Diversifying the Image Retrieval Results

Kai Song[1], Yonghong Tian[1], Tiejun Huang[2], Wen Gao[1,2]
[1]Institute of Computing Technology, Chinese Academy of Science, Beijing, 100080, China
[2]School of Electronics Engineering and Computer Science, Peking University
{ksong, yhtian, tjhuang, wgao}@jdl.ac.cn

ABSTRACT
In the area of image retrieval, post-retrieval processing is often used to refine the retrieval results to better satisfy users' requirements. Previous methods mainly focus on presenting users with relevant results. However, in most cases, users cannot clearly present their requirements by several query words. Therefore, relevant results with rich topic coverage are more likely to meet users' ambiguous needs. In this paper, a re-ranking method based on topic richness analysis is proposed to enrich topic coverage in retrieval results. Furthermore, a quantitative criterion called *diversity scores* (DS) is proposed to evaluate the improvement. Given a set of images, topics that are rarely included in the set are scarce topics, as oppose to rich topics that are widely distributed among the set. Scarce topics contribute more than rich topics do to the DS of images. Five researchers are invited to evaluate the re-ranked results both in topic coverage and relevance. Experimental results on over 20,000 images demonstrate that our proposed approach is effective in improving the topic coverage of retrieval results without loss of relevance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, search process*;

General Terms
Algorithms, Measurement, Performance,

Keywords: Image retrieval, topic coverage, diversity score, topic richness, re-rank

1. INTRODUCTION
The area of image retrieval has been extensively studied in recent years [10]. Lots of techniques including relevant feedback [6][8] and cross-modal correlation analysis [3][5][9][14] are introduced to bridge the semantic-gap between low-level visual features and high-level semantic features of images. A comprehensive strategy is presented to evaluate image retrieval algorithms [11].

Nevertheless, most of the current image search engines tend to provide users a list of retrieval results with respect to the relevance score of each image to the query. This scheme is effective when users' needs are clear and they mainly concern the precision and recall in the results. Nevertheless, in most cases, users cannot describe their requests only by several query words accurately. Therefore their actual requirements are ambiguous. For example, the top retrieval results are often dominated by a set of closely related images on some specific topics and users are often stuck in the situations: the topic coverage of retrieval results is too limited to meet the various needs of the users. Such frustrated retrieval experiences are nightmares of end-users.

It is a sagacious alternative to present users a set of results containing various topics related to the queries. As reported in [2], most people said they preferred the retrieval results with broad and interesting topics. Previous works on diversifying document retrieval results has showed great promise in web search engines [13] and product recommendation [15]. In the literature of image retrieval, several approaches have been proposed to achieve such target. Goh *et al* [4] exploit a SVM-based active learning algorithm, which incorporates diversity [1] for image retrieval. A two-scale image retrieval scheme using meta-information feedback is another attempt [7]. The early methods mainly focus on low-level visual feature similarities, either by clustering and picking up the top results in each cluster or calculating the angle of two visual feature vectors. Nonetheless, it seems that the previous methods have their intrinsic drawbacks due to the semantic gap. Furthermore, none of the works propose a *quantitative criterion* to measure topic coverage of image retrieval results. Thus, it is difficult to perform the evaluation and improve the retrieval results.

All these investigations motivate us to develop our work that contains the following two aspects: (1) An effective re-ranking method to improve the topic coverage of retrieval results, especially in top retrieval results; (2) a quantitative criterion called *diversity score* (DS) to evaluate the topic coverage of image retrieval results. The proposed re-ranking approach is based on *topic richness* (TR) scores of images, which quantitatively measure the contributions made by each image to the improvement of topic coverage of retrieval results. The score is computed by analyzing the degree of mutual topic coverage between an image pair. In addition, diversity score is used to evaluate topic coverage of retrieval results, which is calculated by the following intuition: the more scarce topics an images contains, the higher score it obtains. Five researchers are invited to evaluate the re-ranked results both in topic coverage and relevance. The experimental results demonstrate that our proposed approach is effective in evaluating the topic coverage of image retrieval results and the re-ranking method outperforms the clustering method in improving the topic coverage of retrieval results significantly without the loss of relevance.

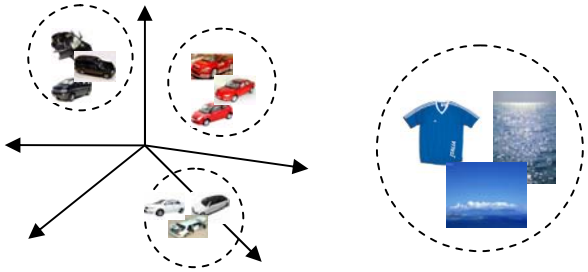
The rest of the paper is organized as follows. In Section 2, we briefly discuss the intrinsic drawbacks of clustering techniques in improving topic coverage of image retrieval results. In Section 3,

MM'06, October 23?7, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

we introduce our proposed re-ranking method as well as our evaluation strategy in detail. Experiments and evaluations are reported in Section 4. Conclusion and discussions are made in Section 5.

2. PROBLEMS OF CLUSTERING AND PICKING STRTEGY

Intuitively, we can enrich the topic coverage of image retrieval results by clustering the images with respect to the low-level feature similarities and picking the top results in each cluster. This paradigm is based on the hypothesis that topic-related images tend to be closely associated together in visual feature space. Nevertheless, most topic-related images are scattered in visual space and can hardly clustered together. Take the images of cars with different colors for instance (Figure 1(a)), they are scattered in different visual clusters. Furthermore, images in the same cluster may contain different topics (Figure 1(b)), the cluster of the color blue may contain sky, sea, or the uniform of Italian soccer team. Therefore, only choosing the top images in each cluster will result in enormous loss in topic coverage.



(a) (b)

Figure 1. (a) Image of cars in different colors are scattered in different visual clusters; (b) cluster of blue color contains different topics.

3. TOPIC COVERAGE IMPROVMENT

Our proposed re-ranking approach is based on *topic richness* (TR) scores of images, which quantitatively measure the contributions made by each image to the improvement of topic coverage of retrieval results. Topic richness score is computed by analyzing the degree of mutual topic coverage between an image pair. In addition, *diversity score* (DS) is used to evaluate topic coverage of retrieval results, which is calculated by the following intuition: in a set of images, topics that are rarely included in the set are scarce topics, as oppose to rich topics that are widely distributed among the set. The more scarce topics an image contains, the more important it is. Therefore, the higher score it obtains.

3.1 Topic Richness Analysis

We use the image set $I=\{i_1, i_2, \ldots, i_n\}$ to denote the retrieval results generated by a certain image search engine. We assume that all the images in I are annotated by several words. Since annotation is not the main concern in this paper, we investigate the state of art of annotation and topic discovery in [3][5][9][12]. The method in [9] requires no parameter tuning, no clustering, no user-determined constant and has been proved to be effective in automatic image annotation. Therefore, we adopt it as our annotation strategy and assume it is effective and accurate. Let $W_k (1 \le k \le n)$ denote the annotation set of image $i_k (1 \le k \le n)$. Each different word in annotations is considered as a different topic. The similarity between an image pair i_j and i_k is defined as:

$$m_{jk} = sim(i_j, i_k) = \begin{cases} \dfrac{|W_k \cap W_j|}{|W_k|} & k \neq j \\ 0 & k = j \end{cases} \quad (1)$$

Where $|\cdot|$ denotes the set cardinality. Note that the similarity defined here is asymmetric, which reflects the asymmetry of topic coverage relations between an image pair. If $m_{jk} \neq 0$, image i_j, i_k are considered as neighbors.

We use *topic richness* score $TR(i_k)$ to denote the richness of topics contained in image i_k with respect to the entire image set I. Our strategy for $TR(i_k)$ score computation lies in the intuition that the higher score an image's neighbor obtains, the higher score it does, which is similar to the link analysis method in PangRank. Therefore, $TR(i_k)$ can be deduced from those of other images that have related topics with it and it is calculated in a recursive way:

$$TR(i_k) = \sum_{j=1, j \neq i}^{n} m_{kj} TR(i_j) \quad (2)$$

m_{jk} is used as the weight of TR score computation because it quantitatively characterizes the topic coverage relations between an image pair. Let $\vec{\omega} = [TR(i_1), TR(i_2), \cdots, TR(i_n)]_{n \times 1}$, matrix $M=(m_{jk})_{n \times n}$ and let it be column-normalized. (2) can be written in a matrix form:

$$\vec{\omega} = M\vec{\omega} \quad (3)$$

Since topic-overlapping does not exist in every image pair, it is possible that the matrix M has all-zero rows. This will cause failure in eigenvector computation. Similar to the random jumping factor in PageRank, a dumping factor c is introduced to overcome the problem:

$$\vec{\omega} = cM\vec{\omega} + \frac{1-c}{n}\vec{e} \quad (4)$$

Where $\vec{e}$ is a column vector with all its n elements equaling to 1. By solving equation (4), we obtain TR score of every image in I.

3.2 The Re-ranking Method

Generally, given a retrieval result in terms of the relevance with the query, images with high TR score are chosen to present in the top retrieval results so that the image set of top results are various in topics. Note that topics that are contained in the chosen images become less important in enriching topic coverage in further steps. Based on the intuition, our re-ranking strategy is to decrease the scores of the images whose topic-related images have already been chosen in the top results.

Step 0. *Initialize set $A = \Phi$, which denotes the retrieval set after re-ranking, set $I = \{i_1, i_2, \cdots, i_n\}$, which denotes the retrieval results generated by a certain search engine,*

Step 1. *Sort all the elements in set I by their TR score in descending order.*

Step 2. *Put the image i_k with the highest TR score from set I to set A. For $j \neq k$, re-calculate the TR score in the following way: $TR(i_j) = TR(i_j) - m_{jk} \cdot TR(i_k)$*

Step 3. Re-sort the images in set I by the updated TR in descending order.

Step 4. Go to Step 2 until top N retrieval results are chosen.

In Step 2, we impose a penalty algorithm to the images that are topic-related with image i_k. The more an image is topic-related with image i_k, the more penalties it obtains.

3.3 Topic Coverage Evaluation

We use diversity score (DS) to measure topic coverage of images retrieval results. The way we compute DS lies in the intuition that images that cover the rare topics of a certain set deserve higher DS. The formal definition of diversity score is given as follows:

Diversity Score: Given a set of images I = {i_1, i_2, ... i_n}, i_k(I≤ k ≤ n) is annotated with m_k different words. We use $DS_I(i_k)$ to denote diversity score of image i_k in set I. It can be calculated as:

$$DS_I(i_k) = \frac{1}{m_k}\sum_{j=1}^{m_k}\frac{1}{N_{t_j}^I} \quad (5)$$

Where $N_{t_j}^I$ denotes the number of images in set *I* that contain topic t_j. Accordingly, diversity score of image set *I* is given as follows:

$$DS(I) = \frac{1}{n}\sum_{j=1}^{n} DS_I(i_k) \quad (6)$$

$DS(I)$ represents the average diversity score of *n* images and is used to evaluate topic coverage of the top *n* retrieval results. The validity of DS in evaluating topic coverage lies in the fact that images with high DS tend to cover scarce topics that can significantly improve topic coverage of retrieval results.

4. EXPERIMENTS

Our experiments are based on 718 annotation words and over 20,000 illustrations extracted from digital books in China-American Digital Academic Library (CADAL) project. For the illustration dataset, our specific consideration lies in that illustration is a typical collection covers repetitive topics because of the cross-topics in digital books. Therefore, our proposed approach is more likely to show the improvements in diversifying the retrieval results compared with other methods.

In our experiments, we choose 20 queries and the top 50 retrieval results of each query are passed to our approach and k-means algorithm separately to re-rank top 20 results that often draw most attentions of users. For k-means algorithm, we set *k*=20 and pick up the top 1 result of each cluster to generate the top 20 results. We compare the re-ranked retrieval results generated by our method with those by the clustering technique (e.g. k-means).

In order to evaluate the effectiveness of the proposed approach, five researchers in the area of image retrieval are invited to evaluate topic coverage in the re-ranked results. They are asked to count the number of topics in the re-ranked retrieval results (both by our method and k-means) of each query. Since the criteria of topic classifying vary from user to user, results are evaluated in the form of relative changes, which is defined as follows:

$$\Delta = \frac{1}{N}\sum_{i=1}^{N}(T_i^{TR} - T_i^{k-means}) \quad (7)$$

Where *N* is the number of users and *N*=5 in our experiments. T_i^{TR} denotes the number of topics in the results re-ranked by our method given by user *i* and $T_i^{k-means}$ denotes the number of topics in the results re-ranked by k-means given by user *i*. Table 1 shows the relative changes in top 20 retrieval results of each query.

Table 1. Relative changes in top 20 results of 20 queries

No.	Relative Changes	No.	Relative Changes
1	+5.2	11	+4.2
2	+4.8	12	+4.0
3	+4.2	13	+4.8
4	+3.0	14	+2.8
5	+3.8	15	+4.8
6	+4.4	16	+4.2
7	+3.2	17	+5.0
8	+5.4	18	+3.8
9	+4.8	19	+3.6
10	+4.6	20	+4.2

From Table 1, we can tell that our proposed re-ranking method based on TR score improves the topic coverage in top 20 retrieval results significantly compared with k-means algorithm. Here we present an example to illustrate the improvement in top 20 results. When we query *forest*, before re-ranking, only 2 topics (*forest landscape and endangered forest*) related with forest are in the top 20 results. However, after re-ranking, 5 topics including *forest coverage in various districts, forest landscape, endangered forests, forest fire and forest insects* are presented in top 20 results.

Although the proposed method achieves significant improvements in topic coverage of the top retrieval results, we cannot simply obtain the improvement at the cost of losing relevance in the results. Therefore, the five researchers are also asked to give the relevance score of each image in top20 results before and after re-ranking (2-relevant, 1-hard to tell, 0-irrelevant). The average relevance score (ARS) of an image set *I* is calculated as follows:

$$ARS(I) = \frac{1}{MN}\sum_{k=1}^{N}\sum_{j=1}^{M} RS_k(i_j) \quad (8)$$

M denotes the number of images in set *I* (*M*=20 in our method). $RS_k(i_j)$ denotes the relevance score of image i_j given by user *k*.

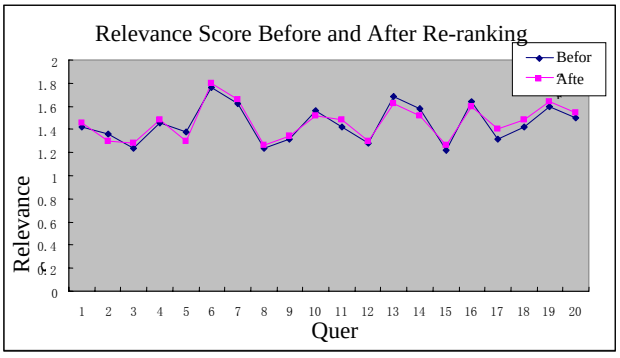


Figure 2. Relevance score before and after re-ranking

Figure 2 illustrates the average relevance score of each query before and after re-ranking. From Figure 2, we can tell that our

proposed approach almost has no influence on the relevance of retrieval results.

We also design an experiment to verify effectiveness of the proposed approach in topic-coverage evaluation. As illustrated in Figure 3, after re-ranking, by k-means or our method, the diversity scores of top 20 retrieval results of each query achieve significant increase and our re-ranking strategy achieves better than k-means method does.

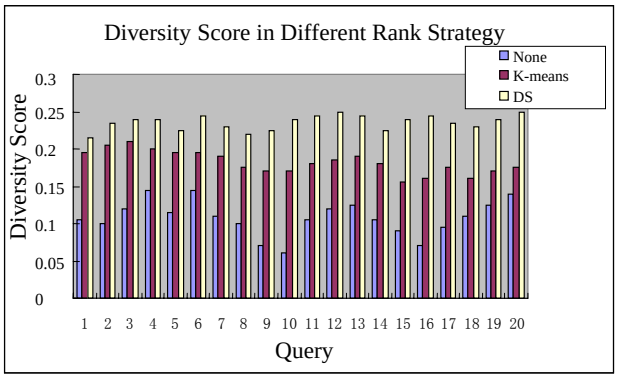


Figure 3. Diversity Score in Different Rank Strategy

Figures 3 shows that the evaluation results of DS accords with the results of users (see table 1), which demonstrates the effectiveness of the proposed approach in evaluating topic coverage of image retrieval results.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we focus on the refinement of the retrieval results by solving two problems: (1) How to improve the topic coverage of the results; (2) How to evaluate the topic coverage of image retrieval results. Accordingly, two contributions are made. First, we introduce a re-ranking method that achieves topic-coverage improvement significantly. In addition, a quantitative method named Diversity Score is proposed to evaluate the topic coverage of image retrieval results. Moreover, to the best of our knowledge, this is the first effort in the area of image retrieval that tries to quantify the topic coverage of image retrieval results.

Experimental results demonstrate that the proposed method is effective in evaluating the topic coverage in image retrieval results and the new ranking algorithm outperforms the existing clustering methods such as k-means.

Post-retrieval processing is a fascinating and rewarding research area in the literature of image retrieval. By presenting our simple and intuitive ideas, we expect more sophisticated thoughts from fellow researchers. Our future work includes applying our method to evaluate and improve topic coverage of video/audio retrieval results and scaling our method to larger dataset.

6. ACKNOWLEDGMENTS

The project in our paper is supported by China-American Digital Academic Library (CADAL) project (No. CADAL2004002) and the Hi-Tech Research and Development Program (863) of China (No. 2003AA119010). We also acknowledge the efforts of the five evaluators.

7. REFERENCES

[1] K. Brinker, “Incorporating diversity in active learning with support vector machines” In *Proceedings of the 20th International Conf. on Machine Learning*, pp. 59--66, 2003.

[2] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. In *Proceedings of the 21st ACM SIGIR* 1998, 335-336

[3] P. Duygulu, K. Barnard, N. Freitas, and D. A.Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary”. *In7th ECCV conference*, volume 4, pp. 97–112, 2002.

[4] K. S. Goh, E. Y. Chang, W. C. Lai, “Multimodal concept-dependent active learning for image retrieval”. In *Proceedings of the 12th annual ACM international conference on Multimedia,* pp. 564-571, 2004.

[5] J. Jeon, V. Lavrenko, and R. Manmatha. “Automatic image annotation and retrieval using cross-media relevance models”. *In 26th ACM SIGIR Conference*, 2003.

[6] D. H. Kim, C. W. Chung, and K. Barnard, "Relevance Feedback Using Adaptive Clustering for Image Similarity Retrieval," *Journal of Software Systems and Software*, Volume 78, Issue 1, October 2005, Pages 9-23.

[7] J. Li, “Two-scale image retrieval with significant meta-information feedback” In *Proceedings of the 13th ACM international conference on Multimedia,* pp. 499-502, 2005.

[8] Y. Lu , C. H. Hu , X. Q. Zhu , H. J. Zhang , Q. Yang, “A unified framework for semantics and feature based relevance feedback in image retrieval systems”, *Proceedings of the 8th ACM international conference on Multimedia*, p.31-37, 2000.

[9] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu. “Automatic Multimedia Cross-modal Correlation Discovery”. In *Proceedings of the 10th ACM SIGKDD Conference*, pp. 653-658, Aug 2004.

[10] Y. Rui and T. S. Huang and S. F. Chang, “Image Retrieval: Past, Present, and Future”, *Journal of Visual Communication and Image Representation*, 1999, Vol. 10, pp.1-23

[11] N. V. Shirahatti and K. Barnard, "Evaluating Image Retrieval" *Computer Vision and Pattern Recognition*, San Diego, CA, pp. I:955-961, June 2005.

[12] B. L. Tseng, C. Y. Lin, M. R. Naphade, A. Natsev and J. R. Smith, “Normalized classifier fusion for semantic visual concept detection”, In *Proceedings of ICIP 2003*: 535-538

[13] B. Y. Zhang, H. Li, Y. Liu, L. Ji, W. S. Xi, W. G. Fan, Z. Chen, W. Y. Ma, “Improving web search results using affinity graph”. In *Proceedings of the 28th annual international ACM SIGIR***,** pp. 504-511, 2005.

[14] X. S. Zhou and T. S. Huang, “Unifying keywords and visual contents in image retrieval”. *IEEE MultiMedia*, 9(2):23--33, 2002.

[15] C-N Ziegler, S.M. McNee, J.A. Konstan, and G. Lausen, ” improving Recommendation Lists Through Topic Diversification”, *In Proceedings of WWW 2005*, pp.22-32, 2005.