UQ-DKE's Participation at MediaEval 2014 Placing Task

Jiewei Cao, Zi Huang, Yang Yang, Heng Tao Shen
School of Information Technology and Electrical Engineering
University of Queensland
Brisbane, QLD, Australia
j.cao3@uq.edu.au, {huang, yang.yang, shenht}@itee.uq.edu.au

ABSTRACT

In this paper, we describe our approach as part of the MediaEval 2014 Placing Task evaluation. We first identify tags that are most indicative of geographic location by calculating the spatial-aware weighting for all tags in the training set. These weighting are applied to a language model-based retrieval framework. To address the geo-tagging problem, we find the most similar training item and propagate its location to the test item. Base on last year's experience, we further improve the accuracy by utilizing the geo-location correlation of images/videos uploaded by the same user.

1. INTRODUCTION

The MediaEval 2014 Placing Task requires participants to assign geographical coordinates (latitude and longitude) to Flickr images or videos (we denote them by *Flickr items* for description simplicity), we refer to [3] for a detailed description. Firstly, we identify spatial-aware tags in the training set using a tag selection method based on Ripley's K statistic [6]. To address the geo-tagging problem, we apply a language model-based document retrieval model to find the most similar training item and propagate its location to the test item. Here, we consider each Flickr item's tags (title and description are excluded) as a *document*. Usually, a document contains 5 to 10 tags and the tag's order is disregarded. Given a test item, a query is constructed by using its tags and then retrieve the most relevant document from training set. The spatial-aware tag weighting is applied to give different weighting for each tag in the query. Experiments show that spatial-aware weighting efficiently improved the accuracy. Base on last year's experience [1], we further improve the accuracy by exploiting the geo-correlation between test items within the same user collection[1].

2. METHODOLOGY

2.1 Data Pre-processing

A total of 5,025,000 geo-referenced Flickr items are provided as training data. For language model-based approach, we treat each Flickr item's tags as a document. Other surrounded texts, such as title or description, are not used in our approach. We carried out two preliminary filter steps on this training set. First, items without tags were removed. Second, we converted all tags to lowercase and special characters were removed. Finally, this resulted in a pre-processed training set with 4,148,564 items. Unless specified, this pre-processed training set is used in the following experiments.

[1] `http://www.flickr.com/help/collections/`

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

2.2 Spatial-aware Tag Weighting

We use a Ripley's K statistic based tag selection method [4] to select the most spatial-aware tags by analyzing the spatial distribution of tags. Specifically, equation (1) was applied to calculate the weighting for each tag t. Given a set Q_t contains the locations of the images/videos which tag t has been assigned, and $N_t = |Q_t|$ is the total number of occurrences of tag t, we have:

$$s(t) = logN_t \cdot \frac{\sum_{p \in Q_t}(|\{q|q \in Q_t, q \neq p, d(p,q) \leq \lambda\}|)^w}{N_t^2}, \quad (1)$$

where $d(\cdot)$ is the distance function. The weighting $s(t)$ is similar to "tf-idf": the first part $log(N_t)$ will prefer tag with large frequency; the second part will downgrade the $s(t)$ if tag t spreads all over the world and vise verse. Specifically, when $w = 1$, if all the images with tag t cluster in a small region (controlled by λ), the second part will near to 1, otherwise, near to 0. In practice, Q_t doesn't need to contain all the items with tag t. For example, if there are more than 1 million Flickr items have t, we can only sample 5000 or so of them, which will be sufficient enough to calculate the weighting. For each tag in the training set, we calculate its spatial-aware weighting by equation (1).

2.3 Retrieval Model

We use the framework proposed by [5] which combines the language model and inference network as our retrieval model. This model provides a set of structured query operators [7] to express complex concepts, each of which can be considered a query node in an inference network. Bayesian Smoothing with Dirichlet priors [8] is applied to avoid a zero probability when a query contains a tag that doesn't occur in the training documents. Given a test item, we use the calculated spatial-aware weighting to assign different weighting to the tags in the query, and then retrieve the most relevant training item and propagate its location to the test item.

2.4 Collection Geo-correlation

To address the data sparsity issue of training data, [2] jointly estimated the geo-locations of all of the test items, where each test item was treated as "virtual" training data

Figure 1: Images and their tags in a collection named *Brisbane Trip 2014* created by the user.

Table 1: Percentage (%) of correctly detected locations and median error distance (in km) of each run in kilometer.

Within	10m	100m	1km	10km	100km	1000km	Median Error(km)
Baseline	1.09	4.87	17.06	34.22	43.00	56.59	380.38
Run 1	1.07	4.98	19.57	41.71	52.46	63.61	51.07
Run 3	1.08	5.05	20.23	43.68	56.03	69.08	27.32

and consequently boosted the performance of the algorithm. On the other hand, [1] proposed a method that utilize the geo-correlation between test items within the same user collection.

Flickr users can organize their images and videos by assigning them to different collections (or albums). Intuitively, items within the same collection would be highly geo-correlated. Take Figure 1 as an example, a user shared his images during a trip to Brisbane, and organized them into a collection named *Brisbane Trip 2014*. As we can see, not every images in this collection is well tagged because user only tagged the images he loved or interested in, and leaving others un-tagged or poorly tagged. Moreover, it's difficult for us to predict their location by the image itself because none of them contain particular landmark or landscape. Images/videos with completely different tags or visual content could be considered as taken in the same location or nearby, if they were within the same user collection. For tag-based geo-tagging approaches, a poorly tagged query item will result in a bad estimation. However, if this item belongs to a user collection which contains one or more images/videos with well estimated location (usually well tagged or contain landmark), then we can use the centroid location of this collection as the estimation for the poorly estimated one.

In this paper we adopted similar strategy as last year we did to find test items within the same collection, please refer [1] for details. Given a test item with no tag, we use the most frequent location of well estimated test items within the same collection as the finial estimation.

3. RESULTS AND DISCUSSION

There are five different test sets and we chose test 5 whose size is 510,000. Following [4], we set $w = 1$ and $\lambda = 40km$ in equation (1) to favor tags that occur centered around a small number of locations. We set $\mu = 5$ for Dirichlet Smoothing because the average document length is around 5 in our case, which means there are 5 tags in each document on average.

We submitted two runs (run 1 and run 3) and the results of our experiments are shown in table 1. Run 2 is omitted which requires only visual and audio cues can be used. Baseline approach used the same retrieval model as run 1, but the spatial-aware tag weighting were not applied. Both baseline approach and run 1 assigned a default location, e.g., New York City $(40.7127, -74.0059)$ in our case, for test items that without tag, whereas run 3 utilized the collection geo-correlation as discussed in section 2.4. As we can see, both spatial-aware tag weight and collection geo-correlation can help improve the geo-tagging accuracy. In this paper, we have set fixed values for w, λ and μ and avoided tailoring these values to the problem. However, we believe there is potential for improvement in the results through the optimal selection of these parameters for the particular data.

4. REFERENCES

[1] J. Cao. Photo set refinement and tag segmentation in georeferencing flickr photos. In *MediaEval*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[2] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 43–48. IEEE, 2012.

[3] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the 3rd ACM International Workshop on Geotagging and Its Applications in Multimedia*, 2014.

[4] O. Van Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt. Spatially aware term selection for geotagging. *IEEE Trans. Knowl. Data Eng.*, 26(1):221–234, 2014.

[5] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40(5):735–750, 2004.

[6] B. D. Ripley. *Spatial statistics*, volume 575. John Wiley & Sons, 2005.

[7] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.

[8] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001.