Anchor Concept Graph Distance for Web Image Re-ranking

Shi Qiu[1], Xiaogang Wang[2,3], and Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
{qs010,xtang}@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

ABSTRACT

Web image re-ranking aims to automatically refine the initial text-based image search results by employing visual information. A strong line of work in image re-ranking relies on building image graphs that requires computing distances between image pairs. In this paper, we present Anchor Concept Graph Distance (ACG Distance), a novel distance measure for image re-ranking. For a given textual query, an Anchor Concept Graph (ACG) is automatically learned from the initial text-based search results. The nodes of the ACG (*i.e.*, anchor concepts) and their correlations well model the semantic structure of the images to be re-ranked. Images are projected to the anchor concepts. The projection vectors undergo a diffusion process over the ACG, and then are used to compute the ACG distance. The ACG distance reduces the semantic gap and better represents distances between images. Experiments on the MSRA-MM and INRIA datasets show that the ACG distance consistently outperforms existing distance measures and significantly improves start-of-the-art methods in image re-ranking.

Categories and Subject Descriptors

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

Keywords

Web image search re-ranking; anchor concept graph; semantic gap

1. INTRODUCTION

The fast growing number of web images spurs the rapid development of image search engines, such as Google, Bing, and Flickr. Due to the success of text retrieval techniques in searching web pages and their efficiency, most web image search engines return images based on textual information including surrounding texts, titles, URLs, and user-given tags. However, text-based image search often suffers from noise in the textual information as well as the discrepancy between textual and visual contents, yielding unsatisfactory results. For example in Figure 1, when users search for the 'panda' animal, unexpected images of cars will also come out in the text-based search results. Researchers therefore employ visual information to re-rank the initial text-based search results. The aim is to boost images relevant to the textual query to top ranks, and lower down irrelevant ones without further user intervention.

A variety of methods have been proposed for web image search re-ranking. Topic models [1, 2] learn the latent topics among returned images and perform re-ranking based on the probability that each image belongs to the dominant topic. Such methods are effective in handling object-like queries, but may fail on general web images or when relevant images undergo significant variations because of their low discriminative power. Classification-based methods first train a discriminative model such as SVM [16] and boosting [15] which are further used to predict the relevance scores of returned images. The drawback of this type of methods is that they adopt pseudo-relevance feedback (PRF) to acquire training samples which are usually unreliable. There are also approaches [6] that train query-relative classifiers and do not involve PRF. However, they are shown to suffer from the overfitting problem [7].

Recently, graph-based methods have drawn increasing attention in image re-ranking [12, 4, 9, 7], as they are able to capture manifold structures underlying imagery data and provide a nice framework to integrate the initial text-based ranking information and visual consistency between images [17]. These methods mostly build on the assumption that the relevant images form a compact cluster in the search results. They typically involves (1) constructing a graph by computing distances of image pairs; (2) detecting confident samples (compact clusters) based on the graph; and (3) propagating the scores of confident samples over the graph. Therefore, the key to the success of graph-based methods lies in a proper distance measure between images, as it determines the structure of the graph which the rest of the methods build on. A good distance measure would close the semantic gap and reveal the manifold structure formed by images to be re-ranked. Despite of its importance in re-ranking algorithms, the choice of distance measure for graph construction is not well solved yet. Jing and Baluja [5] used the portion of matched SIFT features as the similarity between an image pair. It is effective only in ranking images of particular types of objects. Most methods [12, 9, 7] adopt distance between low-level features such as histograms of visual words or color moments. These features generally do not correlate well with the semantics of images, especially when the returned images are diverse in visual content.

In this paper, we propose the Anchor Concept Graph Distance (ACG Distance), an effective distance measure for graph construction in image re-ranking. A graphical illustration is shown in Figure 1. Instead of directly computing distances with low-level features, our proposed distance is computed using a semantic representation of the initial search results called Anchor Concept Graph (ACG). The ACG is automatically learned from the initial search results. Its nodes (called Anchor Concepts) are a set of concepts relevant

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502186 .

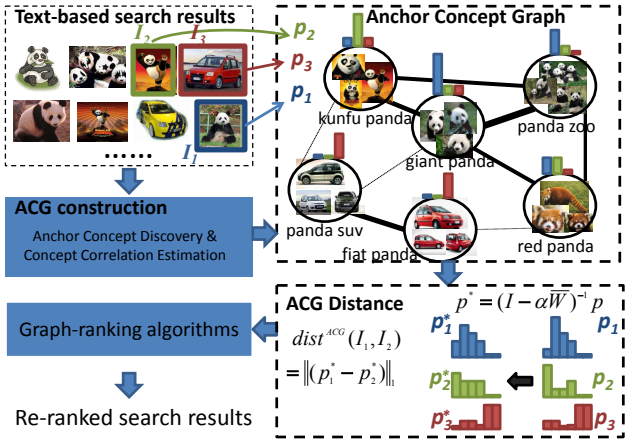

Figure 1: Illustration of the proposed ACG distance for web image search re-ranking. An anchor concept graph is learned from the initial text-based search results of a given query, *e.g.*, panda. Images are projected to the anchor concepts. Image distances are computed using the smoothed concept projections over the anchor concept graph.

to the returned search results. The correlation of these anchor concepts are represented by edges of the ACG. The visual features of images are then projected to the nodes of the ACG, forming a high-level representation that encodes the association of an image with the anchor concepts. The projection vector is further smoothed with a diffusion process which propagates information along the edges of the ACG and the distance between two smoothed projection vectors is defined as the ACG distance of two images. Since the ACG models the structure of initial search results at the semantic level, our ACG distance better reflects the semantic relevance of images and the image graphs constructed from it are more effective in image re-ranking. We demonstrate the effectiveness of our approach on two benchmark datasets and show that ACG distance leads to consistent and significant improvement of state-of-the-art image re-ranking methods.

2. ACG CONSTRUCTION

2.1 Anchor Concept Discovery

Given a textual query q, *e.g.*, "panda", in conjunction with the returned images $\mathcal{I}_q$ and their surrounding texts $\mathcal{T}_q$, a set of anchor concepts, such as "giant panda", "kungfu panda", and "panda suv", can be automatically discovered by employing both textual and visual information. These anchor concepts can more accurately model the visual and semantic content of the diverse images to be re-ranked. The anchor concepts are discovered through query expansions. The key idea is to identify expanded query keywords that occur frequently in visually similar images, as those keywords are more likely to correlate with the visual contents and thus can be viewed as descriptive concepts. The details are provided in Algorithm 1. The learned concepts are denoted as $\mathcal{C}_q = \{c_i\}_{i=1}^{M_q}$. In this paper, we set K, T, and M_q to 15, 3, and 25, respectively.

2.2 Estimating Concept Correlation

The anchor concepts are not isolated. Some anchor concepts are more semantically related, *e.g.*, "giant panda" and "kunfu panda", while some are less related, *e.g.*, "giant panda" and "panda suv". Such correlations are useful in determining image similarities as explained in Section 3.2. We adopt the Google Kernel to estimate correlations among anchor concepts. Google Kernel was first proposed by Sahami *et al.* [10] to measure the similarity of two short texts (typically a short text contains a number of keywords) at the semantic level. For a anchor concept c_i, a set of Google snippets $S(c_i)$ is obtained from the Google web search. A Goolge snippet is a short text summary generated by Google for each search result item with query c_i. We collect the snippets of the top N search result items, and they provide richer semantic context for the anchor concept c_i. We can more robustly determine the similarity between c_i and c_j by computing the textual similarity between $S(c_i)$ and $S(c_j)$ using the traditional term vector model and cosine similarity. We represent the learned correlations by a matrix W, where W_{ij} is the correlation value of c_i and c_j, and W_{ii} is set to 1.

Algorithm 1 Concept Discovery through Query Expansion

Require: Query q, image collection $\mathcal{I}_q$, surrounding texts $\mathcal{T}_q$.
Ensure: Learned concept set $\mathcal{C}_q = \{c_i\}_{i=1}^{M_q}$.
1: **Initialization:** $\mathcal{C}_q := \emptyset$, $r_I(w) := \mathbf{0}$.
2: **for all** images $I_k \in \mathcal{I}_q$ **do**
3: Find the top K visual neighbors, denote as $\mathcal{N}(I_k)$
4: Let $W_{I_k} = \{w_{I_k}^i\}_{i=1}^{T}$ be the T most frequent words in the surrounding texts of $\mathcal{N}(I_k)$.
5: **for all** wordS $w_{I_k}^i \in W(I_k)$ **do**
6: $r_I(w_{I_k}^i) := r_I(w_{I_k}^i) + (T - i)$.
7: **end for**
8: **end for**
9: Combine q and top M_q words with largest $r_I(w)$ to form $\mathcal{C}_q$.

3. RE-RANKING WITH ACG DISTANCE

3.1 Concept Projection

Once the ACG is built, we project images to be re-ranked to the M_q anchor concepts on the ACG, and obtain a high-level representation of images called concept projection. A concept projection of an image is a M_q dimensional vector that encodes its association with each of the anchor concepts. It reduces the semantic gap as it allows images with large difference in visual features to map to similar projection vectors. The projection is done using an multi-class SVM which requires an off-line training stage. Training samples for the anchor concepts are collected by a second round of querying to the search engine. The top N returned images are retained as positive samples for each anchor concept[1]. As the anchor concepts are much less ambiguous than the original queries, their search results are compact and compact enough to serve as training samples. The multi-class SVM is then learned with these training samples. When the training is finished, the learned model is applied to images to be re-ranked, and the probabilistic outputs of the SVM [14] are used as the concept projections.

3.2 ACG distance

The ACG distance for images is then calculated with the concept projection on the anchor concept graph. It is straightforward to define the L_1-distance between two concept projections as their distance. However, this distance is not optimal in that it treats each dimension of the concept projection independently, and the correlations between anchor concepts are ignored. Consider an extreme case that I_1 is an image ideally associated with the anchor concept of "giant panda" (thus its concept projection is a vector with an one in the dimension of "giant panda" and zeros elsewhere). I_2 and I_3 are images ideally associated with "kungfu panda" and

[1] N is fixed to 300 in our experiments. We observe the performance of our approach is not sensitive to the value of N.

Table 1: Performance of different re-ranking methods on the MSRA-MM and the INRIA datasets. Values in the parentheses are the NDCG@k / MAP improvements over initial search results. M: results on the MSRA-MM dataset. I: results on the INRIA dataset.

	Initial	IB[3]	Bayesian [12]	QRel [6]	RWalk [4]		KernelTopN [9]		KernelBVLS [9]	
					original	ACG	original	ACG	original	ACG
NDCG@10 (M)	0.582	0.594(0.012)	0.592(0.010)	–	0.597(0.015)	**0.621(0.039)**	0.598(0.016)	**0.647(0.065)**	0.645(0.063)	**0.657(0.075)**
NDCG@50 (M)	0.556	0.570(0.014)	0.565(0.009)	–	0.563(0.007)	**0.586(0.030)**	0.564(0.008)	**0.610(0.054)**	0.587(0.031)	**0.623(0.067)**
NDCG@100 (M)	0.536	0.554(0.018)	0.559(0.023)	–	0.550(0.014)	**0.580(0.044)**	0.560(0.024)	**0.607(0.071)**	0.580(0.044)	**0.617(0.081)**
MAP (I)	0.570	0.616(0.046)	0.658(0.088)	0.649(0.079)	0.637(0.067)	**0.650(0.080)**	0.661(0.091)	**0.674(0.104)**	0.674(0.104)	**0.689(0.119)**

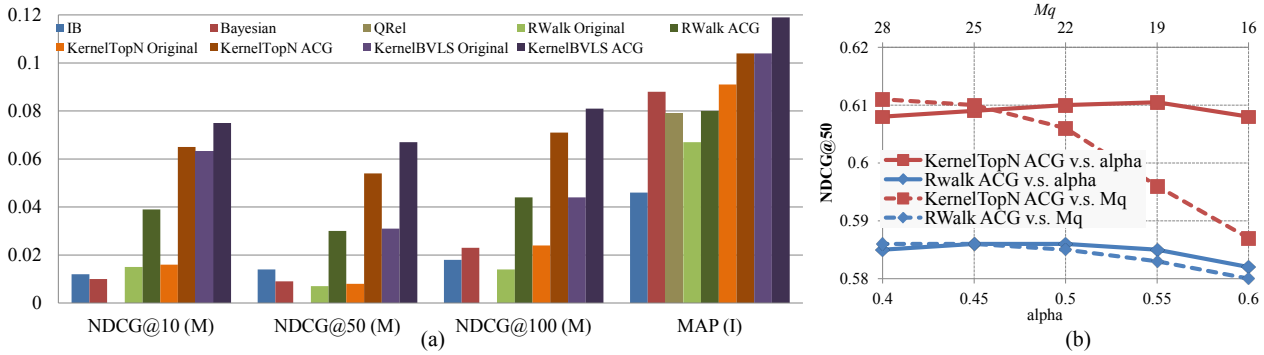

Figure 2: (a) The NDCG@k and MAP improvements over the initial text-based search results on the MSRA-MM and the INRIA datasets. M: results on the MSRA-MM dataset. I: results on the INRIA dataset (b) The performance on the MSRA-MM dataset when changing the value of parameters α and M_q.

"panda suv" respectively. When using L_1-distance, I_2 and I_3 have equal distance to I_1. However, it is clear to us that I_2 should be more similar to I_1 because they are both related to the panda animal while I_3 is related to vehicle. This limitation can be remedied by a smoothing operation on the concept projection before computing the L_1-distance. Concretely, we consider a diffusion process [17] on the ACG that gradually propagate association with one anchor concept (*e.g.* "giant panda") to other correlated concepts (*e.g.* "kungfu panda"). Let p_1 denote the concept projection of I_1, the smoothed concept projection is given by

$$p_1^* = \sum_{n=0}^{\infty} \alpha^n \bar{W}^n p_1 = (I - \alpha\bar{W})^{-1} p_1, \quad (1)$$

where $\bar{W}$ is the column-normalized correlation matrix, *i.e.*, $\bar{W} = WD^{-1}$ and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{M_q} W_{ji}$. α $(0 < \alpha < 1)$ is a damping factor that controls the diffusion rate.

The ACG distance is calculated as $\|p_1^* - p_2^*\|_1$. After dropping constant terms, we have

$$dist^{ACG} = \|(I - \alpha\bar{W})^{-1}(p_1 - p_2)\|_1 \quad (2)$$

The ACG distance can be used with any existing re-ranking methods that rely on pairwise image distances. As the semantic gap is reduced, the manifold structure underlying the returned images are better captured by the ACG distance. State-of-the-art methods can be significantly improved by the ACG distance as shown in the following section.

4. EXPERIMENTS

We conduct experiments on two benchmark datasets for web image re-ranking: INRIA [6] and MSRA-MM V1.0 [13], with 68 and 352 queries respectively. We report the mean average precision (MAP) [6] for the INRIA dataset, and NDCG@k [9] for MSRA-MM as its images are labelled with multiple relevance levels. In Section 4.1, we first combine the ACG distance with existing algorithms and show that the image re-ranking performance can be significantly improved. Several state-of-the-art re-ranking methods are also compared. Section 4.2 compares the ACG distance with other measures directly computed from low-level visual features and/or textual features. It shows that the ACG distance can better capture image semantics. The parameter α is fixed as 0.5 in both experiments. The sensitivity to α is investigated in Section 4.3.

4.1 Comparison with Existing Methods

We combine the ACG distance with state-of-the-art methods by replacing their graph construction component. We use the ACG distance to compute pairwise image distances, and keep the rest of the methods unchanged. We test on three re-ranking approaches: random walk (RWalk) [4], kernel-based re-ranking by taking top N images as confident samples (KernelTopN) [9], and kernel-based re-ranking by detecting confident samples based on bounded variable least square (KernelBVLS) [9]. The parameters involved in these methods are selected according to the original papers. We experiment on both INRIA and MSRA-MM and summarize the performance of the initial text-based search result, the three original approaches, and their counterparts with our ACG distance in Table 1. Three other methods, Information Bottleneck (IB) [3], Bayesian Visual Re-ranking (Bayesian) [12] and Query-relative Classifier (QRel) [6] are also included for comparison.

From Table 1, it is clear that our ACG distance consistently improves upon the state-of-the-art methods on both datasets. Figure 2 (a) shows the improvement of above methods over the initial search results. We can see that the ACG distance is very effective. On MSRA-MM, the NDCG@50 improvements over the initial search result are originally 0.007, 0.008, and 0.031 by using RWalk, KernelTopN and KernelBVLS, and the improvements are boosted to 0.030, 0.054, and 0.067 by using the ACG distance. Over 100% relative improvement are obtained. Some re-ranking results are shown in Figure 4.

4.2 Comparison of Different Distances

Next, we compare the ACG distance with other distance/similarity measures which are based on low-level visual features and/or textual features. Five different measures are used for graph construction in two re-ranking methods, RWalk [4] and KernelTopN [9]. The measures under comparison are (1) SpmSIFT: L_2-distance between spatial pyramids of the bag-of-words SIFT descriptors used

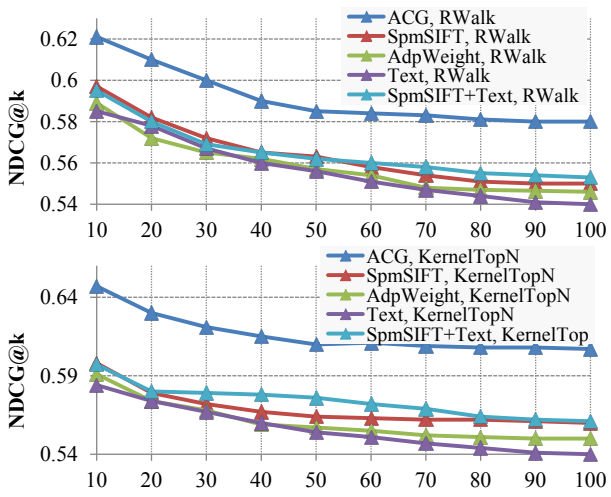

Figure 3: The Performance of four similarity/distance measures on the MSRA-MM dataset. Top: results using RWalk[4]. Bottom: results using KernelTopN [9].

in [9], (2) AdpWeight: distances of multiple visual features with adaptive weighting proposed in [11], (3) Text: cosine similarity between L_2-normalized word histograms extracted from the surrounding texts with tf-idf weighting [8], (4) SpmSIFT+Text, the weighted combination of SpmSIFT and Text, where the weight is optimally tuned, and (5) our ACG distance. The first three represent distance/similarity measures based on local visual feature, global visual feature, and textual feature respectively.

The re-ranking result is given in Figure 3. We can see that our ACG distance consistently outperforms the other measures. Moreover, it is observed that the performance gap increases with the ranking position, indicating that the graph constructed by our ACG distance works more robustly in propagating the relevance information along the semantic manifold. We therefore conclude that the ACG distance is more effective in reducing the semantic gap.

4.3 Parameter Sensitivity

We investigate how the free parameters α in Eqn. 2 and M_q, the number of anchor concepts, will affect the re-ranking performance. We change the value of α and M_q and use the corresponding versions of ACG distance to re-rank images. The NDCG@50 on MSRA-MM by two methods are reported in Figure 2 (b). It can be observed that the performance of ACG distance is robust against small changes in α, and that adding more concepts will always increase the performance. However, the performance gain diminishes quickly after M_q reaches around 25.

5. CONCLUSIONS

In this paper, we propose the ACG distance for web image search re-ranking. The proposed distance is computed over an Anchor Concept Graph which is automatically learned from the initial search results and well models the semantic structure of images to be re-ranked. Our ACG distance better computes image similarities by reducing the semantic gap. It is applicable to a variety of re-ranking approaches that rely on constructing graphs on images. We conduct experiments on two public benchmark datasets, and show that the proposed distance consistently outperforms commonly used distance/similarity measures, and significantly improves the start-of-the-arts methods in web image search re-ranking.

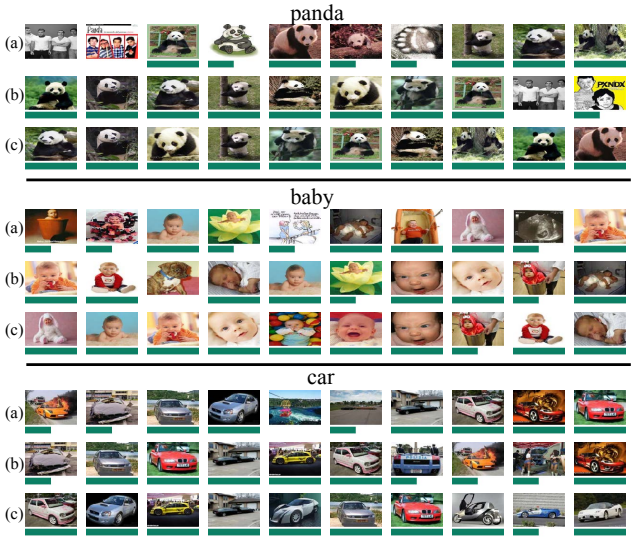

Figure 4: Top 10 re-ranked images for three queries in the MSRA-MM dataset. The green bar underneath each image indicates its groundtruth relevance level. (a) Initial text-based search. (b) KernelTopN [9] with original distance measure. (c) KernelTopN with ACG distance.

6. REFERENCES

[1] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, 2005.

[2] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *CVPR*, 2008.

[3] W. Hsu, L. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *ACM MM*, 2006.

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

[5] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *TPAMI*, 30:1877–1890, 2008.

[6] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web image search results using query-relative classifiers. In *CVPR*, 2010.

[7] W. Liu, Y. Jiang, J. Luo, and S.-F. Chang. Noise resistant graph ranking for improved web image search. In *CVPR*, 2011.

[8] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[9] N. Morioka and J. Wang. Robust visual reranking via sparsity and ranking constraints. In *ACM MM*, 2011.

[10] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW*, 2006.

[11] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang. Intentsearch: Capturing user intention for one-click internet image search. *TPAMI*, 34:1342–1353, 2012.

[12] X. Tian, L. Yang, J. Wang, X. Wu, and X.-S. Hua. Bayesian visual reranking. *TMM*, 13:639–652, 2010.

[13] M. Wang, L. Yang, and X.-S. Hua. Msra-mm: Bridging research and industrial societies for multimedia information retrieval. Technical report, Microsoft Research Asia, 2009.

[14] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *JMLR*, 5:975–1005, 2004.

[15] R. Yan and A. Hauptmann. Co-retrieval: A boosted reranking approach for video retrieval. In *Proc. CIVR*. Springer, 2004.

[16] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proc. CIVR*, 2003.

[17] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *NIPS*, 2004.