7-2010

Co-reranking by mutual reinforcement for image search

Ting YAO

Tao MEI

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Co-reranking by Mutual Reinforcement for Image Search

Ting Yao [1], Tao Mei [2], Chong-Wah Ngo [3]
[1] University of Science and Technology of China, Hefei 230027, P. R. China
[2] Microsoft Research Asia, Beijing 100190, P. R. China
[3] City University of Hong Kong, Kowloon, Hong Kong
tingyao.ustc@gmail.com; tmei@microsoft.com; cwngo@cs.cityu.edu.hk

ABSTRACT

Most existing reranking approaches to image search focus solely on mining "visual" cues within the initial search results. However, the visual information cannot always provide enough guidance to the reranking process. For example, different images with similar appearance may not always present the same relevant information to the query. Observing that multi-modality cues carry complementary relevant information, we propose the idea of co-reranking for image search, by jointly exploring the visual and textual information. Co-reranking couples two random walks, while reinforcing the mutual exchange and propagation of information relevancy across different modalities. The mutual reinforcement is iteratively updated to constrain information exchange during random walk. As a result, the visual and textual reranking can take advantage of more reliable information from each other after every iteration. Experiment results on a real-world dataset (MSRA-MM) collected from Bing image search engine shows that co-reranking outperforms several existing approaches which do not or weakly consider multi-modality interaction.

Categories and Subject Descriptors

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

General Terms

Image search, co-reranking, graph model.

1. INTRODUCTION

With the advance of Web 2.0 technology, multimedia content creation and distribution are much easier than ever before. This has led to the explosive growth of community-contributed media data, as well as the surge of research activities in visual search. Due to the great success of text document retrieval, most existing image search systems only rely on the surrounding text associated with the images. However, visual relevance cannot be merely judged by text-based approaches as the textual information is usually too noisy to precisely describe visual content or even unavailable.

To address this problem, visual search reranking, which is defined as reordering the ranked visual documents based on the initial search results or some auxiliary knowledge to improve search performance, has received increasing attention [8, 10, 11, 15, 16, 17, 20, 21, 23, 28]. The research on visual search reranking has proceeded along three dimensions from the perspective of how external knowledge is exploited [16]: 1) self-reranking [8, 10, 11, 15, 20, 21, 23], which focuses on detecting recurrent patterns in the initial search results without any external knowledge; 2) example-reranking [17, 28], in which the query examples are provided to mine the relevant patterns to the given query; and 3) crowd-reranking [16], which aims to mine relevant visual patterns from visual search results of multiple search engines. In summary, most of existing approaches first detect the *dominant visual patterns*, and then perform *reranking* based on the following two assumptions: 1) the visual documents (e.g., images or video shots) with the dominant patterns should be ranked higher; 2) the visual documents with similar visual appearance should be ranked closely.

Based on the above assumptions, Fig. 1 shows the reranking examples to the query "car." The image ranked list in (a) illustrates the text baseline results, the lists in (b) and (c) present the results from the random walk-based reranking method based on the visual and the textual cues, respectively [10], and the image ranked list in (d) indicates the results obtained by jointly leveraging both visual and textual cues (by this work). We can observe that in the visual based results (b), since "road" and "car" often co-occur in the same image and share similar visual appearance, the image patches within the road area are also detected as the dominant pattern in (b). As a result, the first and the second images in (b) are both ranked closely and highly, although the second one is irrelevant to the query "car." Likewise, when text-based reranking approach is applied in (c), the word "BMW" is detected as the textual dominant pattern due to its high frequency. As a result, the images with the keyword "BMW" (including the "BMW" logo) are ranked highly in (c). Therefore, solely using the visual or textual cues cannot always achieve satisfying reranking results. A more effective reranking approach would jointly explore visual and text cues since relevant information mined from different features can complement each other. As shown in Fig.1 (d), a better reranking result could be obtained if both

CIVR'10, July 5-7, Xi'an, China.

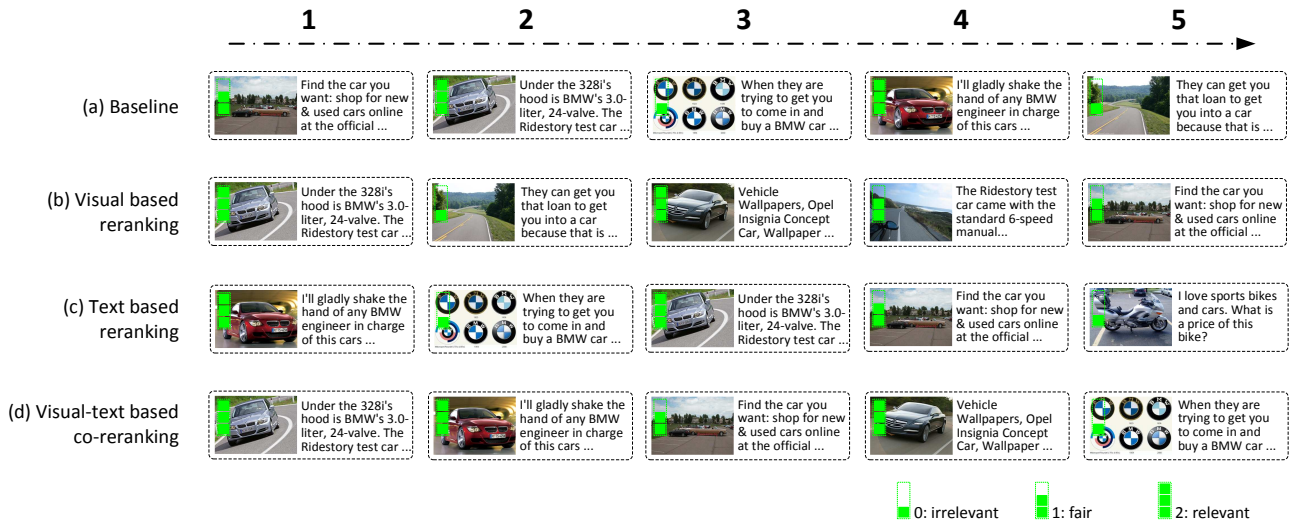


Figure 1: Top 5 image search results with the query "car." (a) Text baseline collected from a popular image search engine [3]; (b) random walk reranking results based on the visual cues [10]; (c) random walk reranking results based on the textual cues [10]; (d) co-reranking results by combining the visual and text cues (this paper). [Best viewed in color].

cues can mutually reinforce each other and thus provide the more reliable relevant information. In (d) the images containing both dominant visual pattern "road" and the dominant textual pattern "BMW" are ranked highly.

Motivated by the above observations, this paper presents a novel reranking approach, called *co-reranking*, which aims to mine and leverage the interrelationship between the visual and textual cues. We assume that there is a mutually reinforcing relationship between visual and textual cues which could be reflected in the rerankings. Specifically, the more similar visual appearances will enhance the relevance of documents with dissimilar textual descriptions, while dissimilar textual descriptions will reduce the relevance of documents with similar visual appearances. In our work, co-reranking is modeled in a random walk like framework. Two coupling random walks are proposed to combine the visual and textual cues, aiming at reinforcing the mutual exchange and propagation of information relevancy across different cues. As a result, the visual and textual reranking can take advantage of more reliable information from each other. The assumption is that the documents with both similar visual appearances and textual descriptions are to be ranked closely.

It is worth noting that multimodal search (i.e., combining textual, visual, audio, and other context information for search and reranking) has attracted intensive attention in recent years. The existing approaches to multimodal search mainly adopt a linear combination [6] or probabilistic models [2, 4, 21]. However, most of these approaches use the multiple cues independently and neglect the reinforcement relation among them. In contrast, co-reranking mutually takes the advantage of textual and visual cues from each other for simultaneous improvement of text and visual search.

The rest of the paper is organized as follows. We review the related work on visual search reranking and (re)ranking with the multiple cues in Section 2. Section 3 gives the detailed descriptions on the proposed co-reranking approach. Experiments are reported in Section 4, followed by the conclusions in Section 5.

2. RELATED WORK

There exists rich research on visual search reranking in recent years. We first give a brief survey on the related works about image and video search reranking. On the other hand, image search by combining the multimodal cues has attracted increasing attention recently. We also discuss representative works in this topic.

2.1 Visual Search Reranking

The research on image and video search reranking has proceeded along three dimensions from the perspective of the external knowledge used [16]: *self-reranking* which requires no external knowledge, *example-reranking* which is based on the user-provided query examples, and *crowd-reranking* which exploits the online crowdsourcing knowledge.

The first dimension, i.e., self-reranking, aims to improve the initial performance by only mining the initial ranked list without any external knowledge [8, 10, 11, 15, 20, 21, 23]. For example, Hsu *et al.* formulate the reranking process as a random walk over a context graph, where video stories are nodes and the edges between them are weighted by multimodal similarities [10]. Fergus *et al.* first perform the visual clustering on initial returned images by probabilistic Latent Semantic Analysis (pLSA), learn the visual object category, and then rerank the images according to the distance to the learned categories [8].

The second dimension, i.e., example-reranking, leverages a few query examples (e.g., images or video shots) to train the reranking models [17, 28]. The search performance can be improved due to the external knowledge derived from these examples. For example, Yan *et al.* and Schroff *et al.* view the query examples as pseudo-positives and the bottom-ranked initial results as pseudo-negatives [28]. A reranking model is then built based on these samples by Support Vector Machine (SVM). Liu *et al.* use the query examples to discover the relevant and irrelevant concepts for a given query, and then identify an optimal set of document

pairs via an information theory [17]. The final reranking list is directly recovered from this optimal pair set.

The third dimension, i.e., crowd-reranking, is characterized by mining relevant visual patterns from the crowdsourcing knowledge available on the Internet. For example, a recent work first constructs a set of visual words based on the local image patches collected from multiple image search engines, explicitly detects the so-called salient and concurrent patterns among the visual words, and then theoretically formalizes the reranking as an optimization problem on the basis of the mined visual patterns [16].

However, it is observed that most of existing reranking methods mainly exploit the visual cues from the initial search results. Even if they tried to leverage multimodal cues, they deal with different kinds of features independently. In other words, the mutual enforcement or connection between different modalities for reranking has not been fully exploited yet. To address the issue, in this paper we leverage visual and textual information via mutual reinforcement.

2.2 Ranking/Reranking with Multiple Cues

The literature of combining multiple cues for search is extensive in recent years [15, 19, 20, 21, 24, 26]. For example, Wang *et al.* give the reviews on multimedia content analysis by using audio and visual clues, in which the advances in using audio and visual information are jointly discussed for accomplishing classification, indexing, retrieval, summarizing, and browsing [26]. Mei *et al.* present a video search scheme, in which the multimodal fusion and reranking module leverages the visual, text and concept modality aiming to enhance the video search performance [19]. Lin *et al.* propose a web image retrieval reranking based on a probabilistic model which evaluates the relevance of the HTML document linking to the image and assigns a probability of relevance by a Bayesian-based relevance model. The relevance model is built automatically through a web text search engine [15]. Schroff *et al.* present a multi-modal ranking approach by employing text, web metadata, and visual features [21]. The images are first reranked using a Bayesian posterior estimator trained on the surrounding texts, and then the top-ranked images are used as training data. A SVM-based visual classifier is learned to improve the ranking performance. Wang *et al.* present a retrieval method for object images by exploiting online text and visual resources [24]. With the rich human compiled text and image data, the text models are built based on Wikipedia [27], which is the biggest free text encyclopedia on the Internet. Meanwhile, the image models are built by Caltech data sets and Flickr [9], which provide clean and immense number of images, respectively. Finally, the text and image models are combined to produce the image ranking function.

However, all of the mentioned works deal with the multiple cues independently and neglect the connections among them. In fact, the multiple cues can mutually influence each other and therefore are able to improve reranking performance. In the proposed co-reranking method, we explore the mutual exchange and reinforcement of visual-textual cues in an iterative process.

3. CO-RERANKING APPROACH

The basic idea of the proposed co-reranking for image search is that the relationship between the visual and textual cues can mutually influence each other. The exploration of this interactive enforcement of these two cues can lead to a more effective search performance.

The overview of the proposed co-reranking approach is shown in Fig. 2. Given a textual query, an initial ranked list of visual documents is obtained by text-based search technique. We regard each document as composed of visual and textual parts, denoted as i_i and t_i in Fig. 2. The rank list is then organized into two graphs based on visual and textual description. Each node in a graph carries a score based on the initial ranking, and there is a one-to-one correspondence, marked by dotted line in Fig. 2, between visual and textual description of a document. Mutual relationship between these two graphs is reinforced by using visual similarity as edge weight to link textual description and vice versa. The length of edge presents the degree of the similarity. The shorter the edge, the larger the similarity. Co-reranking is performed by iterative propagation of visual (textual) edge weights on textual (visual) description through separate random walks on two graphs. In other words, the information exchange among visual description is governed by textual similarity and vice versa. During random walks, the visually and textually consistence patterns are expected to receive higher scores as a result of mutual information propagation, while similar patterns will also get closer to each other because of mutual reinforcement from different cues. When the random walks converge, a reranked list can be generated by re-ordering the documents according to their joint visual-textual scores.

For example, we can see that the two nodes i_2 and i_4 are adjacent in the text space, also the nodes i_2 and i_4 are similar based on their visual features. With the iterative process performs, the relevance score of i_2 and i_4 become close. On the other hand, although the nodes i_2 has the similar visual appearances to i_5, the textual descriptions of their corresponding nodes t_2 and t_5 are dissimilar, thus the relevance score of i_2 and i_5 become far away.

3.1 Problem Formulation

Suppose we have a document set χ with N documents to be ranked, where $\chi = \{d_1, d_2, ..., d_N\}$ and d_i $(i = 1, ..., N)$ denote the document list and the i^{th} visual document in χ, respectively. Let v_{T_j} and v_{I_j} denote the textual and visual initial ranking scores for the j^{th} document, respectively. Two graphs which depict the visual and textual parts of χ are formed, shown in the Fig. 2 (b). Co-reranking is then formulated as a problem of two random walks on these graphs in Eq. (1).

$$\begin{cases} r_{T_j}^{(t)} = \omega_1 \sum_i r_{I_i}^{(t-1)} p_{I_{ij}} + (1-\omega_1) v_{T_j} \\ r_{I_j}^{(t)} = \omega_2 \sum_i r_{T_i}^{(t)} p_{T_{ij}} + (1-\omega_2) v_{I_j} \end{cases} \quad (1)$$

where $r_{T_j}^{(t)}$ and $r_{I_j}^{(t)}$ denote the relevance score of text and image of the j^{th} document at the t^{th} iteration, respectively. The score range of $r_{T_j}^{(t)}$ and $r_{I_j}^{(t)}$ is $(0, 1]$. $p_{T_{ij}}$ indicates the transition probability from text t_i to t_j, while $p_{I_{ij}}$ indicates the transition probability from image i_i to i_j. ω_1 and ω_2 are weighting parameters $(0 \leqslant \omega_1, \omega_2 \leqslant 1)$. In Eq. (1), the first term represents the information exchanged from neighboring nodes. Note that mutual exchange is imposed by using visual score $r_{I_j}^{(t-1)}$ to guide new textual score $r_{T_j}^{(t)}$ and vice versa. The second term is the initial textual and visual score.

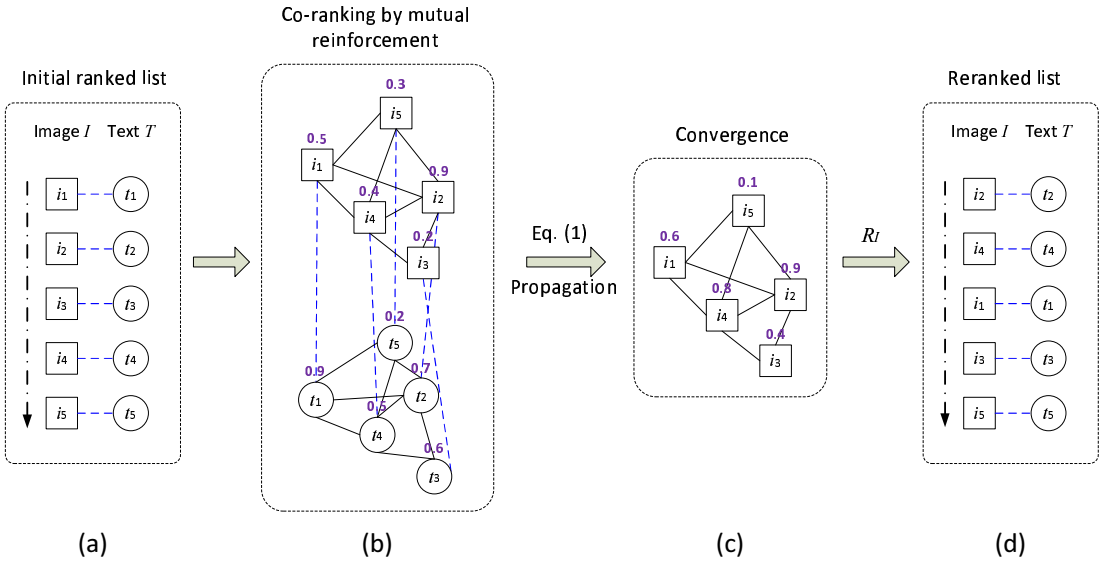


Figure 2: The approach overview of co-reranking for image search: (a) initial ranked list obtained by the text-based search; (b) co-reranking by mutual reinforcement of visual and text information via two graphs; (c) convergence after propagations in (b); (d) reranked list. (Note that the square and circle nodes indicate the visual and textual parts of an image, respectively, the relevance score is shown above each node, and the length of an edge indicates the similarity between two nodes.)

3.2 Initial Ranking Score

In this section, we will show how to compute the initial text and image ranking scores v_{T_i} and v_{I_i}.

Initial Textual Ranking Score. The textual cues in the image search include the surrounding texts and image captions. The standard stemming and stop word removal [19] is performed in the preprocessing. In addition, HTML-tags and domain-specific stop words (such as "html" or "jpg") are ignored. We extract top L terms of the rest terms in the documents for each query and calculate their term frequency (tf) for each document as the textual feature. To compute textual similarity, we use the cosine distance which is widely adopted in information retrieval [1]. Let $\mathbf{D}_i$ denote the L dimension vector of the tf of the i^{th} document. The k^{th} element of the $\mathbf{D}_i$ is represented as d_{ik}. Then, the similarity between the i^{th} document and the j^{th} document is defined by

$$s_{T_{ij}} = \frac{\sum_{k=1}^{L} d_{ik} d_{jk}}{\sqrt{\sum_{k=1}^{L} d_{ik}^2 \sum_{k=1}^{L} d_{jk}^2}} \tag{2}$$

In general, as most popular image search engines build only upon text information for the initial ranked list, we directly use the initial ranked score as the textual initial scores and form a row vector $\mathbf{V}_T \equiv [v_{T_i}]_{1\times N}$. The v_{T_i} is given by

$$v_{T_i} = \frac{N-i}{N}, \quad i = 0, 1, ..., N-1 \tag{3}$$

Initial Visual Ranking Score. We follow the approach in [18] and adopt scale-invariant feature transform (SIFT) descriptor with a Difference of Gaussian (DoG) interest point detector for extracting the images' visual patterns. The interest point is referred to as local salient patch, each associated with a 128-dimensional feature vector. We further use K-means to cluster the similar patches into "visual words," and use Bag-of-Word (BoW) to represent each image as it has proven to be effective for object and scene retrieval [13, 14, 16, 22]. We use the cosine distance to calculate the image similarity $s_{I_{ij}}$ similar to the text which is described in Eq. (2).

For the visual initial ranked list, we consider estimating the visual clustering density based on the initial results. A straightforward implementation is to first perform K-mean clustering, and then make a linear combination of cluster scores and initial scores. This kind of combination is widely used in multimodal video search systems [7] and video search reranking [11]. The visual initial ranked scores can also be formed as a row vector $\mathbf{V}_I \equiv [v_{I_i}]_{1\times N}$ and the element v_{I_i} is formulated as the Eq. (4).

$$v_{I_i} = \lambda \times v_{I_i}^c + (1-\lambda) \times v_{T_i} \tag{4}$$

where λ $(0 \leqslant \lambda \leqslant 1)$ is the tradeoff parameter. v_{T_i} is the initial relevance score obtained in Eq. (3) and $v_{I_i}^c$ is the score of cluster which the i^{th} image belongs to, it is defined by the average of initial relevance scores in the cluster.

3.3 Problem Solution

In this section, we discuss the solution to the problem presented in section 3.1. We re-write the Eq. (1) in the following matrix form:

$$\begin{cases} \mathbf{R}_T^{(t)} = \omega_1 \mathbf{R}_I^{(t-1)} \mathbf{P}_I + (1-\omega_1)\mathbf{V}_T \\ \mathbf{R}_I^{(t)} = \omega_2 \mathbf{R}_T^{(t)} \mathbf{P}_T + (1-\omega_2)\mathbf{V}_I \end{cases} \tag{5}$$

where $\mathbf{R}_T^{(t)} \equiv [r_{T_i}^{(t)}]_{1\times N}$ and the $\mathbf{R}_I^{(t)} \equiv [r_{I_i}^{(t)}]_{1\times N}$. $\mathbf{P}_T$ and $\mathbf{P}_I$ are the N-by-N transition matrices for text and image, respectively. Owing to the similar calculation process of the two transition matrices, we take text as an example. The

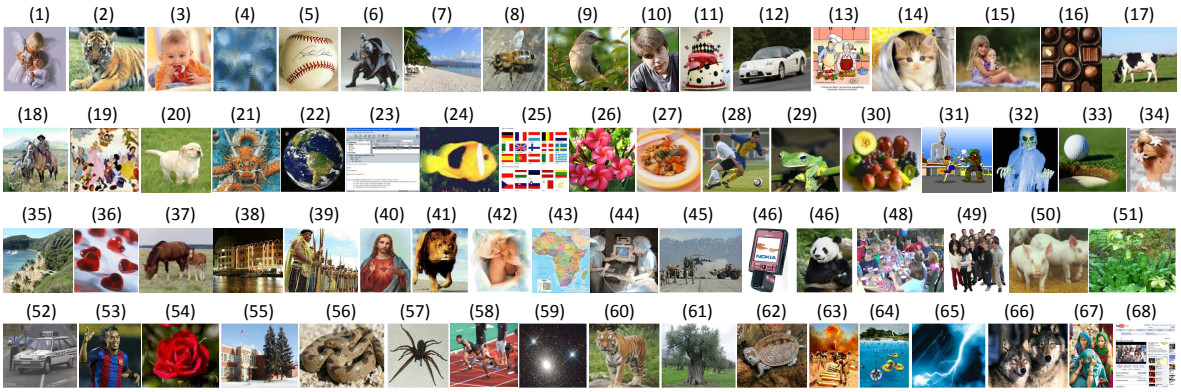


Figure 3: The exemplary relevant image thumbnails for the 68 queries in MSRA-MM dataset [25].

element $p_{T_{ij}}$ is given by

$$p_{T_{ij}} = \frac{s_{T_{ij}}}{\sum_k s_{T_{ik}}} \tag{6}$$

where $s_{T_{ij}}$ is the similarity between t_i and t_j, as defined in Eq. (2).

The iteration of $\mathbf{R}_I^{(t)}$ converges to a fixed point $\mathbf{R}_I^{\infty}$, which is proven as the follows

Proof:

$$\begin{aligned}
\mathbf{R}_I^{(t)} &= \omega_2(\omega_1\mathbf{R}_I^{(t-1)}\mathbf{P}_I + (1-\omega_1)\mathbf{V}_T)\mathbf{P}_T + (1-\omega_2)\mathbf{V}_I \\
&= \omega_2\omega_1\mathbf{R}_I^{(t-1)}\mathbf{P}_I\mathbf{P}_T + \mathbf{A} \\
&= \omega_2\omega_1(\omega_2(\omega_1\mathbf{R}_I^{(t-2)}\mathbf{P}_I + (1-\omega_1)\mathbf{V}_T)\mathbf{P}_T + \\
&\quad (1-\omega_2)\mathbf{V}_I)\mathbf{P}_I\mathbf{P}_T + \mathbf{A} \\
&= (\omega_2\omega_1)^2\mathbf{R}_I^{(t-2)}(\mathbf{P}_I\mathbf{P}_T)^2 + \omega_2\omega_1\mathbf{A}\mathbf{P}_I\mathbf{P}_T + \mathbf{A} \\
&= ... \\
&= (\omega_2\omega_1)^t\mathbf{R}_I^0(\mathbf{P}_I\mathbf{P}_T)^t + \\
&\quad \mathbf{A}(\mathbf{U} + \omega_2\omega_1\mathbf{P}_I\mathbf{P}_T + ... + (\omega_2\omega_1\mathbf{P}_I\mathbf{P}_T)^{(t-1)})
\end{aligned} \tag{7}$$

where $\mathbf{A} = \omega_2(1-\omega_1)\mathbf{V}_T\mathbf{P}_T + (1-\omega_2)\mathbf{V}_I$, and matrix $\mathbf{U}$ is an identity matrix which diagonal elements are 1 and the others are 0. According to Eq. (6), we note that the $\mathbf{P}_I$ and $\mathbf{P}_T$ have been row normalized to 1. For $(0 \leqslant \omega_1, \omega_2 \leqslant 1)$, we can derive that

$$\mathbf{R}_I^{\infty} = \lim_{t\to\infty} \mathbf{R}_I^{(t)} = \mathbf{A}(\mathbf{U} - \omega_2\omega_1\mathbf{P}_I\mathbf{P}_T)^{-1} \tag{8}$$

Eq. (8) is the unique solution. □

4. EXPERIMENTS

4.1 Data

We conducted experiments on the MSRA-MM Dataset [25], which consists of 68 representative queries collected based on the query log of Microsoft Bing Search [3] [1]. These queries cover a wide variety of categories, including objects, people, event, entertainments, and location. For each query, about top 900 images along with the surrounding texts are collected. As a result, the dataset contains 60,257 images in total. The rank orders of these images are obtained as the initial ranked lists. The surrounding texts of each image are first translated to English if they are from non-English sources. Then, the stemming and stop word removal [19] is performed, which results in 86,734 unique words in total. Fig. 3 shows the exemplary relevant image thumbnails for these queries.

4.2 Methodologies

In the dataset, each image to the corresponding query was manually labeled on a scale of 0-2: (0) "irrelevant," (1) "fair," and (2) "relevant." We adopt Normalized Discounted Cumulative Gain (NDCG) as the performance metric since it is widely used to deal with multiple relevance levels [12]. Given a query q, the NDCG score at the depth d in the ranked documents is defined by:

$$NDCG@d = Z_d \sum_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1+j)} \tag{9}$$

where r^j is the rating of the j^{th} document, Z_d is a normalization constant and is chosen so that a perfect ranking's $NDCG@d$ value is 1.

In our experiments, we used top 500 images in the initial search results for reranking, since it is typical that there are very few relevant images after the top 500 search results [16]. Empirically, the number of visual words is set to 2,000 [22] and the tradeoff parameter λ is set to 0.90 considering that the cluster score plays the main role for search relevance [11]. To estimate the initial visual ranking score, the K-Means clustering is performed and the cluster number is set to 20.

To demonstrate the effectiveness of the proposed co-reranking method, we compared with the following two reranking methods. In the two approaches, we selected the parameters achieving the best performance.

- Random walk reranking [10]. A representative graph-based reranking method which only uses one type of

[1] The queries include: (1) angel, (2) animals, (3) baby, (4) backgrounds, (5) baseball, (6) batman, (7) beach, (8) bees, (9) birds, (10) boy, (11) cake, (12) car, (13) cartoon, (14) cat, (15) children, (16) chocolates, (17) cow, (18) cowboys, (19) disney, (20) dogs, (21) dragons, (22) earth, (23) email, (24) fish, (25) flags, (26) flowers, (27) food, (28) football, (29) frogs, (30) fruit, (31) games, (32) ghosts, (33) golf, (34) hairstyles, (35) hawaii, (36) heart, (37) horses, (38) hotels, (39) indians, (40) jesus, (41) lion, (42) love, (43) maps, (44) medical, (45) military, (46) nokia, (47) panda, (48) party, (49) people, (50) pigs, (51) plants, (52) police, (53) ronaldinho, (54) rose, (55) school, (56) snakes, (57) spider, (58) sports, (59) stars, (60) tiger, (61) trees, (62) turtles, (63) war, (64) waterparks, (65) weather, (66) wolves, (67) women, (68) youtube.

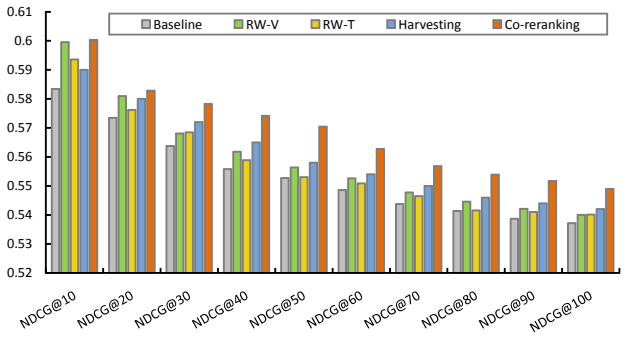


Figure 4: Comparison of reranking approaches in terms of NDCG.

feature to perform reranking. In the following comparison, we use textual and visual features individually, named RW-T (Random Walk-Textual) and RW-V (Random Walk-Visual) in the next.

- Harvesting reranking [21]. A typical reranking method based on both text and visual information. It is also built based on the Pseudo Relevance Feedback (PRF) framework [5] widely used in text search field, which assume the top-ranked results are much more relevant than the low-ranked results in general.

4.3 Evaluations

4.3.1 Evaluation of Reranking Performance

Fig. 4 shows the experimental results. We selected the weighting parameters ω_1 and ω_2 which achieve the best performance, and more details will be analyzed in the next subsection. We can see that the proposed co-reranking outperform the others. Moreover, it can be observed that:

- The improvements of the proposed co-reranking over text-based and visual-based random walk reranking methods indicate that exploiting both text and visual cues can benefit reranking a lot.
- The superiority of the proposed co-reranking to harvesting reranking indicates that text and visual cues can mutual reinforce each other, rather than react individually. It is reasonable to leverage their interrelationship in reranking.
- The text-based random walk method did not outperform the visual-based random walk method, harvesting reranking, as well as the co-reranking. The main reason is that there exists much noise and irrelevant information in texts associated with the images, and directly using such text to describe the image content will lead to the unsatisfying search performance.

Furthermore, the performance improvements are consistent and stable, i.e., most queries are improved compared to the initial ranked lists and have better performance than the other methods, as shown in Fig. 5. From this figure, we can also find different queries are sensitive with different type of cues. For example, the queries "animal" and "youtube" are more sensitive with text cue since the query words are specific and the visual appearance is of high diversity; while the queries "heart" and "rose" are sensitive with

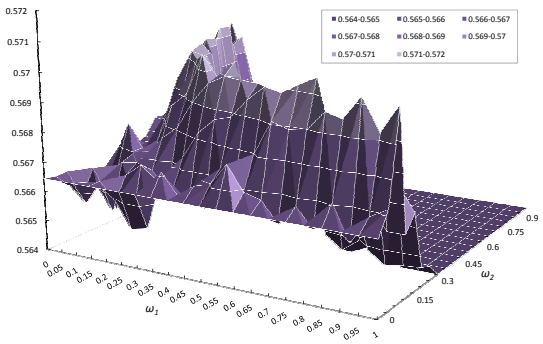


Figure 7: Performance with different ω_1 and ω_2 measured by $NDCG@50$.

visual cue since the relevant results of these queries share relative high visual similarity and specific in visual feature space. On the other hand, for the queries "women" and "backgrounds" which neither specific in text space nor in visual space, the co-reranking method revealed advantage over the other methods which only use one cue.

Fig. 6 shows the top 12 images of different reranking approaches for the query "wolves" and "cat." We can easily see the proposed co-reranking method gets the most satisfying reranking results.

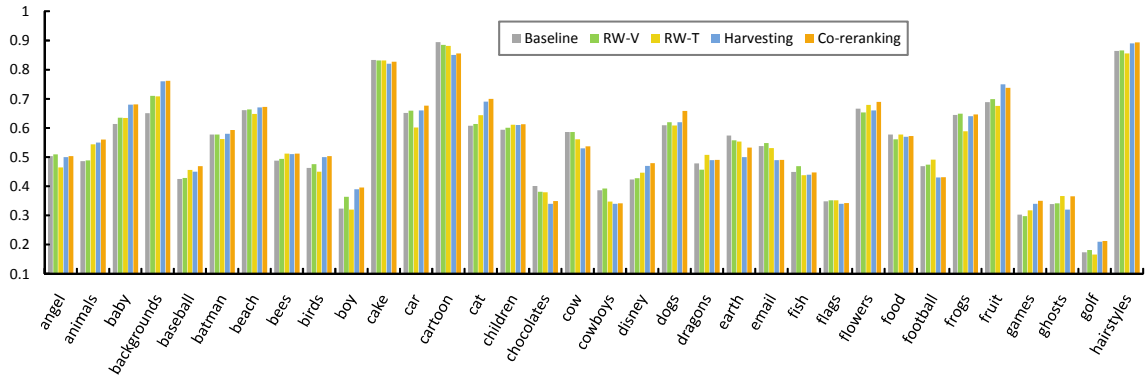
4.3.2 Evaluation of Weighting Parameters

We also investigated the performance of co-reranking method with different weighting parameters ω_1 and ω_2 in Eq. (1). Fig. 7 shows the performance of the co-reranking method with different ω_1 and ω_2 in terms of $NDCG@50$. From the figure, we can see that the performance surface is convex as ω_1 and ω_2 increase.

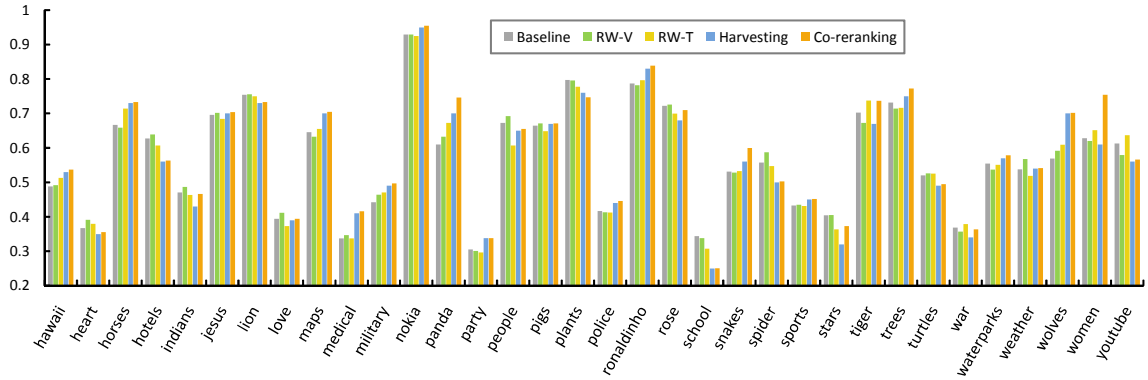
- From the Eq. (1), we can find that when ω_2 goes to 0, the reranking process relies entirely on the image ranking results; when (ω_1, ω_2) is set to $(0, 1)$, the reranking process relies entirely on the text cues mined in the initial search results; when (ω_1, ω_2) is set to (1, 1), the reranking process relies entirely ignore the initial ranked list, only based on the visual and text cues mined in initial search results. Therefore, basically a relatively larger ω_1 and smaller ω_2 would be more suitable for a worse initial search result. It can be concluded that the weighting parameter ω_1 and ω_2 can be set according to the performance of initial search results.
- As shown in Fig.7, the performance increases when ω_1 and ω_2 increase simultaneously and arrives at the peak at $\omega_1 = 0.15$ and $\omega_2 = 0.75$. From this observation, we can conclude that all the initial search results, visual and textual cues play important roles in the reranking.

5. CONCLUSIONS

In this paper, we have explored the mutual exchange and reinforcement of visual-textual cues as a co-reranking problem for image search. Under our formulation, the result of random walk from one modality is iteratively exchanged to constrain the random walk of another modality. This leads



(a) Reranking performance on 1–34 queries in the MSRA-MM dataset.



(b) Reranking performance on 35–68 queries in the MSRA-MM dataset.

Figure 5: Performance of each query measured by $NDCG@50$.

to gradual discovery of dominant and similar visual patterns, which are helpful for image reranking. Our experimental results on MSRA-MM dataset have also demonstrated that co-reranking outperforms several existing approaches which treat each modality independently for reranking. Future work includes the extension of co-reranking to video domain where more modalities can be explored for mutual reinforcement. In addition, the incorporation of external knowledge for co-reranking is also another issue worth further investigation.

6. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] K. Barnard and D. Forsyth. Learning the semantic of words and pictures. In *Proceedings of IEEE International Conference on Computer Vision*, pages 408–415, 2001.

[3] Bing. http://www.bing.com/.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of ACM Special Interest Group on Information Retrieval*, pages 127–134, 2003.

[5] J. Carbonell, Y. Yang, R. Frederking, R. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence*, pages 708–714, 1997.

[6] Z. Chen, W.-Y. Liu, M. L. F. Zhang, and H. Zhang. Web mining for web image retrieval. 52(10):831–839, 2001.

[7] K. M. Donald and A.F.Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *ACM International Conference on Image and Video Retrieval*, pages 61–70, 2005.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1816–1823, 2005.

[9] Flickr. http://www.flickr.com/.

[10] W. Hsu, L. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *ACM International Conference on Multimedia*, pages 971–980, 2007.

[11] W. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Proceedings of the ACM International Conference on Multimedia*, pages 35–44, 2006.

[12] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of ACM Special Interest Group on Information Retrieval*, pages 41–48, 2000.

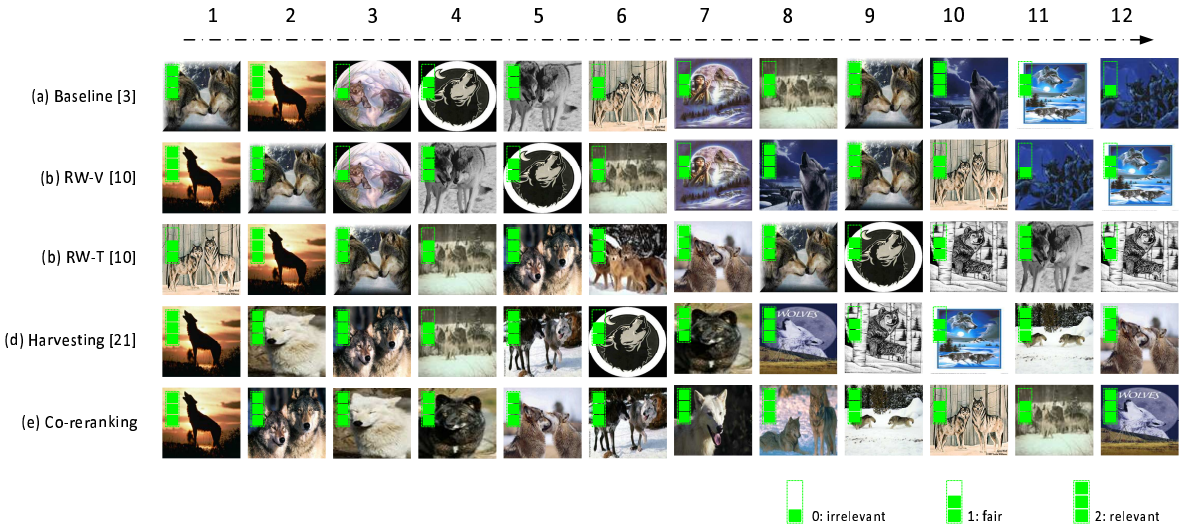
[13] Y. Jing and S. Baluja. Pagerank for product image search. In *Proceedings of International World Wide Web Conference*, pages 307–316, 2008.

[14] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001.

[15] W. H. Lin, R. Jin, and A. Hauptmann. Web image retrieval re-ranking with relevance model. In *IEEE/WIC International Conference on Web Intelligence*, pages 13–17, 2003.

[16] Y. Liu, T. Mei, and X.-S. Hua. CrowdReranking: Exploring Multiple Search Engines for Visual Search Reranking. In *Proceedings of ACM Special Interest Group on Information Retrieval*, pages 500–507, 2009.

[17] Y. Liu, T. Mei, X. Wu, and X.-S. Hua. Optimizing video



(a) The query "wolves."

(b) The query "cat."

Figure 6: Examples of different methods. [Best viewed in color].

search reranking via minimum incremental information loss. In *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, pages 253–259, 2008.

[18] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1150–1157, 1999.

[19] T. Mei, X.-S. Hua, W. Lai, L. Yang, and et al. MSRA-USTC-SJTU at TRECVID 2007: High-Level Feature Extraction and Search. In *TREC Video Retrieval Evaluation Online Proceedings*, 2007.

[20] A. Natsev, A. Haubold, J. Tešić, L. Xie, and R.Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *ACM International Conference on Multimedia*, pages 991–1000, Augsburg, Germany, 2007.

[21] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[23] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *ACM International Conference on Multimedia*, pages 131–140, 2008.

[24] G. Wang and D. Forsyth. Object image retrieval by exploiting online knowledge resources. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[25] M. Wang, L. Yang, and X.-S. Hua. MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval. In *Microsoft Technical Report, MSR-TR-2009-30*, 2009.

[26] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia content analysis-using both audio and visual clues. *Signal Processing Magazine*, 17(6):12–36, Feb 2000.

[27] Wikipedia. http://www.wikipedia.org/.

[28] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *ACM International Conference on Image and Video Retrieval*, pages 238–247, 2003.