

Heuristic Label Set Relevance Learning for Image Annotation

Kai Zhou[†] and Feng Tian^{††},

Northeast Petroleum University, School of Computer and Information Technology, Daqing, China

Summary

Automatic annotation can automatically annotate images with semantic labels to significantly facilitate image retrieval and organization. Traditional web image annotation methods often estimate specific label relevance to image, and neglect the relevance of the assigned label set as a whole. In this paper, A novel image annotation method by heuristic relevance learning is proposed. Label relevance are formulated into a joint framework. Measures that can estimate the relevance are designed, and the assigned label set can provide a more precise description of the image. To reduce the complexity, a heuristic algorithm is introduced, thus making the framework more applicable to the large scale web image dataset. Experimental results demonstrate the general applicability of the algorithm.

Key words:

Image semantic annotation; label set relevance; heuristic learning.

1. Introduction

Web image annotation has received broad attentions in recent years. Given a set of annotated images as training data, many methods have been proposed in the literature to find most representative labels to annotate an uncaptioned image. Most learning based methods about image annotation focus on learning a mapping between images and words given a number of training images. Compared with the potentially unlimited vocabulary existing in the web scale image databases, only a very limited number of concepts can be modeled^[1]. By leveraging numerous Web pages, search based approaches seem to be a promising way to solve this problem^[2-5]. Wang et al. proposed a search-based annotation system^[2]. Content based image retrieval is used to retrieve a set of visually similar images from a large-scale web image set. Text-based label search is used to obtain a ranked list of candidate annotations for each retrieved image, then the top ones in the ranked label lists are annotated. Liu et al. propose a label ranking approach which is able to rank the labels that are associated with an image according to their relevance levels^[3]. Li et al. introduce an approach that learns the relevance scores of labels by a neighborhood voting method^[4]. Given an image and one of its associated labels, the relevance score is learned by accumulating the votes from the visual neighbors of the image. They then further extend the work to multiple visual spaces. They learn the

relevance scores of labels and rank them by neighborhood voting in different feature spaces, and the results are aggregated with a score fusion or rank fusion method. Different aggregation methods have been investigated, such as the average score fusion, Borda count, and RankBoost. The results show that a simple average fusion of scores is already able to perform close to supervised fusion methods, like RankBoost. Recent studies indicates that the correlation between the labels can improve learning quality^[5-6]. But these methods often rely on complicated learning algorithms and are not easy to model the correlations between labels when extending to large number of labels. Intuitively, such information is helpful for us to better understand image content. Large scale web image annotation require the relevance learning method has both effectiveness and efficiency. In this paper, we propose a novel annotation algorithm which assigns not the common label set but the relevant and correlative label set. a web image annotation method based on label set relevance is proposed. A heuristic and iterative method to quickly discover the label set is proposed. To effectively estimate the information capability of a label, we utilize the change of posterior distributions of other labels after this label is added to the candidate label set. We further use the relevance between the test image and the label to filter out those labels that are "informative" but not relevant to the test image. To determine the correlation of a label to the candidate label set, we model the label and the candidate label set as two vectors in the semantic space which correlation can be measured by the cosine similarity.

2. Problem formulation

In this paper, we use S to denote a vocabulary, and S_q is a label set with q labels, the relevance of S_q to image I is defined as $R(I, S_q)$, internal correlation of S_q is defined as $C(I, S_q)$. We want to assign the most relevant and internal correlative label set to I . We use $F(I, S_q)$ to denote the objective. Thus, the optimization problem is: $S_q^* = \underset{S_q \subset S}{\operatorname{argmax}} F(I, S_q) = \underset{S_q \subset S}{\operatorname{argmax}} (\eta \cdot R(I, S_q) + (1 - \eta) \cdot C(I, S_q))$

where S_q^* is the ideal label set to I , η is a controlling parameter. Since the number of possible label set is

exponential with respect to the size of the label vocabulary. Thus we propose a heuristic algorithm to get an approximate solution. That is, searching the label which is most relevant to I and correlative to S_{q-1} at q -th iteration. Given a label denoted as s , we have $R(I, S_q) \approx R(I, S_{q-1}) + R(I, s)$, where $S_q = S_{q-1} \cup \{s\}$, $R(I, s)$ denotes the relevance of s to I . Similarly, the internal correlation of S_q with respect to image I can be estimated by $C(I, S_q) \approx C(I, S_{q-1}) + C(S_{q-1}, s)$, where $C(I, S_{q-1})$ denotes the internal correlation of S_{q-1} with respect to image I , and $C(S_{q-1}, s)$ denotes the relevance of label s to S_{q-1} with respect to image I . Then, we select the label which maximizes $F_s(I, s)$ at q -th iteration and added it into S_{q-1} , where $F_s(I, s)$ integrates the relevance information of s added to S_{q-1} . Thus, we have

$$s^* = \underset{s \in S/S_q}{\operatorname{argmax}} F_s(I, s) = \underset{s \in S/S_{q-1}}{\operatorname{argmax}} (\eta \cdot R(I, s) + (1 - \eta) \cdot C(S_{q-1}, s)) \quad (1)$$

3. The relevance of label set

Based on the information theory, we assume that the less ambiguous a label set is, the more information it has.. There are two scenarios for which one would want to suggest a new label to the user. The first scenario is if the current label set has more than one meaning. Resolving this type of ambiguity is non-trivial, as there exist many different ways a label set can appear ambiguous. Examples of ambiguity are word-sense ambiguity (e.g. label set {"apple", "photo"} can describe a fruit or a computer, while {"apple", "computer"} or {"apple", "fruit"} can be more discriminant). The second scenario is if the current label set is not sufficiently specific (e.g. a label set {"car"} is not specific since there are various brand like "Chevrolet" or "Ford"). Labels "nice" or "cool" added into will not give more relevant information to the image. Note that, from the perspective of the whole label set, the meaning of a label's relevance to one image indicates how much relevant information it added to the label set, which can be reflected from the ambiguity changes. We make a basic assumption about the meaning of ambiguity: A set of labels is ambiguous if there exist one label such that adding one or the other gives rise to very different distributions over the remaining labels. Thus, given the label "apple", adding the labels "fruit" or "computer" leads to very different means; and the other labels we see in this context are likely to. When a label w is added in, we can compute the KL divergence of the posterior distributions $P(s' | S_{q-1})$ and $P(s' | S_q)$, where $s' \in S$. The greater the

divergence value is, more chance w is selected into the label set. Thus, we have:

$$\begin{aligned} R(I, s) &\propto f(KL(P(S | S_{q-1} \cup \{s\}) \| P(S | S_{q-1}))) \\ &= f(KL(P(S | S_q) \| P(S | S_{q-1}))) \end{aligned} \quad (2)$$

where $f(\cdot)$ should be a monotonically increasing function, and KL divergence can be computed by:

$$\begin{aligned} KL(P(S | S_q) \| P(S | S_{q-1})) &= \sum_{s' \in S} P(s' | S_q) \log \frac{P(s' | S_q)}{P(s' | S_{q-1})} \\ \text{where } P(s' | \Omega) &= \frac{P(\Omega | s') P(s')}{P(\Omega)} = \frac{P(s') \prod_{s_1 \in \Omega} P(s_1 | s')}{\sum_{s_2 \in S} P(s_2) \prod_{s_1 \in \Omega} P(s_1 | s_2)}. \end{aligned}$$

To abbreviate notation, let Ω be S_q or S_{q-1} . The conditional probability of the label can be easily computed by label co-occurrence. Since there have large differences in the frequency of different labels, estimating their prior probability directly by their frequency simply to will include bias, so we compute conditional probability and prior probability as:

$$\begin{aligned} P(s_i | s_j) &= \frac{co(s_i, s_j)}{\sum_{k=1}^q co(s_j, s_k)}, \\ P(s_i) &= \frac{\sum_{j=1}^q co(s_i, s_j)}{\sum_{j=1}^q \sum_{k=1}^q co(s_j, s_k)}. \end{aligned}$$

where $co(\cdot)$ denotes the co-occurrence number. Considering the relevance between labels and image, formula(2) is rewritten as:

$$R(I, s) \propto f(KL(P(S | S_{q-1} \cup \{s\}) \| P(S | S_{q-1}))) \cdot P(s, I) \quad (3)$$

where $P(s, I)$ can be estimated with the expectation over the nearest neighbors of I as follows,

$$\begin{aligned} P(s, I) &= \sum_{i=1}^k P(s, I | I_i) P(I_i) = \sum_{i=1}^k P(s | I_i) P(I | I_i) P(I_i) \\ &\approx \frac{1}{k} \sum_{i=1}^k \delta(s, I_i) \cdot sim_{visual}(I, I_i) \end{aligned}$$

where $\delta(s, I_i)$ denotes whether I_i has label s , $sim_{visual}(I, I_i)$ denotes the visual distance of samples.

4. Label set internal correlation

In fact, label do not appear only in pairs. To simplify computation, we have assumed that conditional co-occurrences are independent in the estimation of label information capability in Section 3. However, this method may induce deviation because it does not consider the

correlation of label set. Considering the following scenario, label set {"sky", "sun"} has been assigned to image, then the label "rain" may be added in since it has correlation with "sky". So the label set should be correlative as a whole, namely, each label should be relevant to all other labels in the label set. The $C(S_{q-1}, s)$ is proposed to measure the relevance between S_{q-1} and label s . Thus, we have $c(S_{q-1}, I) = sim_{text}(S^v, I^v)$, where both the label set and label are represented by vector. A concept combination process is used to combine all the labels in the candidate label set to get the "label set vector" S^v . To get the vector of combined "label set vector", we need to construct the vector of each label first. The label to object matrix which reflects the relationship between labels and images is described as $D_{|S| \times n}$, where $|S|$ is the vocabulary size, n is the number of images in the training set. Then, the label to label matrix $E_{|S| \times |S|} = DD^T$ is obtained, which describes labels' semantic correlation. We normalize the matrix E by:

$$E_{ij} = \frac{E_{ij}}{E_{ii} + E_{jj} - E_{ij}}$$

where E_{ij} denotes the co-occurrence frequency of s_i and s_j . Thus, the E 's i -th row vector E_i can be regarded as the neighborhood vector of s_i , and the semantic correlation can be estimated by corresponding neighborhood vector E_i and E_j . We also use a heuristic concept combination process to construct the vector of the label set. Given two labels:

$$s_i^v = \langle E_{i,1}, \dots, E_{i,|S|} \rangle,$$

$$s_j^v = \langle E_{j,1}, \dots, E_{j,|S|} \rangle.$$

The combined label vector is

$$S^v_{s_i \cup s_j, k} = E_{ik} + E_{jk} \quad (4)$$

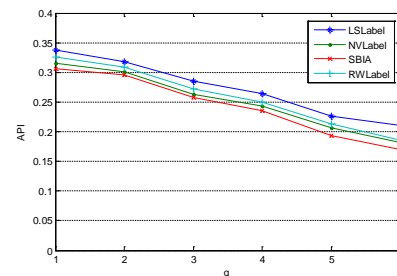
Then, the vector of the combined label set is normalized. Thus, the combined "label set vector" S^v is a combination of $(q-1) \times 1$ "neighborhood vector", i.e., $(\dots((s_1 \cup s_2) \cup s_3) \cup \dots \cup s_{q-1})$. Then the cosine similarity is applied to measure the semantic similarity of "label set vector" S^v and label s_i .

5. Experiments

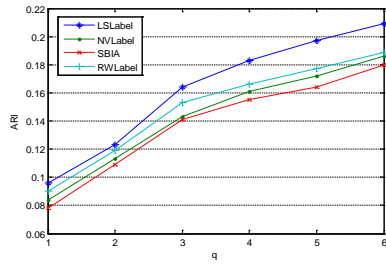
We conduct extensive experiments on dataset NUS-WIDE which contains 269,648 images and 5,018 labels^[7]. All of images are crawled from the image sharing website FLICKR. We split NUS-WIDE into a training set with

150,000 images and a test set with the remaining images. We evaluate our models with standard performance measures such as . average precision per label (ARL) and average precision per image (ARI). We extract different types of features commonly used for image search. We use two types of global image descriptors: Gist features, and color histograms with 16 bins in each color channel for RGB, LAB, HSV representations. Local features include SIFT as well as a robust hue descriptor, both extracted densely on a multiscale grid or for Harris-Laplacian interest points. Each local feature descriptor is quantized using k-means on samples from the training set. Images are then represented as a bag-of-words histogram. All descriptors but Gist are L1-normalised and also computed in a spatial arrangement. We compute the histograms over three horizontal regions of the image, and concatenate them to form a new global descriptor. To limit color histogram sizes, here, we reduced the quantization to 12 bins in each channel. Note that this spatial binning differs from segmented image regions, as used in some previous work. This results in 15 distinct descriptors, namely one Gist descriptor, 6 color histograms and 8 bag-of-features. To compute the distances from the descriptors we follow previous work and use L2 as the base metric for Gist, L1 for global color histograms, and L2 for the others.

We compare our algorithm LSLabel with the following state of the art web image annotation algorithms, i.e., SBIA(Search based Method)^[2], RWLabel^[3] and NVLabel^[4]. Figure 1 shows the curve of API and ARI with q varied from 1 to 6. According to the results, LSLabel archives encouraging improvements, and the API and ARI are greatly improved. For example, when q is fixed as 6, LSLabel has an improvement about 16% over NVLabel, 24% over SBIA, 14% over RWLabel in terms of API. In terms of ARI it has an improvement about 12% over NVLabel, 16% over SBIA, 11% over RWLabel. The reasons are that, it is difficult for SBIA to properly determine the number of clusters. NVLabel usually assign common labels which has larger frequency in the neighborhood. This is the same to RWLabel in which the labels has the most correlation to other labels are assigned to the test image. However, the LSLabel prefers the relevant label set as a whole.



(a) Curve of API



(b) Curve of ARI

Fig. 1 Annotation performance with q varied from 1 to 6

To compare the average precision and recall in terms of label, we fix the number of labels for image as 6 and obtain the APL and ARL. Figure 2 shows the results. We can observed that LSLLabel has an improvement 17-31% over other algorithms, and for the average recall of label, LSLLabel has an improvement about 17-35% over the other algorithms. Because these algorithms prefer the labels which appear frequently in the visually similar images. However, the LSLLabel prefers the relevant label set as a whole, and the label set which is correlative and has great relevance information to image is assigned.

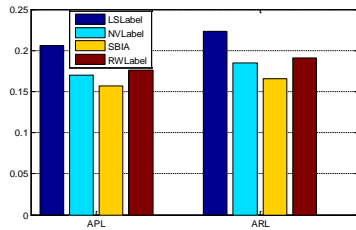


Fig.2 Comparison of APL & ARL on NUS-WIDE dataset

Figure 3 illustrates the time cost per image with varied scale of training set from 40K to 200K and labels for each image fixed as 6. The results shows that LSLLabel is a little

slower than SBIA, and the most efficient is Neighbor Voting, the worst is RWLabel.

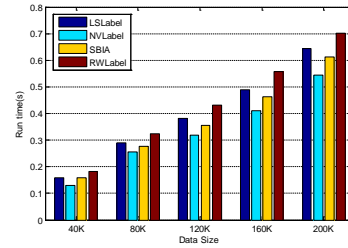





Fig.3 Comparison of run time on NUS-WIDE dataset

Figure 4 shows examples of F1 scores for some individual labels. Table 4 gives some real examples of annotation results and labels italic font are matching .

6. Conclusion

In this paper, a new method is proposed by learning "label set" relevance in a joint framework. The "label set"-to-image relevance is computed based on the posterior probability computation using KL divergence and label set's internal correlation is computed by label correlation analysis. We further solve the label set selection problem in a heuristic and iterative manner, and improve the method's efficiency by learning from neighborhood label set to reduce the computational complexity, thus making the framework more applicable to the large scale online annotation environment. Experiments show that the proposed method achieves excellent and superior performance.

Table 1: Annotation results generated by different methods (labels in italic font are matching)

			
Manual Labels	Easy parking urbnsim urban architecture	Allegra portrait girl effects colors	cloud lake preserve country forest poplar
LSLabel	<i>architecture parking building urban road car</i>	<i>girl portrait face boy colors kid</i>	<i>lake forest water nature country grass</i>
RWLabel	<i>urban city road architecture building car</i>	<i>girl kid happy pretty face portrait</i>	<i>lake water boat country sun grass</i>
SBIA	<i>urban city tree building car road</i>	<i>face girl woman boy pretty model</i>	<i>lake water fish grass nature tree sun</i>
NVLabel	<i>building factory architecture city road car</i>	<i>girl face hair colors boy kid</i>	<i>grass water tree sun cloud bank leaf</i>

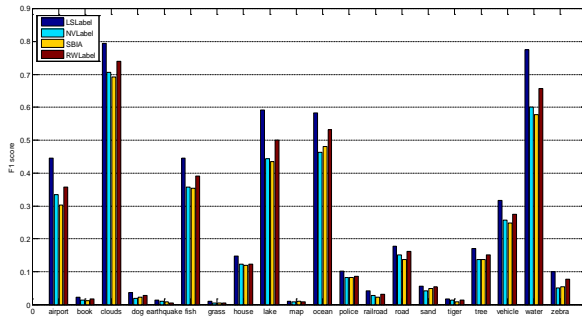


Fig. 4 Comparison of F1 for some labels in NUS-WIDE



Kai Zhou received the B.S. degrees in Computer Science from Northeast Petroleum University in 2006. Her main research interests include image annotation, cross media analysis, multimedia mining, virtual reality and pattern recognition.

Acknowledgments

This work is supported by Youth Foundation of Northeast Petroleum University (NO: 2013NQ120). We also would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

- [1] Wang Meng, Ni Bingbing and Hua Xian-Sheng, "Assistive tagging: a survey of multimedia tagging with human-computer joint exploration," *ACM Computing Surveys*, vol. 44, no. 4, pp.1-24,2012.
- [2] Xin-jing Wang and Lei Zhang, "Annotating images by mining image search results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30,no 11, pp. 1919-1932, 2008.
- [3] Dong Liu and Xian-Sheng Hua, "Tag ranking," *Proceeding of ACM International Conference on World Wide Web*, pp.340-351, 2009.
- [4] Xirong Li, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, pp.1310-1322, 2009.
- [5] H. Wang and H. Huang, "Image annotation using multi-label correlated Green's function," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-8,2009.
- [6] H.Wang and H. Huang, "Multi-label feature transform for image classifications," *Proceedings of European Conference on Computer Vision*, pp. 793-806, 2010.
- [7] Tat-Seng Chua and Jinhui Tang, "NUS-WIDE: a real-world web image database from national university of singapore," *Proceedings of ACM Conference on Image and Video Retrieval*, pp.1-9, 2009.