Content-Based Image Annotation Refinement[1]

Changhu Wang
Department of EEIS, University of Science and Technology of China
wch@ustc.edu

Feng Jing*, Lei Zhang*, Hong-Jiang Zhang†
* Microsoft Research Asia
† Microsoft Research Advanced Technology Center
{fengjing,leizhang,hjzhang}@microsoft.com

Abstract

Automatic image annotation has been an active research topic due to its great importance in image retrieval and management. However, results of the state-of-the-art image annotation methods are often unsatisfactory. Despite continuous efforts in inventing new annotation algorithms, it would be advantageous to develop a dedicated approach that could refine imprecise annotations. In this paper, a novel approach to automatically refining the original annotations of images is proposed. For a query image, an existing image annotation method is first employed to obtain a set of candidate annotations. Then, the candidate annotations are re-ranked and only the top ones are reserved as the final annotations. By formulating the annotation refinement process as a Markov process and defining the candidate annotations as the states of a Markov chain, a content-based image annotation refinement (CIAR) algorithm is proposed to re-rank the candidate annotations. It leverages both corpus information and the content feature of a query image. Experimental results on a typical Corel dataset show not only the validity of the refinement, but also the superiority of the proposed algorithm over existing ones.

1. Introduction

As a promising solution to improve many applications, including Web image search and desktop photo management, image annotation has become a core research topic in Content-Based Image Retrieval (CBIR).

Most existing image annotation approaches can be classified into two categories, classification-based methods and probabilistic modeling-based methods. Classification-based methods treat keywords (concepts) as classes and employ trained classifiers to annotate an input image based on classification results. Many representative classifiers have been used, such as the two-dimensional multi-resolution hidden Markov models (2D MHMMs) [17], support vector machine (SVM) [7][10][25], Bayes Point Machine [5], and Mixture Hierarchical Model [3][4].

Probabilistic modeling-based methods attempt to infer the correlations or joint probabilities between images and annotations. As the pioneering work, Mori *et al*. proposed a method for annotating image grids using co-occurrences in 1999 [20]. In [8], Duygulu *et al*. proposed a novel approach that treated image annotation as a machine translation problem. A statistic machine translation model was used to "translate" textual keywords to visual keywords, i.e. image blob tokens obtained by clustering. Another way of capturing co-occurrence information is to introduce latent variables to model hidden concepts in images. The representative work includes Gaussian Mixture Model, Latent Dirichlet Allocation Model (LDA) and correspondence LDA [2]. Inspired by the relevance language models, several relevance models have been proposed recently, including Cross-Media Relevance Model (CMRM) [12], Continuous Relevance Model (CRM) [13][16], and Multiple Bernoulli Relevance Model (MBRM) [9].

Since 2006, motivated by Web search technologies in many commercial systems, several search-based image annotation methods [18][23][24] have been developed, using Web-scale image database and unlimited vocabulary.

Despite the continuous efforts placed on image annotation, results of existing image annotation methods are still unsatisfactory. Alternatively, it would be advantageous if a dedicated approach could refine current annotation results.

Jin *et al*. [14] have done pioneering work on annotation refinement using a generic knowledge-based WordNet [19]. From a small candidate annotation set obtained by an annotation method, irrelevant annotations are pruned using WordNet. The basic assumption is that highly correlated annotations should be reserved and non-correlated annotations should be removed. We will call this assumption "*majority should win*" in the remainder of this paper. For example, Figure 1(a) has the correct keywords "beach, people, sand, desert" and noisy keyword "snow".

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

1-4244-1180-7/07/$25.00 ©2007 IEEE


(a) beach people sand desert snow


(b) beach people sand desert snow

Figure 1: Two examples of image annotations including both correct and noisy keywords.
According to the assumption "*majority should win*", the two images will have the same refinement results without considering the image content.

Since the correlation between "beach" and "sand" is greater than "snow" and "sand" based on WordNet, the noisy keyword "snow" will be discarded. However, 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 4.1.3) decreases compared with the original annotation method. There are at least two main reasons for the unsatisfactory performance. On the one hand, although WordNet contains additional generic knowledge about word relationships, it has two limitations. One is that it is independent of the dataset and therefore does not reflect the characteristics of the specific image dataset. The other limitation is that it cannot deal with the annotations that do not exist in the lexicon of WordNet. On the other hand, once the candidate annotations have been decided, the annotation refinement process is independent of the original query image. The underlying strategy that reserves the majority of correlated candidate annotations will be inappropriate for queries with relatively poor initial annotation results.

Recently, Wang *et al.* [22] proposed a novel annotation refinement algorithm to try to resolve the issues in [14]. In [22], an algorithm using Random Walk with Restarts (RWR) was 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. However, although Wang's work resolved the aforementioned first issue in [14], it was still implicitly based on the assumption *majority should win* and the refinement process was still independent of the original query image. Therefore, the algorithm is sensitive to the size of candidate annotation set. That is, if the size is too small, the refinement process will be less useful; while if the size is too large, there might be other noisy candidate annotations that will let the assumption *majority should win* fail. For example, Figure 1(b) has the correct keyword "people, snow" and noisy keywords "beach, sand, desert". According to the assumption *majority should win* in [14] and [22], correct keyword "snow" will be discarded, while all noisy keywords will be reserved. Therefore, being independent of the content feature of the query image, existing annotation refinement algorithms [14][22] will obtain the same results for Figure 1(a) and (b), in which the refined results of Figure 1(b) will be incorrect.

In this paper, a novel content-based image annotation refinement (CIAR) algorithm is proposed. For a query image, an existing image annotation method is first employed to obtain a set of candidate annotations. Then, the annotations are refined by re-ranking the candidate annotations and reserving the top ones. In this work, the annotation refinement process is formulated as a Markov process and the candidate annotations are defined as the states of a Markov chain. To solve the aforementioned issues in [14] and [22], a query-biased Markov chain (QBMC) with a query-biased transition matrix (QBTM) is dynamically constructed based on the query image, using both content feature of the query image and corpus information. Since QBMC is particularly constructed for the query image using all the corpus information, the refinement results of Figure 1(a) and (b) will be different and more relevant to each image. Therefore, our proposed annotation refinement algorithm not only is free of the first issue in [14], but also resolves the second issue in [14] and [22] to some extent.

The rest of the paper is organized as follows. Section 2 presents Markov process formulation of annotation refinement process. Section 3 describes the content-based image annotation refinement algorithm in detail. Extensive experimental results are shown in Section 4. We conclude in Section 5.

2. Formulating refinement process as a Markov process

This section provides some insights into existing image annotation refinement algorithms by re-formulating them as Markov processes, and describes the proposed content-based image annotation refinement idea.

2.1. Image annotation refinement problem

Assume that the original query image is I_q, and there are totally N^+ keywords in the lexicon space Ω. Also assume that N candidate annotations $\{w_i | w_i \in \Omega, i=1,\ldots,N\}$ out of N^+ could be obtained using a certain annotation algorithm. From the statistical view, the aim of image annotation refinement is to refine the conditional probability $p(w_i|I_q)$ so that more accurate annotations will have higher probabilities. As a result, the annotations with highest probabilities could be reserved as the final annotations.

2.2. Insights into existing refinement algorithms

Let us define the candidate annotations $\{w_i | w_i \in \Omega, i=1,\ldots,N\}$ as the states of a first-order homogeneous Markov chain. Let $p(w_i|w_j)$ be the one-step transition

probability from w_j to w_i, and let P be the transition matrix of this Markov chain. Then we can explain the previous annotation refinement works [14][22] in terms of Markov random process with different kinds of transition probabilities.

Let us first look at the WordNet-based method (WNM) [14]. Its basic assumption is that highly correlated annotations should be reserved and non-correlated annotations should be removed, which can be described as follows:

$$p(w_i \mid I_q) \triangleq \alpha \sum_{j=1}^{N} p(w_i \mid w_j) \tag{1}$$

where $p(w_i|w_j)$ is calculated by the semantic similarity of w_i and w_j in WordNet. α is a normalization constant.

From Equation 1 we can see that, in WNM, $p(w_i|I_q)$ merely reflects the assumption *majority should win*. The annotation probability $p(w_i|I_q)$ of a state w_i is independent of annotation probabilities of other states. Moreover, the right hand side of Equation 1 is independent of query image I_q, which makes the content feature of I_q ignored in the whole refinement process.

In the Random Walk with Restarts Model (RWRM) [22], Wang *et al*. utilize the corpus or training set information to calculate $p(w_i|w_j)$. The description equation of RWRM can be given as follows:

$$p^{(t+1)}(w_i \mid I_q) = (1-c)\sum_{j=1}^{N} p(w_i \mid w_j) \cdot p^{(t)}(w_j \mid I_q) + c p_v(w_i) \tag{2}$$

where $p_v(w_i)$ is the confidence information of w_i obtained once the candidate annotations are identified. c is a constant that represents the probability of restarting the random walk process. $p^{(t)}(w_i|I_q)$ is the probability that the annotation is in the state w_i at time t. $p(w_i|I_q)$ is defined as the stationary probability of state w_i in this Markov chain, given by the following formula:

$$p(w_i \mid I_q) = \lim_{t \to \infty} p^{(t)}(w_i \mid I_q) \tag{3}$$

Although RWRM tries to utilize the confidence information of the initial annotation method in the refinement process, the refinement process is still independent of I_q and is also based on the assumption *majority should win*, since $p_v(w_i)$ has been fixed before refining. Therefore, the performance of RWRM is mainly decided by two factors: the assumption *majority should win* and the performance of initial annotation method. On the one hand, because of the assumption *majority should win*, the algorithm is sensitive to the size of candidate annotation set. That is, if the size is too small, the refinement process will be less useful; while if the size is too large, there might be other noisy candidate annotations that will let the assumption *majority should win* fail. On the other hand, the algorithm is restricted by the performance of initial annotation method.

2.3. The idea to refine image annotation based on content analysis

To fully utilize the correlations between different labels, we assume that each label in lexicon is related with other labels given a query image, described as follows:

$$\begin{aligned} p(w_i \mid I_q) &= \sum_{j=1}^{N^+} p(w_i, w_j \mid I_q) \\ &= \sum_{j=1}^{N^+} p(w_i \mid w_j, I_q) \cdot p(w_j \mid I_q) \quad (i \in [1, N^+]) \end{aligned} \tag{4}$$

with constrains:

$$\sum_{i=1}^{N^+} p(w_i \mid I_q) = 1 \tag{5}$$

and

$$\sum_{i=1}^{N^+} p(w_i \mid w_j, I_q) = 1 \quad (j \in [1, N^+]) \tag{6}$$

where $p(w_i|w_j, I_q)$ is the probability of w_i being an annotation of I_q given that w_j is already an annotation of I_q.

Let us define $\{w_i \mid w_i \in \Omega,\ i=1,\ldots,N^+\}$ as the states of a first-order homogeneous Markov chain. Equation 4 is just the transition formulation of this Markov chain, and $p(w_i|w_j, I_q)$ corresponds to the one-step transition probability from w_j to w_i.

Using a certain annotation algorithm, N candidate annotations $\{w_i \mid w_i \in \Omega,\ i=1,\ldots,N\}$ out of N^+ could be obtained first. By merging all other N^+-N labels w_i ($i=N+1,\ldots,N^+$) to be one state in the Markov chain, denoted as w_{N+1}, the Markov chain with N^+ states can be simplified to be a chain with $N+1$ states, described as follows:

$$p(w_i \mid I_q) = \sum_{j=1}^{N+1} p(w_i \mid w_j, I_q) \cdot p(w_j \mid I_q) \quad (i \in [1, N+1]) \tag{7}$$

In Equation 7, for $i, j \in [1,N]$, $p(w_i|w_j, I_q)$ represents the probability of w_i being an annotation of I_q when w_j is already an annotation of I_q. For $j \in [1,N]$, $p(w_{N+1}|w_j, I_q)$ is the probability of I_q annotated by other labels not in the candidate set, when w_j is already an annotation of I_q.

Let us compare Equation 7 with Equation 1 and 2. One significant difference is that the transition probability from w_j to w_i is denoted as $p(w_i|w_j, I_q)$, instead of $p(w_i|w_j)$. Therefore, we can see that the content feature of query image is important information for the refinement process, which is inappropriate to be ignored. The transition matrix P will be named as query-biased transition matrix (QBTM). The Markov chain described in Equation 4 and 7 will be named as query-biased Markov chain (QBMC).

Therefore, we propose to utilize the constructed query-biased Markov chain (QBMC) to solve the image annotation refinement problem. Different from existing algorithms, QBMC is constructed based on both candidate annotations and content features of the query image. That is, transition probabilities between any states are influenced by the query image I_q, and the constructed Markov chains are

different for different query images, even for the same candidate annotation set.

In this Markov model based formulation, the conditional probability of w_i is decided by all labels' conditional probabilities $p(w_j|I_q)$, which are all needed to be refined further. Therefore, we can rewrite Equation 7 as an iterative process as follows:

$$p^{(t+1)}(w_i \mid I_q) = \sum_{j=1}^{N+1} p(w_i \mid w_j, I_q) \cdot p^{(t)}(w_j \mid I_q) \quad (i \in [1, N+1]) \quad (8)$$

Based on this formulation, a content-based image annotation refinement (CIAR) algorithm is proposed and will be described in detail in the next section.

3. Content-based image annotation refinement algorithm

Given a query image, content-based image annotation starts from a set of candidate annotations obtained by an existing image annotation algorithm. Then, the query-biased transition probability matrix (QBTM) is constructed for the query image using both the content feature of the query image and the corpus information. Thereafter, the content-based image annotation refinement algorithm (CIAR) is used to re-rank the candidate annotations. Finally, the top ranked annotations will be reserved as the final annotations.

3.1. Candidate annotation identification

Most existing image annotation algorithms can serve for identifying a set of candidate annotations for the query image. The Cross-Media Relevance Model (CMRM) [12], which was also used in [22], is chosen because of its effectiveness.

CMRM is based on the relevance model [15] 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. For details of CMRM please refer to [12].

3.2. Query-biased transition matrix (QBTM)

We first design a similarity measure between annotations named as "query-biased co-occurrence similarity" (QBCS). Then, we construct the aforementioned query-biased transition matrix (QBTM) from QBCS.

Let $sim(w_i,w_j)$ be the QBCS between two candidate annotations w_i and w_j ($i,j \in [1,N]$). Besides counting the co-occurrence number of the two annotations, we also utilize the similarity information between the query image and the images annotated by both of w_i and w_j in training set. $sim(w_i,w_j)$ can be calculated by the following formula:

$$sim(w_i, w_j) = \sum_{J \in T_{ij}} sim(J, I_q) \quad (9)$$

where $sim(J,I_q)$ denotes the visual similarity between the query image I_q and a training image J. T_{ij} is the image set annotated by both w_i and w_j in training set. There are several ways to compute $sim(J,I_q)$. An image J can be represented by a global feature vector g or l region-based feature vectors r_i ($i \in [1,l]$). For global feature, $sim(J,I_q)$ can be obtained as follows:

$$sim(J, I_q) = \frac{1}{\sqrt{2^k \pi^k \mid \Sigma \mid}} \exp\{-(g - g_q)^T \Sigma^{-1} (g - g_q)\} \quad (10)$$

where g and g_q are the global feature vectors of J and I_q, k is the dimension of feature vector, and $\sum$ is a diagonal matrix with the diagonal elements being the variances of each feature dimension in the training set.

For region-based features, motivated by the work of [16], we use the following formula to calculate the similarity between two images:

$$sim(J, I_q) = \prod_{a=1}^{n_A} \frac{1}{n_B} \sum_{b=1}^{n_B} \frac{1}{\sqrt{2^k \pi^k \mid \Sigma \mid}} \exp\{-(r_a - r_b)^T \Sigma^{-1} (r_a - r_b)\} \quad (11)$$

where n_A and n_B are the numbers of regions of I_q and J. r_a and r_b are the a^{th} region feature vector of I_q and the b^{th} region feature vector of J.

Since all the candidate annotations are identified from the training set, for a candidate annotation w_i, there will always be some images annotated with it. Hence, $sim(w_i,w_i)$ and $\sum_{i=1}^{N} sim(w_i, w_j)$ are both positive.

Based on QBCS, the $(N+1)\times(N+1)$ query-biased transition matrix (QBTM) P is defined as follows:

$$P_{ji} = p(w_i \mid w_j, I_q) = \begin{cases} \dfrac{sim(w_i, w_j)}{\beta} & (i, j \in [1, N]) \\ 1 - \dfrac{\sum_{k=1}^{N} sim(w_k, w_j)}{\beta} & (i = N+1, j \in [1, N]) \\ \dfrac{1}{N+1} & (j = N+1) \end{cases} \quad (12)$$

where β is given as follows:

$$\beta = \max_{j \in [1,N]} \left(\sum_{k=1}^{N} sim(w_k, w_j) \right) + \varepsilon \quad (13)$$

where ε is a constant satisfying $0<\varepsilon<<1$.

We have the following lemmas about the proposed QBTM P:

Lemma 1: *The query-biased transition matrix (QBTM) P is non-negative and row-stochastic.*

Proof: (Please refer to the appendix.)

Therefore, P meets the requirements of a transition probability matrix.

Lemma 2: *The query-biased transition matrix (QBTM) P is aperiodic and irreducible.*

Proof: (Please refer to the appendix.)

3.3. Content-based image annotation refinement algorithm (CIAR)

Let $r^{(t)}$ be the probability vector of all N+1 states w_i ($i \in [1,N+1]$) at time t. It is an $(N+1)\times 1$ vector with its ith element equal to $p^{(t)}(w_i|I_q)$. Therefore, we can rewrite Equation 8 as follows:

$$r^{(t+1)} = P^T r^{(t)} \quad (14)$$

Our goal is to calculate the stationary probability vector r^S of the query-biased Markov chain, in which the ith element denotes the probability of annotating I_q with w_i: $p(w_i|I_q)$. Therefore, we need to prove that the iteration of Equation 14 can converge to one and only r^S.

Before proving this, let us recall some basic facts about Markov process [6].

1) An irreducible, aperiodic, finite, and homogeneous Markov chain is ergodic.

2) An ergodic Markov chain has one and only stationary probability vector, which is just the principle eigenvector of its transition matrix.

Based on above facts of Markov chains and lemmas aforementioned in Section 3.2, we have the following theorem:

Theorem 1: *The proposed query-biased Markov chain with query-biased transition matrix P has one and only stationary probability vector r^S. And, r^S is the principle eigenvector of P.*

Proof: (Please refer to the appendix.)

Based on Theorem 1, we can obtain r^S by calculating the principle eigenvector of P. When the number of states N is large, we can also use power method [11] to calculate r^S. The ith element of r^S is just $p(w_i|I_q)$, the probability that w_i is the final annotation of I_q.

Once $p(w_i|I_q)$ ($i \in [1,N]$) is obtained, the top m annotations with highest probabilities can be chosen as the final annotations.

4. Experimental results

4.1. Experimental design

To evaluate the proposed algorithm, three annotation refinement methods were compared: WordNet-based method (WNM) [14], Random Walk with Restarts model (RWRM) [22], and the proposed content-based image annotation refinement (CIAR) algorithm. Since the three image annotation refinement algorithms are all based on an existing image annotation algorithm, to be fair, all of them use Cross-Media Relevance Model (CMRM) [12] as the initial annotation algorithm.

4.1.1 Data set

The proposed algorithms were evaluated on Corel dataset from [8]. Corel dataset is a basic comparative dataset for recent research work in image annotation. There are 5,000 images from 50 Stock Photo CDs in this dataset. Each CD includes 100 images on the same topic. Segmentation using normalized cuts [21] followed by quantization ensures that there are 1 to 10 blobs for each image. Each image is annotated with 1 to 5 words and there are 374 words and 500 blobs in the dataset. Details of the above process are described in [8].

4.1.2 Parameter selection

We divided the dataset into 3 parts – with 4,000 training set images, 500 evaluation set images and 500 testing set images. The evaluation set is used to find optimal system parameters. After fixing the parameters, we merged the 4,000 training set images and 500 validation set images to make a new training set. This corresponds to the training set of 4,500 images and the test set of 500 images used in [8]. We use the same parameter settings as [12] when we implement the CMRM algorithm. There are several similarity measures compared in WNM [14] and the JCN measure is proved to be the best one. The hybrid of these measures is orthogonal to our framework and we choose the JCN measure when we implement WNM method. In RWRM, the restart parameter c is empirically set to be 0.3 according to [22].

4.1.3 Evaluation measures

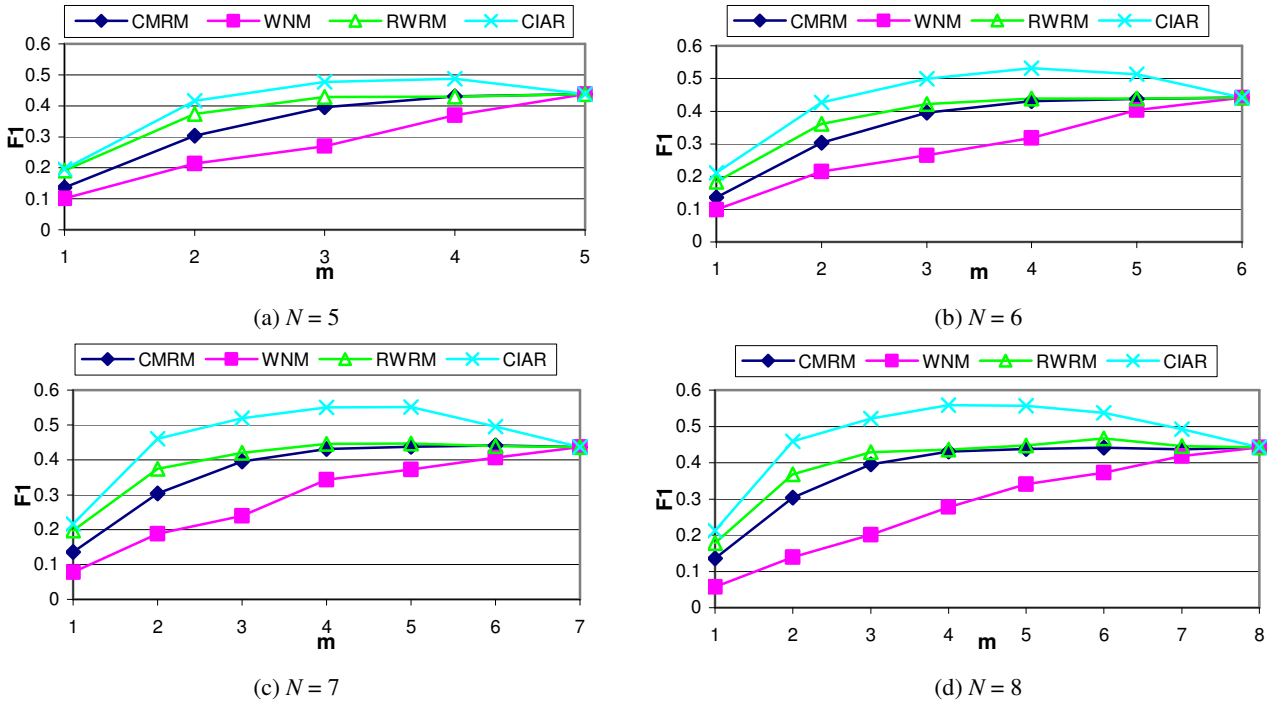
F_1 value was used as the performance measure. It is defined as: 2×Precision×Recall / (Precision+Recall). Recall of a word w_i is defined as the number of images correctly annotated with w_i divided by the number of images that have w_i in the ground truth annotation. Precision of w_i is defined as the number of correctly annotated images divided by the total number of images annotated with w_i. F_1 value is averaged over the subset of the 49 words with best performance as in [14] and [22].

4.2. Experimental results and analysis

Figure 2 shows the comparison results of different algorithms with the size of the candidate annotations N and the number of the final annotations m both evaluated. With N fixed in each sub-image in Figure 2, results of different m (from 1 to N) are shown.

Since the candidate annotations of WNW, RWRM, CIAR and CMRM for each test image are alike, their performances tend to be consistent when m is approaching to N. Several conclusions could be drawn from Figure 2.

First, WNW performs worst among the three annotation refinement algorithms, and has apparently worse performance than the original annotation algorithm CMRM, which is consistent with the results of [22]. There are mainly two reasons for the poor performance of WNM. The first reason is that the similarities between annotations only depend on WordNet, which may be not proper for image annotation refinement problem. There are 49 out of 374 words of the Corel dataset that either do not exist in



(a) $N = 5$ (b) $N = 6$

(c) $N = 7$ (d) $N = 8$

Figure 2: Performance comparison of different algorithms

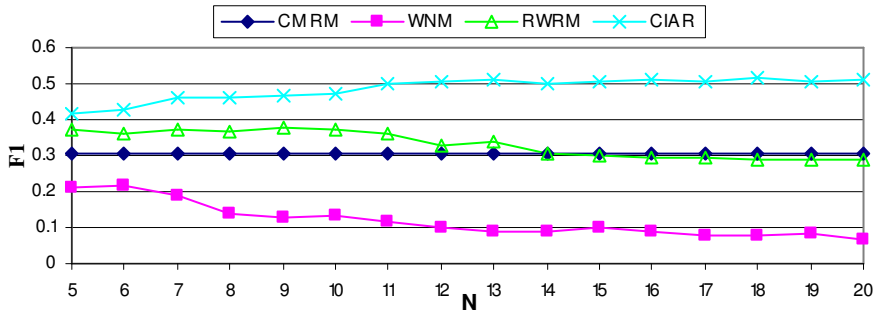


Figure 3: Performance comparison of different algorithms with m fixed to be 2

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, 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. The second reason is due to the assumption *majority should win*. It is risky or misleading when the majority of the candidate annotations are noisy words. Therefore, by leveraging the information of corpus or training set, both RWRM and CIAR consistently outperform WNM to large extent.

Second, although RWRM greatly outperform WNM, it is only slightly better than CMRM. The main limitation of RWRM is its independence of the original query image and the content information of training set. The ignorance of content information will result in the same problem as WNM, when the majority of the candidate annotations are noisy words.

Third, CIAR greatly outperforms other algorithms. The formulation of Markov process for annotation refinement and the dependency of content information of the original query image guarantee its effectiveness in the image annotation refinement problem.

Finally, unlike the curves of other algorithms, which reach their peaks when m reaches N, the curve of CIAR reaches its peak when m is 4 for each N from 5 to 8. It is consistent with the fact that the average ground truth annotations in the testing set is just about 4. It is also a reflection of the properness of CIAR.

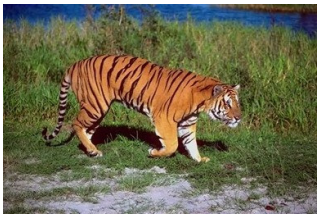
Images			
Image ID	108019	109012	163068
CMRM Annotations	grass, albatross, wings, cat	water, sky, tree, people	tree, water, sky, grass
CIAR Annotations (N=10, m=4)	grass, cat, tiger, forest	tree, water, people, snow	birds, tree, grass, water

Figure 4: Sample annotations before and after refinement using CIAR.

Figure 3 shows the results in a different view, in which N is changed along the horizontal with m fixed to be 2. Notice that if m is fixed, the performance of CMRM is independent of N, since CMRM is the original annotation method without further refinement. From Figure 3, another two conclusions could be drawn.

First, the performances of both WNM and RWRM are decreasing as N is increased. The main reason is that both of the two algorithms are based on the assumption *majority should win*. Thus, both of WNM and RWRM are sensitive to the size of candidate annotation set. The larger N is, the noisier the words may be. These two algorithms only work well when majority annotations are accurate, which is difficult to be achieved since the existing annotations are far from perfect.

Second, an increasing trend can be seen on the curve of CIAR method, as N is becoming larger. This is an illustration that CIAR method is not based on the assumption *majority should win*, and the increase of candidate annotation size cannot deteriorate the performance of CIAR, but provide more chances to search for related annotations.

In Figure 4, we provide sample annotations before and after refinement using the proposed CIAR algorithm. They show that CIAR algorithm can propagate more accurate annotation results compared with the original annotation method.

5. Conclusions

In this paper, we have presented a novel approach to automatically refining the original imprecise annotations of a query image. First, an existing image annotation method is employed to retrieve a set of candidate annotations. Then, the candidate annotations are re-ranked and only the top ones are reserved as the final annotations. By formulating image annotation refinement process as a Markov process and defining the candidate annotations as the states of a Markov chain, a content-based image annotation refinement (CIAR) algorithm is proposed to re-rank the candidate annotations, leveraging both the corpus information and the content feature of the query image. Experimental results show not only the validity of the refinement, but also the superiority of the proposed algorithm over existing ones.

Appendix A

In the appendix, we will prove Theorem 1 mentioned in Section 3.3. The proof of Theorem 1 will be given, after the proofs of 3 lemmas.

Lemma 1: *The query-biased transition matrix (QBTM) P is non-negative and row-stochastic.*

Proof: Based on Equation 9 and 12, we know that each element in matrix P is non-negative, and sum of each row of matrix P is 1. That means P is a non-negative and row-stochastic matrix.

Lemma 2: *The query-biased transition matrix (QBTM) P is aperiodic and irreducible.*

Proof: Since all the diagonal elements of matrix P are positive (see Section 3.2), according to the periodicity theory of Markov chain, the periodicity of each state is 1. That is, all the states are aperiodic. Therefore, the query-biased transition matrix P is aperiodic.

From Equation 7, we can see that, besides the first N states representing N candidate labels, we add the (N+1)-th state to represent "other possible labels". Equation 12 and 13 show that all the transition probabilities between the first N states and the (N+1)-th state are positive. Therefore, all the N+1 states in the proposed Markov chain are connected, which means that the transition matrix P is irreducible.

Lemma 3: *If the transition matrix P of a finite and homogeneous Markov chain MC is aperiodic, and irreducible, then iterative calculation $r^{(t+1)}=P^T r^{(t)}$ converge to the principle eigenvector of P. (Assume $r^{(0)}$ is a positive and normalized vector.)*

Proof: Lemma 3 can be proved based on the following

basic facts about Markov process [6]:

1) An irreducible, aperiodic, finite and homogeneous Markov chain is ergodic.

2) An ergodic Markov chain has one and only stationary probability vector, which is just the principle eigenvector of transition matrix.

Theorem 1: *The proposed query-biased Markov chain with query-biased transition matrix P has one and only stationary probability vector* r^S. *And,* r^S *is the principle eigenvector of P.*

Proof: According to lemma 1 and 2, P is an aperiodic and irreducible transition probability matrix. According to lemma 3, we know that the iterative calculation $r=P^T r$ converge to the principle eigenvector of P, which completes the proof.

References

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

[2] D. Blei and M. I. Jordan. Modeling annotated data. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Toronto, Canada, July 28 - August 01, 2003).

[3] G. Carneiro and N. Vasconcelos. A database centric view of semantic image annotation and retrieval. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Salvador, Brazil, August 15 - 19, 2005).

[4] G. Carneiro and N. Vasconcelos. Formulating Semantic Image Annotation as a Supervised Learning Problem. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cvpr'05) - Volume 2 - Volume 02 (June 20 - 26, 2005).

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

[6] E. Cinlar. Introduction to Stochastic Processes. Prentice Hall, Inc. 1975.

[7] C. Cusano, G. Ciocca, and R. Schettini. Image Annotation using SVM. In Proceedings of Internet imaging IV, Vol. SPIE 5304. 2004.

[8] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In Proceedings of the 7th European Conference on Computer Vision-Part IV (May 28 - 31, 2002).

[9] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, volume 2, pages II--1002--II--1009, 2004.

[10] Y. Gao, J. Fan, H. Luo, X. Xue, and R. Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers. In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006).

[11] G. H. Golub and C. F. Van Loan. Matrix Computations. The Johns Hopkins University Press, 1996.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in informaion Retrieval (Toronto, Canada, July 28 - August 01, 2003).

[13] J. Jeon and R. Manmatha. Automatic Image Annotation of News Images with Large Vocabularies and Low Quality Training Data. In Proceedings of the 12th Annual ACM international Conference on Multimedia, 2004.

[14] Y. Jin, L. Khan, L. Wang and M. Awad. Image annotations by combining multiple evidence & wordNet. In Proceedings of the 13th Annual ACM international Conference on Multimedia (Hilton, Singapore, November 06 - 11, 2005).

[15] V. Lavrenko and W. Croft. Relevance-based language models. In Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (New Orleans, Louisiana, United States, 2001).

[16] V. Lavrenko, R. Manmatha and J. Jeon. A model for learning the semantics of pictures. In Advances in Neural Information Processing Systems (NIPS'03), 2003.

[17] J. Li and J. Z. Wang. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. IEEE Trans. Pattern Anal. Mach. Intell. 25, 9 (Sep. 2003), 1075-1088.

[18] X. Li, L. Chen, L. Zhang, F. Lin, and W. Y. Ma. Image Annotation by Large-Scale Content-based Image Retrieval. In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006).

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

[20] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[21] J. Shi and J. Malik. Normalised cuts and image segmentation. In IEEE Conf. on Computer Vision and Pattern Recognition, pages 731-737, 1997.

[22] C. Wang, F. Jing, L. Zhang, and H. J. Zhang. Image Annotation Refinement using Random Walk with Restarts. In Proceedings of the 14th Annual ACM international Conference on Multimedia (Santa Barbara, CA, USA, October 23 - 27, 2006).

[23] C. Wang, F. Jing, L. Zhang, and H. J. Zhang. Scalable Search-Based Image Annotation of Personal Images. In Proceedings of the 8th ACM international Workshop on Multimedia information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006).

[24] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma. AnnoSearch: Image Auto-Annotation by Search. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (June 17 - 22, 2006).

[25] C. Yang, M. Dong, and J. Hua. Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (June 17 - 22, 2006).