EMPLYING PLSA MODEL AND MAX-BISECTION FOR REFINING IMAGE ANNOTATION

Dongping Tian$^{1,2}$, Wenbo Zhang$^{1,2}$, Xiaofei Zhao$^1$ and Zhongzhi Shi$^1$

1 The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
2 Graduate University of Chinese Academy of Sciences, Beijing 100049, China
{tiandp, zhangwb, zhaoxf, shizz}@ics.ict.ac.cn

ABSTRACT

We present a new method for refining image annotation by fusing probabilistic latent semantic analysis (PLSA) with max-bisection (MB). We first construct a PLSA model with asymmetric modalities to estimate the posterior probabilities of each annotating keyword for an image, and then a label similarity graph is built by a weighted linear combination of label similarity and visual similarity. Followed by the rank-two relaxation heuristics over the constructed label graph is employed to further mine the correlation of the keywords so as to capture the refining annotation, which plays a critical role in semantic based image retrieval. The novelty of our method mainly lies in two aspects: exploiting PLSA to complete the initial semantic annotation task and implementing max-bisection based on the rank-two relaxation algorithm over the weighted label graph to refine the candidate annotations generated by the PLSA. We evaluate our method on the standard Corel dataset and the experimental results are competitive to several state-of-the-art approaches.

Index Terms— Refining image annotation, PLSA, EM, max-bisection, image retrieval

1. INTRODUCTION

Automatic image annotation (AIA) is a challenging problem in computer vision, which plays a crucial role in semantic based image retrieval. The aim of AIA is to automatically assign some keywords to an image that can well describe the content in it. And the representative work includes [1-4].

In recent years, many methods have been developed for refining AIA. As a pioneer work, Jin et al.[5] utilize WordNet to estimate the semantic correlation among the annotating keywords. This method, however, can only achieve a fairly limited effect as it totally ignores the visual content of the images. Wang et al. [6] apply random walk with restarts to refine candidate annotations by integrating word correlation with the original candidate annotation confidence together. Followed by they propose a content based approach by formulating the annotation refinement as a Markov process [7]. In [8], Jin et al. employ randomized weighted-max cut algorithm to image annotation refinement. Subsequently, they extend the work by proposing a new method for knowledge based image annotation refinement [9] in a deterministic polynomial time, in which they investigate various semantic similarity measures between keywords and fuse the outcomes of all these measures together to make a final decision using Dempster-Shafer evidence combination. More recently Liu et al.[10] rank the image tags according to their relevance with respect to the associated images by tag similarity and image similarity in a random walk model. Xu et al.[11] come up with a new graphical model termed as regularized latent Dirichlet allocation (rLDA) for tag refinement. Zhu et al.[12] put forward an efficient iterative approach for image tag refinement. In addition, several nearest neighbor-based method have also been proposed in the most recent years [13,14].

As briefly reviewed above, most of these approaches can achieve promising performance and motivate us to explore better image annotation methods with the help of their excellent experiences and knowledge. So in this paper we present a new method for refining image annotation by integrating PLSA with max-bisection (PLSA-MB). First, a PLSA model with asymmetric modalities is constructed to estimate the scores of all the annotating keywords. Next, a label similarity graph is built based on the initial annotations, especially the weights of edges are calculated by a novel weighted linear combination of label similarity and visual similarity of the candidate annotations. So far, the refining image annotation problem can be completely viewed as a graph partitioning problem. Thus the max-bisection over the built label graph is implemented based on the rank-two relaxation heuristics to further mine the correlation among annotation keywords. We evaluate our method on the Corel-5k dataset and the experimental results are comparative to several state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 introduces the PLSA model. In section 3, the construction of label similarity graph is first introduced, and then the rank-two relaxation heuristics for solving the max-bisection over
the graph is elaborated. Section 4 presents the experimental results on the Corel5k dataset. Section 5 gives the conclusions and future work.

2. PLSA MODEL

PLSA is a statistical latent class model which introduces a hidden variable (latent aspect) $z_k$ in the generative process of each element $x_j$ in a document $d_i$. Given this unobservable variable $z_k$, each occurrence $x_j$ is independent of the document it belongs to, which corresponds to the following joint probability:

$$P(d_i, x_j) = P(d_i) \sum_{k=1}^{K} P(z_k | d_i)P(x_j | z_k)$$

(1)

The model parameters of PLSA are the two conditional distributions: $P(x_j | z_k)$ and $P(z_k | d_i)$. $P(x_j | z_k)$ characterizes each aspect and remains valid for documents out of the training set. $P(z_k | d_i)$ is only relative to the specific documents and cannot carry any prior information to an unseen document. An EM algorithm is used to estimate the parameters through maximizing the log-likelihood of the observed data.

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, x_j) \log P(d_i, x_j)$$

(2)

where $n(d_i, x_j)$ is the count of element $x_j$ in document $d_i$. The steps of the EM algorithm can be described as follows.

E-step. The conditional distribution $P(z_k | d_i, x_j)$ is computed from the previous estimate of the parameters:

$$P(z_k | d_i, x_j) = \frac{P(z_k | d_i)P(x_j | z_k)}{\sum_{k=1}^{K} P(z_k | d_i)P(x_j | z_k)}$$

(3)

M-step. The parameters $P(x_j | z_k)$ and $P(z_k | d_i)$ are updated with the new expected values $P(z_k | d_i, x_j)$:

$$P(x_j | z_k) = \frac{\sum_{i=1}^{N} n(d_i, x_j)P(z_k | d_i, x_j)}{\sum_{i=1}^{N} \sum_{m=1}^{M} n(d_i, x_m)P(z_k | d_i, x_m)}$$

(4)

$$P(z_k | d_i) = \frac{\sum_{j=1}^{M} n(d_i, x_j)P(z_k | d_i, x_j)}{\sum_{j=1}^{M} n(d_i, x_j)}$$

(5)

If one of the parameters ($P(x_j | z_k)$ or $P(z_k | d_i)$) is known, the other one can be inferred by using fold-in method, which updates the unknown parameters with the known parameters kept fixed, so that it can maximize the likelihood with respect to the previously trained parameters. Given an unseen image visual feature $v(d_{new})$, the conditional probability distribution $P(z_k | d_{new})$ can be inferred with the previously estimated model parameters $P(v | z_k)$, then:

$$P(w | d_{new}) = \sum_{k=1}^{K} P(w | z_k)P(z_k | d_{new})$$

(6)

From equation (6), a candidate set of annotations with confidence scores can be easily obtained.

3. REFINING IMAGE ANNOTATION

As a latent aspect model, PLSA has been successfully applied in AIA [4, 15]. However, since all the annotations are calculated independently in PLSA model and the correlation among them is not fully exploited, which inevitably results in some ambiguity and inconsistency in the process of image annotation. In order to combine the prior confidence of candidate annotations and word correlation together, we present a two-stage refining image annotation approach illustrated in Fig.1. More details of it will be described in the following subsections.

3.1. Label Graph Construction

To construct the label graph, each candidate is transformed to a vertex and the pairwise label similarity is served as the weight of the corresponding edge. The most common methods include WordNet [16] and normalized Google distance (NGD) [17]. However, both of them build word correlation only based on textual descriptions and the visual information of images is not utilized at all, which also plays a crucial role in precise image annotation. So in this paper, the pairwise annotation similarity is calculated by a weighted linear combination of label similarity and visual similarity, which can effectively avoid the phenomenon that different images with the same candidate annotations would obtain the same refinement results after the secondary refining processing. The label similarity between $w_i$ and $w_j$ is defined as follows:

$$s_i(w_i, w_j) = \exp(-d(w_i, w_j))$$

(7)

where $d(w_i, w_j)$ represents the distance between two labels $w_i$ and $w_j$ and it is defined similarly to NGD as:

$$d(w_i, w_j) = \max \left( \log f(w_i), \log f(w_j) \right) - \log f(w_i, w_j)$$

(8)

where $f(w_i)$ and $f(w_j)$ are the numbers of images containing labels $w_i$ and $w_j$ respectively, $f(w_i, w_j)$ is the number of images containing both $w_i$ and $w_j$, and $G$ is the total number of images in the dataset. From the point view of labels associated with an image, the visual similarity between labels $w_i$ and $w_j$ is given as below:

$$s_v(w_i, w_j) = \exp\left( -\frac{1}{K \times K} \sum_{x \in \Gamma_{w_i}, y \in \Gamma_{w_j}} \frac{||x - y||^2}{\sigma^2} \right)$$

(9)
where $\Gamma_w$ is the representative image collection of label $w$, $x$ and $y$ denote image features corresponding to the respective image collections of label $w_t$ and $w_s$, $\sigma$ is the radius parameter of the Gaussian kernel function. To benefit from each of the two similarities described above, a weighted linear combination of label similarity and visual similarity is defined as follows:

$$s_{ij} = s(w_i, w_j) = \lambda s_v(w_i, w_j) + (1 - \lambda)s_e(w_i, w_j)$$

(10)

where $\lambda \in [0, 1]$ controls the weight for each measurement.

3.2. Implementing Max-bisection

The max-bisection problem is to find a partition of the vertices of a graph into two equal size subsets that maximizes the sum of the weights of the edges with endpoints in both subsets. Assume that $G=(V,E)$ is an undirected and connected graph. The max-bisection of $G$ is a partition of $V$ into two equally sized sets $(S, S')$ (i.e., a bisection of $V$) that maximizes the sum of the weights of the edges between $S$ and $S'$ (hence implying that $|V|$ should be even). So the max-bisection problem can be formulated as follows.

$$\text{max} \: \frac{1}{2} \sum_{1 \leq i < j \leq n} w_{ij} (1 - x_i x_j)$$

(11)

subject to

$$\sum_{i=1}^n x_i = 0, x_i \in \{-1, 1\}, i = 1, 2, \cdots, n.$$ 

where $w_{ij}$ is the edge weight, $w_{ij}=0$ for $(i, j) \notin E$, and in particular let $w_{ij}$=0, which has the same solution as the following binary quadratic program.

$$\text{min} \: \sum_{1 \leq i < j \leq n} w_{ij} x_i x_j$$

(12)

Its constraint is the same as that in Eq.(11). Note that the unit scalar variables $x_i$ in (12) can be replaced by unit vectors $v_i \in R^2$ (not $R^n$) and the scalar products $x_i x_j$ by the inner products $v^T v_j$. Correspondingly, all the vectors $v_i$ should be on the unit circle. Using polar coordinates, a set of $n$ unit vectors $v_1, v_2, \ldots, v_n \in R^2$ can be represented by means of a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in R^n$ consisting of $n$ angles, i.e., $v_i = [\cos \theta_i, \sin \theta_i]^T, \: i = 1, 2, \ldots, n$. Accordingly, the equation $v^T v_j = \cos (\theta_i - \theta_j), \: \forall \: i, j = 1, 2, \ldots, n$ is satisfied. We adopt the rank two relaxation heuristics [18] and polar coordinates, a new relaxation for max-bisection can be formulated as below:

$$\text{min} \: W \times \cos(\theta_i-\theta_j)/2, \: i, j = 1, 2, \ldots, n.$$ 

(13)

subject to

$$e \times \cos(\theta_i-\theta_j) = 0$$

where $W=[w_{ij}]$, $e$ is the vector of all ones. Suppose that a local or global minimizer $\theta$ has been obtained for (13). In addition, let us assume that $n$ is even and that $\theta$ satisfies $\theta_i \in [0, 2\pi], \: i = 1, 2, \ldots, n$ and $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_n$ after a reordering if necessary. Then to generate a bisection by picking any integer $k \in [1, n/2]$ and let $x_i=1$ if $i \in [k, j+n/2]$, otherwise $x_i=-1$. So far, we can give the procedure of PLSA-MB for refining image annotation as follows.

Algorithm 1: PLSA-MB for refining image annotation

1. Input: training set $T$, unlabeled image $I$, keyword vocabulary $V$, iteration number $N$, initial vector $\theta(0) \in R^n$
2. Output: the refined annotations for image $I$
3. Train PLSA model on image set $T$
4. Estimate the candidate annotations using PLSA model
5. Construct the label graph $G$ using equation (10)
6. Initial: $MB=0$, the iteration number of consecutive but non-improving random perturbations $N$
7. Starting from $\theta(0)$, minimize $f(\theta)(\text{Eq.}(13))$ to get $\theta^*=\arg\min f(\theta)$
8. for $j=1$ to $N$
9. Given $\theta^* \in R^n$ and $\theta_1, \theta_2, \ldots, \theta_n \leq \theta \leq \theta_n$, let $MB=0$
10. for $k=1$ to $n/2-1$
11. Generate a bisection $x \in R^n$ and compute $r(x)(\text{Eq.}(11))$
12. if $r(x) > MB$, then $MB=r(x)$
13. end for
14. if $r(x) > MB$, then $MB=r(x)$
15. end set $\theta^* = \theta^* + \Delta \theta^*$, $\Delta \theta^* \in R^n, j=j+1$
16. end for
17. Calculate the sum of weights of the edges belonging to each partition set $S_1$ and $S_2$, denoted by $w_{ij}$ and $w_{ij}$
18. if $w_{ij} \geq w_{ij}$, then labels in $S_1$ are selected as the results
19. else return the labels in $S_2$ as the refining annotations

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, PLSA-MB is tested on the Corel5k data set which contains 5000 images collected from larger Corel CD set. The data set is further divided into training set and testing set which contains 4500 and 500 images, respectively. Finally, the dictionary contains 300 words that appear in both the training and testing set. Here, each image is first partitioned into 32x32-sized blocks, then five kinds of features, 81-dim grid color moments, 59-dim LBP features, 120-dim Gabor texture features, 37-dim shape features and 512-dim SURF features are extracted [19] from each block so as to generate the bag-of-visual-words with 1000 dimension. In addition, the radius parameter $\sigma$ is set to the median value of all pairwise Euclidean distances between images and $K$ in Eq.(9) is set to 50 by trial and error.

4.1. Evaluation for the Weight $\lambda$

To find the optimal solution for the parameter $\lambda$, the average annotation precision is taken as the rule of evaluation. Fig. 2 depicts the curve of the average precision with the variant $\lambda$. We can clearly see that the performance is better when $\lambda \in (0, 1)$ than $\lambda=0$ or $\lambda=1$ individually. Particularly, $\lambda=0.7$ is the inflexion on the varied curve and hence is the optimal setting according to the better results, which further demonstrates the complementary nature of label similarity and visual similarity.
4.2. Comparison for Different Semantic Metric

To demonstrate the effectiveness of the proposed label similarity calculation, we conduct experiments for the 10 most frequent keywords in the Corel5k. 200 images for each keyword are selected respectively to constitute the final experimental dataset so as to make a fair comparison with the results obtained by WordNet and NGD, and the average annotating precision is used to measure their performance. As can be seen from Figure 3, our method, i.e., the weighted linear combination of label similarity and visual similarity, obviously gives superior precision to the others.

4.3. Refining Image Annotation on Corel5k

We perform our experiment with all 371 words to make a direct comparison with the results obtained by the literatures. The results are shown in Table 1. The left part is the annotation results in which ‘P’ is mean precision, ‘R’ is mean recall and ‘N+’ denotes the number of keywords with non-zero recall value. The right part is the retrieval results in which the mean average precision (MAP) is used to measure the performance. From Table 1, it is easy to see that PLSA-MB is superior or highly competitive to several state-of-the-art methods.

Table 1. Comparison of AIA on Corel5k dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>N+</th>
<th>Retrieval (MAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMRM[1]</td>
<td>0.10</td>
<td>0.09</td>
<td>66</td>
<td>0.17 0.20</td>
</tr>
<tr>
<td>CRM[2]</td>
<td>0.16</td>
<td>0.19</td>
<td>107</td>
<td>0.24 0.27</td>
</tr>
<tr>
<td>MBRM[3]</td>
<td>0.19</td>
<td>0.20</td>
<td>122</td>
<td>0.30 0.35</td>
</tr>
<tr>
<td>PLSA-WORDS[4]</td>
<td>0.14</td>
<td>0.20</td>
<td>105</td>
<td>0.22 0.26</td>
</tr>
<tr>
<td>PLSA-FUSION[15]</td>
<td>0.16</td>
<td>0.22</td>
<td>122</td>
<td>0.26 0.30</td>
</tr>
<tr>
<td>LASSO[13]</td>
<td>0.24</td>
<td>0.29</td>
<td>127</td>
<td>-    -</td>
</tr>
<tr>
<td>JEC[13]</td>
<td>0.27</td>
<td>0.32</td>
<td>139</td>
<td>-    -</td>
</tr>
<tr>
<td>SML[20]</td>
<td>0.23</td>
<td>0.29</td>
<td>137</td>
<td>0.31 0.49</td>
</tr>
<tr>
<td>TGLM[21]</td>
<td>0.25</td>
<td>0.29</td>
<td>131</td>
<td>0.29 0.52</td>
</tr>
<tr>
<td>PLSA-MB</td>
<td>0.26</td>
<td>0.30</td>
<td>132</td>
<td>0.28 0.58</td>
</tr>
</tbody>
</table>

To further illustrate the effect of PLSA-MB proposed in this paper, Figure 4 presents the retrieval results obtained with single word queries on several challenging visual concepts being queries. Each row displays the top five matches to the semantic query “flower”, “coast”, “tiger” and “mountain” from top to bottom respectively. The diversity of visual appearance of the returned images demonstrates that PLSA-MB also has good generalization ability.

5. CONCLUSION AND FUTURE WORK

We present a novel refining image annotation method by integrating PLSA with max-bisection. Particularly, the method for image annotation similarity calculation and the max-bisection, which can effectively avoid the phenomenon that different images with the same candidate annotations would obtain the same refinement results after the secondary refining processing. The experimental results on the Corel5k dataset show that PLSA-MB outperforms several state-of-the-art approaches. In the future, we intend to extend our algorithm to the object recognition problem under the settings that there are only a few labeled but a large number of unlabeled images. Besides, we plan to use different image datasets, especially some real-world images such as Mirflickr, to test the scalability of the proposed PLSA-MB.

6. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.61035003, 61072085, 60933004, 60903141), the National Program on Key Basic Research Project (No.2013CB329502), the National High-tech R&D Program of China (863 Program) (No.2012AA011003) and the National Science and Technology Support Program of China (2012BA107B02).
7. REFERENCES

annotation and retrieval using cross-media relevance model-

learning the semantics of pictures,” In Proc. NIPS, pp.553-


cross-media image indexing,” In PAMI, 29(10):1802-1817,
2007.

combining multiple evidence and wordnet,” In Proc. ACM

refinement using random walk with restarts,” In Proc. ACM


ing graph algorithm for image annotation refinement pro-


rank, content-tag prior and error sparsity,” In Proc. ACM


discriminative metric learning in nearest neighbor models for


heuristics for max-cut and other binary quadratic programs,”

cse.cuhk.edu.hk/~jkzhu/felib.html.
