# Estimating Twitter User Location Using Social Interactions – A Content Based Approach

Swarup Chandra, Latifur Khan
Department of Computer Science
University of Texas at Dallas
swarup.chandra@utdallas.edu

Fahad Bin Muhaya
College of Business Administration
King Saud University
fmuhaya@ksu.edu.sa

*Abstract*—**Microblogging services such as Twitter allow users to interact with each other by forming a social network. The interaction between users in a social network group forms a dialogue or discussion. A typical dialogue between users involves a set of topics. We make the assumption that this set of topics remains constant throughout the conversation. Using this model of social interaction between users in the Twitter social network, along with content-derived location information, we employ a probabilistic framework to estimate the city-level location of a Twitter user, based on the content of the tweets in their dialogues, using reply-tweet messages. We estimate the city-level user location based purely on the content of the tweets, which may include reply-tweet information, without the use of any external information, such as a gazetteer, IP information etc. The current framework for estimating user location does not consider the underlying social interaction, i.e. the structure of interactions between the users. In this paper, we calculate a baseline probability estimate of the distribution of words used by a user. This distribution is formed by using the fact that terms used in the tweets of a certain discussion may be related to the location information of the user initiating the discussion. We also estimate the top K probable cities for a given user and measure the accuracy. We find that our baseline estimation yields an accuracy higher that the 10% accuracy of the current state of the art estimation.**

**Keywords – Twitter, Content-Based Location Estimation, Social Interaction, Data Mining**

## I. INTRODUCTION

Internet social media services, such as microblogging and social networking, which are offered by platforms such as Twitter, have seen phenomenal growth in their user bases. Java et al. [11] described this microblogging phenomenon as early as 2007, noting that users use the service to talk about their daily activities and to seek or share information. This growth has spurred an interest in using the data provided by these platforms for extracting certain information, such as geographic location, from its users. The information obtained can be used to provide users with personalized services, such as local news, local advertisements, application sharing etc. With more than 200 million accounts on Twitter in diverse geographical locations, the short messages or tweets form a huge data set that can be analyzed to extract such geographic information.

Twitter allows its users to specify their geographical location as user information (Meta Data). This location information is manually entered by the user or updated with a GPS enabled device. The feature to update the user location with a GPS enabled device has not been adopted by a significant number of users [2]. Hence, this geographic location data for most users may be missing or incorrect.

There are several drawbacks to relying on a user's manual update of location.

- Users may have incorrect geographic location data. For example, a Twitter user may enter his/her location as 'Krypton'. This may not be the name of a real geographic location.
- Users may not always have a city-level location. Users can input location names vaguely, such as the name of a state (E.g. Arizona) or the name of a country (E.g. U.S). These location names cannot be directly used in determining the city-level location of the user.
- Users may have multiple locations. If a user travels to different locations, he/she might mention more than one location in the Meta-Data of his/her Twitter page. This makes it very difficult to determine their current, singular city-level geographic location.
- Users may have incomplete location data. A user may have specified an ambiguous name that may refer to different locations. For example, if a user specifies a location such as 'Washington', this name can be related to a state name or a city name (Washington D.C). These types of ambiguous names make it difficult to determine the exact user location.

Therefore, the reliability of such data for determining a city-level geographic location of a user is low. To overcome this problem of sparsely available geo-location information of users, we evaluate the Twitter user's city-level geographic location based purely on his/her tweet content along with the content of the related reply-tweet messages. We use a probabilistic framework that considers a distribution of terms used in the tweet messages of a certain conversation containing reply-tweet messages, initiated by the user.

### *A. Motivation:*

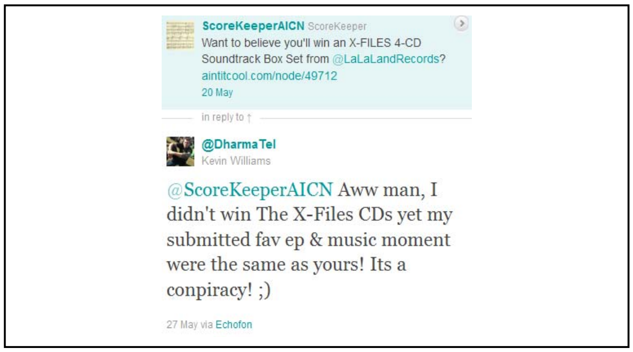
On Twitter, users can post microblogs known as tweets, which can be read by other users. Along with this microblogging service, Twitter also provides a social networking service where a user *(follower)* can "*follow*" another user (*followee*). Each edge of the social network is formed by this "*follow*" relationship. As a *follower*, a user receives all the tweets posted by the *followee*, and in turn can reply to these tweets with a reply-tweet. This reply-tweet is received by the *followee* from the *follower*. This forms the

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

basis of a conversation between two different users. Huberman et al. [14] analyzed more than 300 thousand users and found that reply-tweets and directed tweets constitute about 25.4 percent of all posts on Twitter. This shows that the reply-tweet feature is used widely among Twitter users.

Our intuition is that a conversation between users can be related to a set of topics such as weather, sports, etc., including certain location-specific topics, such as an event related to a city, or a reference to a specific place or an entity in a city. We assume this set of topics remains constant during a conversation. When a user posts a tweet message, it can be seen as the start of a conversation. This conversation can continue when another user posts a reply-tweet to the original tweet. Without detailing the topic of the reply-tweet, it can be assumed that the topic is the same as the original tweet message. Under this assumption, any content of the reply-tweet can be related back to the topic of the original tweet message. For example, consider the tweet messages exchanged by two users in Figure 1. A user posts a tweet message, and another user replies back with a reply-tweet message. Note that the topic of conversation remains the same during the conversation.

So, by combining (1) the above assumption, with (2) the use of tweet content that may have location-specific data, we should obtain a better result than if we considered tweets in isolation, or if we just relied on user-specified location.



ScoreKeeperAICN ScoreKeeper
Want to believe you'll win an X-FILES 4-CD Soundtrack Box Set from @LaLaLandRecords? aintitcool.com/node/49712
20 May

in reply to ↑

@DharmaTel
Kevin Williams

@ScoreKeeperAICN Aww man, I didn't win The X-Files CDs yet my submitted fav ep & music moment were the same as yours! Its a conpiracy! ;)

27 May via Echofon

Figure 1. Initial Tweet and a Reply-Tweet.

### B. *Challenges:*

The use of pure tweet content for estimating the Twitter user location, along with the above mentioned intuition, presents some challenges. These challenges are based on the semantic complexities of the natural language used in tweets.

Some users may use non-standard vocabulary in their tweets. Users from a city may refer to the same location-specific entity with different names. For example, a user from Los Angeles can refer to the name of the city as LA, L.A., or City of Los Angeles, etc. Users may also refer to different locations with the same name. For example, a user from New York can refer to 6th Street as a street name in New York, whereas a user from Austin may refer to the street with the same name in Austin. These examples can dilute the spatial distribution of the terms.

The tweets do not always contain location-specific terms. They may contain a lot of general words from a natural language, as users tweet about general topics in their daily life. Hence, the content of the tweets are considered noisy.

The users may also use different languages to communicate. Twitter allows use of multiple languages. This greatly complicates the estimation of location. Additionally, the tweets may have English (Roman) letters but may not be related to any English word. So, if we choose to consider only tweets in English, the tweets with this merely-English-looking characteristic will be extremely difficult to eliminate. Such tweets will add to the noise of the tweet content.

A tweet can have terms referencing to multiple locations. This reduces the ability to estimate a specific location of the user. When considering a conversation, the topic of a conversation may not remain the same throughout the conversation, as assumed. A change of topic in a reply-tweet may result in multiple location specific terms or dilution of the spatial distribution of terms.

Taking note of these challenges, we use a probabilistic framework to estimate the city-level Twitter user location for each user. The probabilities are based on the contents of the tweet messages with the aid of reply-tweet messages generated from the interaction between different users in the Twitter social network. The use of reply-tweet messages provides better association of words with the user location, thus reducing the noise in the spatial distribution of terms. We also provide the top K list of most probable cities for each user. We find that our estimation of the user location is within 100 miles of the actual user location 22% of the time, as compared to the previous work which had an accuracy of about 10%, using a similar probabilistic framework.

This paper has been organized into the following sections. In section II, we list some related works and describe the difference between our work and that of the others. In Section III, we briefly describe the data set used for the experiment, and we describe the details of the metrics used for measuring the outcome of the experiment. In Section IV, we describe in detail, the intuition and algorithm. In Section V, we describe the details of the experimental setup. In Section VI, we discuss the results obtained from the experiments. Lastly, in Section VII, we conclude the paper with possible future enhancements.

## II. Related Works

The geographic location estimation problem has been studied extensively by researchers who propose various ways to extract user location information from internet social media platforms. These social media platforms include web pages [3] and blogs [6] etc. These works rely on external resources such as gazetteers and databases, to identify the related geographical information. In our work, we do not use any external resource to estimate the geographic location of the user. Also, the work in [13] studies the variation of language usage on Twitter. This can also be used to augment our work to improve the accuracy of predicting user geographic location.

There have been works on: relations between geotags [4], geo-location estimation in search engine query logs [7], user privacy of geotags [8], predicting geographic location on proximity [1], and a study of private information trials [9] using correlations between different publicly available pieces of information to extract private information about a person. A recent work studied the use of location sharing services by users of platforms like Foursquare [10]. They use the user 'Check-In' information to study the mobility characteristics of the users. Another recent work involves location prediction of Twitter users based on his/her social network [5]. The authors mine implicit attributes associated with the user in his/her social network, and predict the user location based on these attributes. Note that this work does not estimate location based purely on content; it uses additional knowledge to evaluate the user's location. These works can be used to augment our work in estimating the geographic location of a user.

The most relevant related work is the content-based approach proposed by Z. Cheng et al. [2] to estimate the geo-location of a Twitter user. In our work, we use a variation of their probabilistic framework. In their approach to the problem of estimating geo-location of a Twitter user, they consider a set of tweets from a set of users belonging to a set of cities across the United States. They estimate the probability distribution of terms used in these tweets, across the cities considered in their data set. This probability distribution is then used to estimate the user location, given the set of terms used by a user, in his/her tweets. However, in their approach toward calculating the probability distribution of terms from the user's tweets, they ignore the relationship between different tweet messages, such as reply-tweet messages. This failure to use such relationship information has an impact on the distribution of terms across the cities considered. They consider the tweet messages as independent entities. In our work, we use the relationship between tweets and related reply-tweets to evaluate the spatial distribution of terms, in order to increase accuracy in estimating the geographic location of the user.

Z. Cheng et al. [2] also describe major refinements to the baseline probability estimation. These major refinements involve use of a model to evaluate spatial variations of terms and to obtain term localization with respect to geographic location. This model has been proposed by L. Backstrom et al. [7] for the geo-location estimation in search engine query logs. The model uses the baseline probability distribution, which is augmented with the model parameters for each word in the distribution. This model can be used to augment our probability distribution model as well.

## III. Data Set and Metrics for evaluation

The data set used for evaluating the probabilistic model is a part of the data set used in [2]. The data set contains only users within the United States. The training set is filtered for valid city name and state name. Other hierarchical addressing is discarded. The test set contains users having more than 1000 tweets. For our experiment, 10,584 training users were considered with 500,000 tweet messages, and 540 test users were considered with 600,000 tweet messages.

For evaluating the quality of the results obtained, we used the following metrics:

i) Error Distance (*ErrDist*): Distance in miles between the actual location and the estimated location

$$ErrDist_i = Estimated\ Location_i - Actual\ Location_i$$

ii) Average Error Distance (*AvgErrDist*): Average of the error distances for all users in the data set.

$$AvgErrDist = (\ \textstyle\sum_{i \in users} (ErrDist_i)\ )\ /\ n$$

iii) Accuracy (*Acc*): Percentage of users with error distance within 100 miles.

$$Acc = |\ \{i \mid i \in users \wedge ErrDist_i \leq 100\}\ |\ /\ n$$

Here, $n$ is the number of users in the data set.

## IV. Location Estimation

As noted earlier, extracting geographic location specific information from the tweet content alone is challenging. With the social interaction model, we use the content of the tweets in any interaction between users, to determine the probability distribution of terms used during the conversation.

### A. Probability Distribution Model (PDM)

This probability distribution technique is as follows. We assume that each user belong to a particular city, and thus his/her tweets also belong to that city. That is, the terms occurring in the user's tweet can be assigned as terms related to the user's city. This forms the basic distribution of terms for the set of cities considered in the complete data set. The probability distribution of term $t$ over the entire data set, for each city $c$, is given as

$$p\ (t \mid c) = |\ \{t \mid t \in terms \wedge t \text{ occurs in city } c\}\ |\ /\ |\ t\ |$$

That is, the number of occurrences of term $t$ for a city $c$ divided by the total occurrences of the term $t$ in the entire dataset. A probability distribution matrix of size $n \times m$ is formed, where $n$ is the size of the term list, i.e., the size of the dictionary, and $m$ is the total number of cities in the data set that are considered for evaluation.

### B. Reply Based Probability Distribution Model (RBPDM)

In the basic calculation of the PDM, the terms in a user's tweet are assigned to the city to which the user belongs. It does not consider the relation between different tweet messages.

Twitter offers a feature, called a *reply-tag*, to tag another user in a tweet. This tag directs the message to the user who is addressed in the tweet. With this in mind, a tweet message can be classified into three different types.

- The first type of tweet message is a general message that a user typically posts on Twitter. These tweet messages do not contain any reply-tag. The terms used in this type of tweet message can be used to form a direct relation to the user's city in evaluating the spatial distribution of terms.
- The second kind of tweet message is one that contains a reply-tag. This type of message, called a *reply-tweet*, is generally used to reply to a certain tweet posted by another user. The reply tweet message will be directed to the user who posted the original tweet message. This tweet will generally contain the reply tag at the beginning of the tweet message.

- The third type of tweet message is one that has a reply-tag, but is not a reply-tweet. It may be a tweet message that is directed to a user, but it need not be a reply to a previous tweet from that user. This message generally may contain the reply-tag in-between the tweet words. It can also be a re-tweet where a user re-posts the tweet message of another user so that his/her followers can receive the tweet message.

The relationship between two tweet messages occurs when the reply-tag in a tweet is taken into consideration. The reply-tweet will have a direct relationship with the original tweet that inspired the reply-tweet message. The PDM distribution ignores all these relationship between tweet messages.

Here, we consider this relationship between different tweet messages while calculating the probability distribution of terms from the data set. This relationship forms the basis of a conversation between different users, i.e., a tweet message and its reply-tweet messages can be considered as a dialogue between the users. So, by making the assumption that the topic of a conversation remains constant throughout the relevant reply-tweet messages, we can relate the terms to the topic of the conversation. The conversation may involve location-specific terms related to the topic. Instead of plainly assigning the terms used in the tweet to the user who posted the tweet, the terms occurring in the complete conversation can be assigned to the user who initiated the conversation since the initiator may initiate a conversation topic involving his/her geographic location. Thus, when a reply-tweet is encountered in the data set, we assign the terms involved in that tweet to the recipient of the reply-tweet rather than to the user who posted the reply-tweet message. With this assignment of terms to different users, and in turn, to different cities to which the user belongs, we evaluate the probability distribution that recognizes the different types of tweet messages and the relationships between them. Hence, the social structure of the network is considered while estimating the geographic location of a user in the Twitter social network. The formal algorithm is given in Figure 2.

The input for Algorithm 1 is a list of cities, a list of users, and a list of tweet messages considered in the training data set. Line 1 in the algorithm considers each tweet in the training set. The tweet contains a set of words. These words are normalized in line 2. The main difference between the PDM and RBPDM is in lines 3 to line 7, where we form a Posting List structure (contains the term frequency and total frequency) to determine the statistics required to calculate the probability in line 10. Here, we check for reply messages, which are tweet messages that begin with a symbol @, followed by a user screen name. In line 2 of the algorithm, the tweet considered is that of *userA*, for instance. We check whether the tweet is a reply-tweet in line 3. If it is, we update the posting list structure instance for *userB*, who is tagged in the reply-tweet; else, we update the posting list structure instance for *userA*, who posted the tweet message. This is explained in detail in the next sub-section. Line 8 builds a dictionary that is later used to evaluate the baseline probability estimate, which is explained in the next paragraph. After evaluating all the tweet messages considered in the training set, in line 10 we obtain the distribution matrix (*ReplyDistribution*), which is of size $c \times w$, were $c$ is the size of the *CityList* and $w$ is the size of the *Dictionary*. The *EvalDistribution* function used to evaluate the distribution matrix uses the equation shown in the PDM sub-section to evaluate the probability of a term $t$ given a city $c$. This subroutine is shown in Figure 3.

Algorithm 1 : Reply Based Probability Distribution Model

**Input**:
*CityList*: List of cities occurring in the Data Set.
*Tweets*: List of tweet messages considered in the Training Set.
*Users*: List of users occurring in the Training Set.
**Output**:
Probability Distribution of terms across *cities* ∈ *CityList*

```
1.  for each  $tweet \in Tweets$
2.      $Terms$ = Normalize words in $tweet$ from
                 $userA \in Users$
3.      if $tweet$ begin with '@$userB$'
4.          $PostingList(Terms, userB)$
5.      else
6.          $PostingList(Terms, userA)$
7.      end if
8.      Update $Dictionary$ with $Terms \notin Dictionary$
9.  end for loop
10. $ReplyDistribution = EvalDistribution(PostingList, CityList, Dictionary, Users)$
11. Return $ReplyDistribution$
```

Figure 2. Algorithm for Probability Distribution Estimation using RBPDM Technique.

Algorithm 2 : EvalDistribution Function

**Input**:
*PostingList*: List of users with term frequency, for each term in the dictionary, and total frequency
*CityList*: List of cities occurring in the Data Set.
*Dictionary*: List of distinct terms obtained from the Data Set.
*Users*: List of users occurring in the Training Set.
**Output**:
Probability Distribution of terms across *cities* ∈ *CityList*

```
for each  $term \in Dictionary$
    for each $city \in CityList$
        $Distribution[term][city] = 0$
        for each $user \in Users$ located in $city$
            $Distribution[term][city]$ +=
                $PostingList.term.user.termFrequency$
        end for loop
        $Distribution[term][city]$ /=
            $PostingList.term.totalFrequency$
    end for loop
end for loop
Return $Distribution$
```

Figure 3. Algorithm for evaluating the Probability distribution (*EvalDistribution*) given in Algorithm 1.

The input for Algorithm 2 is: the *PostingList*, which is formed in Algorithm 1 and contains (1) the list of users for each *term* *Dictionary*, where each *term* is used by the user in his/her tweet, and (2) the corresponding term frequency (Number of occurrences of the *term* for each user) and (3) the total frequency for the *term* (Number of occurrences of the *term* in the total data set); the *CityList*, which is the list of all cities occurring the training data set; the *Dictionary*, which is a list of all distinct terms appearing in the tweet messages in the training set; and *Users*, which is the list of users in the training set. Lines 1 and 2 in the algorithm show the distribution evaluation for each term in the dictionary, for each city. Line 3 initializes the distribution value. In line 4, we consider all the users whose location is the city under consideration. The sum, for all users in the city, of the relevant term frequencies is obtained in line 5. In line 7, the probability for a term given a city, is calculated by dividing the city-wide sum of term frequencies obtained in line 5, by the total frequency of the term in the entire collection. Finally, the whole probability distribution is returned in line 10.

The running time of the complete algorithm to obtain the distribution is in the order of: the number of terms in all the tweet messages, plus the product of the dictionary size, the city list size, and the user list size (per city), considered in the training set. Note that this is the same complexity as that of the algorithm used in [2] for calculating the distribution using PDM.

### C. Term Distribution Estimator

Using the distribution of terms across the cities considered in the data set, the probability of a city $c$ given a term $t$ can be calculated based on maximum likelihood estimation.

$$p(c \mid t) = \max_{\forall c \quad cities} p(\mathrm{t} \mid c)$$

The probability estimate of the user $u$ being located in city $c$ is the total probability of the terms extracted from the user tweet for the city $c$, i.e.

$$p(c \mid u) = \sum\nolimits_{(w \quad terms)} p(c \mid w) * p(w)$$

Using this equation, the probability estimation matrix is obtained by using Algorithm 1 mentioned in [2], which has size $p \times q$, where $p$ is the size of the user list being considered, and $q$ is the size of the city list being considered in the data set. The city-level geographic location estimation is made by selecting the city with the highest probability for that user. A list of top K estimated cities can also be obtained by sorting the probability estimation matrix for each user, and listing the top K most probable cities from it.

## V. Experiment

In order to compare the results of PDM and RBPDM, we use the two probability distribution approaches given in the previous section to estimate the city-level geographic location of a Twitter user. From the data set considered, we use the training set to obtain the probability distribution of words across the cities considered within the United States. We use an experimental setup similar to the one used in [2], in order to provide a comparison between our result and their reported result, for the baseline approach. We obtained the user list and the corresponding locations after parsing through the list of tweet messages obtained for the training set. For tweet term normalization, we do not consider the occurrence of stop words (a standard list of 432), punctuations, progressives, word case (all words are made lower case), or hyperlinks. We also use the Jaccard Coefficient method to check for variations in words encountered and to eliminate these variations. Since tweets typically contain non-standard language, the Jaccard Coefficient method helps to identify the variation better than any stemming technique.

As noted earlier, while encountering a reply-tweet, the terms (after normalization) occurring in the tweets are assigned to the user who is being addressed in the tweet, rather than the user who posts the tweet. E.g. consider the following tweet encountered in the dataset. "*Adam: @SarahM damn good burger!*" Here a user with screen name *Adam*, posts a tweet addressed to a user with screen name *SarahM.* This is a reply-tweet. The terms occurring in this tweet are assigned to *SarahM* rather than to *Adam*. In this experiment, we consider conversations that have a direct reply-tweet to the original message. Any reply-tweet to a reply-tweet is considered part of a conversation with the user who is addressed in the reply-tag. This assumption is made in order to address the change of topic during the conversation.

The probability distribution was obtained for both the PDM term distribution technique and the RBPDM term distribution technique. The term distribution estimator algorithm was thus executed for both these probability distributions, to obtain a list of estimated top K cities, for each user in the test set.

## VI. Experimental Results

The goal of the experiment was to study the behavior of the spatial distribution of words, in order to estimate the geographical location of a Twitter user, by using the information provided by the underlying social network.

From the experimental setup, we obtained a list of 116412 distinct words. We found that about 10.78% of the 540 users in the test set were assigned an estimated location within 100 miles of their actual location, when using the PDM method with the term distribution estimator. The average error distance with this estimation was found to be 1343.17 miles. However, by using the RBPDM method with the term distribution estimator, we found that about 22.01% of the same 540 users were assigned a location within 100 miles to their actual location. The average error distance was found to be 1044.28 miles. The variation in accuracy between the two methods, with respect to the error distance, is shown in Figure 4 (K=1). The better performance of RBPDM against PDM can be attributed to the assumption that during a conversation between two users, the topic of the conversation remains the same. For example, with the error distance of 300 miles, the estimator with PDM yielded an accuracy of 15.86%, and that of RBPDM yielded an accuracy of 29.63%. Every tweet conversation can be viewed according to the spatial distribution of its terms. So, the main work of determining the conversation initiator's geographic location was captured when establishing the spatial distribution of all terms during the estimation of the probability distribution.

The results given above are with the restriction that the accuracy was calculated for the top location (K=1) that was predicted by the estimator. By relaxing such a restriction, we can have a better understanding of the spatial distribution of the terms when estimating the user location with (RBPDM) rather than with (PDM), because we account for the information provided by the underlying social structure of interaction between users. The variation in estimator accuracy, with respect to the change in error distance, for the top K estimated cities, can be seen in Figure 4 (with K=2 and K=5), i.e., the actual location of a user appears in the top K cities of the estimated city list. It can be seen that the term distribution estimators using our method (RBPDM) performed significantly better than those using the PDM method, while considering the top two cities and the top five cities in the estimated list (Figure 4). For instance, in Figure 4, using RBPDM, the accuracy obtained by the estimator with K=2 was 30.926% with the error distance of 100 miles, as compared to the corresponding accuracy of 16.694% for PDM. Similarly, with K=5, the accuracy for RBPDM was 58.88% with an error distance of 300 miles, and that of PDM had an accuracy of 48.58%. This indicates that with an increase in K, accuracy increases with both methods, but the prediction of the estimator using the RBPDM technique is consistently better than that of the estimator using the PDM technique.

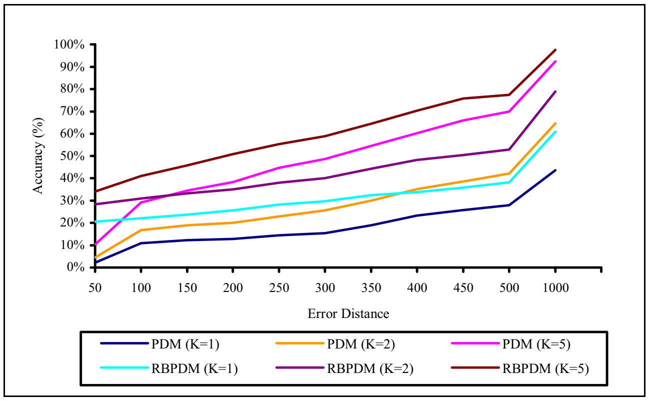


Figure 4. Accuracy vs. Error Distance Increase

On a negative note, the amount of data used in this experiment, to estimate the probability of a user's geographic location, was significantly more limited than the amount used in [2].

## VII. Conclusion

The experiment performed in this paper provides good insight into the problem of estimating a user's geographic location information purely based on the content of the user's publicly available information, while making use of the characteristics of the Twitter communication model. We use the concept of user interactions on Twitter and exploit the relationship between of different tweet message types. From the experimental results, we conclude that associating the tweet content of a conversation, containing reply-tweets, with the initial tweet's user's location (to obtain a spatial distribution of terms), improves the accuracy of estimating a user location.

Further, the quality of this work can be refined by considering a larger data set that takes into account the reply messages for a given tweet. We would also like to see further improvements by combining the information in the underlying social network with additional information, to obtain a more accurate prediction of user location in a social network environment.

## Acknowledgment

We thank Satyen Abrol for his valuable suggestions and his help in conducting the experiment.

## References

[1] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In WWW, 2010.

[2] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010

[3] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In SIGIR, 2004.

[4] S. Lee, D. Won and D. McLeod. Tag-geotag correlation in social network. In SSM `08 Proceeding of the 2008 ACM workshop on Search in social media.

[5] S. Abrol and L. Khan. TweetHood: Agglomorative Clustering on Fuzzy k-Closest Friends with Variable depth for Location Mining. SocialCom/PASSAT 2010: 153-160

[6] C. Fink, C. Piatko, J. Mayfield, T. Finin, and J. Martineau. Geolocating blogs from their textual content. In AAAI 2009 Spring Symposia on Social Semantic Web: Where Web 2.0 Meets Web 3.0, 2009.

[7] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In WWW, 2008

[8] G. Friedland and R. Sommer. Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, D.C.

[9] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Social network classification incorporating link type. In IEEE Intelligence and Security Informatics, 2009.

[10] Z. Cheng, J. Caverlee, K. Lee and D. Sui. Exploring Millions of Footprints in Location Sharing Services. In Proceeding of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona, July 2011.

[11] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. In Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007, 2007.

[12] B. Huberman and D. R. F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14, 2009.

[13] J. Eisenstein, N. Smith and E. Xing, A Latent Variable Model for Geographic Lexical Variation, EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.