TAGRANK - MEASURING TAG IMPORTANCE FOR IMAGE ANNOTATION

Xiao Ling[1]*, Jimin Jia*[2]*, Nenghai Yu*[2]*, Mingjing Li*[3]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MOE-Microsoft Key Lab of MCC and Department of EEIS
University of Science and Technology of China, Hefei 230027, China
[3]Microsoft Research Asia, 49 Zhichun Road, Beijing 100190, China

ABSTRACT

Traditional image annotation approaches are only applicable for datasets with small and limited lexicon. Besides, annotation words are treated equally without considering the importance of each word in the real world. To address these problems, we propose TagRank, a method to model the relative importance of every candidate word. By exploiting tag clusters on Flickr, TagRank could be modeled as random walk with restarts, which incorporates both word frequency and word correlation information. As a result, a ranked annotation vocabulary could be built. By utilizing the tag importance in a real image annotation experiment, we show that TagRank is helpful for improving the performance of image annotation.

Index Terms— TagRank, image annotation, tag importance.

1. INTRODUCTION

Image annotation aims to provide several keywords automatically for a given image to describe its content, which is very useful in image retrieval. Traditionally, there are many methods to deal with automatic image annotation, such as SVM 0, Translation model™ [4], Cross-Media Relevance Model(CMRM) [5], Continuous Relevance Model(CRM) [6] and Graph-based model [8], etc. However, these traditional methods are all built on datasets with limited words. For example, the commonly used dataset Corel only contains about 371 words in total. Some popular words may be not included in these datasets, like "mp3", "ipod", etc. On the other hand, there are abundant resources on the web, providing us with large-scale images and user-labeled tags. Some works [2][3] have been conducted on web image annotation. Nevertheless, the problem is that the surrounding texts used in these methods are very noisy and disordered. Besides, the unlimited vocabulary actually retards the speed and accuracy of image annotation. Therefore it is very meaningful to build a reasonable annotation vocabulary. With this vocabulary, annotations could be produced from it since words in the vocabulary are more suitable for annotation.

We believe annotation words should be associated with an importance measurement, instead of being considered equally. Although the word importance is a subjective matter at first sight, there are still some quantitative criteria. In practice, users would prefer some words rather than others even if these words have the same meaning. For instance, if we submit 'UK' and 'United Kingdom' to Flickr, it will return 1,386,422 and 35,798 images respectively, which indicates that people are prone to choose 'UK' as image tags rather than 'United Kingdom'. Therefore, words in the annotation vocabulary should have some measure of importance. Intuitively, there are several criteria for constructing such a vocabulary: (1) the vocabulary should cover as many aspects as possible since the more aspects the vocabulary covers, the more convincing it is considered as a reasonable vocabulary; (2) words in the dictionary should be commonly used words in the daily life as much as possible since people are more likely to use these words for retrieval in most cases. Obviously, the vocabulary should be built on a large scale dataset. As a large image sharing community, Flickr provides a public platform for user to upload and share their images. These images are associated with a series of tags, due to the human contribution in the social website. Therefore Flickr could be considered as an excellent source for building annotation vocabulary. As the size of tags on Flickr is too large, the vocabulary should be filtered according to the tag importance.

Our work focuses on modeling the relative importance of words based on tags on Flickr. We propose TagRank, a method to compute an importance score for every candidate word. It is called TagRank because it is inspired by the widely known method PageRank[7] in link analysis which measures the importance of every webpage on the Internet. Firstly we choose the most representative words from Flickr.

By combining word frequencies and word correlations, a random walk with restarts model could be built for TagRank. After limited times of iteration, the final steady-state rank of every candidate word could be calculated. As a result, a ranked vocabulary is obtained. To the best of our knowledge, the work on mining the importance of annotation words has never been done before.

The paper is organized as follows. The definition and implementation details of TagRank are introduced in Section 2. The evaluation method and experimental results on 5,000 images are shown in Section 3. Finally, we conclude our work in Section 4.

2. TAGRANK

2.1 Definition

TagRank refers to an overall relative importance of words. In order to measure the importance of every single word, two factors should be taken into consideration. One is the word frequency, which is the number of images tagged by the word. Obviously, higher word frequency means that users are more prone to tag the uploading images with that word. In practice, we adopt frequency values as initial importance for each annotation word, and they could be easily obtained from Flickr. The other factor is the correlation between words. This factor is crucial because even two words with similar word frequencies still have some differences. Suppose word $w1$ and $w2$ have the same word frequency, but $w1$ was correlated to word $w3$ which has higher importance, while $w2$ was correlated to word $w4$ which has lower importance. Correspondingly, the importance of $w1$ should be reinforced by its neighbors, hence $w1$ is considered more important than $w2$.

Unlike the traditional off-line word semantic correlation ontology, like WordNet, the word correlation should be consistent with the users' preference of the tags in practice. There are many web resources providing us clues for word correlations, like PicSearch, Flickr, etc. As an abundant source of images and user-labeled tags, Flickr provides word clusters for most of words. These clusters are represented by a series of tags, which we believe are statistically the most representative words correlated with the query word, according to the user-labeled tags. For instance, if 'sunset' was submitted to Flickr[*], it will return clusters like 'sky', 'clouds', 'sun', 'water'. In our work, we derive word correlations over word clusters on Flickr. The reason we choose Flickr as our source is as follows: firstly the clusters are built on large-scale user-labeled tags; secondly, the clusters are represented by a series of words which are most relevant with the query word.

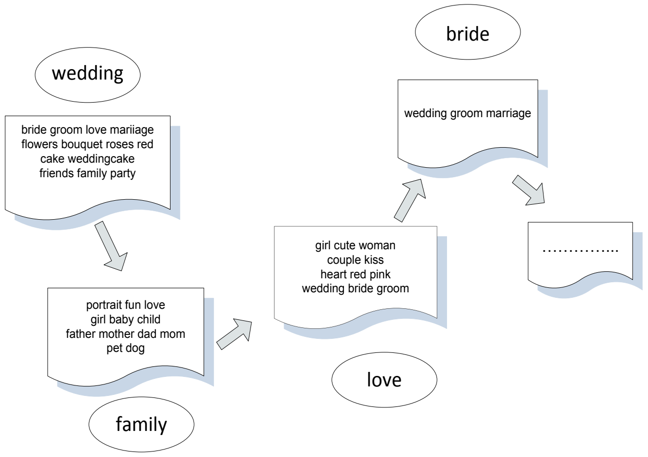
wedding
bride groom love mariiage flowers bouquet roses red cake weddingcake friends family party
portrait fun love girl baby child father mother dad mom pet dog
family
girl cute woman couple kiss heart red pink wedding bride groom
love
wedding groom marriage
bride
…………...

Figure 1 random walk model for TagRank

In PageRank, the importance of a webpage is modeled by the sum of importance of its backlinks. Similarly, the same idea could also be applied for the importance of words, which could be modeled by the sum of importance of its 'backwords'. Here 'backwords' refers to the directional relationship of the query word and tags in its clusters. In the above 'sunset' example, 'sunset' is the backword of 'sky', 'clouds', 'sun' and 'water'. But this may not be true vice versa. That is, 'water' may be not the backword of 'sunset', for 'sunset' may not exist in 'water's clusters.

Denote $R(u)$ the TagRank of word u, B_u the set of backwords of u. Suppose word v is one of backword of u, $v \in B_u$. Besides u, v is the backword of N_v words in all. Suppose $R(v)$ is TagRank of v, therefore the contribution of v for $R(u)$ should be distributed equally among the N_v words. That is $R(v)/N_v$. By summing all contributions of backwords of u in B_u, $R(u)$ could be represented as:

$$R(u) = \sum_{v \epsilon B_u} \frac{R(v)}{N_v} \tag{1}$$

Let's denote $A(u, v)$ as the transitional relationship between word u and word v, and represent it as:

$$A(u,v) = \begin{cases} \frac{1}{N_v} & v\ is\ backword\ of\ word\ u \\ 0 & v\ is\ not\ backword\ of\ word\ u \end{cases} \tag{2}$$

Then formula (1) could be rewriten as:

$$R(u) = \sum_{v \epsilon B_u} A(u,v)R(v) \tag{3}$$

The random walk with restarts model makes sense here. As could be seen in Figure 1, suppose given an image μ, it is firstly annotated with a word, like 'wedding'. Generally, besides 'wedding', other words could also be regarded as reasonable tags for the image. We look up 'wedding's clusters, and find the word 'family' is another available

[*] http://www.flickr.com/photos/tags/sunset/clusters/

annotation word for the image. After that, we choose other related words in 'family' clusters. This process could be repeated recursively. With the help of the random walk with restarts model, TagRank could be computed iteratively in several steps described in section 2.2.

2.2 Computation Steps

Let's denote I the word frequencies on Flickr. Take I as initial TagRank $R^{(0)}$. Then TagRank could be iteratively computed as follows:

1. $R^{(0)} = I$
2. $R^{(t)} = \alpha A R^{(t-1)} + (1-\alpha) I$
3. normalize $R^{(t)}$, so that $\sum_i R^{(t)}(i) = 1$
4. $d = \left\| R^{(t)} - R^{(t-1)} \right\|$
5. repeat step 3 to 6 until $d < C$

C is the threshold constant. Since the transitional matrix A is aperiodic and irreducible, the iterative calculation (step 2) converges to the principle eigenvector of A.

2.3 Implementation

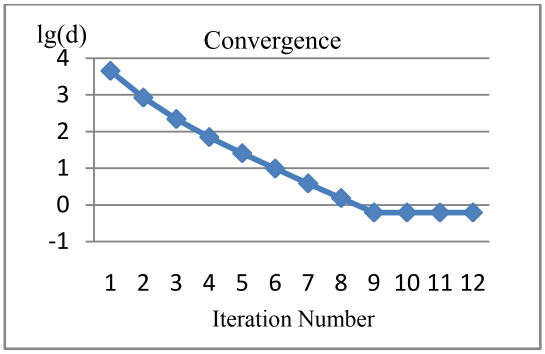

Figure 2 convergence of TagRank

We collect most popular tags[*] on Flickr as seeds. There are 140 words in all. All the seeds are used to search their clusters. Then new words in the clusters are added to the word seeds and search their own clusters. This process was repeated iteratively. We empirically set the iterative depth as 100. In the end, 20,058 different words in all are obtained. These words does not cover the whole vocabulary of Flickr, but at least the most representative words users prefer to tag their images. After iteratively computing, TagRank could converge to a stable state. The convergence of TagRank is shown in Figure 2**Error! Reference source not found.**(here α is empirically set 0.5). As shown in the convergence curve, the convergence speed is relatively quick, converging to a reasonable tolerance in less than 10 steps. Top 50 words are listed in the appendix.

3. EXPERIMENTS

3.1 Experimental method

To evaluate the validity of TagRank, we test the performance in a real image annotation application, where TagRank could be seen as the prior probabilities of words. Here we use a popular graph-based method, manifold ranking [8], to evaluate the performance. In graph-based method, the probabilities of annotation words are propagated between training images and test images according to their similarity. Suppose $F_i^{(t)}$ is a column vector, indicating the probability of ith word as annotation words for all images in the tth iteration step, S represents normalized visual similarities between images, Y_i is the prior probability of ith word as annotation words for all images. If tags of image I contains the ith word, $Y_i(I)$ is 1. Otherwise, $Y_i(I)$ is generally set 0. Then the *(t+1)*th iteration step for word i is :

$$F_i^{(t+1)} = (1-c) S F_i^{(t)} + c Y_i \qquad (4)$$

where c is the weighted coefficient and is manually tuned via experiments . The visual similarity $W(I,J)$ between image I and image J is $\exp(-d^2(I,J)/2\sigma^2)$, $d(I,J)$ is the L1-distance of image I and J. Normalized image similarities S could be written as: $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$. D is diagonal matrix with diagonal element D_{ii} equal to the sum of the ith row of W.

We compare two methods of image annotation: one method not using TagRank(called Manifold); The other one using TagRank, where the prior probability Y_i for test images are set to the TagRank of word i.

About 5,000 images with user-labeled tags are crawled from Flickr. Among the 5,000 images, 4,500 images are used as training set, 500 images are used for test set. The users labeled tags are considered as ground truth annotations. As for the evaluation metric, the commonly evaluation measure precision-recall is used:

$$precision@m = \frac{1}{M} \sum_{j \epsilon I} \frac{correct\_j(m)}{m} \qquad (6)$$

$$recall@m = \frac{1}{M} \sum_{j \epsilon I} \frac{correct\_j(m)}{Nr(j)} \qquad (7)$$

where M is the test image number, *correct_j(m)* is the number of correct annotations in the first m annotations for the *jth* image, *Nr(j)* is the number of ground truth annotation words for the *jth* image .

3.2. Experimental Results

[*] http://www.flickr.com/photos/tags/

The experiment results of precision and recall against the number of annotation words are shown in Figure 3 and Figure 4 respectively. As shown in the comparison bars, the TagRank method performs better in both precision and recall than the common manifold method without TagRank. Particularly in precision measure, the TagRank method obviously outperforms Manifold in the top words. With the test examples from Flickr, it indicates that by considering word importance in the real world, the annotation is more consistent with user-labeled tags.

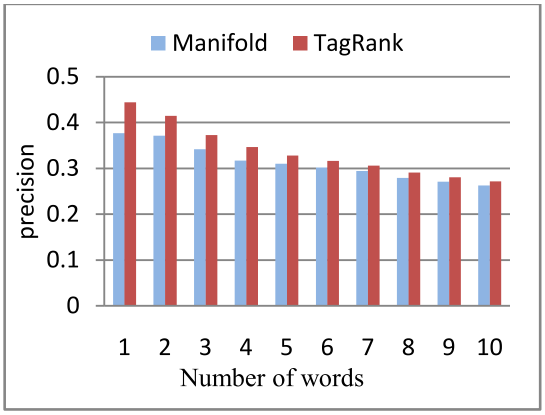

Figure 3 precision@m

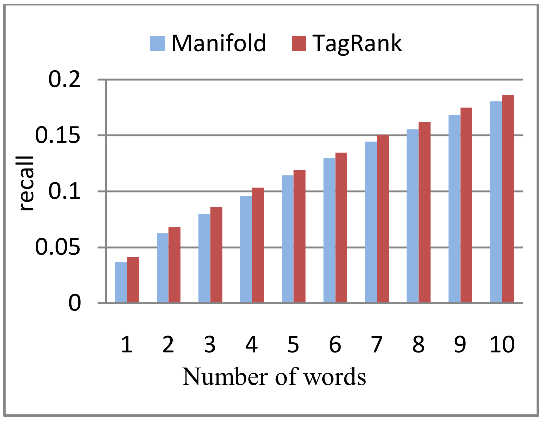

Figure 4 recall@m

4. CONCLUSION

We proposed TagRank to measure the importance of different words on the real-world image sharing community Flickr. By incorporating both the frequency and correlation of different tags, the final rank of each word could be acquired. The experiments show that the performance of image annotation could be improved by considering TagRank.

5. ACKNOWLEDGMENTS

The research is supported in part by National Natural Science Foundation of China (60672056) and Specialized Research Fund for the Doctoral Program of Higher Education (20070358040).

6. REFERENCE

[1] Claudio, C., Gianluigi, C., and Raimondo, S., "Image annotation using SVM", In Proceeding of Internet imaging IV, Vol. SPIE, 2004.

[2] X.J. Wang, L. Zhang, F. Jing, W.Y. Ma. AnnoSearch: Image Auto-Annotation by Search. International Conference on Computer Vision and Pattern Recognition, New York, USA, June, 2006.

[3] C. Wang, F. Jing, L. Zhang, H.J. Zhang. Scalable search-based image annotation of personal images. ACM international workshop on Multimedia information retrieval. 2006.

[4] Duygulu, P., and Barnard, K., "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", In Proceeding of ECCV, 2002.

[5] Jeon, J., Lavrenko, V., and Manmatha, R., "Automatic Image Annotation and Retrieval Using Cross-media Relevance Models", In Proceeding of SIGIR, Toronto, July 2003.

[6] Lavrenko, V., Manmatha, R., and Jeon, J., "A Model for Learning the Semantics of Pictures", In Proceeding of NIPS, 2003.

[7] Page, L., Brin, S., Motwani, R., Wingrad, T. 1998. The pagerank citation ranking: Bringing order to the web. Tech. Rep.. Computer Systems Laboratory, Stanford University,Stanford, CA.

[8] J. Liu, M. Li, W.-Y. Ma, Q.Liu, H. Lu, An Adaptive Graph Model for Automatic Image Annotation. ACM Int'l Conf. on Multimedia Information Retrieval, 2006.

7. APPENDIX – TOP 50 WORDS

sky	city	tree	flowers	sun
blue	clouds	travel	black	family
water	light	nyc	flower	bridge
nature	sea	architecture	ocean	california
beach	portrait	girl	london	france
red	white	people	art	woman
night	trees	summer	yellow	reflection
sunset	landscape	newyork	italy	vacation
bw	wedding	party	orange	europe
green	street	macro	japan	pink