Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Exploiting Contextual Information for Fine-Grained Tweet Geolocation

Wen-Haw Chong, Ee-Peng Lim
Singapore Management University
80 Stamford Road, Singapore 178902
whchong.2013@phdis.smu.edu.sg, eplim@smu.edu.sg

Abstract

The problem of *fine-grained tweet geolocation* is to link tweets to their posting venues. We solve this in a learning to rank framework by ranking candidate venues given a test tweet. The problem is challenging as tweets are short and the vast majority are non-geocoded, meaning information is sparse for building models. Nonetheless, although only a small fraction of tweets are geocoded, we find that they are posted by a substantial proportion of users. Essentially, such users have location history data. Along with tweet posting time, these serve as additional contextual information for geolocation. In designing our geolocation models, we also utilize the properties of (1) *spatial focus* where users are more likely to visit venues near each other and (2) *spatial homophily* where venues near each other tend to share more similar tweet content, compared to venues further apart. Our proposed model significantly outperforms the content-only approaches.

Introduction

In fine-grained geolocation of tweets (Lee et al. 2014; Li et al. 2011), we link tweets to the specific venues from which they are posted, e.g. a restaurant. In this work, we cast fine-grained geolocation as a ranking problem. Given a test tweet, we rank venues such that high ranking venues are more likely to be the posting venue. Tweet geolocation is useful in applications such as location-based advertising, venue recommendation, etc. However the problem is challenging as tweets are short and may not contain any location names or location indicative words, e.g. airport. To mitigate this challenge, we exploit additional contextual information such as posting time and location history.

Empirically we show that while the proportion of geocoded tweets in Twitter is small (Hong et al. 2012; Ahmed, Hong, and Smola 2013), they are posted by a substantial proportion of users, ranging from 30% to 40%. Essentially these users have location history which can be used to personalize geolocation models. Furthermore, location history is useful for incorporating the following user behavioral characteristic that we observed: users are *spatially focused* and are more likely to visit venues that are near each other.

We also consider the property of *spatial homophily*, which means that social media content from the same city/region are more likely to share common words than content from different cities/regions. Spatial homophily has been studied at coarse geographical resolution (Chang et al. 2012; Ahmed, Hong, and Smola 2013). Based on the idea that different kinds of neighborhoods (Cranshaw et al. 2012) exist within a city, spatial homophily may exist at very fine spatial scale, i.e. venues near each other tend to share more similar content than venues further apart in the same city. To account for this effect, we include spatial smoothing in our model.

Henceforth, we propose a spatially smoothed model which exploits posting time and user location history on top of tweet content. We train this model in a learning to rank framework. Over different datasets, our model achieves ranking accuracy improvement from 6% to 60% over a baseline approach using only tweet content.

Location History

We show that although the proportion of geocoded tweets is small, they are posted by a substantial portion of users. We randomly sample 50,000 Twitter users from Singapore for 2014 and same number of users from Jakarta for June to Dec 2016. Table 1 shows the statistics compiled. Clearly, the proportion of geocoded tweets is tiny at 3.22% for Singapore and 4.62% for Jakarta. However these are posted by a substantial proportion of users. For ease of discussion, we denote the set of users who posted at least one geocoded tweet as $\{u\}_g$. Table 1 shows that in Singapore, $\{u\}_g$ constitutes 30.34% of the sampled users. This is much larger than the value of 3.22% if one does a naive inference based on the fraction of geocoded tweets. Similarly in Jakarta, $\{u\}_g$ is substantial at 41.97% of the users. Such proportion characteristics arise because users in $\{u\}_g$ post both geocoded and non-geocoded tweets, with the latter at much larger counts. The last two rows of Table 1 illustrates this. On average, a Singapore user in $\{u\}_g$ post 289.69 geocoded tweets and 4532.98 non-geocoded tweets. A similar bias in tweeting behavior can be observed for Jakarta.

In short, users in $\{u\}_g$ have location history which makes it possible to build personalized geolocation models. With appropriate personalization, one should be able to better geolocate non-geocoded tweets posted by such users.

Table 1: Statistics for 50,000 sampled users from Singapore (2014) and from Jakarta (June to Dec, 2016).

	Singapore	Jakarta
Total Tweets	136,548,216	20,466,019
Geocoded Tweets	4,394,378 (3.22%)	946,432 (4.62%)
Users with geocoded tweets, $\{u\}_g$	15,169 (30.34%)	20,982 (41.97%)
Ave. geocoded tweets / user in $\{u\}_g$	289.69	45.11
Ave. non-geocoded tweets / user in $\{u\}_g$	4532.98	157.48

Spatial Focus and Spatial Homophily

Due to space constraints, we will present the empirical analysis of the two properties in a longer version of this paper.

Spatial Focus. Intuitively, an average user is constrained by geographical, social or personal factors. This leads to venue revisits or the conduct of much activities (e.g. work) in geographically localized regions. For example, one may frequent neighborhoods near the home or workplace. We say that the user is *spatially focused*, i.e. he is more likely to visit venues that are near his previously visited venues.

Spatial Homophily. Users in the same city/region generate more similar social media content when compared to another city/region (Cheng, Caverlee, and Lee 2010; Chang et al. 2012). We refer to this as spatial homophily with respect to locations. In fact, spatial homophily exists at very fine spatial scale as well. We have observed that venues near each other tend to have more similar tweet content, compared to venues further apart in the same city. This is partially contributed by the mentions of local spatial features, e.g. a landmark, or neighborhood characteristics, e.g. a nightlife area will have more tweets about clubbing and partying.

Proposed Model

Let **w** be a test tweet posted by user u at time of day t. To geolocate **w**, we rank candidate venues by:

$$p(v|\mathbf{w}, t, u) \propto p(v|t)p(u|v)\prod_{w\in\mathbf{w}} p(w|v) \qquad (1)$$

which is a product of probabilities modeling posting time, location history and tweet content.

Posting Time. $p(v|t)$ accounts for venue popularity at time of day t. We model time of day t as a continuous variable and estimate $p(v|t)$ in an approach motivated by kernel density estimation (KDE) (Lichman and Smyth 2014). For time of day t, define a time interval of length $T(t)$ which covers t. Denote V as the number of distinct venues, $f(v,t)$ as the number of user visitations to venue v in the interval $T(t)$ and let $f(.,t) = \sum_v f(v,t)$. Given a test tweet with time of day t, we have $p(v|t) = \frac{f(v,t)+\beta}{f(.,t)+V\beta}$ where β is the smoothing parameter.

Location History. To compute $p(u|v)$, we use the location history of u. Since location history are specific to users, it is more intuitive to compute $p(v|u)$ instead of $p(u|v)$. $p(v|u)$ can also be represented by 2-D distributions over geographical space, which is convenient for interpretation and visualization. By Bayes rule, $p(u|v) = p(v|u)p(u)/p(v)$ and assuming constant $p(u), p(v)$, we have $p(u|v) \propto p(v|u)$. Thus the probability term $p(u|v)$ in Equation (1) can be replaced by $p(v|u)$. To model $p(v|u)$, we recap that users are spatially focused in that he is more likely to visit venues spatially near any of his previously visited venues. To capture this idea, we define $p(v|u)$ as

$$p(v|u) \propto exp(-S \cdot \min(d(v, \mathbb{V}_u))) \qquad (2)$$

where $\mathbb{V}_u$ is the set of venues in u's location history, $d(.)$ measures spatial distances and $S \geq 0$ is the spatial smoothing parameter. For large S, $p(v|u)$ decreases faster with increasing distance between v and the nearest venue in $\mathbb{V}_u$. Equivalently the user is more spatially focused.

Tweet Content and Spatial Smoothing. Let W be the vocabulary size of tweet words. We use $c(w,v)$ as the frequency of word w at venue v and $c(.,v)$ to denote $\sum_w c(w,v)$. We compute $p(w|v) = \frac{c(w,v)+\alpha}{c(.,v)+W\alpha}$ where α is the smoothing parameter which can be tuned or set at 1 for Laplace smoothing. To account for the presence of spatial homophily, we spatially smooth $p(w|v)$. For each word w at the ego venue v, we extend the definition of $p(w|v)$ with word frequencies of v's set of spatial neighbors, denoted by $nb(v)$. The spatially smoothed $p(w|v)$ is defined as:

$$p(w|v) = \frac{c(w,v) + \alpha + \frac{\gamma}{|nb(v)|}\sum\limits_{v_i\in nb(v)} c(w,v_i)}{c(.,v) + W\alpha + \frac{\gamma}{|nb(v)|}\sum\limits_{v_i\in nb(v)} c(.,v_i)} \qquad (3)$$

where $0 \leq \gamma \leq 1$ is a weight factor to be set. When $\gamma = 1$, a word w found in every v's neighbor will be equivalent to a single w occurrence in v. Otherwise, the words from neighbors are weighted less than the native words in v.

Learning to Rank

Given a tweet, one desires its posting venue to be ranked high. Thus there is only one relevant venue and the Mean Reciprocal Rank (MRR) is a suitable metric. Given tweet i, let the rank of its posting venue be r_i, where $r_i = 0$ for the top rank. MRR is defined as $(1/N)\sum_{i=1}^{N}(1/(r_i+1))$ where N is the number of test cases.

We optimize our model parameters with respect to MRR via Learning to Rank (LTR). However, it is infeasible to maximize MRR directly (Christakopoulou and Banerjee 2015) via LTR. Instead, one has to approximate MRR maximization by minimizing a proxy loss function. Since maximizing MRR is equivalent to minimizing the sum of multiple 0-1 loss functions, a good proxy should approximate the 0-1 loss well, while retaining sufficient gradient for learning. We thus introduce the log-log loss function from (Christakopoulou and Banerjee 2015) into our models. This loss function has been proposed as a better alternative to logistic loss. For each model, we construct the loss function over venue pairs for minimization.

Loss function. For a posting venue v_i to be ranked high, $p(v_i)$ should be large while $p(v_j)$ should be small for $j \neq i$, i.e. non-posting venues. For computation convenience, we use log probabilities for ranking. Let $z(v_i, v_j) = \ln p(v_i) - \ln p(v_j)$ and $R(v_i, v_j) = \ln(1+e^{-z(v_i,v_j)})$. The log-log loss function for a tweet with posting venue v_i is:

$$L(v_i) = \sum\nolimits_{v_j \in \mathbf{V}\text{-}} \ln(1 + R(v_i, v_j)) \quad (4)$$

where **V-** is the set of non-posting venues. This can be all venues, randomly sampled or selected based on heuristics. To obtain the global loss function, one computes and sums $L(v_i)$ over all tweets.

Re-parameterization. With the loss function defined, we can perform gradient descent to minimize it. However there are constraints on the parameters. The smoothing parameters α, β and S are required to be positive. The spatial weight factor γ has to satisfy the constraint $0 \leq \gamma \leq 1$. Instead of constrained optimization, we incorporate the above constraints by re-parameterizing the model as follows: $\alpha = x_\alpha^2$, $\beta = x_\beta^2$, $S = x_S^2$ and $\gamma = (1 + e^{-x_\gamma})^{-1}$. The new parameters are now easily learnt from unconstrained optimization. In this paper, we use stochastic gradient descent.

Experiments

Data For model building and testing, we associate tweets with their posting venues using:

- **Shouts**: Comments authored by users as they check-in to venues in Foursquare, a popular location app. The comments are also referred to as shouts. We process the shouts (Cao et al. 2015; Li et al. 2011) to exclude the app-generated portion.
- **Pure tweets**: Non-geocoded tweets posted by users within 5 minutes of their check-ins. We assume these tweets are being posted from the check-in venues.

We collect data for users from Singapore (SG) and Jakarta (JKT). For Singapore, we collected 1,190,522 Foursquare check-ins from 2014, of which 30% involve shouts. We refer to this dataset as **SG-SHT**. We also collected 90,250 pure tweets and designate the dataset as **SG-TWT**. For Jakarta, the **JKT-SHT** dataset comprises 177,570 check-ins for the period 2015 to mid-2016, of which 49% are shouts. Linking the check-ins to pure tweets, we obtain only 1335 pure tweets. This small number is possibly due to API changes of the Foursquare platform which affected crawling.

Terminology. In this paper, 'tweets' refer to both pure tweets and shouts. Where differentiation is required, we use each term explicitly, i.e. pure tweets or shouts.

Setup. We split the datasets into training, tuning and test sets. Model parameters are learnt from the training set to minimize the loss on the tuning set. We include venues as ranking candidates only if they have at least 5 tweets in the training set. We also filter out stop words and rare words (frequency <4). The test set consists of test cases of tweets, each posted from some venue by a user with location history. On inspection, we noticed 'easy' test cases, where a user repeatedly uses a highly unique word everytime he posts from a certain venue. This makes the unique word highly indicative of the posting venue, leading to high ranking accuracy for such cases. To make the problem more challenging, we filter them from the training set as follows: for each test case with user u and posting venue v, we exclude u's other tweets posted at v from the training set. In other words, our training set does not observe any postings of u from venue v.

For each dataset, we conduct 20 runs where for each run, we sample 5000 tweets for testing/tuning and use the remaining for training. From the sampled set, we use 1000 tweets for tuning and the remainder for testing. Due to various filtering discussed above, the number of test cases per run is less than 4000. The average number of test cases are reported with the results for each experiment.

Models. We compare the following models:

- TFIDF: We represent venues and tweets as TFIDF vectors in terms of content. Given a test tweet, we use cosine similarity to retrieve and rank venues. This is very similar to the method in (Yohei Ikawa and Tatsubori 2012).
- NB: The naive Bayes, content-only approach from (Lee et al. 2014; Kinsella, Murdock, and O'Hare 2011).
- NB+S: This extends the NB model with spatial smoothing. For spatial smoothing, we use $k = 5$ nearest neighbors of each venue to smooth the word probabilities.
- NB+S+T: This uses content with spatial smoothing plus the contextual information of posting time.
- NB+S+T+U: Our proposed model with spatial smoothing, posting time and user location history.

We optimize each model with stochastic gradient descent. To account for local optimal, we randomly initialize and train 10 instances per model. We then select the instance with the highest tuning set Mean Reciprocal Rank (MRR) to apply on the test set.

Results on Shouts. Tables 2 and 3 present results for Singapore (SG-SHT) and Jakarta shouts (JKT-SHT) respectively. Both tables exhibit similar trends. TFIDF performs the worst. From the NB model onwards, MRR improves as we add spatial smoothing and additional contextual information to the models. For adjacent models, e.g. NB vs NB+S, we have also conducted significance testing with the Wilcoxon signed rank test. The differences between models are statistically significant at p-value of 0.05, except for the case of NB vs NB+S in Table 3 (p-value=0.067).

Comparing NB and NB+S, spatial smoothing improves MRR slightly. The improvement is small but consistent across different runs. This may be due to the limited strength of spatial homophily at fine granularities. We also note that prior work on coarse-grained geolocation (Cheng, Caverlee, and Lee 2010) had reported limited improvement from spatial smoothing. However another more probable explanation is that even without smoothing, we are already capturing much of the spatial homophily effect. Recall that this means venues near each other have more similar content. In the NB model, we are modeling the venue content directly anyway, thus implicitly accounting for spatial homophily in a downstream manner.

For both cities, substantial improvement comes from adding posting time and location history. For example,

Table 2: Ave. MRR for SG-SHT. On average, there are 2626.2 test cases and 10814.5 venues to rank per run.

Model	MRR	Improvement over NB
TFIDF	0.03571	-
NB	0.09592	-
NB+S	0.09622	0.31%
NB+S+T	0.09899	3.20%
NB+S+T+U	0.10224	6.59%

Table 3: Ave. MRR for JKT-SHT. On average, there are 975.9 test cases and 2713.75 venues to rank per run.

Model	MRR	Improvement over NB
TFIDF	0.04193	-
NB	0.13414	-
NB+S	0.13439	0.19%
NB+S+T	0.14564	8.58%
NB+S+T+U	0.14712	9.68%

NB+S+T provides 3.2% improvement over NB in Table 2. For Jakarta in Table 3, the corresponding improvement is 8.58%. Thus venue popularity with time of the day plays a role. Adding user location history helps to increase MRR even more, with NB+S+T+U being consistently the best performing model in both tables. This shows that location history is highly useful contextual information. Also recap that our modeling approach captures the idea that users are spatially focused in being more likely to visit venues that are near each other. The experiment results further validate this.

Results on Pure Tweets. Table 4 displays the results for training and testing on Singapore pure tweets (SG-TWT). The trend is similar to previous experiment on shouts. TFIDF performs poorly. Spatial smoothing again provides only slight improvement over the NB model, although it is statistically significant over 20 paired runs. The inclusion of time and location history provides very sharp improvement. NB+S+T+U again has the highest MRR with over 60% improvement from NB.

Typically, MRR is not compared across experiments that rank different number of items. However here, we can make certain statements by comparing Tables 4 and 2. In Table 4 for pure tweets, we rank fewer venues, but obtain mostly lower MRR than Table 2 for shouts. Since we have fewer venues to rank, the task should have been easier, resulting in a higher MRR. The lower MRR thus implies that it is more challenging to rank venues for pure tweets than shouts. One possible reason will be that pure tweets are about more diverse topics not related to the posting venue. Obviously this will impact ranking accuracy.

If the contents of pure tweets are not highly indicative of venues, then contextual information such as posting time and user location history become relatively more important. This is illustrated by the huge gains in MRR as we move from model NB to NB+S+T / NB+S+T+U. The percentage improvement is much larger in Table 4 than the case for shouts in Table 2.

Table 4: Ave. MRR for SG-TWT. On average, there are 2061.9 test cases and 2783.55 venues to rank per run.

Model	MRR	Improvement over NB
TFIDF	0.02059	-
NB	0.05539	-
NB+S	0.05565	0.46%
NB+S+T	0.07603	37.26%
NB+S+T+U	0.08986	62.24%

Conclusion.

We have proposed a model for fine-grained geolocation, which exploits contextual information such as posting time and location history. In addition for model design, we incorporate intuitive properties such as spatial homophily and spatial focus. Our proposed model is able to achieve a large improvement in ranking accuracy over baselines. Further work can include other contextual information for modeling, e.g. relationships.

Acknowledgments

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative, and DSO National Laboratories.

References

Ahmed, A.; Hong, L.; and Smola, A. J. 2013. Hierarchical geographical modeling of user locations from social media posts. *WWW*.

Cao, B.; Chen, F.; Joshi, D.; and Yu, P. S. 2015. Inferring crowdsourced venues for tweets. *Big Data*.

Chang, H.-W.; Lee, D.; Eltaher, M.; and Lee, J. 2012. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. *ASONAM*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. *CIKM*.

Christakopoulou, K., and Banerjee, A. 2015. Collaborative ranking with a push at the top. *WWW*.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. M. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. *ICWSM*.

Hong, L.; Ahmed, A.; Gurumurthy, S.; Smola, A.; and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the twitter stream. *WWW*.

Kinsella, S.; Murdock, V.; and O'Hare, N. 2011. "I'm eating a sandwich in Glasgow": modeling locations with tweets. *SMUC*.

Lee, K.; Ganti, R. K.; Srivatsa, M.; and Liu, L. 2014. When twitter meets foursquare: tweet location prediction using foursquare. *MobiQuitous*.

Li, W.; Serdyukov, P.; de Vries, A. P.; Eickhoff, C.; and Larson, M. 2011. The where in the tweet. *CIKM*.

Lichman, M., and Smyth, P. 2014. Modeling human location data with mixtures of kernel densities. *KDD*.

Yohei Ikawa, M. E., and Tatsubori, M. 2012. Location inference using microblog messages. *WWW (Companion Volume)*.