Retagging Social Images Based on Visual and Semantic Consistency[*]

Dong Liu[†], Xian-Sheng Hua[‡], Meng Wang[‡], Hong-Jiang Zhang[§]
[†] School of Computer Sci. & Tec., Harbin Institute of Technology, Harbin, 150001, P.R.China
[‡] Microsoft Research Asia, Beijing, 100190, P.R.China
[§] Microsoft Advanced Technology Center, Beijing, 100190, P.R.China
dongliu.hit@gmail.com,{xshua,mengwang,hjzhang}@microsoft.com

ABSTRACT

The tags on social media websites such as Flickr are frequently imprecise and incomplete, thus there is still a gap between these tags and the actual content of the images. This paper proposes a social image "retagging" scheme that aims at assigning images with better content descriptors. The refining process is formulated as an optimization framework based on the consistency between "visual similarity" and "semantic similarity" in social images. An effective iterative bound optimization algorithm is applied to learn the optimal tag assignment. In addition, as many tags are intrinsically not closely-related to the visual content of the images, we employ a knowledge-based method to differentiate visual content related from unrelated tags and then constrain the tagging vocabulary of our automatic algorithm within the content related tags. Experimental results on a Flickr image collection demonstrate the effectiveness of this approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Tagging, Tag Refinement, Retagging

1. INTRODUCTION

Online media repositories allow users to upload their media data and annotate them with freely-chosen tags. Despite the high popularity of tagging social images manually, the tags are often *imprecise*, *biased* and *incomplete* for describing the content of the images, which have significantly limited the performance of social image search and organization [1].

In this work, we propose an optimization framework to improve the tag quality based on the following two observations in real-world social images. First, consistency between visual similarity and semantic similarity, that is, similar images often reflect similar semantic theme, and thus are annotated with similar tags. Second, user-provided tags, despite imperfect, still reveal the primary semantic theme of the image content. In our approach, the improved tag assignments are automatically learned by maximizing the consistency between visual similarity and semantic similarity while minimizing the deviation from initially user-provided tags. This is actually using information from different channels to complement each other in a collective way.

However, the first observation mentioned above is mainly applicable for "content related" tags. That is, those tags that have high correspondence with the visual content. If we introduce "content unrelated" tags into the above optimization framework, we may even degrade the performance of the scheme. Accordingly, we propose a method to filter out those content unrelated tags to ensure that the quality of content related tags can be significantly improved.

2. TAG FILTERING

We perform tag filtering by taking advantage of lexical and domain knowledge. First, from the part-of-speech of words, we only consider nouns. Thus we restrict ourselves in the noun set of WordNet lexicon [2], which contains $114,648$ noun entries and the tags that are out of this set will not be considered.

Then we further analyze the noun tags and adopt an automatic process to detect visual property of the tags. We first empirically select a set of high level categories including "organism", "artifact", "thing", "color" and "natural phenomenon" as a taxonomy of our domain knowledge in vision field. Then the detection process can be implemented based on the WordNet lexicon which contains an implicit structure among words. For each noun entry in WordNet lexicon, we traverse along the path that is composed of hypernyms of the given word until one of the pre-defined visual categories is matched. If the match succeeds, the word is decided as content-related, and otherwise it is decided as content-unrelated.

3. TAG REFINEMENT

Denote by $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ a social image collection. All unique tags appearing in this collection are $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$. The initial tag membership for the whole image collection can be presented in a binary matrix $\widehat{\mathbf{Y}} \in \{0,1\}^{n\times m}$ whose element $\widehat{Y}_{ij}$ indicates the membership of tag t_j with respect to image x_i. To represent the refinement

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

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

results, we define another matrix $\mathbf{Y}$ whose element $Y_{ij} \geq 0$ denotes the confidence score of assigning tag t_j to image x_i. Denote by $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\top$ the confidence score vector of assigning the tags to the i-th image. Let $\mathbf{W}$ denote a similarity matrix whose element W_{ij} indicates the visual similarity between images x_i and x_j, which can be directly computed via $W_{ij} = exp(-\| x_i - x_j \|^2/\sigma^2)$. The semantic similarity of two images is defined based on their tag sets. We introduce the tag similarity matrix $\mathbf{S}$, in which the element $S_{ij} \geq 0$ indicates the tag similarity between tags t_i and t_j. In this work, we adopt Lin's similarity measure [4] and define the semantic similarity of images by a weighted dot product, i.e.,$\mathbf{y}_i^\top \mathbf{S} \mathbf{y}_j = \sum_{k,l=1}^{m} Y_{ik} S_{kl} Y_{jl}$.

According to our consistency assumption, the visual similarity is expected to be close to semantic similarity, i.e., $W_{ij} \approx \mathbf{y}_i^\top \mathbf{S} \mathbf{y}_j$. We then consider the assumption that user-provided tags are relevant with high probability. Here we introduce the minimization of $\sum_{j=1}^{n} \sum_{l=1}^{m} (Y_{jl} - \alpha_j \widehat{Y}_{jl})^2 exp(\widehat{Y}_{jl})$, where α_j is a scaling factor. Formally, we can summarize the above two assumptions into the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Y},\boldsymbol{\alpha}} \quad \mathcal{L} = & \sum_{i,j=1}^{n} (W_{ij} - \sum_{k,l=1}^{m} Y_{ik} S_{kl} Y_{jl})^2 \\ & + C \sum_{j=1}^{n} \sum_{l=1}^{m} (Y_{jl} - \alpha_j \widehat{Y}_{jl})^2 exp(\widehat{Y}_{jl}) \\ & s.t. \ Y_{jl}, \alpha_j \geq 0, \ i,j = 1,2,\ldots,n, \ k,l = 1,2,\ldots,m \end{aligned} \tag{1}$$

where C is a weighting factor to modulate the two terms.

In this work, we propose an efficient iterative bound optimization method to obtain its solution, which is analogous to [5]. The algorithm can be seen in our previous work [6].

4. EMPIRICAL STUDY

We collect a Flickr image collection consisting of $50,000$ images and $106,565$ unique tags. For each image, we extract 428-dimensional features, including 225-dimensional blockwise color moment features generated from 5-by-5 fixed partition of the image, 128-dimensional wavelet texture features, and 75-dimensional edge distribution histogram.

We first perform tag filtering on the Flickr dataset and obtain $4,556$ content-related tags. Then the proposed tag refinement method is evaluated within the filtered vocabulary. The radius parameter σ in visual similarity estimation is set to the median value of all the pairwise Euclidean distances between images. The parameter C in Eq. 1 is empirically set to 10. We compare the following three methods:

- Baseline, i.e., keeping the original tags.
- Content Based Annotation Refinement (CBAR). We adopt the method proposed in [3].
- Our tag refinement method.

Note that actually CBAR and our tag refinement method both produce confidence scores for tags. Therefore, for these two methods we rank the tags of each image based on their confidence scores and then keep the top m tags where m is the number of the original tags. The ground truths of the tags are voted by three volunteers. If a tag is relevant to the image, it is labeled as positive, and otherwise it is negative. However, manually labeling all the image-tag pairs will be too labor-intensive, and thus here we randomly select $2,500$ images as the evaluation set. We adopt precision/recall/F1-measure as performance measurements. But a problem is that the estimation of recall measurement needs to know the full set of relevant tags for each image, but in our case it is unknown. So here we adopt an alternative strategy. We gather the tags obtained by CBAR and our method as well as the original tags for each image, and then the positive tags among them are regarded as the full relevant set.

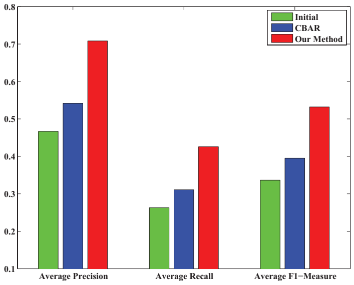

Figure 1: Performance comparison of the baseline and the two tag refinement methods.

Fig. 1 shows the precision, recall and F1-measure measurements obtained by the three methods, averaged over all evaluation images. We can see that CBAR and our method both outperform the baseline method, but the improvement of CBAR is limited. This is due to the fact that visual information has not been sufficiently explored in CBAR. Our tag refinement method performs much better than the baseline and the CBAR methods.

5. CONCLUSION

In this paper, we have introduced an image retagging scheme that aims at improving the quality of the tags associated with social images in terms of content relevance. Experiments on real-world social image dataset have demonstrated its effectiveness.

6. REFERENCES

[1] D. Liu, X. S. Hua, L. J. Yang, M. Wang and H. J. Zhang. Tag Ranking. In *Proceeding of ACM International World Wide Web Conference*, 2009.

[2] C. Fellbaum. Wordnet: An Electronic Lexical Database. *Bradford Books*, 1998.

[3] C. Wang, F. Jing, L. Zhang and H. J. Zhang. Content-Based Image Annotation Refinement. In *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[4] D. Lin. Using Syntatic Dependency as a Local Context to Resolve Word Sense Ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

[5] Y. Liu, R. Jin and L. Yang. Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[6] D. Liu, M. Wang, L. Yang, X.-S. Hua and H. J. Zhang. Tag Quality Improvement for Social Images. In *International Conference on Multimedia & Expo*, 2009.