TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 6544 - 6547

Location Prediction Techniques for Twitter Data

U. Satish Kumar[1], M. Raja Suguna[2]

[1]UG Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai
[2]Assistant Professor, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai
satishuppalapati789@gmail.com[1], suguna.raj89@gmail.com[2]

Article Info
Volume 82
Page Number: 6544 - 6547
Publication Issue:
January-February 2020

Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *01 February 2020*

Abstract

Twitter is a social networking service, where users interact with a short message called tweets. Twitter becomes a rich source of information such as trends of a particular topic or events in the society. Research has been carried out to mine the sentimental analysis of tweets revealing the polarity of tweets shared. Besides predicting user's location can be useful to Crisis Management during emergencies and Natural Disasters and to Cyber Crime. It is Challenging to predict Tweeter User location because it is mostly disclosed. Three types of locations such as user home locations, tweet locations, and mentioned locations can be predicted for a given tweet and the owner of the tweet. In this paper, a detailed study on twitter location prediction techniques in the literature have been carried out.

Key Words*: Twitter, Tweets, Home Location, Tweet Location, Mentioned Location, Location Prediction*

1. Introduction

Now days Peoples are exponentially using Social Networking sites for sharing their views, messages, images, videos, opinion about any important events. Twitter is the most powerful American Microblogging site, where registered user can post a short message called Tweet of 140-character length. Users are encouraged to tweet more frequently but with a limited Character. Users can stay in touch friends, follow Celebrities (Politicians, Actors, Comedians, and Industrial Persons), share a message. Friendship in this site is not mutual, i.e There is no need to follow a person, who are following us. Many Ground-breaking News are reaching the more people even more faster than any other media.

With increasing use of GPS enabled devices such as Smart Phones, tablets, users post the geo-tagged tweets containing their latitude and longitude co-ordinates of their locations. Three types of locations such as mentioned location, tweet location; home location is associated with each tweet. Many Real time application can be developed which makes use of these location. For example, Target Marketing, Focussing similar interested people, identifying regions with outbreak of particular brand of products, identifying region of emergency or disaster, summarizing regional topic.

Home Location, Tweet location and mentioned location on tweet can be extracted from the twitter site component such as Twitter Context, Twitter Content and the Twitter Network.

2. Overview of Twitter

Tweet Content

Tweet is a small message of 140 characters length. Users post their mood, opinion, events occurring nearby, places visited, images in the form of a tweet. A user can also retweet the other user's tweet. Both tweets and retweets are pushed to the followers of the user. They can use the symbol "@" for mentioning other users in their content and this message is also shared to that mentioned user.

January-February 2020
ISSN: 0193-4120 Page No. 6544 - 6547

Twitter Context

It contains the timestamp, the time when the tweet is posted. User also "geo-tag" (current location) their tweets by using their GPS enabled devices. User Profile also contains the real home location. Timestamp, geo-tags, user profile information constitutes the twitter context.

Twitter Network

Twitter Network Constitutes the type of friendship or connection between the registered users. If user "A" follows user "B", then A is a friend of B, but B is not a friend of A. If both the users follow each other, then "A" and "B" are mutual friends. Many distant and strange persons also become mutual friends in the Twitter network. A user can also subscribe to another user twitter account. Below fig 1, shows the overview of three components of a tweet.

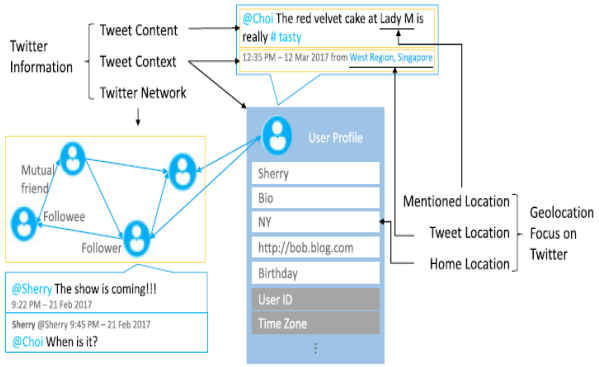

Figure 1: Overview of Twitter Components [24]

Home Location Prediction

In Twitter Content users may mention their home location as the content. Particular location users may use some kind of local words in their tweets. Therefore by finding out the local words mapping with the tweet, by word centric method and location centric method, home location of tweet can be predicted.

Laere et al. [1] proposes Kernel Density Estimation where a word is mapped to a two dimensional probability distribution on the earth surface and another method Ripley's K statistic which measures deviation of a word set from the spatial co-ordinates.

Ren et al. [2] proposes a method to filter out the local words used by the users in their tweets using Inverse Location Frequency filter and Remote Word Filter. According to this method, local words are distributed in fewer locations and their Inverse Location Frequency will be high. Yamaguchi et al. [3] proposes method based on KL Divergence and works with streaming tweets, where user location is dynamically updated on newly arriving tweets by the user. M. McPherson [5] proposed a Homophily in Social Networks, which states that similar people will communicate more than dissimilar people. The mutual friends in twitter network will communicate more, based on this fact, predicting the home location of a user can also been done by predicting its friend's home location. Davis et al. [6] employs the Homophily proposed by [5], for predicting the home location of the users and uses the same technique by [2]. Rodrigues et al. [7] predicts the home location by using the Potts model proposed by [8] which maximizes global home co-locations between mutual friends. All the above mentioned works [5], [6], [7] predicts the home location of twitter users by using the Twitter Network and they do not use the co-ordinates of the home location, they considered home location as the discrete set of objects. J. Mahmud et al [9] models Time Zone Classifier where the twitter posting time is recorded as GMT day. A binned GMT day is then mapped to time of equal parts. Thus the tweets are viewed as a distribution of tweet posting times. This Time Zone Classifier Reveals the locations of user based on their Time Zones.

Thus the Home location of a Tweet is predicted by all the three twitter components such as Twitter Context, Twitter Content and Twitter Network.

3. Tweet Location Prediction

Tweet location prediction contains only that particular tweet as an input to the model where as in home location prediction large number of tweets is given as input to the model. Hence more number of parameters are needed to predict the tweet location.

R. Priedhorskyet al [10] uses Gaussian Mixture Model to predict the tweet location. This model uses both the spatial Distribution of words along with n-grams. The usage n-grams increase the accuracy of tweet location prediction. Chong and Lim et al [11] find that users with more similar tweet content history may be more similar in their venue visitation history. Collaborative filtering is adopted to propagate visitation information to users without location visiting history based on the similarity of historical tweet content. They provide us a new view that useful information can be obtained even from users without following or followed relationship.

4. Mentioned Location Prediction

J. Laffertyet al [12] used Named Entity for predicting mentioned location using Random Field Algorithm which gave state of art results. L. Ratinov et al [13] used

January-February 2020
ISSN: 0193-4120 Page No. 6544 - 6547

linguistic features like capitalizations and Parts -0f-speech tags to train their model for classification. Ritter et al. [15] build a Classifier which makes use of capitalization and Parts of Speech Tags and used Brown clustering [14] to find out variations clusters of words and rebuild the NER pipeline. Similarly, Liu et al. [16], [17] corrected informal words (e.g.,"gooood" to "good") and trained a normalization model for tweets before performing NER.

5. Comparison of Benchmark Models

In this Section, some of the benchmarks model for predicting the location of a tweet is analysed. All model needs a dataset with or without metadata to train their respective Classifier. These Metadata constitutes time zone information, user profile information, information about user specified location, tweet post time and so on.

Deep Geo Model

Lau, et al [19] used a Character level Recurrent Convolutional Network Model for training the dataset. The preprocessed Character matrix is given as input to LSTM Layer then its output State matrix is fed as an input to Convolutional layer followed by max-pooling layer generating sub word features. An Attention Network is modeled to combine all sub word features to generate the Single Vector. All of these vectors along with the text features and metadata features fed into deep dense layers for Classification.

Fujixerox

Miura, Y., et al [20] used a variant of Fast Text Model. FUJIXEROX is one of the team of W-NUT Geo-Tag Task. The Text is represented with its aggregate of n-gram words. The n-gram words are Sub Word Features. The model also make use of the user profile information as a vector user Specified location Vector. All of these three vectors are concatenated with the time-zone embedding vector which is the fed as an input to the Dense layer network for Classification.

Convolutional Neural Network Model

Huang, B. and Carley [21] trained the CNN Model with Four vectors namely user name, user specified location information, user profile information and tweet content along with the four one hot vectors user language, tweet language, tweet post time and the time zone. All these eight vectors Concatenated and fed s a input to the dense layer of CNN for Classification.

Naïve Bayes Model

Chi, et al [22] proposed a Naïve Bayes Model for Classification. They include many features such as hashtags, location Indiative words and so on.

5. Results and Discussion

All benchmarks Models Discussed above are Compared for their Accuracies. Table 1, gives Comparison of Model with Metadata and Table 2.

Table 1: Comparison of Models without Metadata [23]

Without Metadata				
Model	**Acc_1**	**Acc_2**	**Mean**	**Median**
Naive Bayes*	0.146	-	5338.9	3424.6
FUJIXEROX	0.168	0.566	4441.5	**1900.5**
CNN Model	0.207	0.581	5106.8	2687.6
DeepGeo$^+$	0.202	**0.597**	4805.5	2500.9

Table 2: Comparison of Models with Metadata [23]

With Metadata				
Model	**Acc_1**	**Acc_2**	**Mean**	**Median**
FUJIXEROX	0.409	-	1792.5	69.5
DeepGeo	0.428	-	-	-
CSIRO	0.436	-	2538.2	74.7

6. Conclusion

A detailed study on the techniques for predicting the location of Tweet in the Twitter Site has been carried out. The location of the user who posting the Tweet can be predicted by using the twitter Components such as Twitter Context, Twitter Content, Twitter Network. All benchmarks Model are also compared for their Accuracy of Predicting location of Tweet.

References

[1] O. V. Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt, "Spatially aware term selection for geotagging," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 221–234, Jan. 2014.

[2] K. Ren, S. Zhang, and H. Lin, "Where are you settling down: Geo-locating twitter users based on tweets and social networks," in Proc. Asia Inf. Retrieval Symp., 2012, pp. 150–161.

[3] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa, "Online user location inference exploiting spatiotemporal correlations in social streams," in Proc. ACM Conf. Inf. Knowl. Manage., 2014,pp. 1139–1148.

January-February 2020
ISSN: 0193-4120 Page No. 6544 - 6547

[4] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from? inferring home locations of twitter users," in Proc. Int. Conf. Weblogs Social Media, 2012, pp. 511–514.

[5] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annu. Rev. Sociology, vol. 27, no. 1, pp. 415–444, 2001.

[6] C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de LArcanjo, "Inferring the location of twitter messages based on user relationships," Trans. GIS, vol. 15, no. 6, pp. 735–751, 2011.

[7] E. C. Rodrigues, R. Assunc, G. L. Pappa, D. R. R. Oliveira, and W. M. Jr, "Exploring multiple evidence to infer users' location in twitter," Neuro computing, vol. 171, no. C, pp. 30–38, 2016.

[8] S. Z. Li, Markov Random Field Modeling in Image Analysis, ser. Advances in Pattern Recognition. Berlin, Germany: Springer, 2009.

[9] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 47:1–47:21, 2014.

[10] R. Priedhorsky, A. Culotta, and S. Y. Del Valle, "Inferring the origin locations of tweets with quantitative confidence," in Proc.ACM Conf. Comput. Supported Cooperative Work Social Comput.,2014, pp. 1523–1536

[11] W. Chong and E. Lim, "Tweet geolocation: Leveraging location, user and peer signals," in Proc. ACM Conf. Inf. Knowl. Manage., 2017, pp. 1279–1288.

[12] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. Int. Conf. Mach. Learn., 2001, pp. 282–289

[13] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in Proc. Conf. Comput. Natural Language Learn., 2009, pp. 147–155.

[14] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," Comput. Linguistics, vol. 18, no. 4, pp. 467–479, 1992.

[15] A. Ritter, S. Clark, Mausam, and O. Etzioni, "Named entity recognitionin tweets: An experimental study," in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 1524–1534.

[16] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entitiesin tweets," Proc. Annu. Meeting Assoc. Comput. Linguistics, 2011, pp. 359–367.

[17] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named entity recognition for tweets," ACM Trans. Intell. Syst. Technol., vol. 4, no. 1, 2013, Art. no. 3.

[18] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee Twiner: Named entity recognition in targeted twitter stream," in Proc. ACMSIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 721–730.

[19] Lau, J. H., Chi, L., Tran, K.-N., and Cohn, T. (2017). End-to-end network for twitter geolocation prediction and hashing. arXiv preprint arXiv:1710.04802

[20] Miura, Y., Taniguchi, M., Taniguchi, T., and Ohkuma, T. (2016). A simple scalable neural networks based model for geolocation prediction in twitter. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 235–239

[21] Huang, B. and Carley, K. M. (2017). On predicting geolocation of tweets using convolutional neural networks. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (Springer), 281–291

[22] Chi, L., Lim, K. H., Alam, N., and Butler, C. J. (2016). Geolocation prediction in twitter using location indicative words and textual features. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 227–234

[23] Huang Chieh-Yang, Tong Hanghang, He Jingrui, Maciejewski Ross Location Prediction for Tweets, Frontiers in Big Data volume 2, 2019, 5 pages https://www.frontiersin.org/article/10.3389/fdata.2019.00005 ,2624-909X

[24] X. Zheng, J. Han and A. Sun, "A Survey of Location Prediction on Twitter," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 9, pp. 1652-1671, 1 Sept. 2018.doi: 10.1109/TKDE.2018.2807840