Placing User-generated Content on the Map with Confidence

Suradej Intagorn
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292 USA
intagorn@usc.edu

Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292 USA
lerman@isi.edu

ABSTRACT

We describe a method that predicts the location of user-generated content using textual features alone. Unlike previous methods for geotagging text documents, our proposed method is not sensitive to how we discretize space. We also discover that spatial resolution has an impact on the prediction accuracy, which allows us to trade-off the spatial resolution of the predicted location against our confidence about its accuracy. Our method can be used to estimate the error in document's predicted location, enabling us to filter out poor quality predictions. We evaluate the proposed method extensively on user-generated content collected from two different social media sites, Flickr and Twitter. Our evaluation examines its performance on the geotagging task and with respect to different parameters. We achieve state-of-the-art results for all three tasks: location prediction, error estimation and result ranking and also provide a theoretical explanation of the effect of spatial resolution factor on geotagging accuracy. Our findings provide valuable insights into the design of geotagging systems and their quality control.

Categories and Subject Descriptors

H.4.0 [**INFORMATION SYSTEMS APPLICATIONS**]: General

General Terms

Theory,Algorithms

Keywords

Geographic information retrieval, Geolocation

1. INTRODUCTION

We describe a method that places on the map short text documents, such as photos and posts shared in social media, by predicting their location and also estimates the quality of the prediction by giving its confidence and prediction error. The method starts by estimating the spatial probability density of document terms using a training set of geotagged documents. Then, given a new document, it predicts as its location one that maximizes the confidence of the prediction. We show that this method is as simple as the popular alternative, Naive Bayes, but leads to better performance on the location prediction task and moreover, allows us to directly estimate the quality of predictions.

Previous methods used to automatically identify the location of text documents fall into two main classes. The first uses gazeteers to extract place names from text. The second class of methods for geotagging documents uses statistical models, such as classification, to associate terms with places. This approach can more accurately estimates geographic location of social media content than the gazetteer-based approaches [5]. However, one challenge faced by these methods is that of scale: changing scale can significantly affect spatial statistics and classification results. While a variety of solutions were proposed (e.g., using entropy at all spatial scales [6]), they are often *ad hoc* or limited by the assumptions they make (e.g., assuming a specific form of the probability distribution). Our approach addresses these challenges with non-parametric estimation of spatial density distribution.

SIGSPATIAL'14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666433 .

2. PREDICTION CONFIDENCE

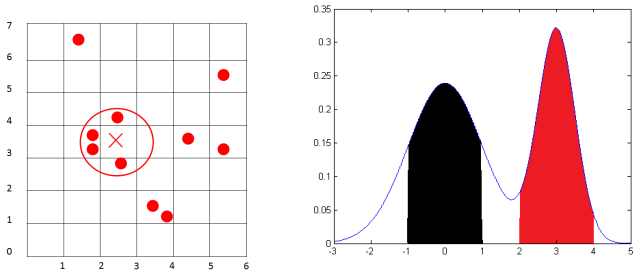
Given a new document with some textual features or terms, such as a tag 'california', we can predict its location and moreover estimate the quality of the prediction. We integrate PDF of the features over some region of radius r, and return as document's predicted location the center of the region that maximizes this value, which we call *confidence*. The radius r controls the *error* of the prediction and gives its *spatial resolution*. The confidence represents our certainty about the prediction at that spatial resolution, i.e., the likelihood that document's true location falls within the region of radius r centered on the predicted location. There exists a trade-off between resolution scale and the degree of confidence. We can increase the confidence of the prediction by increasing the resolution scale r. For example, we can increase confidence to 100% by setting r very large.

To estimate confidence from discrete samples, we calculate the likelihood of finding a sample within a circle of radius r (spatial resolution) by counting the fraction of training samples that fall within the circle.

$$Confidence(x, r, w) = N_w(x, r)/N_w$$

Here, N_w is the number of samples containing a feature w, and $N_w(x, r)$ is the number of samples with a feature w within circle of radius r centered at location x. Note that we can raise the confidence by increasing the spatial resolution: since N_w remains constant, but $N_w(x, r)$ increases monotonically with r.

Figure 1(a) illustrates the calculation. Here, red points are samples of feature w_1. Given a new document with a single w_1, the confidence of the predicted location x at the specified spatial res-



(a) (b)

Figure 1: Illustrations of confidence estimation from (a) samples and from (b) the probability density function.

olution is 40%, which is the fraction of training samples that fall within the circle centered at x.

In practice, the training set of documents containing terms W, e.g., photos with tags {'california','john','2007', 'iphone','hbd'}, may be very sparse. As a result, we may not have enough samples to predict their location with high confidence. Instead, we assume that we can approximate the probability density function of terms by combining the individual PDFs of each term. If we know the true PDF that generated the document with terms W, we can compute the prediction confidence at location x with spatial resolution r as:

$$Confidence(x_0, r, W) = \int_R f_W(x)\,\mathrm{d}x. \tag{1}$$

Here x_0 is the location at which we want to measure prediction confidence, and r is the given spatial resolution. The function $f_W(x)$ is the probability density function of documents with terms W. A detailed discussion of $f_W(x)$ estimation is provided in the next section. The region R is the set of locations with distances to x_0 that are less than r: $R = \{x \in X | dist(x_0, x) \leq r\}$.

Another view of confidence, $Confidence(x_0, r, W)$, is that it is the probability mass of $f_W(x_0)$ within a circle of radius r centered at x_0. Depending on the specific density distribution, the predicted location will vary with r. We illustrate this point with a one-dimensional example in Figure 1(b). The example can be easily generalized to two-dimensional geographic space.

The predicted location is the center of a segment of length $2r$ that gives the highest confidence, i.e., area under the PDF curve in this one-dimensional example. The predicted location depends significantly on the choice of r in this example. When r is small, e.g., $r = 0.01$, we will predict location at $x = 3$, because this location results in the largest probability mass at that value of r. However, when $r = 1$, location $x = 0$ is preferred over $x = 3$, because the probability mass at that location (black area) is larger than at $x = 3$ (red area).

We calculate the integral of $f_W(x)$ numerically using the Riemann sum. The accuracy of the sum depends on the grid size parameter used to discretize space. We empirically investigate this dependence in Section 3.

2.1 Confidence-based Location Prediction

The most likely location of a document containing features W is the location x that maximizes confidence:

$$x_{predict} = \arg\max_{x \in X} Confidence(x, r, W)$$

The set X of possible locations can be computed by discretizing space into small grids, or we can assume that X is a set of locations of training samples of terms W. The confidence can be computed from integral of $f_W(x)$ described in the previous section.

In practice, we do not know the true density $f_W(x)$. We can only approximate it. Below we describe a simple method that is easy to implement and relatively fast in practice. Improving accuracy of PDF estimation is the subject for future research. There are two important choices for estimating $f_W(x)$. The first choice is how to model the combined PDF of all terms $w \in W$. We use the *mixture model* to approximate $f_W(x)$ as the weighted sum of individual PDFs $f_w(x)$:

$$f_W(x) = \sum_{w \in W} \lambda_w f_w(x), \tag{2}$$

where λ_w is the weight of term w, which represents the importance of that term. For example, it is obvious that the term 'california' should have the highest weight in a photo tagged {'california', 'iphone', 'ramen'}, because it is the most geographically indicative term. Thus, we can incorporate the feature selection into this model by giving such terms a higher weight. For simplicity, in this section, we assume that all terms are weighted equally, and $\lambda_w = 1/|W|$. The second modeling choice is how to describe, $f_w(x)$, the spatial distribution of each term w. In this work, $f_w(x)$ is estimated by the histogram method.

2.2 Prediction Error

When placing documents on the map, it is useful to give the error associated with the predicted location. We employ the confidence prediction framework to estimate the location error. First, we estimate the probability mass around the predicted location, as described above. If the predicted location has moderate probability mass within a small spatial resolution r, it is likely that this predicted location has low error. If another location has the same probability mass but at a larger spatial resolution, we can infer that this prediction has a higher error than the first prediction.

We can dynamically adjust the resolution scale to optimize prediction confidence. We start by setting the error to minimum spatial resolution r. If the probability mass at this error bound is higher than threshold, we report the estimated error as r. If not, we increase the spatial resolution until the probability mass exceeds threshold. The confidence error estimation method is:

x = predicted location
e = estimated error
while $Confidence(x, r, W) < \theta$ **do**
  e = r
  $r = r + \Delta$
end while

The estimated error increases at every iteration until the confidence value exceeds the threshold. The algorithm is guaranteed to terminate if $\theta < 1$ because $Confidence()$ is a monotonically increasing function of r and always reaches 1 for large enough r.

2.3 Adaptive Resolution Prediction

The most appropriate level of granularity at which a document should be localized is generally not known beforehand and must be estimated [7]. For example, if only tag available for a document is 'lax', we should set the spatial resolution, r, to a landmark level instead of a city level. Adaptive resolution prediction provides a solution to this problem by searching for the smallest spatial resolution that make the prediction confidence high enough:

x = predicted location
r = spatial resolution
while $Confidence(x, r, W) < \theta$ **do**
  $x = \arg\max_{x \in X} Confidence(x, r, W)$

$r = r + \Delta$

end while

*Weighted confidence prediction:*We can improve the performance of the location prediction algorithm by giving more weight to terms that are more predictive of a document's location. One popular procedure for identifying place names is to measure how well-localized the probability density function $f_w(x)$ is. We can estimate how well-localized the term w is by calculating the uncertainty of x using continuous entropy:

$$H_w(x) = \int_X \hat{f}_w(x) \ln \hat{f}_w(x) \, \mathrm{d}x. \quad (3)$$

As in any nonparametric smoothing application, the statistical properties of estimators such as $\hat{f}_w(x)$, dependent on the choice of the bandwidth parameter h, and will, in turn, affect estimated entropy. An inappropriate value of h may lead to an estimator with a large bias or variance or both. The optimal bandwidth can be obtained by using the Least Squares Cross Validation method (LSCV) [8].

We integrate feature selection into the confidence prediction framework called *weighted confidence prediction*. Instead of using uniform weights in the mixture model Eq. 2, we give high entropy features (above some threshold) a low weight, because they are probably not indicative of places.

3. EVALUATION

We used Flickr and Twitter data to evaluate the performance of the proposed geotagging method. The Flickr data came from the placing task competition in the MediaEval2013 workshop. [1] We used the third test set with 53,000 documents in our experiments. We filtered out documents that did not contain any textual information, because our method does not have anything to go on in predicting their location. After filtering, there were 45,742 documents in the test set. The Twitter data set contained tweets we collected using twitter4j, a Java library for the Twitter API. After filtering, there were 44,289 documents in the Twitter test set

We evaluate how well the proposed location prediction method performs with respect to different parameter values and also how accurate its predictions are compared to the baseline. The five experiments presented below summarize our findings. All experiments were carried out on both Flickr and Twitter data sets. Following convention [2], we report the bandwidth and spatial resolution (error bound) in the units of degrees and location errors in kilometers. While the spatial extent of one degree is location-dependent, for simplicity we take it to be approximately 100km.

The first experiment tested the effect of the histogram discretization parameter, the bandwidth h, on the accuracy of the proposed method. We hypothesize that bandwidth does not affect prediction accuracy very much. As long as bandwidth is relatively smaller than the spatial resolution scale r, predictions made using different bandwidths should be approximately the same. This is because bandwidth only affects the value of the Riemann sum used to calculate the confidence, not the predicted location itself.

The results confirm our hypothesis. At the $r = 100$km error level, prediction accuracy is 0.6 and approximately same regardless of grid size. The result can be interpreted that the algorithm can predict 60 of 100 documents with less than 100km error. Below the 50km error level, finer bandwidths produce more accurate results, probably due to better accuracy of the Riemann sum. Thus, we can choose $h \to 0$ to obtain a good accuracy at small level without sacrificing the accuracy at larger error bounds. However, small grid size can increase computational time due to larger search space and Riemann sum computation.

In the second experiment we test the effect of spatial resolution of the prediction on its accuracy. We hypothesize that we can maximize accuracy at different spatial granularity levels by changing the error bound r. There is no optimal error bound for the prediction: for example, we may want to sacrifice accuracy at larger spatial resolutions to obtain better accuracy at finer resolutions. Theoretically, the error bound r leads to the maximum accuracy at that resolution, because we are maximizing confidence, or the mass of the PDF in a circle of radius r. Results in both data sets confirm our hypothesis about the relationship between the spatial resolution, r, and prediction accuracy at some error levels. Thus, if we want more accurate predictions at a finer-grained level, for example, at a city-level rather than country-level, we should set the error bound lower. For example, to maximize prediction accuracy at 10km, we have to set the spatial resolution to $r = 0.1^\circ$ (10km). However, if we are interested in getting the country-level of the prediction right, we should set the error bound at a larger value, e.g., $r = 5^\circ$ (500km).

In the third experiment we compare variations of the proposed location prediction method to the Naive Bayes (NB) baseline. Specifically, we compare the *fixed* error bound confidence-based location prediction method , to term-weighted, to adaptive confidence prediction with the baseline. The baseline has two parameters: grid size and the smoothing parameter. We set the grid size to $1^\circ \times 1^\circ$ and smooth densities using the add-one method. The fixed confidence prediction has two parameters: bandwidth, h, which we set to 0.01°, and the spatial resolution, r, which we set to $r = 0.5$.

The weight confidence prediction has two additional parameters: weights and weight threshold. We use weight to $w_i = 1$ to high entropy features and other features $w_i = 2$. We rank the features according to their entropy and set the first 20% to be high entropy features in both data sets. Summation of weights in a document is then normalized to be 1.

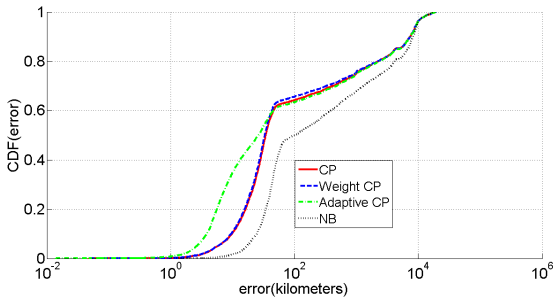
The adaptive confidence prediction has three parameters: bandwidth (h), maximum spatial resolution (r_{max}) and confidence threshold (θ). We set bandwidth, h, to 0.01°, the maximum spatial resolution to $r_{max} = 0.5$, and confidence threshold to $\theta = 0.2$.

Figure 2 reports results of this experiment in the two data sets. The proposed confidence prediction method outperforms the baseline significantly on Flickr data at all granularity levels. For Flickr data, the accuracy of the normal confidence prediction is significantly higher than baseline. The weighted mixture model leads to only a slight improvement. A possible reason for this is that non-place tags are uniformly distributed, reducing their weights may not change the mode's location. However, these tags can change the calculated confidence values.

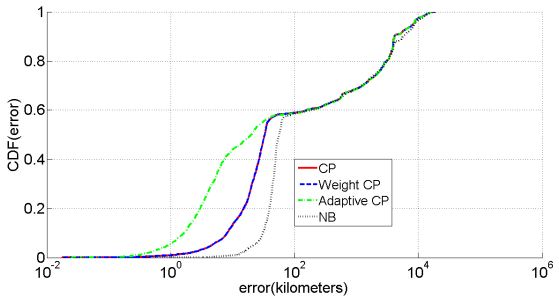
Next, in the forth experiment , we compare error estimates produced by the Naive Bayes's top n method (the variation method) [4] and the proposed method. We use the Kendall-Tau correlation suggested in [4] to compute the correlation between the actual errors the estimated errors. Higher correlation means better estimation. Alternatively, we can view error estimation as a decision problem, i.e., decide whether location error is lower than a specified bound or not? We set the boundary decision to the median of each method to have the same number of positive and negative classes. Accuracy [1] is used to evaluate performance.

For Flickr data, the proposed confidence prediction framework is better than baseline according to both metrics. Error estimation can be used to rank the quality of prediction results. In some situations, we may improve prediction quality by sacrificing coverage, in other words, by filtering out documents whose locations we cannot precisely predict, we retain a small fraction of documents

[1] http://www.multimediaeval.org/mediaeval2013/



(a) Flickr



(b) Twitter

Figure 2: Evaluating the performances of four different methods: normal, weight, adaptive confidence prediction and add-one smoothing Naive Bayes classifier on the prediction error

Method	Data Set	Correlation	Accuracy	Median(km)
Variation	Twitter	0.2460	0.5033	56
Confidence	Twitter	0.5528	0.8067	33
Weight	Twitter	0.5556	0.8097	33
Variation	Flickr	0.2912	0.5066	102
Confidence	Flickr	0.3864	0.6789	34
Weight	Flickr	0.4221	0.6797	33

Table 1: Evaluating the performances of three different methods: normal confidence, weight confidence and Variation method on the predicted error estimation task. The Kendall-Tau correlation and accuracy metric is used to evaluate the performance.

whose predicted locations we believe to be close their actual locations. In the final experiment, we evaluated this aspect of the error estimation task.

We sorted locations predicted by each method based on their estimated errors. We sorted predictions in an ascending order of their estimated errors, with lowest error predictions at the top of the list. Then, we select the top 75% results to see the performance improvements. The baseline variation has the lowest improvement on both data sets. The state-of-the-art Probability ratio [3] has a comparable performance to the proposed method in both data sets.

4. CONCLUSION

In this paper, we propose a framework for predicting the location of a document that allows us to estimate the confidence of the prediction and its error. This allows us to determine whether the document is *placable* on the map or not: if the confidence is low, and error is high, then the document cannot be accurately geotagged. Our framework also allows us to predict at different levels of spatial granularity. Using calculated confidence values, we can adaptively find the granularity level appropriate for a given document's features.

Method	Data Set	Acc	Acc 75%	Improve
Variation	Twitter	0.5805	0.6676	15.0
PR Ratio	Twitter	0.5805	0.7305	25.8
CP	Twitter	0.5892	0.7501	27.3
Weight CP	Twitter	0.5892	0.7508	27.4
Variation	Flickr	0.4990	0.5617	12.5
PR Ratio	Flickr	0.4990	0.6028	20.8
CP	Flickr	0.6420	0.7719	20.5
Weight CP	Flickr	0.6546	0.7983	21.9

Table 2: Comparing the performances of four different methods: normal confidence, weight confidence, Variation and Probability Ratio method on the result ranking task. All accuracies are reported at the error level $= 100km$. Acc column reports the accuracy before filtering. $Acc75\%$ column reports the accuracy after filtering 25% of results.

Confidence prediction framework shows promising results . We evaluated its performance in two different social media data sets on three different tasks: location estimation, error estimation and result ranking. We showed that the proposed method outperforms the baseline on all three tasks.

Acknowledgments

This material is based upon work supported by the AFOSR (contract FA9550-10-1-0569) and by DARPA (contract W911NF-12-1-0034)

5. REFERENCES

[1] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3121–3124. IEEE, 2010.

[2] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 761–770, New York, NY, USA, 2009. ACM.

[3] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *J. Artif. Intell. Res.(JAIR)*, 49:451–500, 2014.

[4] C. Hauff, B. Thomee, and M. Trevisiol. Working notes for the placing task at mediaeval 2013. In *MediaEval 2013 Workshop, Barcelona, Spain*, 2013.

[5] S. Kinsella, V. Murdock, and N. O'Hare. 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. ACM, 2011.

[6] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1):1, 2009.

[7] O. Van Laere, S. Schockaert, and B. Dhoedt. Georeferencing flickr resources based on textual meta-data. *Information Sciences*, 238:52–74, 2013.

[8] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60. Crc Press, 1994.