Geographical Coordinates Estimation by Tag/Image Information

Sheng Cheng  
University of Illinois at Urbana-Champaign  
Champaign, IL 61801  
scheng41@illinois.edu

Wutong Gao  
University of Illinois at Urbana-Champaign  
Champaign, IL 61801  
wutongg2@illinois.edu

Chengrui Zhu  
University of Illinois at Urbana-Champaign  
Champaign IL 61801  
cz8@illinois.edu

Xiangdong Yu  
University of Illinois at Urbana-Champaign  
Champaign, IL 61801  
xyu72@illinois.edu

ABSTRACT
Location coordinates detection according to the visual/text information is widely used in this age, like geo-visual, geotagging and LGTA. We propose a two-level LDA scheme applied on Geo-tagging that can demonstrate relatively high accuracy compared with the tradition Standard Language Model (LM), Search Similarity (SS) and LM/SS hybrid. Our method is evaluated on a dataset from a task of MediaEval2013. LDA shows much more improvement with the comparison with same distance threshold, especially when the distance is short. It means estimation with LDA is much more precise not only within a long range, but also in a small area. Meanwhile, we present a Geo-visual ranking method that can estimate geographic information using only the visual content of those images. Geo-visual ranking combines both ranking and classification. It can classify images that visual similarity to the geo-location and generated similarity with other query images. Both geo-visual ranking and geo-tagging methods demonstrate outstanding performance with high prediction accuracy result on geographic location.

Keywords
Geo-tagging, geo-visual, LDA, Learning Model, Similarity Search

1. INTRODUCTION
Nowadays, many websites and apps like Flickr and Instagram provide information composed of both structured and unstructured metadata. The structured metadata is photos or videos provided by cameras. The main unstructured metadata is in form of tags, which are mainly some short texts like descriptions of images from users. These tags can do a lot of things: one is to help us to mine location metadata of images. With the availability of location metadata, users can retrieve similar photos or videos which were taken at the same location before. Moreover, users can analyze the correlation between location metadata occurrence of the tags to discover some geographic knowledge.
Geographical coordinates estimation by tag/image is to estimate location of images using information of tags and images, which can be used to track accurate location for lost rescue and historical site exploration. In recent works, there are several methods to estimate geographical location. Belgium researchers [1] used two methods, language model and similarity search to estimate where a given photo or video was taken. Several kinds of information are available to estimate geological location, including visual features, user profiles and tags. In their work, since information provided by visual feature and user profile is not enough to do estimation, only tags are used to estimate geographic location. To estimate the location of a photo object based on its tags, they use three natural strategies. First, gazetteers are used to find the locations of those tags that correspond to toponyms. However, this strategy is particularly challenging in practice since that if there is no capitalization occurs in tags, it is difficult to identify the toponyms. It is also due to the ambiguity of toponyms. The limited amount of context information makes it impossible to recognize the ambiguity. Second, Standard language modeling approaches can determine the most likely area for a given resource and resolve a classification problem of georeferencing, which partition the locations on earth into a finite number of areas. This method eliminates the problem of the first method how to determine which tags are toponyms, or any form of disambiguation. Nevertheless, there is another drawback that it cannot result in a precise area. The more areas in the partition, the higher the probability of classification error is. Third, similarity search is used to estimate the location of a given resource as a weighted average of the locations of the most similar objects in their training set. In this case, the performance of the method is limited since it treats relevant tags in the same way as others.
In this paper, we try to estimate geographical location by using both Geo-tagging method and Geo-visual ranking.
For Geo-Tagging method, a two-level LDA scheme is used to filter out non-geographic terms. LDA is the abbreviation of latent Dirichlet allocation, which is a local generative topic model allowing to explain the similarity of observation groups. Comparing to the above two method, language modeling and similarity search, the LDA method can achieve more accurate results that can find lost victim faster and more precise. Also, we introduce a method based on visual feature, called Geo-visual ranking, which is an offline training, based on the result of Geo-tagging. Geo-visual combines classification and ranking using candidate images extract from tag. Thus, we use not only tags to estimate geographic location, but also visual features to make result more precise.

The remainder of the paper is organized as follows. In body paragraph, we first introduce the approaches and theoretical analysis of Geo-tagging based on LDA. Then, we compare the results among two kinds of LDA, tmax and tmean, language modeling, similarity search and hybrid of these two methods. After that, we discuss how Geo-visual ranking do ranking and classification. In the conclusion, we conclude the contribution of our team and future work that we can do.

2. RELATED WORK

2.1 Latent Geographical Topic Analysis (LGTA)

The paper [2] was aimed to comparing and discovering geographical topic from GPS-associated documents. Three different way of modeling geographical topics are compared and proposed in the paper, including location-driven model, text-driven model, and a novel joint model called LGTA which combines location and text information. LGTA is a location-text joint model combining geographical clustering and topic modeling. In this model, the topic would be generated from region so that the words in the same space will be highly possibly clustered into same geographical topic. The model would firstly discover geographical topics with the Gaussian distribution for sample location.

\[ p(l_{d}|\mu_r, \Sigma_r) = \frac{1}{2\pi \sqrt{\det(\Sigma_r)}} \exp\left(-\frac{(l_{d} - \mu_r)^T \Sigma_r^{-1}(l_{d} - \mu_r)}{2}\right) \]

Then, comparing geographical topics:

\[ p(z|l, \Psi) \propto p(l|z, \Psi)p(z|\Psi) \]
\[ \propto \sum_{r \in R} p(l|\mu_r, \Sigma_r)p(z|r)p(r|\alpha) \]

The model was tested from Flickr website including Landscape, Activity, Manhattan, National park, Festival, Car, and Food. The result shows that the LGTA model well performed on finding regions of interests and providing effective comparisons of different topics.

2.2 Two alternative location estimation using geo-tagged images

There are many other existing approaches regarding to estimating locations using geo-tagged images, most of which use clustering-based classification to solve the problem. One of the approaches [5] utilized mean shift clustering, a technique for estimating the modes of an underlying probability distribution from a set of samples, to find and characterize locations. The approach uses visual features (interest points) as well as textual features (image tags) for estimating the location of an image. The experiment is conducted over 33 million images taken by Flickr users and a set of k landmarks are selected to build models, based on both Bayesian classifiers and linear Support Vector Machines (SVMs), with SVMs perform slightly better than Bayesian classifiers. Two spatial resolutions are considered: the metropolitan-area scale (100 kilometers) and individual-scale landmark scale (100 meters) and the results show great improvements. Another approach [5] is a visual feature-based method. For a given image, 120 images from the training set that are most similar to this image will form a set of nearest neighbors – a probability map. Then, mean-shift clustering is performed to determine the major mode of the distribution and the mode of the cluster with the highest cardinality is the estimated location.

2.3 Geo-visual Study

Based on the work of Geo-tagging, we want to make full use of the information about image itself. There is several information that we can get from images such as the content of image, the texture of the image and the features of images. For image texture [5], color and texture information can be put into one histogram. The feature of images can be extracted by the first layer of neural network. The hardest thing is to extract the information from the images. However, if given the training label, it is easy to judge whether the images belong to one class. The common method is the 1-NN, which finds the nearest neighbor in the training dataset. Based on the location of the nearest image, the location of test image can be predicted.

3. METHODOLOGY

3.1 Geo-tagging

3.1.1 Estimate using language models

The first approach [6] is to assign each test image to the area with the highest probability. This method uses a unigram language model, and the probability for each area a given the tags t of test image x is:

\[ P(a | x) \propto P(a) \cdot \prod_{t \in x} P(t | a) \]

Where P(a) is the prior probability, and it can be estimated using maximum likelihood:
\[ P(a) = \frac{|X_a|}{\sum_{b \in A} |X_b|} \]

Where for area \( a \in A \), \( X_a \) represents the set of training images belonging to the area. Notice that if no information is given for the test image, it would be assigned to the area that contains the largest number of training images. As for the computation of \( P(t \mid a) \), some smoothing needs to be done. This is to avoid the case when some tags of the test images are not associated with any training image and thus \( P(t \mid a) \) becomes zero. Several smoothing methods to choose are: Laplace smoothing, Jelinek-Mercer smoothing and Bayesian smoothing. In practice, Bayesian smoothing has the best performance, and \( P(t \mid a) \) can be represented as:

\[
P(t \mid a) = \frac{O_{ta} + \mu (\sum_{a' \in A} O_{a'a'})}{(\sum_{a' \in A} O_{a'a'}) + \mu}
\]

Where \( O_{ta} \) is the number of occurrences for tag \( t \) in area \( a \), and \( V \) is the vocabulary which contains the tags determined by Chi-squared feature selection. For the choice of \( \mu \), which takes a value in \([0, +\infty]\), it is set to be 1750 for optimal performance. Thus the problem of finding the area to assign is now to find:

\[
a_x = \arg \max_{a \in A} P(a) \cdot \prod_{t \in x} P(t \mid a)
\]

Once the test image is assigned to an area, its location can be estimated using the medoid of the area:

\[
med(a) = \arg \min_{x \in a} \sum_{y \in a} d(x, y)
\]

Where \( d(x, y) \) is the geodesic distance between two training images within the same area. The medoid is used instead of center of gravity because the medoid is more robust to outliers.

### 3.1.2 Estimate using similarity searches

Instead of first clustering the training set and then calculate the probability of each area, a more direct approach [3] would be to select \( k \) images from the training set that are most similar to the test image and estimate the location by averaging the locations of the \( k \) images. Thus the location of test image \( x \) can be represented as:

\[
loc(x) = \frac{1}{k} \sum_{i=1}^{k} sim(x, y_i) \cdot loc(y_i)
\]

Where \( y_1, y_2, \ldots, y_k \) are the \( k \) images in the training set most similar to \( x \), and \( \alpha \) is a parameter that determines the influence of the most similar image on the estimated location. For the similarity measure function \( sim(x, y_i) \), the Jaccard method will be used:

\[
sim(x, y) = \frac{|x \cap y|}{|x \cup y|}
\]

Generally, the Jaccard method can be combined with other similarity measures or visual features.

In the location estimation function, the locations are represented as Cartesian coordinates, not longitude/latitude pairs. Thus it is necessary to first convert the location of each image to Cartesian coordinates. After the location is estimated, the Cartesian coordinates can be converted back to longitudes and latitudes.

Another parameter besides \( \alpha \) that has a major impact on the performance of this method is \( R_x \), the set of training images taken into consideration when computing the \( k \) most similar images. Although we can set \( R_x \) to be the entire training set, there are other choices of \( R_x \) which may possibly be better for the performance. For example, a threshold can be set for the similarity and only images that are similar enough are considered. Note that if the number of images in \( R_x \) after threshold is less than \( k \), it is necessary to consider all the images in \( R_x \), regardless of their similarities.

### 3.1.3 Estimate using a hybrid approach

This approach [3] can be seen as a combination of the last two: an area is first determined using the language model in 2.1, then the similar search in 2.2 is applied on the found area. For the similarity search, if no images in the area have similarity above the threshold, the medoid of the area can be used instead. Here an important parameter is the number of clusters. If a test image has no common tags with the selected cluster, the next-finest cluster will be chosen to get a suitable area.
3.1.4 Estimate using two-level LDAs

Based on the previous approaches, we propose an approach that utilize two-level LDAs [4] to filter out the tags that are considered “non-geographical”, in order to improve the results. For the first LDA process, we apply on each clustered area to generate a local topic distribution. Then, a second LDA is applied to the entire training set to generate a topic-area distribution. We finally calculate the topic entropy and use a threshold to generate the bag-of-excluded-words. The proposed approach consists two steps.

1) Offline analysis step

For this step, we will work on the training data and generate a geographical tag model. In general, this step comprises three parts.

A. Filtering

Because all the initial training images are randomly crawled from Flickr, some images would contain “machine-generated” tags or even no tags at all. Thus in this step, we would need to filter out those non-human-generated tags and also eliminate those images with no tags. The filtered images will be used as the training data.

B. Spatial clustering & Local LDA

For this step, we will first cluster all the training images based on their geographical locations (longitude and latitude values), using the k-means method. After clustering, we will get K (K is set to 500 in our approach) geographical clustered areas, which will each contain approximately 1000 images on average. We will apply local LDA (we assume 50 topics and 10 tags for each topic) on each of the clustered areas and generate the topic distribution of each local area. We can represent each area as

\[ A_i = \{u_j, \{w_j\}, \{\tau_{jk}\}\} \]

where \( u_j \) is the id of area \( j \), \( w_j \) is the images in this area and \( \tau_{jk} \) is the local distribution of the k-th topic.

C. Global LDA & Generate bag-of-excluded-words

After we compute the local LDA for each area, we will go on to generate the global topic-area distribution by applying LDA to the entire training set. The number of topics we set is 150 and the number of tags per topic is 20. For each topic generated, we will calculate its distribution over all areas, which will end up like a histogram. From the topic distribution we can calculate the entropy for each topic and get the topics that have entropy exceeding 36, which will then form the bad-of-excluded-words (BoEW). Notice that even though most of the words in the BoEW are real “non-geographical” tags, there could be words that are mistakenly put into the BoEW because they might have a span of large geographical area.

2) Online estimation step

After a geographical tag model is generated, we can then perform the location estimation for each image in the test set, which can also be separated into three parts.

First of all, we need to filter the test images to remove the machine-generated tags and tags present in the BoEW. Then, we will compute the similarity (Jaccard similarity) between tags of each test image and tags of each topic of each local LDA:

\[ S_{ik} = \frac{|t_i \cap \tau_{jk}|}{|t_i \cup \tau_{jk}|} \]

There are two ways to assign each test image to the most similar area. The first way is to select the area which contains the topic that has the highest similarity with the test image, referred to as Tmax. The second way is to compute the mean similarity of all topics that have at least one common tag with the test image and then assign to the area with highest mean similarity, referred to as Tmean. Finally, we can perform location estimation for each of the test image: first determine k training images with highest similarity in the assigned area, and then use the center of gravity weighted by the similarity as the estimation of location for the test image. If a test image does not belong to any of the areas, we will set its location to be the centroid of the largest area.

3.2 Geo-Visual

The shortcoming of the previous method is that once the nearest training image is incorrect, the prediction of location will be wrong. Therefore, we provide a new method not just focus on the similarity of the images, but the similarity of the topics. This method is a ranking based classification method for location prediction.

The main idea of the ranking geo-visual method is that based on the images selected by tags, we use k-NN to find the images candidate. We then decide the location of images based on the ranking of topics instead of the ranking of images.

3.2.1 Image selection

It is unwise to use whole dataset to find the nearest neighbor. Since we have gotten the tags clustering and processing in the geo-tagging part, it is nature to choose the appropriate training candidate based on that tags. For instance, in the geo-tagging, we get the topic about “France”, “Pairs” and “Eiffel tower” Tower, so we can choose the dataset based on that. This is an online training process. It needs to train the images as well as its tags by the data collected in Flicker. However, offline training is also suitable for our model. Instead of using the images from Flicker, we collect the data based on location information. The advantage is that the training images are typical and easily get the image features because, if training online, it requires high accuracy of topic clustering. A data that belongs to a cluster appearing on an irrelevant cluster will pollute the quality of training dataset.

3.2.2 Location extraction
If given the candidate training dataset as well as the unknown images in this topic, we calculate the distance between the unknown image with training dataset. This is a computer vision problem, it can be solved by the SIFT algorithm [6], which calculate the minimum L2 distance between keypoints. To save calculation time, we set two parameters to limit the number of images selected. The first one is the number of nearest neighbors(D). If the training candidate is small, we will directly choose D images. If the dataset is too large, another parameter S is used to limit the number of images for calculation. If meeting that parameter, the model will stop test and directly choose the D nearest neighbors.

3.2.3 Location ranking
To get more reliable prediction, we use the image topic of tags for ranking. This will prevent the problem that an image is close to an image just based on image feature, but its topic is closer to another. Therefore, we will re-rank the images based on the formula below:

\[ Score(g,q) = \sum_{e \in g} sim(e,q) \]

where the \( sim \) functions is the similarity calculation we mentioned above. \( q \) is the image needed to test. Image \( e \) is an image in group \( g \).

4. EXPERIMENT SETUP
In this section, we analyze our experiments on both Geo-Tagging and Geo-Visual, including our dataset selection, methodology implementation, and result comparison.

4.1 Geo-tagging
4.1.1 Data Set
For this method, we downloaded 800 million tagged images from Yahoo Flickr dataset. These datasets contained image IDs, ground truth longitude/latitude, and image tag information. Randomly 500,000 training set images and 5,000 test set images were selected from the entire data set. The image information was cleaned and filtered by using R. We set 300 clustering areas. [9]
For local LDA, we generated 20 topics with 4 words per topic. Meanwhile, for global LDA, 100 topics and 10 words per topic. And the BoEW entropy threshold is 20.
For language model and similarity search, we set 100 words for each cluster.
We iterated 1000 times for training set, and assign the area ID corresponding to the max/mean values for the test set.
Our methods, two-level LDA scheme were used to compared the three different clustering methods, LM, SS and the hybrid with same parameters set, and the accuracy were calculated by the different distance setup by 1 km, 10 km, 100 km, and 1,000 km.
The distance different is calculated by the Harersine formula:

\[ a = \sin^2(\Delta \varphi/2) + \cos \varphi_1 \cdot \cos \varphi_2 \cdot \sin^2(\Delta \lambda/2) \]
\[ c = 2 \cdot \tan2(\sqrt{a}, \sqrt{1-a}) \]
\[ d = R \cdot c \]

,where R is the earth mean radius, 6371 km.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Our Approach</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc(1 km)</td>
<td>60.4%</td>
<td>48.56%</td>
</tr>
<tr>
<td>Acc(10 km)</td>
<td>65.46%</td>
<td>53.02%</td>
</tr>
<tr>
<td>Acc(100 km)</td>
<td>70%</td>
<td>58.34%</td>
</tr>
<tr>
<td>Acc(1000 km)</td>
<td>78.1%</td>
<td>68.88%</td>
</tr>
</tbody>
</table>

Figure 1. Geo-Tagging Approach Result Table

4.1.2 Result Discussion
We choose the accuracy of 1km, 10km, 100km and 1000km for evaluation compared with the work of Hybrid SS and LM. From the comparison in figure 1, we noticed that as we expected, LDA method showed a significant improvement on location estimation. Particularly with low distance error, Tmax’s accuracy is 4 times greater than the SS’s, almost 5 times greater than the LM’s and 2 times greater than the Hybrid’s when the error allowance is 1km.
The main reason of the significant better performance is because of the two-level LDA filtering steps, machine tags filter and BoEW filter, which deeply removed the non-geographical tag from the row image datasets. Meanwhile, LDA is a probabilistic model with interpretable topics in natural language processing, which providing more guaranteed similarity of different terms with different topics.
Under the comparison of our approaches, the max similarity method is over around 10% than the mean similarity one.

4.2 Geo-visual
4.2.1 Data Set
For geo-visual ranking method, we make a small demo for this method. We train 8 classes offline, whose names are also the tags of dataset. Each one belongs to different location. And then we select a topic which at least contains one of the tags. We put the images through the model. And the accuracy of the test result is 75%. We use the location of image that is the nearest neighbor in the top ranking clustering. The predict location is pretty close to the actual location.

4.2.2 Result Discussion
We demonstrated partial result from the test set information in figure 2. But the prediction will be wrong when the background changed dramatically. The drawback of this method is obvious. Whether this method can be widely used highly depending on the number of classes we have.
In other words, it can only correct the location of images that have clear location tags. Another drawback is about time. Because the size of images is different in Flicker, the test data needs pre-processing before fed into model. Calculating the distance of two images also cost a lot of time.

![Figure 2. Test Image Visualization Selected Result Sample](image)

5. Conclusion and Future Work

We tried two segmentation information approaches to predict the location on image sets. One is the geo-tagging method using two-level LDA filtering, and the other is the geo-visual method. Compared to previous geo-tagging schemes including Language Model, Similarity Search and hybrid of both methods, LDA demonstrated much better accuracy improvement with same error allowance. The geo-visual ranking approach also give a good performance on location coordinate estimation result.

For geo-tagging future work, we can try to perform better analysis of the source of errors and include additional information of input images. Some of the parameter selections to be considered for better accuracies are: (1) For local LDAs, the number of topics and tags per topic we choose (2) The threshold of entropy for each topic (3) The number of geographical clustered areas. For the addition information, we can utilize the author information as well as external geotag data collected from the internet, in order to increase coverage. For geo-visual ranking method, there are still several potential improvements available for the future work. The visual recognition can be better regardless of the contrast and brightness of the images. We should find a method to generate more classification based on the number of images to reach more detail information.

The last future potential approach is Geo-tagging and geo-visual can be mutually enhanced each other overcoming the possibility of low compatibility of two approach by adapting the idea of heterogeneous information network concept.

6. REFERENCE: