Image Annotation Based on Recommendation Model

Zijia Lin
School of Software, Tsinghua University
1 Tsinghua Park
Beijing, 100084, P.R.China
linzj07@mails.tsinghua.edu.cn

Guiguang Ding
School of Software, Tsinghua University
1 Tsinghua Park
Beijing, 100084, P.R.China
dinggg@tsinghua.edu.cn

Jianmin Wang
School of Software, Tsinghua University
1 Tsinghua Park
Beijing, 100084, P.R.China
jimwang@tsinghua.edu.cn

ABSTRACT

In this paper, a novel approach based on recommendation model is proposed for automatic image annotation. For any to-be-annotated image, we first select some related images with tags from training dataset according to their visual similarity. And then we estimate the initial ratings for tags of the training images based on tag ranking method and construct a rating matrix. We also construct a trust matrix based on visual similarity with a k-NN strategy. Then a recommendation model is built on both matrices to rank candidate tags for the target image. The proposed approach is evaluated using two benchmark image datasets, and experimental results have indicated its effectiveness.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Verification

Keywords

image annotation, recommendation model, TrustWalker, retrieval

1. INTRODUCTION

The exponential growth of Web images has created a compelling need for innovative methods to retrieve and manage them. As one of the promising solutions, automatic image annotation has recently attracted numerous research attentions, whose ambition is to find several tags that can well represent the visual content when given a to-be-annotated image with no or few tags. Though many effective algorithms and models have been proposed for this challenge, e.g. *JEC* [4], automatic image annotation is still far from practical and needs further research.

In this paper, image annotation problem is transformed into personal item recommendation problem, where images are modeled as users and tags as items. Then the process of image annotation is to recommend tags for to-be-annotated images.

To apply any recommendation model, the so-called "cold start" problem must be firstly taken into consideration, because to-be-annotated images have no or few tags. Here we adopt a trust-based and item-based recommendation model named *TrustWalker* [2] as our basic model, which recommends tags mostly according to the trust between images, without requiring that the to-be-annotated image should have any existing tags. Furthermore, the initial ratings for tags of each training image, and the trust between images, need to be estimated. Then a model developed from *TrustWalker* can be applied to recommend tags for to-be-annotated images. The whole process of our proposed approach, *IARM* for short, is shown in **Figure 1**, and it can be roughly separated into Rating Estimation and Recommendation Model.

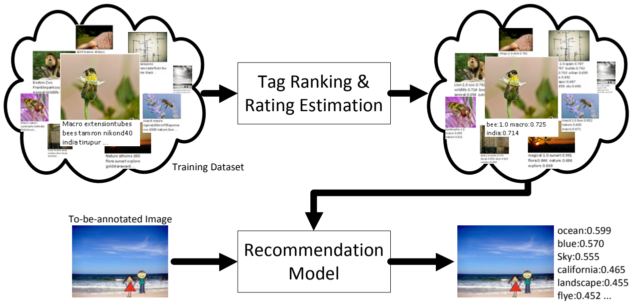

Figure 1: Process of our proposed *IARM*

2. RATING ESTIMATION

To estimate the initial ratings for tags of any training image, first we adopt the tag ranking method proposed by D. Liu [1], hoping to rank higher the tags that can better represent the visual content. The tag ranking method firstly employs a probability model to estimate a relevance score for each tag of image $\boldsymbol{v}$. Then a tag graph is constructed, where the edge weight is calculated with the exemplar similarity and semantic similarity between tags. Finally, a random walk is performed over the tag graph for obtaining stable tag scores, according to which all tags are ranked.

Inspired by the performance evaluation measurement named *NDCG* [1], the rating of any image $\boldsymbol{v}$ on tag $\boldsymbol{j}$, is initially estimated as $r_{v,j} = \frac{log\left(\frac{N}{N_j}\right)+1}{log_2(rk_{v,j}+1)}$, where $\boldsymbol{rk_{v,j}}$ is the rank position of the tag, $\boldsymbol{N}$ is the number of all images and $\boldsymbol{N_j}$ is the number of images containing tag $\boldsymbol{j}$. Here the numerator is the inverse image frequency of tag $\boldsymbol{j}$, which is introduced for punishing frequent but meaningless tags, e.g. "image". Furthermore, since existing tags are mostly relevant with visual content, mapping to higher ratings in recommendation model, we finally modify $\boldsymbol{r_{v,j}}$ into:

$$r_{v,j} = N_v \frac{log\left(\frac{N}{N_j}\right)+1}{log_2(rk_j+1)} + 0.5 \tag{1}$$

where $\boldsymbol{N_v}$ is a normalizing factor to map all ratings into [0, 1.0], and 0.5 acts as a score base. Then a rating matrix $\boldsymbol{R}$ can be constructed with all estimated ratings.

3. RECOMMENDATION MODEL

In basic *TrustWalker* model [2], to predict the rating of user $\boldsymbol{u_0}$ on item $\boldsymbol{i_0}$, the algorithm will firstly search the trust set of $\boldsymbol{u_0}$. If some trusted users have rated the target item, their ratings will be directly returned. As for those having not rated $\boldsymbol{i_0}$, the algorithm may return their ratings on some similar item, or further search their trust sets and return other users' ratings on $\boldsymbol{i_0}$ or similar items. Then all returned ratings are merged to be the predicted

SIGIR'11, July 24-28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

rating. The process of searching and selecting ratings can be simulated with several random walks, the step of which is equivalent to the depth of searching. Candidate items then are ranked according to their predicted ratings, and the top $\mathcal{K}$ will be recommended for $\boldsymbol{u_0}$.

To better fit the challenge of automatic image annotation, we have made some improvements from basic *TrustWalker*, including restraint on the search range, introduction of real-valued trust degree, etc. When annotating a given image $\boldsymbol{I_0}$, firstly we select $\mathcal{N}$ related images with tags from training dataset based on visual similarity, acting as a restricted search range and their tags being candidates. Such a restriction helps to reduce tag noise and computing complexity. Then the trust degree $\boldsymbol{tr_{I_m,I_n}}$ from image $\boldsymbol{I_m}$ to image $\boldsymbol{I_n}$ is calculated with the Gaussian kernel function [1]. That is, $tr_{I_m,I_n} = exp\,(-\frac{||I_m-I_n||^2}{\sigma^2})$, where $||\boldsymbol{I_m} - \boldsymbol{I_n}||$ is the visual distance, and $\boldsymbol{\sigma}$ is a radius parameter. Then each of the $(\mathcal{N} + \mathbf{1})$ images can respectively construct its own trust set with $\mathcal{M}$ of its most trusted images within the $\mathcal{N}$ related images. Finally a trust matrix $\boldsymbol{Tr_{(\mathcal{N}+1)\times(\mathcal{N}+1)}}$ consisting of real-valued trust degree can be constructed. To predict the rating $\boldsymbol{\check{r}_{I_0,t_k}}$ of image $\boldsymbol{I_0}$ on candidate tag $\boldsymbol{t_k}$, several random walks over the trust matrix are performed, as in basic *TrustWalker*. By calculating and merging the probability for rating selection, we can finally get the probability for the random walk to stop at image $\boldsymbol{v}$ and return its rating on tag $\boldsymbol{j}$, namely, $\boldsymbol{P(v,j)}$. Then $\boldsymbol{\check{r}_{I_0,t_k}}$ can be calculated as:

$$\check{r}_{I_0,t_k} = \sum_{\{(v,j)|B_{v,j}\ \&\ j\in Top3\}} P(v,j)r_{v,j} \qquad (2)$$

where $\boldsymbol{B_{v,j}}$ is a boolean variable denoting whether image $\boldsymbol{v}$ owns tag $\boldsymbol{j}$, $\boldsymbol{r_{v,j}}$ is the tag rating, and $\boldsymbol{j} \in \boldsymbol{Top3}$ means that for any related image, we just merge the returned ratings of the top 3 most similar tags, hoping to reduce the negative effects of dissimilar tags. Then we rank all candidate tags based on their corresponding predicted ratings, and take the top $\mathcal{K}$ as annotating result.

4. EXPERIMENT

Our proposed approach, *IARM*, is evaluated with two benchmark image datasets, namely, Ground Truth Database of University of Washington (*UW*) [5] and the real-world *MIRFlickr* [6]. In *UW*, there are quite few noisy tags, and many similar images are of the same scene in different views. In *MIRFlickr*, however, the existing tags are fairly noisy, including much metadata and many meaningless tags, and the images are quite distinct from each other. In both datasets, we respectively take 1/10th of the images to form our test sets and the remaining are used as training datasets. And all tags in training datasets are pre-processed with WordNet, for stemming and kicking out strange words.

As for baselines, we take the *LNP* proposed by Wang [3] and the baseline *JEC* proposed by Makadia [4]. Moreover, we extract six kinds of features for each image, that is, ColorLayout, Gabor, Sift, ScalableColor, EdgeHistogram and Tamura. When calculating visual distance between images, we normalize and merge distances of different features with equal weights. As for experimental parameters, the threshold for random walk is set as 0.001, and $\boldsymbol{\sigma}$ in Gaussian kernel function is set as the average visual distance. In *UW*, we set $\mathcal{N}$ as 20, $\mathcal{M}$ as 5, while in *MIRFlickr*, both are respectively set as 40 and 10. And in *UW*, the top 5 tags of each algorithm are selected as its annotating result, while top 10 in *MIRFlickr*. The performances of all three approaches are measured with *top N precision rate* [7], which measures the correctness of the top *N* annotations for an image. *Top N precision rate* is defined as: $p(N) = \frac{1}{M}\sum_{i\in I_t}\frac{correct\_i(N)}{N}$, where $\boldsymbol{I_t}$ is the test image set, $\boldsymbol{M}$ is the size of $\boldsymbol{I_t}$, and $\boldsymbol{correct\_i(N)}$ is the number of correct annotations in the top *N* annotations for image $\boldsymbol{i}$. For *UW*, the original tags of each test image act as the judging standard, while for *MIRFlickr*, volunteers fluent at English are asked to determine the correctness of each annotation, without knowing which algorithm it comes from.

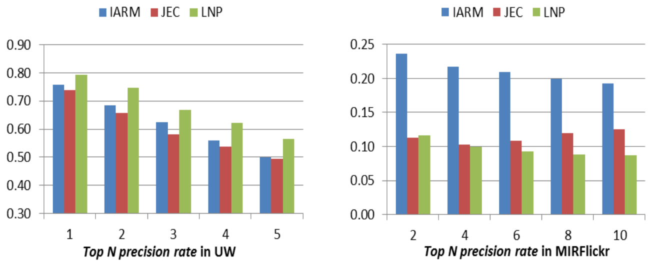

Figure 2: ***Top N Precision Rate*** **in dataset** ***UW*** **and** ***MIRFlickr***

As shown in **Figure 2**, in dataset *UW*, *LNP* achieves the best performance, and our proposed *IARM* achieves the second, while *JEC* the worst. However, in dataset *MIRFlickr*, *IARM* performs much better than the other two, while *LNP* achieves the worst performance. Moreover, because of the considerable frequent noisy tags and the distinct images, all three algorithms that adopt the k-NN strategy perform much worse in *MIRFlickr*. By comparing the performances on both datasets, we can draw following conclusions: (1) Our proposed *IARM* is effective and relatively more robust; (2) Automatic image annotation is far from practical, as illustrated by performances in real-world *MIRFlickr*.

5. CONCLUSION

In this paper, we propose a novel approach *IARM* based on recommendation model for automatic image annotation. The proposed *IARM* is evaluated using two benchmark image datasets and experimental results have indicated its effectiveness, which suggests that other mature theories in recommendation problem may worth further research for being applied in image annotation.

6. ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No. 60972096, 90924003, 61005045)

7. REFERENCES

[1] D. Liu, X. Hua, L. Yang, M. Wang, H. Zhang. 2009. Tag Ranking. *ACM WWW'09.*

[2] M. Jamali, M. Ester. 2009. TrustWalker: A Random Walk Model for Combining Trust-based and Item-based Recommendation. *ACM SIGKDD '09.*

[3] F. Wang, C. Zhang. 2008. Label Propagation through Linear Neighborhoods. *IEEE TKDE.*

[4] A. Makadia, V. Pavlovic, S. Kumar. 2008. A New Baseline for Image Annotation. *ECCV'08.*

[5] http://www.cs.washington.edu/research/imagedatabase/

[6] M. J. Huiskes, M. S. Lew. 2008. The MIR Flickr Retrieval Evaluation. *ACM MIR'08.*

[7] X. Rui, M. Li, Z. Li, W. Ma, N. Yu. 2007. Bipartite graph reinforcement model for web image annotation. *ACM MM'07*