Active discovery of location-based twitter data

Rambabu Thakkalapelli

December 2016

Abstract
Social media platforms collect data from individuals who share their interests, activities, professional careers etc. Many social networking sites and communities help to bring people together based on their common interests. Our study focuses on the data collected from the social networking sites and tries to locate an individual geographically using relevant data. This research analyzes different methods and approaches for determining where an individual is geographically located based on his/her own social connections.

1 Introduction
Recent research shows that a large amount of data collected from social media from different public sectors such as health care could turn out to be helpful for the people if the data is mined correctly for the relevant information. Due to the large amounts of data, and the rate limits on the downloads imposed by the host sites, it is often not practical to process all of it. We discuss an approach towards the problem of discovering large amounts of data with a particular location by maximizing the amount of task-relevant data retrieved when the full data set cannot be searched.

Social science researchers discuss the importance and different approaches of associating the huge amount of social media data particular to geographic location, and to maximize the results. Our problem statement relates to locating and maximizing the amount of data collected for a specific geographic location. For a given amount of social media data, we are studying how to maximize the user data to a specific geographic location using social connections, graphs, coreness and degree of the graph etc.

The rest of this paper discusses the problem as follows: Section 2 provides the related work done so far. Section 3 describes the data used for the research and data extraction and analysis. Section 4 explains different graph features used to study the graphs and connectivity between the users. approaches and methods used in stating the approach for the problem i.e maximizing the identified data to a specific geographic location using network features. Section 5 explains the detailed study of the classification methods defined and evaluates and runs the defined problem approach. Section 6 discusses the results of the active learning
to boost performance in the research and Section 7 is the conclusions the future scope of the research.

2 Related Work

As studies show, most of the analysis of geographic location data from social media is Twitter-based. Due to the recent changes in policies of information sharing, the data that is freely available is comparatively low in amount and not very detailed. There is a lot of research going on regarding the procedure of how to get the relevant information from the large amount of data when the data cannot be searched completely. We stick to the location information i.e city, state, country etc and try to find other data/users to geotag to that particular location.

Davis Jr et al. (2011) studied the problem of active search, using the social connections of significant users in a particular area. They described the approach using a voting algorithm, in view of the majority of each user’s friends’ locations (with friends characterized as sets of Twitter users who follow each other) started with 5 random users as sources and finished with 61,400 users. They discussed that the users that have location data are 40 percent of the given users. The described study results show that the technique enhances the ability to locate the user by upto 45 percent.

Chen et al. (2013) approache the problem of estimating the location based on the interest of the users. The approach is described in three phases: interest detection, mapping from the location function(construction of hidden relationships between users interest) to interest, and location estimation. The results show that the model is independent of language but may vary based on culture or user behavior in different regions.

Chandra et al. (2011) describes the experimental approach in two models. The first model studies the probability of each term from all tweets assuming that tweets from each user belong to the city of that user. The second model studies the terms from all the responses to each user’s tweets and weighs the conversations between different users. The results show that the average error distance using the first and second model are estimated as 1343.17 miles and 1044.28 miles.

Hecht et al. (2011) studied two classifiers, ‘Count’ and ‘Calgari’. ‘Count’ helps to identify the number of times each word occurs in a certain set of tweets, and uses the multinomial naive bayes model. ‘Calgari’ is a variable that functions to show that people from a given location are more likely to use certain words as compared to the general population. The study discovered that 72.7 percent of the times, the first model can predict the country of the Twitter user. This is possible when the data is distributed evenly over the countries and states. The test results shows that, the performance of first classifier was 4 percent better than that of second model. For the second test, using random data to predict the country, the second model’s accuracy was 88.6 percent. In the second test, ‘Calgari’ results were 16 percent higher than that of ‘count’. In the
third test, user’s are classified using the their respective state for uniform data. In this test second model achieved the accuracy of 30.28 percent. In the fourth test, users were classified based on their state, but the data was randomised, the accuracy was 27.31 percent for the first classifier.

Sadilek et al. (2012) shows that the friendship prediction works on the similarities between two users, and identifying the amount of overlapping data between the two users. The overlapping data is defined in various terms like time and locations of the user’s tweet; the vocabulary from the tweets, and their friends’ lists. The location prediction uses dynamic Bayesian network that works on two forms of nodes. The first node gives the users’ friends’ location, and the second node gives the location of the user. Even though the Bayesian network model performed better than the others, the conclusion states that it is hard to use geographic distance alone to model social relationships.

Cheng et al. (2010) estimates user’s city location using a probabilistic framework based on the user’s tweets, without having any information about the geographic location. The proposed approach describes having three features: (i) This approach uses tweet content rather than location information such as IP information or private log-in information; (ii) A classification component that helps in identifying tweets based on a particular location; (iii) a lattice-based neighbor approach in estimating the location of the user in descending order of confidence. The result states that 51 percent of studied Twitter users’ location estimates to within 100 miles of their actual location.

3 Data

We collected data focusing on the specific twitter users (as a problem to approach) i.e students at our university. We collected the data with our University registrar’s help as they were interested in learning from the social media about student satisfaction. Synergistically, their interest in our approach towards the problem motivated our approach and interest in the research.

Figure 1 describes the work flow of our experiments we retrieved from maximizing social media data specific geographic location.

For this data, each tweet is extracted in json format that contains fields indicating screen name of the user, whether a tweet is a retweet, number of retweets for a particular tweet, text of the tweet, date and time of a given tweet, language of the tweet, latitude and longitude (if enabled) of the tweet, tagged users in a text of the tweet, number of followers and friends for a user and the tweet, among other information.

We collected the data in the following two phases.

3.1 Phase I

we collected data from the Twitter API Makice (2009), using following queries as the streaming API allows
Figure 1: Work flow of the experiments described in this paper.

- All tweets collected from within a given geographical location that surrounds our university. The Twitter API streams the tweets that: (a) were geotagged within this radius of location OR (b) the users that declare their home region in the given location that falls in the radius of our university location. As the radius varies the downloaded data and the tweets that contains satisfying (a) or (b).

- All tweets that are related to our university; that are tagged such as “our-university” or “#our-university our-mascot,” or #our-university.”

We collected 300 distinct user profiles with the tweets in this phase. The collected user profile consists of some users that are not related to our university. To make sure we get only user profiles from the university especially students, we examined the profiles manually using following terms. We started rating the users. Since this is a subjective task, we used following rules to make it a consistent and fair process.

- The user profile that says the user is currently enrolled student with our university.
- The user had current pictures of the events at university’s campus.
- The user tweets about the current events or about the weather conditions or classes being cancelled etc.
• The user talks about the fest or about tiger pride etc.

• The profile that stated the user were a student of our university.

• The user wearing our university gear in their pictures or sharing any kind of university pictures in the tweets.

• The user following our university’s official pages such as RITTigerPride, BrickCityHomeComing, SpringFest, OrientationWeek etc.

• The user having many friends that are self-identified as the students from our university.

• Users having their own websites that link up using university’s accounts or user profiles that have university mail id as the account contact mail.

3.2 Phase II

We continued to collect data for another month profiles from the two Twitter Makice (2009) streams described above. We then added to set the profiles of all users mentioned in any of the tweets from those users labeled as students in Phase I. This new list yielded 2,705 new profiles, from which we sampled 745 to label, using the same process as in Phase I.

We built a meteor web application to label the collected data. We used the following described 'Rate User Location' meteor web application to label the sampled 745 twitter users.

3.2.1 Building Meteor Application

To build the 'meteor' Strack (2012) application for the project, we installed 'meteor' and added required packages such as ‘npm’ etc. Figure 2 is the developed meteor web application to rate the twitter users. The application contains a text field to enter the name of the user in order to do the survey of rating the twitter users based on the location. We used a CSV file to upload the required data of the user-names with five radio buttons to rate the given user from 1 to 5 rating the location to a specific area. We used 'Mongo' Strack (2012) collection to store the user-names in the database. Each user-name displayed in the application contains a hyperlink and when clicked navigates to the new window and opens the twitter account of that particular user. We defined ratings as definitely not RIT student, probably not RIT student, cannot tell, probably RIT student, definitely RIT student. This information is used to rate the location of the user.

We rated every twitter user on a scale of 1 to 5 using the metrics definitely not RIT student, probably not RIT student, cannot tell, probably RIT student and definitely RIT student. We scaled the ratings to ‘s’ and ‘n’ using 1 labeled as ‘s’ and 2,3,4,5 are labeled as ‘n’ respectively.

Initially we rated 745 twitter users and compared the results from the annotation program using old ratings and new ratings. The old ratings consist of
"s" and "n" where as the new ratings are ‘1’ to ‘5’ which later scaled to the old labels for comparison.

Old labels data showed 211 of the labeled profiles as RIT students and 534 as non-students. New labels data showed 165 of the labeled profiles as RIT students and 580 as non-students.

### 3.2.2 Cohen’s Kappa - Inter Annotation

We worked with an inter-rater Kulkarni et al. (2009) agreement program for multiple annotators using the Cohen’s Kappa algorithm. We used ratings as categories and twitter users as the subjects in the annotation program. The input to the Cohen’s Kappa algorithm is the matrix form of the subjects and categories. The output of the program is the Kappa value that tells the percentage of the agreement of the raters in rating the twitter users.

Cohen’s Kappa Berry and Mielke (1988) "is a measure of agreement between two raters who each classify N items into M mutually exclusive categories. It is calculated using the relative observed agreement among raters and the hypothetical probability of chance of agreement. The hypothetical probability is calculated using the observed data to calculate the probabilities of each observer randomly choosing each category”.

We used the old and new labels of the 745 twitter users. Two individuals rated the old and new label data respectively. Hence the Kappa value is calculated for two raters that rated 745 twitter users for 2 categories. Figure 3 describes
the results of Kappa value calculated using old and new labels.

Figure 3: Results - Kappa calculation for 745 twitter users labels by 2 raters

4 Baseline Classification Using Network Features

We analyze our network-analysis Andre Lobato Ramos (2015) based approach for discovering new users from a particular geographic region. Using at-graph we find other users that are connected to a particular user that has connection with "@" approach. We analyzed the at-graph labeled users that have user connections that are labeled as students using following graph features.

Figure 4: Andre Lobato Ramos (2015) At-Graph showing all labeled users and their connections - Andre Lobato Ramos work.
Using Active Learning to Boost Performance

In this section, we follow the process Andre Lobato Ramos (2015) of adding nodes greedily to identify how likely they are to be students.

- Consider a baseline at-graph 'G' that has the training set that contains all the users identified as students.
- Consider a test set to be 'T'
- Consider p to be equal to |T|.
- From the graph G, consider the fraction of users that are labeled as not a student as 'f'
• condition check and repeat when \(|P| > f\),
  
  – Consider a logistic regressor \(m\) and train it on \(G\).
  – consider \(\hat{k}\) is equal to \(\text{arg max}\ \{m(k) | k \in T\}\).
  – Add \(\hat{k}\) to \(G\).
  – Remove \(k\) from \(T\).

6 Results

We used logistic regression here, rather than the other learning models, because it alone had a non-trivial ROC curve to work with. The experiments are repeated 6 times and all the averages from the experiments are calculated respectively. Figure 8 explains the calculated average Precision, Figure 9 explains the calculated average Recall and Figure 10 explains the calculated average F1 score of logistic regression, single decision tree, gradient boosted trees, and Random forest labeled twitter users data in increasing the number of labeled users in 500 scale. Figure 11 shows the ROC curve for the logistic regression using the 2704 labeled twitter users data.

![Graph features showing for 1000 number of users for different classifiers.](image-url)
Figure 7: Precision, Recall and F1 score of logistic regression, single decision tree, gradient boosted trees, and Random forest. for 1000 number of users.

Figure 8: Precision of logistic regression, single decision tree, gradient boosted trees, and Random forest. Every unit is 500 no. of users in x axis.
Figure 9: Recall of (left-to-right) logistic regression, single decision tree, gradient boosted trees, and Random forest. Every unit is 500 no. of users in x axis.
Figure 10: F1 score of (left-to-right) logistic regression, single decision tree, gradient boosted trees, and Random forest. Every unit is 500 no. of users in x axis.

Figure 11: The ROC curve for the logistic regression (AUC=.8).
7 Conclusion

We performed our analysis on the twitter users and collected data from the twitter using the geographical location and graph properties. Our results are preliminary and shows that greedily adding user friends/connections as nodes to the graph improved the precision with high false negatives. The results are achieved by retraining the classifier with the greedy added nodes. We measured the average values of the precision, recall and f1 scores for all the 2704 labeled users and graphs are plotted for increase in 500 users interval. The results are not so satisfactory but can be improved with the addition of more relevant data.

In future similar research work can be done using Reddit data. The labeling of the data can also be done in various scales. The data collected from the streaming API’s can be filtered with more improved filters/queries. For the classification of data, in future the scope can be improved using artificial neural networks. The meteor web application can be updated so that when the data is labeled, the classification tasks can be linked to the application and the results and graphs might be shown on the application.

References


