Temporal Modelling of Geospatial Words in Twitter

Bo Han[1], **Antonio Jose Jimeno Yepes**[2], **Andrew MacKinlay**[2,3], **Lianhua Chi**[2]
Hugo AI[1]
IBM Research Australia[2]
University of Melbourne[3]
bhan@hugo.ai, {antonio.jimeno, admackin, lianhuac}@au1.ibm.com

Abstract

Twitter text-based geotagging often uses geospatial words to determine locations. While much work has been done in word geospatiality analysis, there has been little work on temporal variations in the geospatial spread of word usage. In this paper, we investigate geospatial words relative to their temporal locality patterns by fitting periodical models over time. The model jointly captures inherent geospatial locality and periodical factor for a word. The resultant factorisation enables better understanding of word temporal trends and improves geotagging accuracy by only using inherent geospatial local words.

1 Introduction

Automatically inferring geographical location of social media data has become increasingly popular, as geospatial information plays a vital role in applications such as advertising, influenza detection and disaster management. Due to a lack of abundant reliable geospatial information (Cheng et al., 2010), various text-based geotagging methods have been proposed (Cheng et al., 2010; Eisenstein et al., 2010; Wing and Baldridge, 2014).[1] The main idea is to leverage geospatial words such as dialects and location names embedded in Twitter text to infer geographic locations. For instance, *yinz* is primarily used in Pittsburgh, *#auspol* is a popular hashtag in Australia, and *@TransLink* is frequently mentioned by Vancouver users.

Social media data often comes as a stream, and its contents and topics change over time. This implies geospatial words in social media can be further distinguished based on their temporal patterns of locality (i.e. how location-indicative a word is). Some time-invariant geospatial words (e.g. *CalTech*) are consistently associated with a location, while other words like *#ALTA216* are only transiently associated with a location (e.g. during the conference).

In this paper, we investigate the word locality pattern over time. First, we bin streaming tweet data into a series of sliding time windows and calculate the locality estimators in each window. The time-indexed locality estimators are then fed into periodical models which jointly capture the inherent geospatial locality and periodical factor. We show that by removing the periodical factor, we can obtain improved geotagging accuracy. Furthermore, we demonstrate the utility of fitted model parameters in explaining some intuitive observations for reoccurring words.

2 Background

Text-based geotagging is often formulated as a classification task (Cheng et al., 2010) which involves predicting a location from a set of predefined geographical partitions. It exploits words in tweets on the grounds that some of them carry geographical information. The accuracy is often measured by the agreement of the predicted locations with true oracle locations (Wing and Baldridge, 2011).

Earlier work in geotagging exploits language models identified from textual data from different locations (Cheng et al., 2010; Wing and Baldridge, 2011; Kinsella et al., 2011). It selects the location with the most similar language model relative to the input tweet text. However, these methods also include irrelevant words such as stop words (e.g. *the*) and common hashtags (e.g. *#iphone*), meaning they capture imperfect geospatial signals.

Cheng et al. (2010) improved language model-based methods by augmenting local words for

[1] Indeed there is a way that a user can turn on their location sharing options which offers accurate location information, however, the ratio of such users are pretty low (Cheng et al., 2010).

Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay and Lianhua Chi. 2016. Temporal Modelling of Geospatial Words in Twitter. In *Proceedings of Australasian Language Technology Association Workshop*, pages 133−137.

geotagging. Words with sharply-peaked frequency distributions with respect to location are categorised as local words, and only local words are used in geotagging. Furthermore, ranking geospatial words by their locality in the decreasing order has been proposed (Chang et al., 2012; Laere et al., 2013; Han et al., 2014), however the categorisation of local and non-local words is still binary. Hierarchical models and regularisation have also shown to be effective in geotagging (Ahmed et al., 2013; Wing and Baldridge, 2014).

With much progress in identifying and utilising geospatial words, the temporal variance of geospatial words has not been under studied. In this paper, we study the impact of this temporal aspect for geospatial words.

3 Temporal Geospatial Word Modelling

To analyse the temporal pattern of geospatial words, we first define a fixed-length sliding time window. The collected data within a time window is then aggregated for computing locality variances for each word. The same calculations are performed for each consecutive time window and the location scores are collectively incorporated in a periodical model, and the geospatial words are then ranked and categorised based on this model. Top ranked words are assumed to be consistently location-indicative over time and should therefore be preferred when building geotagging models.

3.1 Measurement of Word Locality

The locality variances of a word are computed on basis of time windows (e.g. one week). For each word found in a time window, we obtain a list of locations (i.e. GPS coordinates) from tweets containing the word. Then we draw random samples of paired locations without replacement (non-exhaustive to improve tractability), and compute the distances between paired locations following Cook et al. (2014), yielding a list of paired location distances. The mean and median of these distances serve as locality variances and are used in subsequent experiments. Permanent location-indicative words should have consistently low locality variances as they are likely to occur in geographically close regions in most time windows. The metric of median distance reduces the influence of outlier locations (e.g. caused by people mentioning their home city while travelling).

3.2 Sinusoidal Modelling

We assume that a word's observed geospatial pattern is jointly influenced by its inherent locality and periodical factor. A general sinusoidal model is applied to capture both factors in Equation 1.

$$f(t) = \underbrace{C}_{\text{Inherent locality}} + \underbrace{\alpha \sin(\omega t + \phi)}_{\text{Periodical factor}} \quad (1)$$

$f(t)$ denotes the geospatial locality variance over time and t is the time window index. C, which is constant with respect to time, models the inherent permanent locality, while the periodic factor is captured by the time-dependent $\alpha \sin(\omega t + \phi)$ in Equation 1.

A smaller C suggests the corresponding word is more inherently location-indicative since the periodic effects are factored out into the other term.

The time component $\alpha \sin(\omega t + \phi)$ is dependent on the time window index t and parameterised by the amplitude α, angular frequency ω and phase ϕ. α represents the maximum impact of periodic component on a word's locality, with a larger value suggesting the word is strongly time-dependent.

ω denotes the frequency of this periodic component. It is inversely proportional to the period, which is important for categorising the geospatial patterns of a word. Ideally, for transient geospatial words, lower locality variances will occur in a tight cluster of time windows giving a large period (and hence low ω), while recurring geospatial words will have a smaller period corresponding to lower locality variances appearing at more regular intervals.

ϕ is the phase of the wave and reflects the point, such as a day of the week, within a time window and it is crucial for curve fitting.

4 Experiments and Discussion

4.1 Datasets

We collected 10% 2014 Twitter stream data from Gnip.[2] Tweets are lowercased and non-English data is removed according to Gnip-provided language code. We use `APR-DEC` geotagged tweets as the training data and `JAN-MAR` users (by aggregating all their tweets) for test.[3] To ensure the

[2] https://gnip.com/

[3] Because we had a major data collection interruptions between March and April, we chose the larger of two datasets for training and the other for test.

quality of test data, we only include users who have more than 10 English tweets that are within 150 km of a city centre according to the geotag coordinates of the tweets, with at least 80% of these tweets having the same closest city. The closest city is stored as the user's true location.[4] The datasets after applying the above process are shown in Table 1.[5]

Datasets	Data size
`Train(APR-DEC)`	45.4M tweets
`Test(JAN-MAR)`	373K users

Table 1: Filtered Twitter dataset

4.2 Fitting Sinusoidal Models

We estimate parameters for Equation 1 using *Non-linear Least Square* for each word. The initial values for important factors are set as follows:[6]

- C is set to the mean of the "10%-trimmed" (5th-95th percentile) locality variance numbers.
- α is determined by taking the root mean square of the 10%-trimmed locality variance numbers and dividing by $\sqrt{2}$.
- ω is set to the dominant frequency in *Discrete-Time Fourier Transformation*.

We further check the model validity by checking p-values for each C, and only keeping geospatial words with models that are significant at $p < 0.05$. We also only keep words that occur at least 20 times in at least 20 time windows due to efficiency considerations and because rare words have less impact in terms of analysis and building statistical models.

4.3 Impact on Geotagging

One advantage of temporal modelling of geospatial words is that the geotagging accuracy is expected to improve by teasing out transient and recurring location-indicative words. We set a benchmark multinomial naive Bayes classifier with the following settings to test the impact.[7]

- WHOLE: This baseline uses the whole training data collection period to calculate the locality median score in training without specifying time windows, which is roughly equivalent to conventional word-based non-temporal geotagging model.
- SIN-MEAN: This setting uses a fourteen-day sliding time window with one day as the sliding step.[8] Random location pairs are generated three times with each sample size equal to 20. Mean numbers are used to fit the sinusoidal model and the initial parameters are estimated as described in Section 4.2.
- SIN-MEDIAN: Similar to SIN-MEAN, but we use medians as locality variances to reduce the negative impact of outliers.

We then rank words by the C value from (3.2) and evaluate the accuracy of the geotagging models produced by using the top n words. The experiment range starts from the top 5K with a next 5K increment up to 40K.[9] As shown in Figure 1, we found (1) As expected, the more features, the better the performance of the geotagger; (2) With the same number of features, both SIN-MEAN and SIN-MEDIAN outperform WHOLE consistently by 2.0-4.7% error reduction (with an absolute accuracy 5.3-9.3%); (3) The differences between WHOLE and each of the SIN methods are all statistically significant ($p < 0.0001$). In contrast, the difference between SIN methods is indistinguishable. Overall this indicates that using only permanent location indicative words (as determined by C) allows us to build more accurate models for text-based geotagging.

4.4 Post Analysis and Discussions

The fitted model parameters also imply some interesting patterns for geospatial words. As explained in Section 3.2, C provides an indication of inherent locality (as shown in Table 2).

[4] We adopted the 3709 city partitions of Han et al. (2014)

[5] Pavalanathan and Eisenstein (2015) mentioned geotagging results are influenced by demographics and where the true location source being used. These influencing factors are less of our concern when analysing the impact of temporal factors.

[6] The remaining parameters for data fitting are set as follows: $\phi = 1$, $maxIter = 500$.

[7] The objective is to analyse the impact of temporal geospatial word modelling instead of building state-of-the-art geotagging systems. Advanced models including a regularised logistic regression as discussed by Wing and Baldridge (2014) are preferred in building geotaggers for better accuracy.

[8] In theory, we can choose other window sizes, however shorter time windows would produce sparse location distributions and unreliable samples.

[9] The total feature number is around 43K.

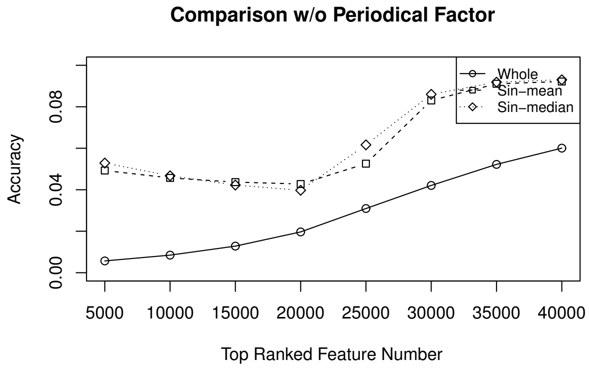

Figure 1: Geotagging performance comparisons (over 373K test users)

$C \geq 10^5$	$C \leq 10^{-3}$
siya	*@zeling97*
#thankyoulord	*@seandasheepdog*
@armorogod	*#xiomaraforugirlperu*
@aquariusquotez	*@parazetsalive*
@shainedawson	*@mishecollins*

Table 2: Examples of large and small C values

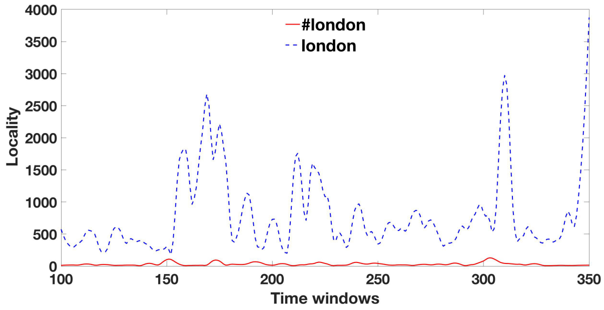

Figure 2: Locality comparison: locations versus location hashtags

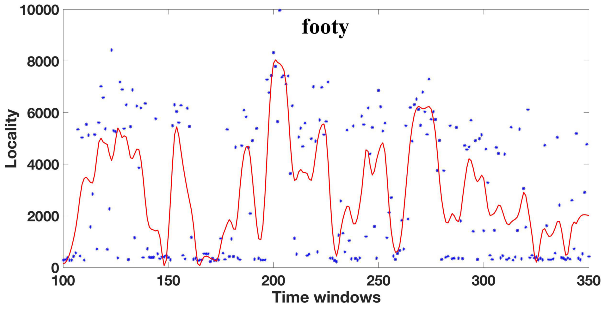

Figure 3: Recurring pattern for local sports

We found words with lower C values are often user names and hashtags which are used in specific regions. For instance, *#telpadmovienight* and *#xiomaraforugirlperu* are popular in Philippines and Peru. Large C values are associated with popular people (e.g. *siya* and *@shainedawson*) or common expressions (e.g. *#thankyoulord*).

Metropolitan city names are often not consistent local words over time, as they might be mentioned across the world. In contrast, the hashtag versions seem to be more permanent location indicative (e.g. Figure 2).[10] Some local sport terms (e.g. Figure 3) share (modest) recurring locality patterns over time as captured by the temporal modelling. Theoretically, it is possible to have location indicative words that are local to one location at certain times, while local to other locations at other times, however, we have not found such examples in our selected features.

The current modelling approach also has several limitations. Some infrequent location-indicative words may be missed due to the minimum frequency threshold in Section 4.2 and the relatively small datasets we used for experiments. Furthermore, we may obtain unreliable fitted parameters due to insufficient data points, because as few as 20 locality variances may be available. Using a training data set which covers a longer time span would partially mitigate this problem and help learn parameters for words with longer periods. We could also increase the time window size (e.g. to a month) to incorporate more words, however this may make it difficult to capture monthly recurring words.

5 Conclusion and Future Work

In this paper, we investigated the temporal variance of geospatial words over time. Specifically, a sinusoidal model is applied to jointly capture the inherent permanent locality along with time-dependent component representing changes which occur in word usage with respect to particular locations over time. This model factors out permanent location indicative words which are shown to be able to improve geotagging accuracy. The fitted parameters also confirm intuitive geospatial locality patterns. In general, we believe such temporal modelling of geospatial words benefits both predictive tasks and geospatial data analysis.

Sinusoidal models are effective in capturing strict recurring patterns, however, this assumption

[10] All figures are smoothed using "loess" with *span* = 0.06 (Cleveland and Devlin, 1988).

is often not satisfied due to the nature of periodical patterns and noise in the data. In future work, we plan to experiment with more advanced periodical models for data fitting, e.g. employ a number of superposed periodical models, instead of using only the dominant frequency model. Inspired by Dredze et al. (2016) that periodical patterns have impacts on geotagging results, we also plan to compare test results based on datasets collected over different periods.

References

Amr Ahmed, Liangjie Hong, and Alexander J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd International Conference on the World Wide Web (WWW 2013)*, pages 25–36, Rio de Janeiro, Brazil.

Hau-wen Chang, Dongwon Lee, Eltaher Mohammed, and Jeongkyu Lee. 2012. @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 111–118, Istanbul, Turkey.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010)*, pages 759–768, Toronto, Canada.

William S Cleveland and Susan J Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610.

Paul Cook, Bo Han, and Timothy Baldwin. 2014. Statistical methods for identifying local dialectal terms from gps-tagged documents. *Dictionaries*, 35:248–271.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for twitter: Timing matters. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069, San Diego, California, June. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287, Cambridge, USA.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal Artificial Intelligence Research (JAIR)*, 49:451–500.

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*, pages 61–68, Glasgow, UK.

Olivier Van Laere, Jonathan Quinn, Steven Schockaert, and Bart Dhoedt. 2013. Spatially-aware term selection for geotagging. *IEEE Transactions on Knowledge and Data Engineering*, 99:221–234.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2138–2148, Lisbon, Portugal.

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 (ACL HLT 2011)*, pages 955–964, Portland, USA.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348, Doha, Qatar, October. Association for Computational Linguistics.