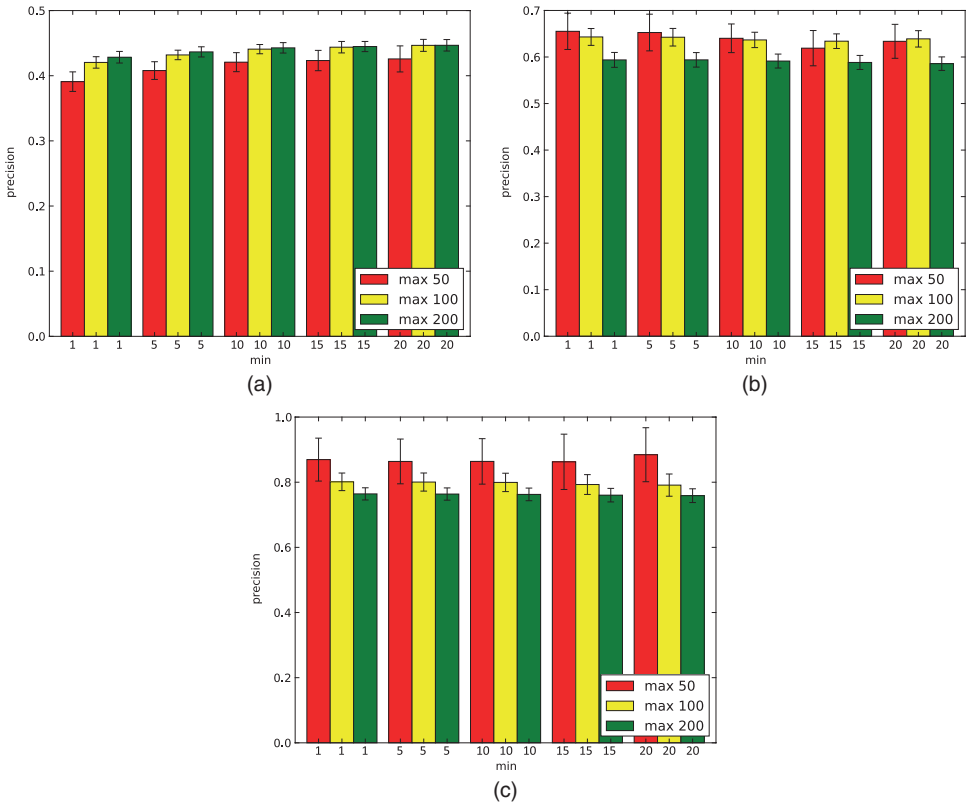


Figure 4 Precision of the location inference method considering [min, max] friends intervals

friends inside the interval resulted in a significant loss in recall, which reached a maximum value of 0.44 when using the interval [20,200].

Although these results show that the low confidence of user votes with too few or too many friends is disrupting the process, there is still another problem. In general, the median number of different cities among the location of the friends of an user was four, and the median of friends was seven. As a consequence, in some cases, the location was inferred using only one or two votes, or a tie was identified. This problem can be avoided by requiring a minimum number of votes before assigning a location to a user. A test was performed using the original data with a minimum of 2, 3, 5 and 10 votes. All the results showed an average precision of approximately 45%. However, the average recall went down from 67% for 2 votes to 22% with 10 votes. Given the very small difference in terms of precision for these votes, and its significant increase over a minimum of one vote (results in the first line of Table 1), we decided to consider a minimum of two votes for each user in order to make an inference.

After this process, we created a third experiment configuration combining all the positive choices made during experimentation: a minimum of two votes that can be given only by users with a number of friends within a defined range. These results are reported in Figures 4c and 5c. As expected, the greater the size of the interval, the lower the precision and the higher the recall. Intervals varying up to 50 friends reach an average

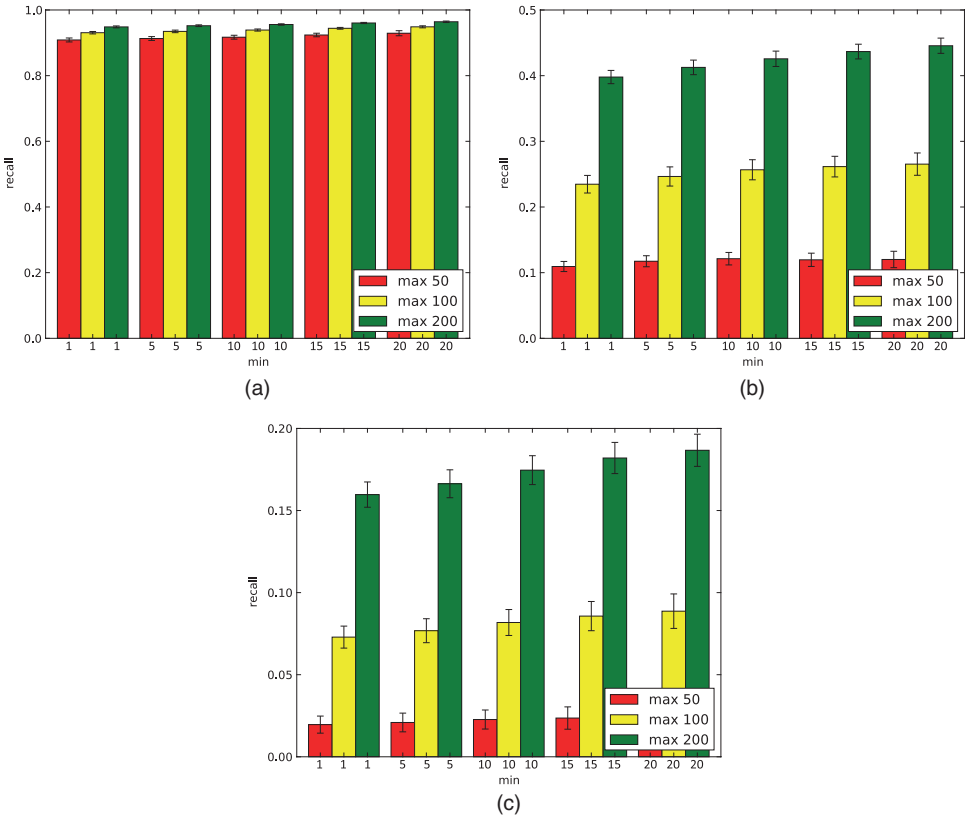


Figure 5 Recall of the location inference method considering [min, max] friends intervals

precision above 86% while intervals that consider up to 200 friends are close to a 80% precision. However, the average recall is lower than 3% for small intervals, and is on average 18% for bigger ones.

As in most contexts where data inference is required, the quality (precision) of the inference and recall moves in opposite directions. In this special case, considering that we only have location data for around 40% of Twitter users, a recall of 18% with an accuracy of 80% might really enrich data: this represents a 45% increase in the number of locatable users.

5 Conclusions and Future Work

Popular interest in various topics, such as sports events, breaking news, and political events among others, ranging from the trivial to the globally important, is currently being detected through large numbers of personal short messages disseminated through online social networks. As a result, there is much interest in assessing the repercussions of subjects related to daily activities by analyzing in real time what goes on in these social networks. For geography-related phenomena, knowing the approximate location of the

message sources is an important asset; however, the position of most users is not known, sometimes because the user has chosen not to declare it, and sometimes because locations provided by the user are intentionally cryptic or inaccurate.

In order to expand the availability of locatable messages from social networks, we devised a simple method to infer user locations based on the location of related users. In this article we applied this idea to Twitter users, so that in the future message streams can be filtered and analyzed for the geographic distribution of the instantaneous social impact of daily events. Our initial target is to know more about the repercussions of dengue fever events in Brazil, as part of our Observatory of the Web initiative, but the method described here can be applied to many similar problems. Naturally, as social networks evolve, it is possible that more people get motivated to publish their approximate locations, so that better inferences can be achieved and the resulting analyses can be more precise. Services such as the Observatory of the Web can be promoted as a motivation for users to collaborate in that direction.

Results show that our method can improve the amount of approximately locatable users by up to 45%, with reasonable confidence (recall of 18% with 80% accuracy over the initial figure of 40% locatable users). We demonstrated that users with too few or too many relationships can increase the location uncertainty, and that users with an intermediate number of friends provide better information for location inference. Given the experimental results, we expect these figures to improve if more users can be motivated to publish their location.

In the future, we want to implement more sophisticate methods for location inference, which can be based on Bayesian networks, spreading activation and fuzzy propagation in graphs. We also intend to extend this method towards other social networks, such as Facebook, in which friendship relationships can have a different meaning, and in which other types of message traffic can be analyzed. Associating the temporal series of messages to their geographic origin can also provide important clues regarding the propagation of subjects of popular interest, and give social scientists valuable insights as to the way online social networks behave.

Acknowledgments

This work was partially supported by CNPq, CAPES, Fapemig, and InWeb – Brazilian National Institute of Science and Technology for the Web.

References

Agarwal A, Rambow O, and Bhardwaj N 2009 Predicting interests of people on online social networks. In *Proceedings of the Twelfth IEEE International Conference on Computational Science and Engineering*, Vancouver, British Columbia: 735–40

Ahlers D and Boll S 2008 Retrieving address-based locations from the web. In *Proceedings of the Fifth International Workshop on Geographic Information Retrieval (GIR'08)*, Napa Valley, California: 27–34

Alencar R O and Davis Jr. C A 2011 Geotagging aided by topic detection with Wikipedia. In Geertman S, Reinhardt W, and Toppen F (eds) *Advancing Geoinformation Science for a Changing World*. Berlin, Springer Lecture Notes in Geoinformation and Cartography: 461–78

Alencar R O, Davis Jr. C A, and Goncalves M A 2010 Geographical classification of documents using evidence from Wikipedia. In *Proceedings of the Sixth Workshop on Geographic Information Retrieval (GIR'10)*, Zurich, Switzerland: 1–8

Backstrom L, Sun E, and Marlow C 2010 Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the Tenth International World Wide Web Conference (WWW '10)*, Hong Kong: 61–70

Borges K A V, Davis Jr. C A, Laender A H F, and Medeiros C B 2011 Ontology-driven discovery of geospatial evidence in web pages. *GeoInformatica* 16: in press

Borges K A V, Laender A H F, Medeiros C B, and Davis Jr. C A 2007 Discovering geographic locations in web pages using urban addresses. In *Proceedings of the Fourth International Workshop on Geographic Information Retrieval (GIR'07)*, Lisbon, Portugal

Buscaldi P and Rosso D 2007 A comparison of methods for the automatic identification of locations in Wikipedia. In *Proceedings of the Fourth International Workshop on Geographic Information Retrieval (GIR'07)*, Lisbon, Portugal: 89–91

Cardoso N, Silva M J, and Santos D 2008 Handling implicit geographic evidence for geographic information retrieval. In *Proceedings of the Seventeenth ACM Conference on Information and Knowledge Management (CIKM 2008)*, Napa Valley, California: 1383–84

Crandall D J, Backstrom L, Cosley D, Suri S, Huttenlocher D, and Kleinberg J 2010 Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences USA* 107: 22436–41

Delboni T M, Borges K A V, and Laender A H F 2005 Geographic web search based on positioning expressions. In *Proceedings of the Second International Workshop on Geographic Information Retrieval (GIR'05)*, Bremen, Germany: 61–64

Delboni T M, Borges K A V, Laender A H F, and Davis Jr. C A 2007 Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS* 11: 377–97

Goldberg D W, Wilson J P, and Knoblock C A 2007 From text to geographic coordinates: The current state of geocoding. *URISA Journal* 19(1): 33–46

Gomide J, Veloso A, Meira W Jr., Benevenuto F, Almeida V, Ferraz F, and Teixeira M 2011 Dengue surveillance based on a computational model of spatiotemporal locality of Twitter. In *Proceedings of the Third International Conference on Web Science (ACM WebSci'11)*, Koblenz, Germany

Gonzalez R, Rumín R C, Cuevas A, and Guerrero C 2011 Where are my followers? Understanding the locality effect in Twitter. In *Proceedings of CoRR Workshop on Service Oriented Computing*, Paphos, Greece

Goodchild M F and Hill L L 2008 Introduction to digital gazetteer research. *International Journal of Geographic Information Science* 22: 1039–44

Hall M M and Jones C B 2008 Evaluating field crisping methods for representing spatial prepositions. In *Proceedings of the Fifth International Workshop on Geographic Information Retrieval (GIR'08)*, Napa Valley, California: 9–10

He J, Chu W, and Liu Z 2006 Inferring privacy information from social networks. In Mehrotra S, Zeng D, Chen H, Thuraisingham B, and Wang F-Y (eds) *Intelligence and Security Informatics*. Berlin, Springer Lecture Notes in Computer Science Vol. 3975: 154–65

Janowicz K and Kessler C 2008 The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science* 22: 1129–57

Kwak H, Lee C, Park H, and Moon S 2010 What is Twitter, a social network or a news media? In *Proceedings of the Nineteenth International Conference on World Wide Web (WWW '10)*, Hong Kong: 591–600

Leidner J L 2004 Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the First International Workshop on Geographic Information Retrieval (GIR)*, Sheffield, United Kingdom

Lu L and Zhou T 2011 Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* 390: 1150–70

Lu Q and Getoor L 2003 Link-based classification. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, Washington, D.C.

Machado I M R, Alencar R O, Campos Jr. R O, and Davis Jr. C A 2010 An ontological gazetteer for geographic information retrieval. In *Proceedings of the Eleventh Brazilian Symposium on Geoinformatics*, Campos do Jordão (SP), Brazil: 21–32

Silva M J, Martins B, Chaves M, Cardoso N, and Afonso A P 2006 Adding geographic scopes to web resources. *Computers, Environment and Urban Systems* 30: 378–99

Sui D T 2008 The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems* 32: 1–5

Twaroch F A, Smart P D, and Jones C B 2008 Mining the web to detect place names. In *Proceedings of the Second International Workshop on Geographic Information Retrieval (GIR'08)*, Napa Valley, California: 43–44

Witten I H and Frank E 2005 *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (Second Edition). San Francisco, CA, Morgan Kaufmann

Ziegler C-N and Lausen G 2004 Spreading activation models for trust propagation. In *Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Services*, Taipei, Taiwan: 83–97

Zong W, Wu D, Sun A, Lim E, and Goh D H G 2005 On assigning place names to geographic related web pages. In *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado: 354–62