Optimizing Multimodal Reranking for Web Image Search

Hao Li
Inst. of Computing Technology
Chinese Academy of Sciences
lihao@ict.ac.cn

Meng Wang
National Univ. of Singapore
eric.mengwang@gmail.com

Zhisheng Li
National Univ. of Singapore
lizs@comp.nus.edu.sg

Zheng-Jun Zha
National Univ. of Singapore
zhazj@comp.nus.edu.sg

Jialie Shen
Singapore Management Univ.
jlshen@smu.edu.sg

ABSTRACT

In this poster, we introduce a web image search reranking approach with exploring multiple modalities. Different from the conventional methods that build graph with one feature set for reranking, our approach integrates multiple feature sets that describe visual content from different aspects. We simultaneously integrate the learning of relevance scores, the weighting of different feature sets, the distance metric and the scaling for each feature set into a unified scheme. Experimental results on a large data set that contains more than 1,100 queries and 1 million images demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Search; Reranking; Graph-based Learning

1. INTRODUCTION

Commercial image search engines, such as Google, Yahoo and Bing, usually index web images using textual information, such as images' titles and ALT text and the surrounding texts on web pages. However, frequently the text information does not describe the content of images, and it can severely degrade the web image search performance. Reranking is an approach to boosting image search performance by adjusting search results based on images' visual information [1][2][4][5][6]. Typically, image search reranking is based on two assumptions: (1) the results after reranking should not change too much from the initial ranking list; and (2) visually similar images should be close in ranking lists. These two assumption usually can be formulated as a graph-based learning scheme, where vertices are images and edges indicate the pairwise similarities [1][2][5].

Although many different reranking algorithms have been proposed, existing results show that reranking is not guaranteed to improve performance. In fact, in several cases search performance may even degrade after reranking. One reason is that the second assumption does not hold for the employed feature space. Actually the effective features should vary across queries.

In this work, we propose a web image search reranking approach with multiple modalities. Here a modality is regarded as a description of images, i.e., a feature set. Our proposed scheme integrates multiple modalities in a graph-based learning framework. It simultaneously learns the relevance scores, the weighting of different modalities, the distance metric and the scaling for each modality. The effects of different modalities can be adaptively modulated for each query. Although multiple modalities are involved, there are only two parameters in our algorithm.

2. MULTIMODAL RERANKING

2.1 Formulation

Generally, graph-based reranking can be formulated as a regularization framework as follows

$$\mathbf{r}^* = \min_{\mathbf{r}} Q(\mathbf{r}, \bar{\mathbf{r}}, \mathbf{X}) = \min_{\mathbf{r}} R(\mathbf{r}, \mathbf{X}) + \lambda L(\mathbf{r}, \bar{\mathbf{r}}) \qquad (1)$$

where $\mathbf{r} = [r_1, r_2, \ldots, r_N]^T$ is the ranking scores corresponding to a sample set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$. Here the term $R(.)$ is the regularization term that models the assumption that visually similar images should be close, and the term $L(.)$ is a loss term that estimates the difference between $\mathbf{r}$ and $\bar{\mathbf{r}}$. For the first term, it is usually formulated as

$$R(\mathbf{r}, \mathbf{X}) = \sum_{i,j} w_{i,j} \left\| \frac{r_i}{d_{ii}} - \frac{r_j}{d_{jj}} \right\|^2 \qquad (2)$$

where $\mathbf{W}$ is a similarity matrix in which w_{ij} indicates the visually similarity of $\mathbf{x}_i$ and $\mathbf{x}_j$, and d_{ii} indicates the sum of the i-th row of $\mathbf{W}$.

Now we extend the scheme to obtain our algorithm. First, considering using one modality, we use Mahalanobis distance metric instead of the Euclidean distance metric

$$w_{ij} = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)) = \exp(-||\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)||^2) \qquad (3)$$

Then we consider there are K modalities. Here we linearly combine the regularizer terms, i.e.,

$$R(\mathbf{r}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \alpha) = \sum_{k=1}^{K} \sum_{i,j} \alpha_k w_{k,ij} \left\| \frac{r_i}{d_{k,ii}} - \frac{r_j}{d_{k,jj}} \right\|^2 \qquad (4)$$

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

where $w_{k,ij} = \exp(-\|\mathbf{A}_k(\mathbf{x}_i - \mathbf{x}_j)\|^2)$ and α_k is the weight for k-th modality that satisfies $0 \le \alpha_k \le 1$ and $\sum_{k=1}^{K} \alpha_k = 1$. As previously mentioned, we integrate the learning of the weights into our regularization framework in order to adaptively modulate the impacts of different modalities. Therefore, the regularizer term turns to

$$R(\mathbf{r}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \alpha) = \sum_{k=1}^{K} \sum_{i,j} \alpha_k w_{k,ij} \left\| \frac{r_i}{d_{k,ii}} - \frac{r_j}{d_{k,jj}} \right\|^2 + \xi ||\alpha||^2 \quad (5)$$

For the loss term, usually it estimates the difference between two ranking lists. Here we directly use the square loss. Therefore, our algorithm can be formulated as the following optimization problem

$$\min_{\mathbf{r}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \alpha} \sum_{k=1}^{K} \sum_{i,j} \alpha_k w_{k,ij} \left\| \frac{r_i}{d_{k,ii}} - \frac{r_j}{d_{k,jj}} \right\|^2 + \lambda ||\mathbf{r} - \bar{\mathbf{r}}||^2 + \xi \|\alpha\|^2$$

$$\text{s.t. } 0 \le \alpha_k \le 1, \sum_{k=1}^{K} \alpha_k = 1 \quad (6)$$

We can see that this optimization framework involves the following variables: (1) $\mathbf{r}$, the ranking scores to be estimated; (2) α, the weights for combining K modalities; and (3) $\mathbf{A}_k, (1 \le k \le K)$, the transform matrices for K modalities.

2.2 Solution

We adopt alternating optimization to solve the problem.

First, we consider α and $\mathbf{A}_k (k = 1, 2, ..., K)$ are fixed, then $\mathbf{r}$ can be solved with a closed-form solution.

Second, we consider $\mathbf{r}$, α, and $\mathbf{A}_1, \ldots, \mathbf{A}_{k-1}, \mathbf{A}_{k+1}, \ldots, \mathbf{A}_K$ are fixed, then we derive the derivative of Q with respect to $\mathbf{A}_k$. It can be derived that

$$\frac{\partial}{\partial \mathbf{A}_k} Q(\mathbf{r}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_K)$$
$$= \alpha_k \sum_{i,j} \left(h_{ij}^2 \frac{\partial w_{k,ij}}{\partial \mathbf{A}_k} - w_{k,ij}^T h_{ij} \left(\frac{r_i}{\sqrt{d_{k,ii}^3}} \frac{\partial d_{k,ii}}{\partial \mathbf{A}_k} - \frac{r_j}{\sqrt{d_{k,jj}^3}} \frac{\partial d_{k,jj}}{\partial \mathbf{A}_k} \right) \right)$$

where $h_{ij} = \frac{r_i}{d_{k,ii}} - \frac{r_j}{d_{k,jj}}$, $\frac{\partial d_{k,ij}}{\partial \mathbf{A}_k} = \sum_{j=1}^{N} \frac{\partial W_{k,ij}}{\partial \mathbf{A}_k}$, $\frac{\partial w_{k,ij}}{\partial \mathbf{A}_k} = -2 w_{k,ij} \mathbf{A}_k (\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)})^T (\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)})$ Thus, $\mathbf{A}_k$ can be optimized with gradient descent method.

Finally, considering $\mathbf{r}$ and $\mathbf{A}_k (k = 1, 2, \ldots, K)$ are fixed, then Eq.6 becomes:

$$\min_{\alpha} \sum_{k=1}^{K} \sum_{i,j} a_k w_{k,ij} || \frac{r_i}{d_{k,ii}} - \frac{r_j}{d_{k,jj}} ||^2 + \xi \|\alpha\|^2 \quad (7)$$

$$\text{s.t. } 0 \le \alpha_k \le 1, \sum_{k=1}^{K} \alpha_k = 1$$

We can employ coordinate descent method to solve Eq.7.

We can iterate the optimization of $\mathbf{r}$, $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K$ and α. Since each step decreases the objective in Eq.6 and the value of the objective function is lower bounded by 0, the whole process is guaranteed to converge.

3. EXPERIMENTS AND CONCLUSION

We evaluate our approach with several existing methods on the large-scale image search dataset, MSRA-MM Version 2.0[3], that contains the search results (1,011,738 images in total) of 1,165 queries from Microsoft Bing image search engine. In [3], the queries are manually classified into 10 categories, and each image is labeled with 3 relevance levels (0, 1, and 2). There are 7 feature sets are extracted. We compare the following methods:

(1) Bayesian rerankingBayesianReranking[5]. We concatenate all features into a long vector and then perform the preference strength based method in [5].

(2) Graph-based reranking with concatenated features. That is, we concatenate all the features into a long vector and then perform graph-based reranking.

(3) Proposed multimodal reranking algorithm. For the initial relevance score of i-th ranking position, we estimate it by averaging the ground truth scores at the i-th position of all 1,165 queries.

The methods are denoted as "Bayesian", "Concatenation" and "Multimodal", respectively. For all the involved parameters, we tune them to their optimal values on the 68 queries in of *Ver1* with the performance evaluation metric of average NDCG@100, and then these parameters are fixed in the processing of all queries.

Table 1: Average NDCG@100 comparison for each category of different reranking methods and the original search results. Here *Ver1*, *Animal*, *Cartoon*, *Event*, *Person*, *Object*, *People*, *Scene*, *Time08* and *Misc* are the 10 query categories in MSRA-MM dataset.

	Baseline	Bayesian	Concatenation	Multimodal
Ver1	0.544	0.542	0.556	**0.568**
Animal	0.734	0.775	0.758	**0.791**
Cartoon	0.807	0.859	0.842	**0.865**
Event	0.788	0.779	0.797	**0.811**
Person	0.908	0.916	0.926	**0.940**
Object	0.703	0.723	0.722	**0.745**
People	0.714	0.703	0.718	**0.742**
Scene	0.702	0.766	0.735	**0.792**
Time08	0.830	0.844	0.863	**0.870**
Misc	0.736	0.760	0.771	**0.790**
Mean	0.747	0.770	0.773	**0.795**

Table 1 illustrates the average NDCG@100 measurements of the queries in each category after reranking. We also demonstrate the performance of original search results without reranking. From the table we can see that, in average, all the reranking methods can improve the original search results. Our method performs the best for all categories. Its superiority over the "Concatenation" method demonstrates the effectiveness of our approach of integrating multiple modalities. All the experimental results clearly demonstrate the effectiveness of our approach.

4. REFERENCES

[1] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *ACM MM*, 2007.

[2] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards Relevant and Diverse Search of Social Images. *IEEE Trans. on Multimedia*, vol. 12, no. 8, 2010.

[3] H. Li, M. Wang, and X.-S. Hua. Msra-mm 2.0: A large-scale web multimedia dataset. In *IEEE ICDM Workshops*, 2009.

[4] M. Wang, X.-S. Hua, J. Tang,and R. Hong. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. In *IEEE Trans. on Multimedia*, vol. 11, no. 3, 2009.

[5] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *ACM MM*, 2008.

[6] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi and Y. Song. Unified Video Annotation Via Multi-Graph Learning. In *Trans. on CSVT*, vol. 19, no. 5, 2009.