A Probabilistic Approach to Mining Geospatial Knowledge from Social Annotations

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

User-generated content, such as photos and videos, is often annotated by users with free-text labels, called tags. Increasingly, such content is also georeferenced, i.e., it is associated with geographic coordinates. The implicit relationships between tags and their locations can tell us much about how people conceptualize places and relations between them. However, extracting such knowledge from social annotations presents many challenges, since annotations are often ambiguous, noisy, uncertain and spatially inhomogeneous. We introduce a probabilistic framework for modeling georeferenced annotations and a method for learning model parameters from data. The framework is flexible and general, and can be used in a variety of applications that mine geospatial knowledge from user-generated content. Specifically, we study three problems: extracting place semantics, predicting locations of photos and learning part-of relations between places. We show our method performs well compared to state-of-the-art approaches developed for the first two problems, and offers a novel solution to the problem of learning relations between places.

1. INTRODUCTION

The Social Web sparked a revolution by putting knowledge production tools in the hands of ordinary people. Today on Social Web sites such as Twitter, Flickr, and YouTube, large numbers of users not only create rich content, including photos and videos, but also annotate content with descriptive labels known as *tags*, and georeference it by associating with it geographic coordinates, known as *geo-tags*. The implicit relationships between tags and their locations can be mined to learn what places people talk about on the Social Web and how much attention they give to a place, what the extent of it is, and how it relates to other places.

Several researchers have recently investigated approaches to learning relations between concepts from social metadata. Schmitz [13] applies a statistical subsumption model [12] to learn hierarchical relations between tags. He addresses the challenge of the popularity vs generality problem using tag frequency. Plangprasopchok et al. [10] learn folksonomies by aggregating many shallow individual hierarchies, expressed through the collection/set relations on Flickr. They address many challenges in their approach, for example, differences in the level of expertise and granularity for each user.

In this paper we propose a probabilistic framework for mining geospatial knowledge from social annotations. One challenge in spatial data analysis is the problem of scale. Changing scale can significantly affect spatial statistics [9] and classification results [16]. In the place identification problem studied by [11], researchers addressed this problem by computing statistics at every predefined scale, then used sum of these statistics for significance testing. Methods developed for location prediction [3, 14] also relied on discretizing data. We represent the spatial distribution of a tag as a mixture of Gaussian probability density functions. Such Gaussian mixture models (GMMs) can be estimated directly from data without using discretization parameters. Being probabilistic, the model also presents a natural way to deal with noise and uncertainty.

The proposed probabilistic framework is general and flexible and can be used in a variety of geospatial data mining applications. We evaluate its performance on three distinct tasks. In the first task, *place identification* [11], we classify tags as place names or not place names based on their spatial distributions. In the second task, *location prediction* [3, 14], we attempt to predict locations of photos given their tags. In the third task, we use the probabilistic framework to *learn relations* between places.

2. PROBABILISTIC FRAMEWORK FOR MINING GEOSPATIAL DATA

We focus on analyzing the social photo-sharing site Flickr, which allows registered users to upload photos and videos and annotate them with descriptive labels, known as *tags*. Tags are used to describe the image's subject (e.g., 'animal', 'family'), properties, as well as where the image was taken (e.g.,'california'). In addition to tagging photos and videos, Flickr also allows users to *geo-tag*, or geo-reference, content by attaching geographic coordinates to it.

The implicit relation between tags and locations of photos annotated with those tags can tell us much about how people think of places and relations between them. We model tags and photos in terms of spatial probability distributions or density functions. Spatial distribution of tag w can be written as $P(X|w)$, where X represents locations of photos tagged with w. Then, we can easily model spatial probability distribution of a photo as a superposition of probability distributions of all tags in that photo.

We start by modeling probability distribution of each tag separately. Using a sufficient number of Gaussian components, and by tuning the parameters of each components and adding them linearly, most continuous distributions can be approximated [2]. Moreover, probabilistic models can address the challenges of noise, uncertainty and ambiguity. However, there still remains a challenge to using GMMs to model spatial distribution of tags, as the number

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

of components may not be known beforehand, and using too many components puts us in danger of overfitting the data. The section below proposes a solution to this problem.

Tag Distribution as a Mixture of Gaussians: Gaussian mixture model is a superposition of K Gaussians:

$$P(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \quad (1)$$

where each Gaussian density $\mathcal{N}(x|\mu_k, \Sigma_k)$ is called a component of the mixture with mean μ_k and covariance matrix Σ_k. Parameter π_k, called the mixing coefficient, gives the weight of the kth component, or the fraction of all data points that are explained by that component. It has a value between 0 and 1, and $\sum_{k=1}^{K} \pi_k = 1$.

The Gaussian mixture model is specified by the number of mixture components, each governed by parameters π_k, μ_k, Σ_k. For a given model, we use expectation-maximization (EM) algorithm [4] to estimate model parameters π_k, μ_k, Σ_k. EM is an iterative algorithm with two major steps: expectation (E) and maximization (M) step. The E step estimate $\gamma(z_{nk})$ from the current parameter values where $\gamma(z_{nk})$ can be viewed as as responsibility of component k generate x_n. The M-step updates the values of π_k, μ_k, Σ_k from $\gamma(z_{nk})$ of the previous step. The process continues until convergence. The convergence is usually defined by when the log likelihood function or finding parameters changes below some threshold [2].

How many mixture components K should we use to model each tag? By using more components (increasing the number of model parameters), one can usually get the model to better describe the data. However, this may lead to overfitting, where a very complex model explains every point in the training data set, but it cannot generalize, and therefore, has no predictive power on test data. We can reduce this problem by penalizing model for its complexity (e.g., using AIC or BIC) The BIC is used for model selection in statistics. It avoids overfitting by introducing a penalty term for the number for parameters in the model [8].

Our model selection process is very simple. We estimate model parameters using the EM algorithm to get the maximum likelihood estimate $\mathcal{L}(K)$ with respect to the number of components K. We then choose the values of K that minimizes the BIC value of the model: $K = \arg\min_K BIC(K)$. Thus, each concept has different number of components. About 34 % of the tags in our data set have between one and ten components, 57 % have between 11 and 20 components and 9 % of the tags have more than 20 components (the maximum number of components we consider is 30).

3. APPLICATIONS

One of the main challenges of mining user-generated content is to extract structured knowledge from a set of annotations. This is done by exploiting their usage patterns. Modeling the distribution of a tag as a mixture of Gaussians offers a general and flexible framework for data mining applications in the geospatial domain. In this section, we study the problems of place identification, location prediction, and learning part-of relations between places. The goal of the first task is to learn whether or not a tag refers to a place based on its spatial distribution. The goal of the second task is to geo-reference a photo based on its tags. Our approach exploits the structure in the data by probabilistically modeling the relations between a photo's location and its tags using the Gaussian mixture model. The goal of the third task is to learn relations between places, for example, that 'disneyland' is in 'california'. This approach can be used to learn novel relations that do not exist in official gazeteers, but may reflect common folk knowledge, 'santa monica' is in 'los angeles'. Our proposed algorithm compares spatial distribution of two tags and if the distribution of a geographically broader term, e.g., 'california', substantially overlaps the distribution of the more localized term, e.g., 'disneyland', then we learn that 'california' subsumes 'disneyland'.

Data collection: The data for the experiments was collected from the social photo-sharing site Flickr. We used the Flickr API to retrieve information about more than 14 million geo-tagged photos created by 260K distinct users. These photos were tagged over 89 million times with 3 million unique tags. As a preprocessing step, we filter out tags which were used by fewer than 100 people. To create the training data set for learning GMM parameters of the distribution of a given tag, we sample a single photo from each of the remaining users uniformly from a 100 km grid described in [3]. After these preprocessing steps, the training set contains 2.5M photos with 192K distinct users and 11K distinct tags.

3.1 Place Identification

Rattenbury et al. [11] observed that "place tags exhibit spatial usage patterns that are significantly geographically localized" [11]. While these photos can be found all over the world, they are much more dense around California. Rattenbury et al. [11] proposed a quantitative method to identify place tags. Their solution relied on multi-scale analysis method that they called Scale-Structure Identification. They tackle problem of MAUP by clustering data at many different values of the scale parameter r and combine entropy values at different scales. Their method is used as the baseline in this experiment. We use the same intuition to distinguish between place or non-place tags, but analyze the spatial patterns of the tags using the probabilistic modeling framework.

Model-Based Place Identification: Instead of discretizing data at different scales, we work with a continuous probability density function. We decide whether a tag is well-localized by examining the continuous entropy of $P(X|w)$. The intuition behind our method is as follows. People usually use non-place tags everywhere in the world, for example, people use the tag 'iphone' all over the globe. This tag has very high uncertainty, thus, it is unlikely that we can predict locations of photos tagged 'iphone'. In contrast, the tag 'alcatraz' is highly localized. In other words, it has low uncertainty in the geospatial domain.

Entropy is used to measure uncertainty of a distribution. In this application, we use entropy to estimate the uncertainty of $P(X|w)$, the spatial distribution of the tag w. There are advantages to using entropy in continuous space. First, geographic locations occur in continuous space. To estimate entropy in continuous space, there is no need for discretization parameters. Once entropy is estimated, it can be directly manipulated, e.g., compared to a threshold,without further processing as in discrete entropy method.

The continuous entropy of tag w can be estimated by

$$H\big(P(X|w)\big) \quad = \quad -\int_X P(x|w)\, \log P(x|w)\, dx \quad (2)$$

There is no analytic form for computing the entropy. Instead, we estimate it using the Monte Carlo method [6]. The idea is to draw a sample x_i from the probability distribution $P(x)$ such that the expectation of $\log P(x)$, $-E_{P(x)}\ (\log P(x)) = H(P)$. Monte Carlo method is used because, according to [6], Monte Carlo sampling has highest accuracy for approximation for high enough number of samples; however, it is computationally expensive. A better runtime approximation can be obtained by more sophisticated approaches discussed in [6]. The Monte Carlo approximation of en-

approach	*AUC*	*Max F1*	*Min CE*
baseline	0.6805	0.6807	0.0937
gmm	0.6989	0.7060	0.0904

Table 1: Comparison of the model-based approach (gmm) to baseline on the place identification task.

tropy can be expressed as:

$$H_{MC}(P) = -\frac{1}{n}\sum_{i=1}^{n}\log P(x_i) \quad (3)$$

where x_i is a sample from the distribution $P(x)$. As the number of samples n grows large, $H_{MC}(P) \rightarrow H(P)$. If a tag's entropy is lower than some threshold θ, then $P(x|w)$ is judged to be localized, and we identify the tag w as a place name:

$$H(P(x|w)) < \theta \quad (4)$$

Evaluation: To compare the performance of different methods on the place identification task, we need a ground truth about place names. For this purpose we use GeoNames (http://geonames.org), a geographical database containing over eight million place names from all over the world.

We use standard precision and recall metrics to evaluate the performance of baseline and the proposed model-based method on the place identification task. Precision measures the fraction of tags that were correctly predicted to be place names relative to the number of all tags that were predicted to be place names. Recall measures the fraction of place names in GeoNames that were predicted to be place names. However, precision and recall are sensitive to threshold value; therefore, we compute the performance as the precision-recall curve. Aggregate metrics, including AUC (area under precision-recall curve, maximum f1 score (Max F1), and minimum classification error rate (Min CE), are reported in Table 1. As we can see from these results, our method performs slightly better than baseline. Its true advantage, however, is flexibility, as the same probabilistic framework can be used to address different geospatial data mining tasks, as shown below.

3.2 Location Prediction

We apply the probabilistic model to solve a different problem, namely, find the most likely geographic location of a photo given its tags. Previous researchers framed the location prediction problem as one of classification [3, 14]. Crandall et al. [3], for example, used the mean-shift clustering algorithm to separate photos into distinct clusters, with each cluster representing a class of locations. The tags of the photos in the cluster are used as prediction features. Their method computes the probability of a class and probability of each feature given the class. Then, given a new photo with some tags, they use a naive Bayes classifier to predict the class the photo belongs to. Their method is used as the baseline in this experiment.

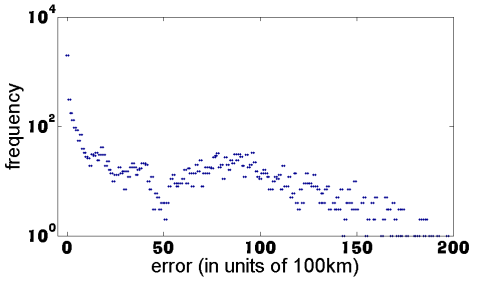
Model-based Location Prediction: We model a photo as a probability density function $P(x)$ in continuous space. Our model-based approach allows us to express the relationship between a photo's location and its tags, whose distributions were learned from the training data. In previous section we have already computed $P(x|w)$ where x is a geographic location and w is a tag. We will show that we can model a probability density function of a photo with this density function.

The assumption in our model is that tag frequency in one photo is one, i.e., a user does not repeatedly use the same tag in a photo. Thus, the probability distribution of each tag in a photo is uniform. Given $W = \{w_i\}$, the set of tags in a photo, then using the assumption above, $P(w_i) = 1/|W|$, where $|W|$ denotes the number of tags in the photo. Probability density function of a photo is a superposition of bivariate Gaussian distributions because each component, $P(x|w)$, is superposition of bivariate Gaussian distributions. The distribution $p(x)$ can be interpreted as probability of the photo's location. Therefore, the best guess of location x of a photo is the mode of $p(x)$, in other words, location x that corresponds to the maximum value of $p(x)$.

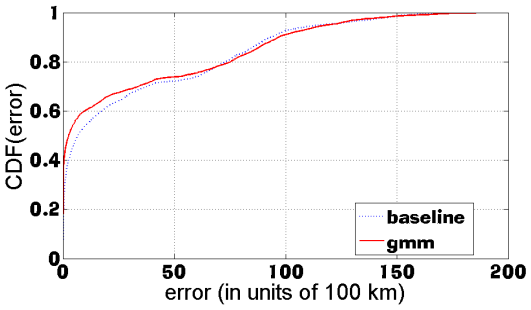
$$x_{predict} = \arg\max_x P(x) = \arg\max_x \sum_{w \in W} P(x|w)P(w)$$

We can ignore the term $P(w)$, because it is constant, independent of the photo's location. However, $P(x)$ of a photo is a non-linear function. To optimize this function numerically, in current implementation, we use Matlab implementation of a numerical method called Nelder-Mead Simplex Method [7].

Evaluation: We use methodology described in [3] to prepare the test data set for evaluating the proposed method. The test set is created by randomly selecting 5000 users, then choosing a photo at random from each user. Thus, the test data set consists of 5000 photos from 5000 users. Finally, photos from the users in the test set are removed from the training set to prevent bias that may come from having the same users in the training and test data sets.



(a) error pdf



(b) error cdf

Figure 1: Comparison of performance of different methods in terms of the error between the photos' predicted and actual locations. (a) Distribution of errors in the test set. (b) Cumulative distribution function of errors produced by the proposed method (gmm) and baseline.

We hide the actual locations of photos in the test set and use our method and baseline to predict their locations. We compare performances between our method and baseline using Geographical distance [15] between predicted (prd) and actual (act) locations by using haversine formula. Figure 1(a) shows the distribution of prediction errors made by the model-based approach as a histogram, where each bin corresponds to a unit of 100 km. The bins

corresponding to the lowest errors have the highest frequency, implying that our method results in small prediction errors most of the time. Figure 1(b) compares the cumulative probability distribution (CDF) of the errors made by the proposed method (gmm) and baseline. The proposed method has higher probability for lower errors, meaning that it produces better predictions than the baseline.

Evaluation: Existing location prediction methods suggest using gazetteers to improve prediction accuracy [1, 5, 14]. The key idea is that place-related terms, e.g., 'goldengate', should have higher weight than non-place-related terms, e.g., 'iphone', in classification.

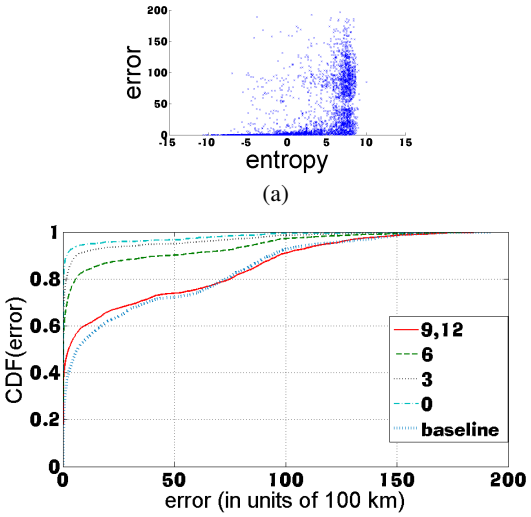


Figure 2: Using tag entropy to improve location prediction from tags. (a) Scatter plot of the prediction error (in units of 100 km) vs entropy of photo's representative (lowest entropy) tag. (b) CDF of prediction errors after filtering out photos whose tags are not well localized. Each line corresponds to a different filtering threshold, i.e., the lowest entropy value for the most localized tag in the photo.

Our model-based method can integrate tag's spatial uncertainty to come up with better, more accurate location predictions. Intuitively, localized tags tend to have higher predictive power; therefore, tags that have low entropy will produce a lower error on the location prediction task. Figure 2(a) validates this assumption. It plots the entropy of the representative (most localized) tag of each photo vs error of the predicted location.

Our model-based method can integrate tag's spatial uncertainty to come up with better, more accurate location predictions. Intuitively, localized tags tend to have higher predictive power; therefore, tags that have low entropy will produce a lower error on the location prediction task. We can use these results to filter out photos the locations of which we cannot confidently predict. We repeat location prediction experiment, keeping photos whose representative tag's entropy value is below some threshold. Figure 2(a) shows the CDF of the prediction error for different values of entropy threshold. The line that shows the CDF of the predicted error for threshold of 12 (which overlaps the results for threshold=9) corresponds to the CDF of the model-based prediction error without filtering (Fig. 1(a)). The figure shows that we can get much more accurate predictions (lower errors) for photos that use well-localized tags, and the more localized the tags, the better the performance. In fact, the performance is much better than baseline.

3.3 Learning Relations between Places

In this section we describe how to use the proposed framework to learn relations between places, for example, 'socal' is *part of* 'california' (represented as 'california'→'socal'). Our approach is inspired by probabilistic subsumption [12]: given two concepts w_1 and w_2, if the occurrences of instances of w_2 can explain most of the instances of w_1, but not vice versa, we say that w_2 *subsumes* w_1, or w_2 refers to a broader concept than w_1. Transposed to the geospatial domain, this implies that the spatial distribution of the more general parent tag, e.g., 'california', should subsume the spatial distribution of the child tag, e.g., 'socal'. Schmitz [13] used probabilistic subsumption to learn broader-narrower relations between Flickr tags, which we use as baseline in our experiments.

Model-based Method for Learning Relations: We can view the problem of learning relations as finding a broader distribution associated with a parent place that can well enough approximate the distribution of the child place. For example, for a given tag w_i with distribution $P(x|w_i)$, the problem is to find $P(x|w_j)$ that can approximate $P(x|w_i)$ well enough, where $w_i \neq w_j$. For $P(x|w_j)$ that can adequately well approximate $P(x|w_i)$, we learn the relation $w_j \rightarrow w_i$. We use *cross entropy* to quantify this intuition.

Cross entropy measures the "difference" in information content between two probability distributions and can be interpreted as the average information for discriminating between them. The cross entropy of distributions P and Q, $H(P,Q)$, is asymmetric, meaning that $H(P,Q) \neq H(Q,P)$. We can use cross entropy to measure information difference between spatial distributions of two tags. The cross entropy of the child tag with respect to its parent should be low compared to the cross entropy of the same child tag with respect to an irrelevant term, since less information is required to differentiate the distribution of the child tag from that of the parent, rather than from the distribution of an irrelevant tag. For example, let tag w_1 ='socal', tag w_2 ='california' and tag w_3 ='malaysia'. Then,

$$H\big(P(x|w_1), P(x|w_2)\big) < H\big(P(x|w_1), P(x|w_3)\big)$$

because 'malaysia' has no spatial relation with 'socal' but 'california' has. Therefore, we can learn a relations between tags as follows: given tag w_1 and w_2, w_2 is a parent of w_1 if and only if

$$H\big(P(x|w_1), P(x|w_2)\big) \leq \theta$$

The parent concept is usually geographically broader than the child concept. If w_1 can approximate w_2 well enough and w_2 can also approximate w_1 well enough, they could potentially be synonyms. Therefore, we need to add two more conditions. First, the child tag distribution cannot approximate the parent tag distribution by the reverse condition. Second, the parent tag is geographically broader than the child tag, which can be measured using its entropy $H(P(x|w))$. This changes the formulations of the relation learning problem. Given tags w_1 and w_2, w_2 is a parent of w_1 if and only if w_1 and w_2 satisfy the following three conditions:

$$H\big(P(x|w_1), P(x|w_2)\big) \leq \theta \quad (5)$$

$$H\big(P(x|w_2), P(x|w_1)\big) > H\big(P(x|w_1), P(x|w_2)\big) \quad (6)$$

$$H\big(P(x|w_2)\big) > H\big(P(x|w_1)\big) \quad (7)$$

where the cross entropy and entropy are defined as:

$$H\big(P(x|w)\big) = -\int_X P(x|w)\, \log P(x|w)\, dx$$

$$H\big(P(x|w_1), P(x|w_2)\big) = -\int_X P(x|w_1)\, \log P(x|w_2)\, dx$$

approach	*model*	*baseline*
U.S. states	0.429	0.437
Contries	0.464	0.318
Continents	0.469	0.123

Table 2: Comparison max F-score of the model-based approach (gmm) to baseline on the inducing relation task.

Entropy in continuous space is defined as the expectation $E_{P(x|w)}$ of $\log(P(x|w))$ with respect to itself, while cross entropy of the distributions $P(x|w_2)$ and $P(x|w_1)$ is defined as the expectation of $\log(P(x|w_2))$ with respect to the distribution $P(x|w_1)$. Each $P(x|w_i)$ is given by Equation 1.

There is no analytic form for computing the cross entropy. Thus, we estimate it using the Monte Carlo method [6]. The idea is same with the approximation of entropy in the place identification section. The Monte Carlo approximation of cross entropy is:

$$H_{MC}(P,Q) = -\frac{1}{n}\sum_{i=1}^{n}\log Q(x_i), \tag{8}$$

where x_i is a sample from the distribution $P(x)$. As the number of samples n grows large, $H_{MC}(P,Q) \rightarrow H(P,Q)$. *Evaluation:* We evaluate our methods on three test sets: U.S. states, countries, and continents. The U.S. states set is seeded by tags such as 'alabama' . The countries set is seeded by tags such as 'germany'. The continents set is seeded by five tags 'africa', 'asia', 'europe', 'northamerica', and 'southamerica'. The different granularity levels of the seed terms illustrate the challenges of mining geospatial knowledge, including the popularity *vs.* generality problem.

We use GeoNames to measure the quality of learned relations. We construct the ground truth as follows. For each seed in the test set, e.g., 'alabama', we identify the corresponding geoid in GeoNames, e.g., geoid corresponding to Alabama, and extract the names of all geoids within it. Then, using methodology described in Section 3.1, we filter these child places, keeping only those places that match Flickr tags in our data. For each child place, we then create a relation $seed{\rightarrow}child$, or "child place is in seed place," for example, 'alabama'→'mobil'.

We use precision and recall to evaluate learned relations. Precision measures the fraction of the learned relations that exist in the ground truth, and recall measures the fraction of relations in the ground truth that the method was able to learn. The maximum F-score is used to quantify the aggregate performance of the method.

Maximum value of F-score (max-f1) attained by the two methods on the data sets are shown in Table 2. The max-f1 score of our method on the U.S. states data set is a little lower than the max-f1 score of baseline. However, on the countries or continents data sets our method performs significantly better than baseline. We conclude that for learning relations between places, our method is more robust than baseline.

4. CONCLUSION

We present a probabilistic framework that models each tag in continuous space as a mixture of Gaussian distributions, the parameters of which can be estimated by analyzing a corpus of geo-referenced and annotated photos. The probabilistic framework is flexible and can used to solve a number of geo-spatial data mining problems. Once the distribution $P(x|w)$ is estimated, it can be used in a number of geospatial applications within one framework. For example, we identify place names by looking for tags that have low entropy: $H(P(x|w)) < \theta$. To predict the location of a photo with tags W, we look for the maximum of the tag distributions: $x_{predict} = \arg\max_x \sum_{w \in W} P(x|w)$. It is also easy to combine these methods to improve results of location prediction by using only spatially informative tags. Similarly, we can learn relations between places represented by tags w_1 and w_2 simply by comparing the cross entropy of their probability distribution. Our method performs better than baseline at different spatial scales. We can continue comparing cross entropy of tags to arrange the learned relations within a taxonomy of places. In summary, the advantages of our framework are its simplicity, flexibility and competitive performance in mining noisy user-generated geospatial data.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-0812677 and CMMI-0753124.

5. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*. ACM, 2004.

[2] C. M. Bishop and S. S. En Ligne. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[3] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[5] C. Gouvêa, S. Loh, L. F. F. Garcia, E. B. Fonseca, and Wendt. Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing. In *Simpósio Brasileiro de Geoinformática-GEOINFO*, 2008.

[6] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *ICASSP*, volume 4. Ieee, 2007.

[7] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *Siam journal of optimization*, 9:112–147, 1998.

[8] A. R. Liddle. Information criteria for astrophysical model selection. *Monthly Notices of the Royal Astronomical Society: Letters*, 377(1):L74–L78, 2007.

[9] S. Openshaw. *The modifiable areal unit problem*. Geo Books Norwich,, UK, 1983.

[10] A. Plangprasopchok, K. Lerman, and L. Getoor. A probabilistic approach for learning folksonomies from structured data. In *WSDM*, Nov. 2011.

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

[12] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *SIGIR*, 1999.

[13] P. Schmitz. Inducing ontology from flickr tags. In *WWW Workshop on Collaborative Web Tagging*, May 2006.

[14] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *SIGIR*, 2009.

[15] R. W. Sinnott. Virtues of the Haversine. *Sky and telescope*, 68:158, 1984.

[16] Y. Yang and G. I. Webb. Discretization for naive-Bayes learning: managing discretization bias and variance. *Machine learning*, 74(1):39–74, 2009.