On Fine-Grained Geolocalisation of Tweets

Jorge David Gonzalez Paule[1], Yashar Moshfeghi[2],
Joemon M. Jose[3] and Piyushimita (Vonu) Thakuriah[4]
[1,3]School of Computing Science, [4]Urban Big Data Centre, University of Glasgow, Glasgow, UK
[2]Department of Computer & Information Sciences, University of Strathclyde, Glasgow, UK
j.gonzalez-paule.1@research.gla.ac.uk,yashar.moshfeghi@strath.ac.uk,
Joemon.Jose@glasgow.ac.uk,Piyushimita.Thakuriah@glasgow.ac.uk

ABSTRACT

Recently, the geolocalisation of tweets has become an important feature for a wide range of tasks in Information Retrieval and other domains, such as real-time event detection, topic detection or disaster and emergency analysis. However, the number of relevant geo-tagged tweets available remains insufficient to reliably perform such tasks. Thus, predicting the location of non-geotagged tweets is an important yet challenging task, which can increase the sample of geo-tagged data and help to a wide range of tasks. In this paper, we propose a location inference method that utilises a ranking approach combined with a majority voting of tweets weighted based on the credibility of its source (Twitter user). Using geo-tagged tweets from two cities, Chicago and New York (USA), our experimental results demonstrate that our method (statistically) significantly outperforms our baselines in terms of accuracy, and error distance, in both cities, with the cost of decrease in recall.

CCS CONCEPTS

•Information systems → Social networking sites; Location based services; Information retrieval;

KEYWORDS

Information Retrieval; Geolocalisation; Fine-Grained; Twitter; User Credibility; Weighted Majority Voting

ACM Reference format:
Jorge David Gonzalez Paule[1], Yashar Moshfeghi[2],
Joemon M. Jose[3] and Piyushimita (Vonu) Thakuriah[4]. 2017. On Fine-Grained Geolocalisation of Tweets. In *Proceedings of ICTIR '17, October 1–4, 2017,Amsterdam, Netherlands, ,* 4 pages.
DOI: https://doi.org/10.1145/3121050.3121104

1 INTRODUCTION

In recent years, social media services such as Twitter have gained increasing popularity within the research community since their data is spatially fine-grained (i.e. at street or neighborhood level). Such a characteristic has provided new opportunities for a broad range of applications in Information Retrieval (IR) including real-time event detection [2], sentiment analysis [3], topic detection [7], and disaster and emergency analysis [1, 8, 11]. However, since only a very small sample of messages in the Twitter stream contain geographical information [5], geo-locating (or geolocalising) individual tweets has become an important yet challenge task. In this paper, we focus on geolocalisation of tweets at a fine-grained level. To tackle this problem, we propose a novel approach to combine evidence gathered from geo-tagged tweets that are similar based on their contents to a given non-geo-tagged tweet.

Several approaches have been proposed in the past to provide fine-grained geolocalisation of tweets, e.g. [9, 13]. These works first create a document for each predefined geographical area by concatenating the texts of the tweets belonging to that area. They then create a vector representation of that area from the generated document using a bag-of-word approach. To geolocate a given tweet, they then find the most similar area to that tweet based on its content-similarity, using the generated vectors [9]. Paraskevopoulos and Palpanas [13], in addition to above, have also considered time-evolution characteristics in their matching algorithm. Although these approaches have provided important insights on how to tackle fine-grained geolocalisation of tweets, due to the noisy nature of Twitter data [19], such an aggregation method could affect the accuracy of matching algorithms and in turn, decrease the accuracy of the geolocalisation.

In this work, we adopted a weighted majority voting algorithm to the problem of fine-grained geolocalisation of tweets. In particular, we estimated the geographical location of a given non-geo-tagged tweet by collecting the geo-location votes of the geo-tagged tweets that are most similar regarding their contents to that tweet. The weights of the votes were calculated based on the credibility of its source, (i.e. Twitter user). We then performed an exhaustive study of different models across two test collections generated based on tweets gathered from two different cities to validate our models. Our experimental results showed significant improvements regarding accuracy and reduction of geographical distance error compared to our baselines.

The rest of the paper is organised as follows. First, we discuss previous research and motivate our work. Second, we introduce our approach for fine-grained geolocalisation of a non-geo-tagged tweet. Finally, we present our experimental setup and discuss our results.

2 BACKGROUND

Several research efforts have identified the problem of geolocalising individual non-geo-tagged tweets. For example, Schulz et al. [18] tackled this problem by exploiting different spatial indicators of a tweet – i.e. tweet text or user profile – and mapping them to different geo-spatial datasets such as DBpedia Spotlight or Geonames. More recently, other works tackled this problem by dividing the geographical space into areas of a given size and then modelled the language for each area [9, 13, 17, 20]. Then, a ranking approach is used to retrieve the most likely area based on the probability that a non-geo-tagged tweet was issued in that area. However, these

studies used a coarse-grained level of granularity – i.e. zip codes to city or country level. In contrast, the problem we aim to tackle is the geolocalisation of Twitter posts at a fine-grained level – i.e. street or neighbourhood level.

An example of previous work on fine-grained geolocalisation is the work by Kinsella et. al. [9]. They attempted to predict location from country level to postal code level. As a result, the accuracy of their model decreases significantly when trying to predict at such fine-grained level. Another example of fine-grained geolocalisation is the work by Paraskevopoulos et. al. [13]. The authors refined the approach proposed by Kinsella et. al. [9] by dividing the geographical space into fine-grained squares of size $1km$. Also, the authors reduced the granularity of time by considering time slots of 4 hours, and computing the number of tweets by time for each candidate location compared with the global activity of the city. In this way, the model promotes short-term events in detriment of long-term events.

Inspired by Paraskevopoulos et. al. [13], we follow the strategy of dividing the city into squares of size $1km$. However, the time dimension is out of the scope of this paper. Thus we consider short-term and long-term events alike. Moreover, the works above perform a concatenation of texts of tweets belonging to a predefined area to represent that area as a single bag-of-word vector. We believe that by concatenating the content of the tweets, relevant information can be missed when predicting a location. In contrast to these works, we consider each tweet individually, representing each area as multiple bag-of-word vectors during the prediction process.

Also, our approach take into account the credibility of tweets. Other works have also considered the credibility of tweets. For example, McCreadie et. al. [11] has considered the idea of assigning a credibility score to tweets but for the disaster and emergency detection task. They have computed the credibility score using regression models with text features and user information. They have used this score to inform the user about the veracity/credibility of events derived from social media. We also incorporate the credibility of tweets in our fine-grained geolocalisation approach. But in contrast to McCreadie et. al. work, we incorporate this score as a weight of each vote in our adopted majority voting approach. The majority voting algorithm is a well known, fast and effective strategy widely adopted for prediction and re-ranking tasks [4, 12, 16]. However, to the best of our knowledge, this is the first time the majority voting is considered to tackle the geo-location of tweets. Next section describes our approach in detail.

3 FINE-GRAINED GEOLOCALISATION

Our proposed approach consists of three steps. First, we create a grid to divide the geographical area into squares of size $1km$ and associate each geo-tagged tweet to an area based on its location. As discussed in Section 2, the grid approach has been widely used in the literature to represent geographical areas at different levels of granularity [9, 13]. Second, we obtained the Top-N content-based similar geo-tagged tweets to a non-geo-tagged using different retrieval models (see Section 4.1). Finally, we combine evidence gathered from the Top-N tweets by adopting a weighted majority voting algorithm where the weight is calculated based on the credibility information of tweets source.

3.1 Combining Evidence using Weighted Majority Voting

In order to combine evidences gathered from the Top-N content-based similar geo-tagged tweets to a non-geo-tagged tweet t_{ng}, we adopted a weighted majority voting algorithm [4, 12, 16] as follows. We represent each element of the Top-N tweets as a tuple (t_i, l_i, s_i) where l_i is the location associated to a geo-tagged tweet t_i posted by the source s_i. We then select the most frequent location within the Top-N set and associate that as the geo-location of a given tweet. In formal definition:

$$Location(t_{ng}) = \underset{l_j \in L}{\operatorname{argmax}} \left(\sum_{i=1}^{N} W_{t_i} * Vote(t_i^{l_i}, l_j) \right) \quad (1)$$

where L is the set of locations (l_j) in the Top-N tweets and $t_i^{l_i}$ is the location of the i-th tweet in the rank. Then, a vote is given to the location l_j by the tweet t_i as follows:

$$Vote(t_i^{l_i}, l_j) = \begin{cases} 1 & t_i^{l_i} = l_j \\ 0 & t_i^{l_i} \neq l_j \end{cases} \quad (2)$$

The vote of the tweet t_i is weighted by:

$$W_{t_i} = \frac{|\{t_{si} \in TN_i \mid distance(t_{s_i}, t_{vi}) \leq 1km\}|}{|TN_i|} \quad (3)$$

where W_{t_i} is based on the credibility of tweet's source s_i. The credibility of tweet's source is calculated as follows. First, we obtain the Top-N content-based most similar tweets for every tweet in a validation set (see Section 4). Second, we calculate the geographical distance (see Section 4) between the tweet in the validation set and each element in its Top-N. Next, for each source s_i we define a set TN_i that contains all the tweets appearing in any of the Top-N rankings (t_{si}) produced for each tweet in the validation set (t_{vi}). Finally, the credibility of source s_i is given by the ratio of all tweets in TN_i placed within less than $1km$ distance from the tweets in the validation set (t_{vi}).

Finally, the location l_j that obtains the highest number of tweet votes is returned as the final predicted geo-location of a given non-geo-tagged tweet.

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup that supports the evaluation of our proposed approach for fine-grained geolocalisation of non-geo-tagged tweets.

Data: Previous studies have shown that geo-tagged and non-geo-tagged data have the same characteristics [6]. Thus, models built from geo-tagged data can be generalised to non-geo-tagged data. We, therefore, experimented over a ground truth sample of English geo-tagged tweets located in two different cities: Chicago and New York City (USA) with 131,273 and 155,114 tweets respectively. Tweets were collected from the Twitter Public stream during March 2016.

To evaluate our approach, we divided our dataset into three subsets. We used the first three weeks of tweets in our collection (i.e. the first three weeks of March) as a training set. We then

randomly divided the last week data into validation and test sets to ensure that they have similar characteristics. Therefore, for Chicago dataset, our training, validation and test sets contained 111,627, 9,823 and 9,823 geo-tagged tweets respectively. For New York dataset, our training, validation and test sets contained 128,746, 13,184 and 13,184 geo-tagged tweets respectively.

4.1 Models

4.1.1 Baseline Models. We implemented our baseline (denoted by "Baseline") inspired by Paraskevopoulos et al. [13] work. To do so, for each of our cities, we first created a grid structure of squared areas with a side length of 1 *km*. For each of these defined squared areas, we created a document by concatenating the text of the tweets associated with each area. We then indexed these documents. As a preprocessing step, usernames and hashtags were preserved as tokens, all hyperlinks were removed from tweets, and re-tweets were preserved in the dataset. Then, we retrieved the most content-based similar document (Top-1) for each non-geo-tagged tweet. As the model returns the Top-1 tweet, the longitude and latitude coordinates of the tweet are returned as the predicted location instead of the squared area associated to the post. We investigated several retrieval models to maximise the performance of our baseline. Five different retrieval models were evaluated: Divergence From Randomness (dfr), Language Model with Dirichlet Smoothing (lmd), IDF (idf), TF-IDF (tf_idf) and BM25 (bm25) using the Apache Lucene[1] implementation. The difference between our baseline and the work by [13] are two-fold. First, we removed stop-words [10] and applied Porter stemming.[2] Second, we also did not consider the time dimension, as described in Section 2.

4.1.2 WMV Models. We also implemented our proposed approach explained in Section 3, denoted by "WMV". We used the same squared areas defined for our baseline models. However, in WMV model, each of these defined squared areas was represented as multiple bag-of-word vectors where each vector represents a single tweet associated with that area. By doing this, we treated each tweet as a single document for the retrieval task. We performed the same preprocessing step applied in our baseline models.

Similarly to our baselines, we investigated the same five retrieval models to maximise the performance of our approach. The results indicated that using IDF gave us the best performance. This is consistent with previous research findings [15].

We apply our weighted majority voting algorithm on top of the retrieval task. We considered the Top-3, -5, -7 and -9 content-based most similar tweets obtained from the retrieval task. The final predicted location is the predefined area that obtain the highest number of votes.

Metrics: To evaluate the effectiveness of our approach, the following metrics are reported. **Average Error distance (km):** we compute the distance on Earth (Haversine formula [14]) between the predicted location and the real coordinates of the tweet in our ground truth. **Accuracy@1km:** the accuracy of the model is measured by determining whether a predicted location lies within a radius of 1km from the real location. **Recall:** we consider Recall as the fraction of tweets in the test set that was geolocalised by our approach regardless of the distance error.

[1] http://lucene.apache.org/

[2] We also tried our baseline without removing stop-words and applying Porter stemming, but resulted in the lower performance and hence we did not report them due to lack of space.

5 RESULTS

Table 1 and 2 shows the average error distance, accuracy, and recall for our approach evaluated on the Chicago and New York datasets respectively. A paired t-test was conducted to assess if the difference in effectiveness between the models is statistically significant. As shown in Table 1 and 2, our approach ("WMV") (statistically) significantly outperforms the best performed baseline (i.e. "Baseline_lmd") in terms of accuracy and error distance, in both cities, across all the investigated values of N for the Top-N tweets, with the cost of decrease in recall.

Table 1: Results for Chicago city dataset. The table presents the Average Error Distance in kilometres (A_Err_km), Accuracy at 1 kilometre (Acc@1km) and Recall for our proposed approach ("WMV") against our Baseline using the Top-N (@TopN) elements in the rank. Significant differences with respect to our best Baseline ("Baseline_lmd") are denoted by $*$ and $**$ where $p<0.05$ and $p<0.01$ respectively.

Chicago			
Model	**A_Err_km**	**Acc@1km**	**Recall**
Baseline_tf_idf	8.100	42.40%	99.97%
Baseline_idf	14.056	13.18%	99.97%
Baseline_dfr	8.586	37.40%	99.97%
Baseline_lmd	6.185	47.79%	99.97%
Baseline_bm25	7.637	41.76%	99.97%
WMV@Top3	3.849**	61.17%**	83.28%**
WMV@Top5	3.669**	62.78%**	79.08%**
WMV@Top7	3.170**	66.82%**	70.41%**
WMV@Top9	2.576**	71.29%**	62.28%**

Table 2: Results for New York city dataset. The table presents the Average Error Distance in kilometres (A_Err_km), Accuracy at 1 kilometre (Acc@1km) and Recall for our proposed approach ("WMV") against our Baseline using the Top-N (@TopN) elements in the rank. Significant differences with respect to our best Baseline ("Baseline_lmd") are denoted by $*$ and $**$ where $p<0.05$ and $p<0.01$ respectively.

New York			
Model	**A_Err_km**	**Acc@1km**	**Recall**
Baseline_tf_idf	7.505	38.39%	99.98%
Baseline_idf	12.755	12.78%	99.98%
Baseline_dfr	7.609	36.28%	99.98%
Baseline_lmd	7.169	37.29%	99.98%
Baseline_bm25	7.460	38.25%	99.98%
WMV@Top3	4.234**	52.33%**	75.84%**
WMV@Top5	4.362**	51.98%**	75.09%**
WMV@Top7	4.008**	54.81%**	67.83%**
WMV@Top9	3.476**	59.23%**	59.94%**

Additionally, our findings show that as the number of voting candidates (i.e. Top-N) increases, our approach achieves lower error distance, higher accuracy but lower recall. Therefore, considering the Top-3 tweets resulted in the best trade-off regarding error distance, accuracy and recall. Also, we observed that our approach performed similarly across both cities despite their geographical and cultural differences. Such similarity in performance suggests that our approach can be generalised and adapted to different cities. Our promising results show the potential of our approach for fine-grained geolocalisation of tweets.

6 CONCLUSIONS

In this work, we proposed an approach for fine-grained geolocalisation of tweets by adopting a weighted majority voting algorithm. The weight of each tweet vote is obtained by calculating the credibility of its source (i.e. Twitter user). Our baseline model is inspired by Paraskevopoulos [13] work, where a grid approach were applied to divide a city into a set of predefined geographical areas of size 1 km. However, in contrast to this work, we did not concatenate the text of tweets into a document to create a single bag-of-word vector to represent a predefined area. Our approach, instead, treats each tweet individually as a single document and represent each area as multiple bag-of-word vectors.

We then indexed these documents and then retrieved the most content-based similar document for each non-geo-tagged tweet. Also, we investigated several retrieval models to find the best performance for our baseline and our proposed approach. For our baseline approach, the geographic location associated with the Top-1 retrieved document is then assigned to the tweet, since each predefined area is only represented by a single document. In our approach, we assign the most voted area of the Top-N content-based similar tweets where N is set to 3, 5, 7 and 9.

To demonstrate the effectiveness of our approach, we conducted an experiment on two datasets of geo-tagged tweets collected from two different cities, Chicago and New York, with 131,273 and 155,114 tweets respectively. The data was collected during March 2016. Our experimental results show that our weighted majority voting approach (statistically) significantly outperforms the best-performed baseline (i.e. "Baseline_lmd") in terms of accuracy and error distance, in both cities, across all the investigated values of N for the Top-N tweets, with the cost of decrease in recall in the two cities of study. Also, we observed that as the number of voting candidates (i.e. Top-N) increases, our approach achieves lower error distance, higher accuracy but lower recall. This behaviour is observed across both datasets which suggest that our approach can be generalised and adapted to different cities.

This shows the power of our proposed approach in predicting geolocation of tweets, and can substantially expand the sample of geo-tagged data at a fine-grained level (i.e. street level or neighbourhood level), helping to a wide range of tasks in information retrieval, including real-time event detection, topic detection and disaster and emergency analysis. In future work, we will investigate whether we can improve recall while maintaining the high accuracy of our approach.

Acknowledgements: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC grant agreement n° 632075.

REFERENCES

[1] Ji Ao, Peng Zhang, and Yanan Cao. 2014. Estimating the Locations of Emergency Events from Twitter Streams. *Procedia Computer Science* 31 (2014), 731 – 739. DOI: http://dx.doi.org/10.1016/j.procs.2014.05.321 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.

[2] Farzindar Atefeh and Wael Khreich. 2015. A Survey of Techniques for Event Detection in Twitter. *Comput. Intell.* 31, 1 (Feb. 2015), 132–164. DOI: http://dx.doi.org/10.1111/coin.12017

[3] Eric Baucom, Azade Sanjari, Xiaozhong Liu, and Miao Chen. 2013. Mirroring the Real World in Social Media: Twitter, Geolocation, and Sentiment Analysis. In *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing (UnstructureNLP '13)*. ACM, New York, NY, USA, 61–68. DOI: http://dx.doi.org/10.1145/2513549.2513559

[4] T.-H. Chiang, H.-Y. Lo, and S.-D. Lin. 2012. A Ranking-based KNN Approach for Multi-Label Classification. In *Proceedings of the Asian Conference on Machine Learning (Proceedings of Machine Learning Research)*, Steven C. H. Hoi and Wray Buntine (Eds.), Vol. 25. PMLR, Singapore Management University, Singapore, 81–96. http://proceedings.mlr.press/v25/chiang12.html

[5] Mark Graham, Scott A. Hale, and Devin Gaffney. 2013. Where in the World are You? Geolocation and Language Identification in Twitter. *CoRR* abs/1308.0683 (2013). http://arxiv.org/abs/1308.0683

[6] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter User Geolocation Prediction. *J. Artif. Int. Res.* 49, 1 (Jan. 2014), 451–500. http://dl.acm.org/citation.cfm?id=2655713.2655726

[7] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering Geographical Topics in the Twitter Stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 769–778. DOI: http://dx.doi.org/10.1145/2187836.2187940

[8] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Comput. Surv.* 47, 4, Article 67 (June 2015), 38 pages. DOI: http://dx.doi.org/10.1145/2771588

[9] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'M Eating a Sandwich in Glasgow": Modeling Locations with Tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC '11)*. ACM, New York, NY, USA, 61–68. DOI: http://dx.doi.org/10.1145/2065023.2065039

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[11] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. EAIMS: Emergency Analysis Identification and Management System. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, New York, NY, USA, 1101–1104. DOI: http://dx.doi.org/10.1145/2911451.2911460

[12] Mawloud Mosbah and Bachir Boucheham. 2015. *Majority Voting Re-ranking Algorithm for Content Based-Image Retrieval*. Springer International Publishing, Cham, 121–131. DOI: http://dx.doi.org/10.1007/978-3-319-24129-6_11

[13] Pavlos Paraskevopoulos and Themis Palpanas. 2015. Fine-Grained Geolocalisation of Non-Geotagged Tweets. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. ACM, New York, NY, USA, 105–112. DOI: http://dx.doi.org/10.1145/2808797.2808869

[14] C. C. Robusto. 1957. The Cosine-Haversine Formula. *The American Mathematical Monthly* 64, 1 (1957), 38–40. http://www.jstor.org/stable/2309088

[15] Jesus Alberto Rodriguez Perez and Joemon M. Jose. 2015. On Microblog Dimensionality and Informativeness: Exploiting Microblogs' Structure and Dimensions for Ad-Hoc Retrieval. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, New York, NY, USA, 211–220. DOI: http://dx.doi.org/10.1145/2808194.2809466

[16] Lior Rokach. 2010. *Pattern Classification Using Ensemble Methods*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

[17] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1500–1510. http://dl.acm.org/citation.cfm?id=2390948.2391120

[18] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. 2013. A Multi-Indicator Approach for Geolocalization of Tweets. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063

[19] Jaime Teevan, Daniel Ramage, and Meredith Ringel Morris. 2011. #TwitterSearch: A Comparison of Microblog Search and Web Search. ACM.

[20] Benjamin P. Wing and Jason Baldridge. 2011. Simple Supervised Document Geolocation with Geodesic Grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 955–964. http://dl.acm.org/citation.cfm?id=2002472.2002593