Image Annotation Refinement using Random Walk with Restarts*

Changhu Wang
Department of EEIS, University of Science and Technology of China
Hefei 230027, China
(86)13581984028
wch@ustc.edu

Feng Jing, Lei Zhang, Hong-Jiang Zhang
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
(86-10)62617711~{6039, 3197, 5791}
{fengjing, leizhang, hjzhang}@microsoft.com

ABSTRACT
Image annotation plays an important role in image retrieval and management. However, the results of the state-of-the-art image annotation methods are often unsatisfactory. Therefore, it is necessary to refine the imprecise annotations obtained by existing annotation methods. In this paper, a novel approach to automatically refine the original annotations of images is proposed. On the one hand, for Web images, textual information, e.g. file name and surrounding text, is used to retrieve a set of candidate annotations. On the other hand, for non-Web images that are lack of textual information, a relevance model-based algorithm using visual information is used to decide the candidate annotations. Then, candidate annotations are re-ranked and only the top ones are reserved as the final annotations. To re-rank the annotations, an algorithm using Random Walk with Restarts (RWR) is proposed to leverage both the corpus information and the original confidence information of the annotations. Experimental results on both non-Web images of Corel dataset and Web images of photo forum sites demonstrate the effectiveness of the proposed method.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms
Algorithms, Experimentation

Keywords
Image annotation refinement, Random Walk with Restarts

1. INTRODUCTION
With the prevalence of Web and digital cameras, there are more and more digital images on personal devices and on the Web. For example, Google Image [1] has indexed more than one billion images. The explosion of digital images necessitates effective image management, browsing and search tools. For Web images, search is the most critical thing. Existing Web image search engines are based on textual information of the images, e.g. file name, ALT text, URL and surrounding text. We can treat the surrounding text as annotations. However, for most of the images, such annotations are usually noisy with irrelevant words. If the imprecise annotations could be refined, the performance of Web image retrieval could be possibly improved. For images on personal devices, how to effectively manage and search them is increasingly important as the number of images grows rapidly. In this case, annotations serve as a key factor in the management and search process. Since images have little textual information, annotation using visual content is required. Therefore, many content-based annotation algorithms have been proposed since 1999. Most existing approaches can be classified into two categories, i.e. classification-based methods and probabilistic modeling-based methods. The methods of the first category try to associate words or concepts with images by learning classifiers [4][5][12]. The probabilistic modeling-based methods attempt to infer the correlations or joint probabilities between images and annotations. The representative work includes Co-occurrence Model [14], Translation Model [6], Latent Dirichlet Allocation Model (LDA) [3], Cross-Media Relevance Model (CMRM) [8], Continuous Relevance Model (CRM) [11], and Multiple Bernoulli Relevance Model (MBRM) [7]. Despite the continuous effect put on image annotation, the results of existing image annotation methods are still unsatisfactory. Consequently, it is necessary to refine the current annotation results.

Jin *et al*. [9] have done pioneer work on annotation refinement using a generic knowledge-based WordNet. From the small candidate annotation set obtained by an annotation method, the irrelevant annotations will be pruned using WordNet [13]. The basic assumption is that highly correlated annotations should be reserved and non-correlated annotations should be removed. However, the experimental results show that although the method can remove some noisy words, many relevant words are also removed. As a result, the F_1 value (see Section 3.1) decreases compared with the original annotation method. There are mainly two reasons for the unsatisfactory performance. First, the algorithm did not fully utilize the original annotation information. More specifically, once the candidate annotations were picked up, the original confidence scores of the annotations were ignored. Second, although WordNet contains additional generic knowledge on the relation of words, it has two limitations. One is that it is independent of the dataset and therefore can not reflect the characteristics of the specific image dataset. The other limitation is that it can not deal with the annotations that do not exist in the lexicon of WordNet.

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

* This work was performed at Microsoft Research Asia.

In this paper, a novel approach to automatically refine the original annotations of both Web and non-Web images is proposed. On the one hand, for Web images, textual information is used to retrieve a set of candidate annotations. On the other hand, for non-Web images that are lack of textual information, a relevance model-based algorithm using visual information is used to decide the candidate annotations. Then, the annotations are refined by re-ranking the candidate annotations and reserving the top ones. To resolve the issues of [9], an algorithm using Random Walk with Restarts (RWR) is proposed to re-rank the candidate annotations. The algorithm not only uses the corpus information by defining a co-occurrence-based similarity, but also leverages the ranking and confidence information of original annotations. Experimental results on both non-Web images of Corel dataset and Web images of photo forum sites demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 describes the annotation refinement algorithm. Experimental results on both Web and non-Web images are shown in Section 3. Finally, we conclude in Section 4.

2. IMAGE ANNOTATION REFINEMENT ALGORITHM

First, a set of candidate annotations for the query image need to be identified. We deal with Web images and non-Web images in different ways. Second, the RWR algorithm is used to re-rank the candidate annotations. Finally, the top ranked annotations will be reserved as the final annotations.

2.1 Candidate Annotations Identification

2.1.1 Web Images

For Web images, there is related textual information of different sources, e.g. file name, ATL text, URL and surrounding text. First, the stop words are removed and stemming is utilized. Then, for each left word w_i, a confidence score is calculated as follows:

$$score(w_i) = \sum_{s \in S} weight(s) \times tf_s(w_i) \quad (1)$$

where S is the set of different sources, $weight(s)$ is the weight of the source s and $tf_s(w_i)$ is the occurrence number of w_i in source s. Finally, the words are ranked according to their confidence scores and only the words with highest scores are reserved as candidate annotations.

2.1.2 Non-Web Images

Since non-Web images have little textual information, annotation by visual content is necessary. To facilitate the annotation refinement process, the annotation algorithm should provide a confidence score for the candidate annotations. Although most of the existing algorithms [3][6][7][8][11] satisfy this requirement, the Cross-Media Relevance Model (CMRM) [8] model is chosen for its simplicity and effectiveness. CMRM is based on the relevance model [10] proposed for information retrieval. It assumes that regions in an image can be described using a vocabulary of blobs. The blobs are generated by clustering the visual features of several regions. Based on a training set of images with annotations, the probability of generating a word given the blobs of an image is estimated. The probability could be used as the confidence score of the word. More specifically, the confidence score of word w_i is defined as:

$$score(w_i) = p(w_i \mid I) \approx p(w_i \mid b_1 ... b_m) \quad (2)$$

where I is the image to be annotated and $b_1 \ldots b_m$ are the blobs of I.

2.2 Annotation Refinement with RWR

In order to fully utilize the confidence scores of the candidate annotations and the corpus information, we reformulate the image annotation refinement process as a graph ranking problem and solve it with the RWR algorithm.

2.2.1 Graph Construction

Each candidate annotation w_i is considered as a vertex of a graph G. All vertices of G are fully connected with proper weights. The weight of an edge is defined based on the "co-occurrence" similarity as follows.

For Web images, each word w_i will be used as a query to query a Web image search engine, e.g. Google Image [1]. The number of search results is denoted as $num(i)$. For two different word w_i and w_j, "w_i w_j" will be used as the query. The number of search results is denoted as $num(i, j)$. The weight of the edge between w_i and w_j can be calculated by the following formula:

$$sim(w_i, w_j) = \begin{cases} \dfrac{num(w_i, w_j)}{\min(num(w_i), num(w_j))} & (num(w_i, w_j) > 0) \\ 0 & (num(w_i, w_j) = 0) \end{cases} \quad (3)$$

For non-Web images, $num(i)$ is defined as the number of images annotated by annotation w_i, and $num(i, j)$ is defined as the number of images annotated by both w_i and w_j.

2.2.2 The RWR Algorithm

The RWR algorithm performs as follows [15]. Assume that there is a random walker that starts from node w_i with a certain probability. At each time-tick, the walker has two choices. One is to randomly choose an available edge to follow. The other choice is to jump to w_j with probability $c \times v(j)$, where v is the restart vector and c is the probability of restarting the random walk [15].

2.2.3 Image Annotation Refinement with RWR

Assume that G is a graph with N vertices w_i ($i \in [1,N]$) constructed as in Section 2.2.1. Let A be the adjacency matrix of G. A is column-normalized to ensure that the sum of each column in A is one. The original confidence scores of candidate annotations are considered as the restart vector v. v is normalized such that the sum of all elements in v is one. The aim is to estimate the steady-state probability of all vertices, which is denoted by u. Let c be the probability of restarting the random walk. It is empirically set to be 0.3 in our implementation. Then the N-by-1 steady state probability vector u satisfies the following equation:

$$u = (1 - c)Au + cv \quad (4)$$

Therefore,

$$u = c(I - (1 - c)A)^{-1} v \quad (5)$$

where I is the $N \times N$ identity matrix.

The ith element $u(i)$ of the steady-state vector u is the probability that w_i can be the final annotation.

There are two methods to decide the final annotations. One way is to choose the top m annotations with highest probabilities. This way will be referred as "Top@m". The other way is to choose all the annotations with probabilities larger than a threshold. We refer to this method as "Threshold".

3. EXPERIMENTS

The proposed algorithm was evaluated on both non-Web and Web images. For non-Web images, the Corel dataset is used and several quantitative results are provided. For Web images, annotation results of some example images randomly selected from photo forum sites are shown.

3.1 Experimental Results of Non-Web Images

The dataset is the same as that of [8]. It was divided into 3 parts: a training set of 4000 images, a validation set of 500 images and a testing set of 500 images [8]. The validation set was used to tune system parameters. After fixing the parameters, the validation set was merged into the training set. In this work, we use the same visual features as [8].

Three methods were considered and compared: the proposed RWR-based method (RWRM), the WordNet-based method (WNM) [9] and the CMRM method (CMRM) [8]. To be fair, both RWRM and WNM used CMRM as the initial annotation algorithm. The JCN measure that has the best performance in [9] was used for WNM.

Precision, recall and F_1 were used as the performance measures. Recall of a word w_i is defined as the number of images correctly annotated with w_i divided by the number of images that have w_i in the ground truth annotation. Precision of w_i is defined as the number of correctly annotated images divided by the total number of images annotated with w_i. F_1 is defined as: 2 × Precision × Recall / (Precision + Recall). All three measures are averaged over the subset of the 49 words with best performance as in [8].

To decide the final annotations, both "Top@m" and "Threshold" strategies as aforementioned in Section 2.2.3 were used and evaluated. Although the size of the candidate annotations affects the final annotation, we have tested it from 5 to 10 and all conclusions are consistent. In this paper it was fixed to 8 due to the space limit.

The precision, recall and F_1 of the "Top@m" results are shown in Figure 1, 2 and 3. Since the candidate annotations of RWRM, CMRM and WNM for each test image are the same, their performances tend to be consistent while m is approaching 8. Although the precision of RWRM is only comparable with that of CMRM, RWRM outperforms CMRM when measured by recall. As a result, the F_1 value of RWRM is higher than that of CMRM. Both RWRM and CMRM consistently outperform WNM. There are two reasons for the poor performance of WNM. One is that it does not utilize the original confidence scores of candidate annotations. The other is that the similarities between annotations only depend on WordNet. There are 49 out of 374 words of the Corel dataset which either do not exist in WordNet lexicon or have zero similarity with all other words using the JCN measure. Moreover, the similarity defined using WordNet is sometimes not appropriate for the annotation refinement problem. For example, "mountain" and "sky" usually appear in a scenery photo together,

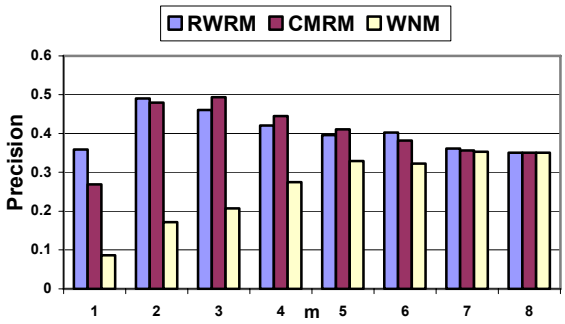

Figure 1. Precision values comparison of Top@m

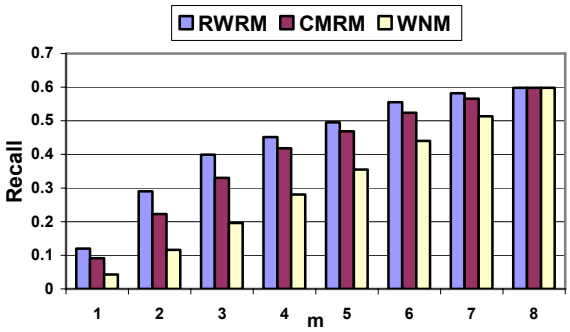

Figure 2. Recall values comparison of Top@m

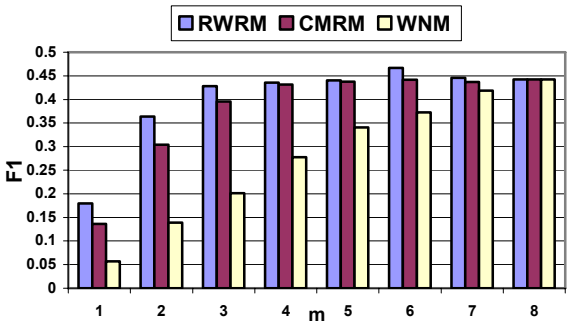

Figure 3. F_1 values comparison of Top@m

Table 1. Performance comparison of "Threshold" results

	Precision	Recall	F_1
CMRM	0.36	0.57	0.44
WNM	0.35	0.56	0.43
RWRM	0.41	0.55	0.47

while "tree" and "flag" seldom simultaneously appear in an image. However, with the JCN measure, the similarities of the above two pairs of words are 0.061 and 0.148 respectively, which is unreasonable. With the proposed "co-occurrence" similarity, the two similarities will be 0.430 and 0.095 respectively, which is more reasonable.

The performances of three algorithms using the "Threshold" strategy are shown in Table 1. The thresholds of WNM, CMRM and RWRM are 0.04, 0.06 and 0.1 which correspond to the best F_1 value on the validation set. Since for different images, the number of final annotations will be different with the "Threshold"

Table 2. Annotation results for five Web images

Image					
Annotations (top 5)	amoer, animal, tiger, safaripark, beekse	eiffel,landscape, tower, paris, night	nature, sky, landscape, sun, moon	beautiful, nature, water, lily, life	nature, landscape, water, art, house
URL	http://www.photosig.com/go/photos/view?id=202162	http://www.photosig.com/go/photos/view?id=68044	http://www.photosig.com/go/photos/view?id=332432	http://www.photosig.com/go/photos/view?id=324661	http://www.photosig.com/go/photos/view?id=19535

strategy, the results of "Threshold" are slightly inconsistent with that of "Top@*m*". Although the recall of RWRM is only comparable with that of CMRM and WNM, the precision of RWRM is much better than that of CMRM (+14%) and WNM (+17%). Therefore, the F_1 value of RWRM is higher than that of CMRM and WNM.

3.2 Experimental Results of Web Images

Due to the difficulty of evaluating the refinement results of Web images quantitatively, illustrative results of several images from photo forum sites, e.g. photosig [2] are used. Images of such sites have rich metadata such as title, category and photographer's comment. In the current implementation, the weights of title, category and comment are simply three, two, and one. An image database is constructed which contains more than two million images from the forum sites. An image search engine named as EnjoyPhoto was built based on the metadata of the images [16]. The co-occurrence similarity is obtained using EnjoyPhoto as the search engine. Table 2 shows the annotation results of five randomly selected images. The results show that the proposed algorithm performs well on Web images.

4. CONCLUSIONS

In this paper, we have presented a novel approach to automatically refining the original imprecise annotations of both Web and non-Web images. On the one hand, for Web images, textual information is used to retrieve a set of candidate annotations. On the other hand, for non-Web images that are lack of textual information, a relevance model-based algorithm using visual information is used to decide the candidate annotations. Then an algorithm based on RWR is proposed to re-rank the candidate annotations, in which both the corpus information and confidence scores of original annotations are leveraged. Experimental results show that the proposed algorithm outperforms an existing annotation refinement algorithm and could effectively improve the original annotation results.

5. REFERENCES

[1] http://images.google.com

[2] http://www.photosig.com

[3] Blei, D. M. and Jordan, M. I. Modeling annotated data. *In Proc. SIGIR*, Toronto, July. 2003.

[4] Chang, E., Kingshy, G., Sychay, G., and Wu, G. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. on CSVT*, 13(1):26– 38, Jan. 2003.

[5] Cusano, C., Ciocca, G., and Schettini, R. Image annotation using SVM. In *Proc. Of Internet imaging IV*, Vol. SPIE, 2004

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

[7] Feng, S. L., Manmatha, R., and Lavrenko, V. Multiple bernoulli relevance models for image and video annotation. In *Proc. of CVPR*, Washington, DC, June, 2004.

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

[9] Jin, Y., Khan, L., Wang, L., and Awad, M. Image Annotations By Combining Multiple Evidence & Wordnet. *Proc. of ACM Multimedia,* Singapore, 2005

[10] Lavrenko, V. and Croft, W. Relevance-based language models. *Proc. of SIGIR*, 2001.

[11] Lavrenko, V., Manmatha, R., and Jeon, J. A Model for Learning the Semantics of Pictures. In *Proc. NIPS*, 2003.

[12] Li, J. and Wang, J. Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(10), Oct. 2003.

[13] Miller, G. A. WordNet: A lexical database for English. *Communication of ACM, 38*, 11 (Nov. 1995), 39–41.

[14] Mori, Y., Takahashi, H., and Oka, R. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM,* 1999.

[15] Page, L., Brin, S., Motwani, R., and Winograd, T. The Pagerank Citation Ranking: Bringing Order to the web. *technical report*, Stanford University, Stanford, CA, 1998.

[16] Zhang, L., Chen, L., Jing, F., Deng, K. F., and Ma, W.Y. EnjoyPhoto—A Vertical Image Search Engine for Enjoying High-Quality Photos. In *ACM multimedia* 2006.