ACTIVE TAGGING FOR IMAGE INDEXING

Kuiyuan Yang[†*], *Meng Wang*[‡], *Hong-Jiang Zhang*[§]

[†]University of Science and Technology of China
[‡] Microsoft Research Asia
[§] Microsoft Advanced Technology Center

ABSTRACT

Concept labeling and ontology-free tagging are the two typical manners of image annotation. Despite extensive research efforts have been dedicated to labeling, currently automatic image labeling algorithms are still far from satisfactory, and meanwhile manual labeling is rather labor-intensive. In contrast with labeling, tagging works in a free way and therefore it has better user experience for annotators. In this paper, we introduce an active tagging scheme that combines human and computer to assign tags to images. The scheme works in an iterative way. In each round, the most informative images are selected for manual tagging, and the remained images can be annotated by a tag prediction component. We have integrated multiple criteria for sample selection, including *ambiguity*, *citation*, and *diversity*. Experiments are conducted on different datasets and empirical results have demonstrated the effectiveness of the proposed approach.

Index Terms— Tagging, active learning.

1. INTRODUCTION

With rapid advances in storage devices, networks and compression techniques, digital images have increased in an explosive way. To effectively manage these data, a promising approach is to annotate them with a set of keywords (may be named labels, concepts or tags in different contexts). By indexing the images with these keywords, manipulation of the images, such as search and browsing, can be easily accomplished.

In this work, we categorize the image annotation approaches into two schemes: *labeling* and *tagging*. Here labeling is referred to as the approach of annotating the data with a fixed concept set which is often called ontology. Given an image and a concept from the ontology, annotators decide whether the image is relevant or irrelevant with respect to the concept. Different from labeling, tagging is ontology-free. Given an image, users can freely provide several keywords (i.e., tags) that are related to the images in their mind.

The labeling scheme, including both manual labeling and learning-based automatic labeling, has been studied for years in research community. Yan et al. have studied the two styles of manual labeling, i.e., annotating an image with multiple concepts simultaneously and annotating multiple images in batch for a given concept [16]. Many research works have also been dedicated to automatic labeling (also known as concept detection [10] or high-level feature extraction [8]). The labeling of each concept is usually regarded as a binary classification problem. First a training set with groundtruth is gathered, and then the models of the concepts are learned. New images can thus be directly predicted by the models. Although ontology-driven labeling has many advantages [9], currently the performances of automatic labeling algorithms are still far from satisfactory. Meanwhile, manual labeling is a labor-intensive and time-consuming process. Hua et al. have proposed that the future trend of large-scale annotation should be leveraging Internet users to contribute efforts [7]. But how to let these users contribute their efforts to the tedious labeling work is a problem.

Different from labeling, tagging is arguably better in terms of user experience because of its free style. Actually nowadays many social media websites such as Flickr [1] and Youtube [2] have adopted this approach. Enormous Internet users provide tags for their data to facilitate the data organization and management. But it will be unimaginable that the users would be willing to annotate whether their data are relevant or irrelevant with respect to hundreds of concepts. Furthermore, recently several gaming-based tagging methods have been proposed, such as ESP game [12], and they can help further improve the user experience of tagging.

Thereby, in this work we focus on the tagging approach, and we propose an active tagging scheme that can reduce the manual effort of users by combining human and computer. As illustrated in Fig. 1, the scheme includes two major components, i.e., tag prediction and sample selection, and it works in an iterative way. In each round, a batch of images is selected according to a set of criteria and then these samples are manually tagged by annotators. Then these images are added to the tagged dataset, and tag prediction can be performed for remained data. So, loosely speaking, the active tagging approach can be regarded as a combination of tagging and active learning. For the labeling approach, active learning [4, 15, 3] has been widely adopted since the labeling of each concept can be viewed as a well-defined binary classification problem. But in tagging the objective is just assigning each image several related keywords, and it focuses more on the tagging precision and is relatively tolerant with the coverage [1]. Therefore, in this work we have developed sample selection criteria that are different from those in the traditional active learning methods. Another difference is that in the traditional active learning approach there will be a re-training and prediction process before selecting the next batch of samples in each iteration, but in our proposed scheme the sample selection does not rely on the tag prediction results (details can be found in the next section). That means the scheme just keeps selecting samples and asking users to provide tags, and the tag prediction can be performed as a final step. This is because the tag prediction will be relatively more time-consuming, and this approach can accelerate the sample selection process, such that users need not to wait in the tagging process.

[*]This work was performed at Microsoft Research Asia.

[1]Coverage if of course important. But for a given image and a set of tags, what we can conclude is just the precision. The coverage can hardly be quantized since there should be enormous (if not infinite) potential tags that are relevant to the image. This should be a specific property of free tagging.

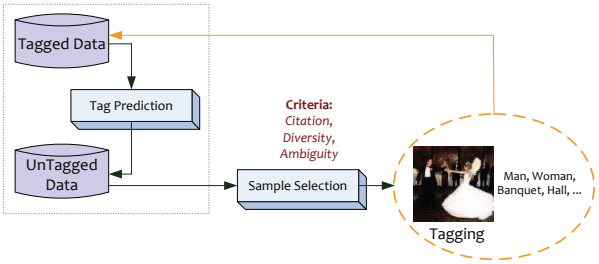

Fig. 1. The illustrative scheme of active tagging

The organization of the rest of this paper is as follows. In Section 2, we introduce the active tagging scheme, including the tag prediction and sample selection components. Empirical results are provided in Section 3. Finally, we conclude the paper in Section 4.

2. ACTIVE TAGGING

As shown in Fig. 1, sample selection and tag prediction are the two main components of the active tagging scheme. Considering we have n images $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, our target will be providing tags for all of these images. We suppose that the first l images have been manually tagged. Let $\mathcal{L} = \{x_1, x_2, \ldots, x_l\}$ and $\mathcal{U} = \{x_{l+1}, x_{l+2}, \ldots, x_n\}$. For each x_i in $\mathcal{L}$, it is associated with a set of tags T_i. Denote by $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ the global tag set (i.e., there are m tags in all). Then the sample selection and tag prediction components aim to move a batch of samples from $\mathcal{U}$ to $\mathcal{L}$ and predict the tags of the samples in $\mathcal{U}$, respectively.

2.1. Tag Prediction

The tag prediction is performed with two random walk processes. First, we construct the image and tag graphs (i.e., similarity matrices). For each image x_i, we find its k-nearest neighbors $\mathcal{N}_i$ based on low-level visual features. We construct an image affinity matrix $\mathbf{W}^I$ defined by $W^I_{ij} = \exp(-\|x_i - x_j\|^2/\sigma^2)$ if $i \neq j$ and $x_j \in \mathcal{N}_i$, and W^I_{ii} is set to 0. The tag graph is constructed by mining Flickr website. Analogous to Google distance [5], we estimate the distance between two tags t_i and t_j as follows

$$d(t_i, t_j) = \exp(-\frac{\max(logf(t_i), logf(t_j)) - logf(t_i, t_j)}{logG - \min(logf(t_i), logf(t_j))}) \quad (1)$$

where $f(t_i)$ and $f(t_j)$ are the numbers of images containing tag t_i and tag t_j respectively, $f(t_i, t_j)$ is the number of images containing both t_i and t_j, and G is the total number of images on Flickr. These numbers can be obtained by performing tag-based search on Flickr. The concurrence similarity between t_i and t_j is then defined as

$$W^T_{ij} = \exp(-d(t_i, t_j)) \quad (2)$$

But it worth mentioning that the tag graph construction step is flexible and it can be easily replaced by other methods, such as using the word similarity in WordNet [6] or developing it based on Flickr distance [14].

We then let $\mathbf{S}^I = \mathbf{D}^{I^{-1/2}}\mathbf{W}^I\mathbf{D}^{I^{-1/2}}$ where $\mathbf{D}^I$ is a diagonal matrix with its (i,i)-th element equals to the sum of the i-th row of $\mathbf{W}^I$. Analogously, let $\mathbf{S}^T = \mathbf{D}^{T^{-1/2}}\mathbf{W}^T\mathbf{D}^{T^{-1/2}}$ where $\mathbf{D}^T$ is a diagonal matrix with its (i,i)-th element equals to the sum of the i-th row of $\mathbf{W}^T$. Define an $n \times m$ matrix $\mathbf{Y}$ where $Y_{ij} = 1$ if x_i is manually tagged with t_j, and otherwise $Y_{ij} = 0$. The image-level random walk process then works by iterating $\mathbf{F} = \alpha\mathbf{S}^I\mathbf{F} + (1-\alpha)\mathbf{Y}$ until convergence, where F_{ij} can be regarded as the initial relevance score of x_i with respect to t_j. This random walk method (also named manifold ranking [17] or label propagation [13]) has been widely applied in many different applications [18]. After obtaining $\mathbf{F}$, we further perform a tag-level random walk. It can refine the relevance scores by leveraging the relationship of the tags. Analogous to the process of image-level random walk, we iterate $\mathbf{R} = \alpha\mathbf{S}^T\mathbf{R} + (1-\alpha)\mathbf{F}'$ until convergence, where $\mathbf{F}'$ is the transpose of $\mathbf{F}$ and R_{ij} is the final relevance score of x_i with respect to tag t_j. For each image x_i, we rank the relevance scores R_{ij}, and then we select the first r tags, which r is set to the average number of tags per image in $\mathcal{L}$.

2.2. Sample Selection

For sample selection, we adopt the following three criteria: *ambiguity*, *citation* and *diversity*. First we define the *ambiguity* measure of an image by analyzing the tags of its neighbors. Suppose there are k_i manually tagged images in the neighborhood of x_i, and there are m_i tags associated with these images. Denote by $P_1, P_2, \ldots, P_{m_i}$ the appearance probabilities of these tags (for example, there are 10 tagged samples in the neighborhood and 2 of them contain tag "apple", then its appearance probability is $1/5$). The ambiguity measure of x_i is defined as

$$ambiguity(x_i) = 1 - \frac{k_i}{K} - \frac{k_i}{K}\frac{\sum_{j=1}^{m_i} p_j logp_j + (1-p_j)log(1-p_j)}{m_i} \quad (3)$$

The rationality of the above heuristic definition of the *ambiguity* measure comes from the following facts:

1. From Eq. (3) we can see that the *ambiguity* is approximately decreasing with k_i, i.e., the number of manually tagged samples in the neighborhood. This is understandable since the tags of an image are propagated from its neighbors. If $k_i = 0$, i.e., no image is manually annotated in the neighborhood, then the *ambiguity* measure achieves its maximum value 1.
2. Given a fixed k_i, the *ambiguity* measure is increasing with the sum of the appearance entropies of the tags. If all the tags appear with probabilities of $1/2$, then the *ambiguity* measure achieves the maximum value 1, i.e., the same to the case that no image is tagged in the neighborhood.

The *citation* measure of an image is defined as the number of images that has taken it as neighbors, i.e.,

$$citation(x_i) = \frac{\sum_{j=1}^{n} I(x_j \in \mathcal{N}_i)}{n} \quad (4)$$

where $I(.)$ is the indicator function ($I(true) = 1$ and $I(false) = 0$). This criterion aims to select the images with high citation values, which are expected to help predict the tags of more images.

The *diversity* criterion has been widely applied in the traditional active learning approach [4, 15]. It enforces the selected samples to be diverse and keeps their variety, such that they will not be constrained in a more and more restricted area. Given a kernel K, the angle between two samples x_i and x_j is defined as

$$cos(< x_i, x_j >) = \frac{|K(x_i, x_j)|}{\sqrt{K(x_i, x_i)K(x_j, x_j)}} \quad (5)$$

We adopt Gaussian kernel, and the diversity measure for sample x_i can thus be estimated as

Input:
$\mathcal{L} = \phi$; /*manually tagged set*/
$\mathcal{U} = \{x_1, x_2, \ldots, x_n\}$; /*untagged set*/
AT; /*number of active learning iterations */
h; /*batch size for sample selection */

Output:
T_i; /*tagging results for i-th image,$1 \leq i \leq n$*/

Functions:
TagPrediction($\mathcal{L}, \mathcal{U}$);
/*tag prediction component,see Section 2.1*/

InfoComputation($\mathcal{L}, \mathcal{U}$);
/*computation of the *informativeness* measure for each image in $\mathcal{U}$,see Section 2.2*/

SampleSelection($\mathcal{L}, \mathcal{U}, h$);
/*select a batch of samples with greatest *informativeness* measures in $\mathcal{U}$*/

Begin:
for $t = 1, 2, \ldots, AT$
InfoComputation($\mathcal{L}, \mathcal{U}$)
S = SampleSelection($\mathcal{L}, \mathcal{U}, h$)
Manually tag the samples in S, and move set S from $\mathcal{U}$ to $\mathcal{L}$;
end
TagPrediction($\mathcal{L}, \mathcal{U}$);

Fig. 2. Pseudo-code of the proposed active tagging process

$$diversity(x_i) = 1 - \max_{x_j \in \mathcal{L}} \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}) \quad (6)$$

We linearly combine the three criteria to form an *informativeness* measure, based on which the sample selection can be performed.

$$informativeness(x_i) = \alpha \times ambiguity(x_i) + \beta \times diversity(x_i) + (1 - \alpha - \beta) \times citation(x_i) \quad (7)$$

where α,β are the weights of *ambiguity* and *diversity* respectively. The detailed implementation of the active tagging is illustrated in Fig. 2. As previously mentioned, we have not performed the tag prediction in each iteration in order to accelerate the sample selection process.

3. EXPERIMENT

We conduct experiments on two datasets. One is from Flickr [1] which contains $41,513$ images, and the other is from LabelMe [11] which contains $47,759$ images. For each image in the two datasets, we extract 353-dimensional features, including 225-dimensional block-wise color moment features generated from 5-by-5 partition of the image and a 128-dimensional wavelet texture features.

For Flickr dataset, we select ten most popular tags, including *cat*, *automobile*, *mountain*, *water*, *sea*, *bird*, *tree*, *sunset*, *flower* and *sky*, and use them as query keywords to perform tag-based search with "ranking by interestingness" option. Then the top $5,000$ images are collected together with their associated information, including tags, uploading time, etc. But many of the raw tags are misspelling and noisy, so we use a pre-filtering to remove these meaningless tags. Specifically, we match each tag with the entries in a Wikipedia thesaurus, and only the tags that appear in the thesaurus are kept, and we further remove the tags that appear less than 5 times. In this way, we retain 817 tags and each image is associated with 6.2 tags on average. For LabelMe dataset, we regarded the annotated keywords as tags. There are 697 tags in total and 3.36 tags per image in average. Several exemplary images from the two datasets and the associated tags are illustrated in Fig. 3 and Fig. 4.

We regard the existing tags as groundtruth, and then evaluate the active tagging scheme. It is noteworthy that this strategy will of course lead to an underestimation of the performance since the existing tags are not complete and several correctly predicted tags in our algorithm may have not been listed, but it is still reasonable to use them for a comparison of different methods. For each image, we compute the precision and recall measures for its manually added or predicted tags, which are defined as

$$Precision = \frac{\#\{\text{Predicted tags} \cap \text{Original tags}\}}{\#\{\text{Predicted tags}\}} \quad (8)$$

$$Recall = \frac{\#\{\text{Predicted tags} \cap \text{Original tags}\}}{\#\{\text{Original tags}\}} \quad (9)$$

Of course for manually tagged examples the precision and recall measures are both 1. We then compute the F-score of the image as $2 \times Precision \times Recall/(Precision + Recall)$. We average the F-scores of all images and it is adopted as the performance evaluation metric of this work. The radius parameter σ for computing image similarity is set to the median value of the Euclidean distances of all connected image pairs, and the weights α and β are both empirically set to $1/3$. The following three strategies are compared for tagging the two datasets:

1. Fully manual method. In this method, we do not perform tag prediction, and thus only the manually tagged images are taken into account.
2. Random sample selection. We randomly select images for manual tagging and then perform tag prediction.
3. The proposed active tagging method. In each round, 500 samples are selected according the criteria introduced in Section 2 for manual tagging.

The results for Flickr and LabelMe datasets with different numbers of manually tagged images are illustrated in Fig. 5 and Fig. 6, respectively. From the results we can clearly see the effectiveness of the tag prediction and sample selection components. The relative improvement from random sample selection to the proposed active tagging approach of the LabelMe dataset is smaller than the Flickr dataset. This is because many images of the LabelMe dataset come from video clips, and so they are rather similar to each other. This means that there are many near-duplicate images in the LabelMe dataset, and thus the random sample selection can already achieve very good performance and the improvement brought by the proposed sample selection approach is relatively limited.

4. CONCLUSION

In this paper we have proposed an active tagging scheme for image indexing. It can be viewed as a combination of tagging and active

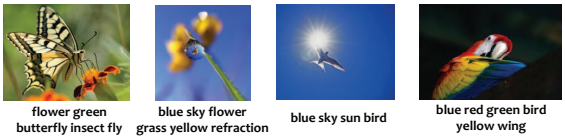

Fig. 3. Exemplary images from the Flickr dataset and the associated tags

Fig. 4. Exemplary images from the LabelMe dataset and the associated tags

learning. Tag prediction and sample selection are the two main components of the scheme. The sample selection component aims to select the most informative samples for manual tagging, and then the tags of remained samples are predicted. Experiments are conducted on Flickr and LabelMe datasets, and empirical results have demonstrated the effectiveness of the proposed approach.

This work can still be extended in different directions. The tag prediction and sample selection algorithms can be further optimized, and the scheme itself can also be extended. We can integrate a tag recommendation component to further facilitate the manual tagging for users. We can also choose gaming-based tagging methods. For example, if we integrate this work and the ESP game [12], we can easily develop an "active ESP" system, which is able to perform image selection and tag propagation instead of just randomly selecting images for users.

5. REFERENCES

[1] Flickr. http://www.flickr.com.

[2] Youtube. http://www.youtube.com.

[3] S. Ayache and G. Quénot. Evaluation of active learning strategies for video indexing. *International Workshop on Content-Based Multimedia Indexing*, 2007.

[4] K. Brinker. Incorporating diversity in active learning with support vector machines. In *International Conference on Machine Learning*, 2003.

[5] R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 2007.

[6] C. Fellbaum. *Wordnet: An electronic lexical database*. Bradford Books, 1998.

[7] X. S. Hua and G. J. Qi. Online multi-label active annotation: towards large-scale content-based video search. In *ACM Multimedia*, 2008.

[8] W. Kraaij and P. Over. TRECVID-2005 high-level feature task: Overview. In *TRECVID (http://www.nlpir.nist.gov/projects/tvpubs/tv6.papers/tv6.hlf.slides-final.pdf)*.

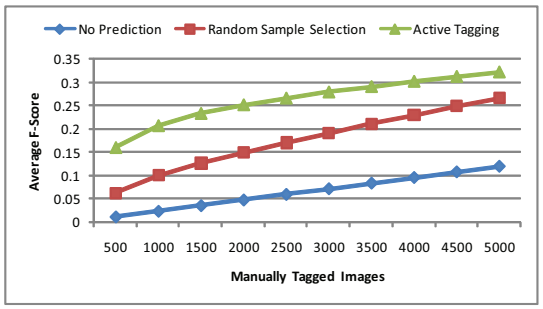

Fig. 5. The comparison of the three tagging methods for Flickr dataset

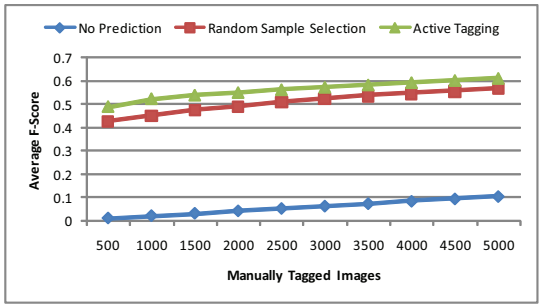

Fig. 6. The comparison of the three tagging methods for LabelMe dataset

[9] M. Naphade, J. R. Smith, J. Tesic, S. Chang, W. Hus, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia Magazine*, 13(3), 2006.

[10] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *ACM Multimedia*, 2004.

[11] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *Internal Journal on computer vision*, 77, 2008.

[12] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in computing systems*, 2004.

[13] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 20(1), 2008.

[14] L. Wu, X. S. Hua, N. Yu, W. Y. Ma, and S. Li. Flickr distance. In *ACM Multimedia*, 2008.

[15] Y. Wu, I. Kozintsev, J.-Y. Bouguet, and C. Dulong. Sampling strategies for active learning in personal photo retrieval. In *International Conference on Multimedia & Expo*, 2006.

[16] R. Yan, A. Natsev, and M. Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *CVPR*, 2008.

[17] X. Yuan, X. S. Hua, M. Wang, and X. Wu. Manifold-ranking based video concept detection on large database and feature pool. In *Proceedings of ACM Multimedia*, 2006.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances of Neural Information Processing*, 2004.