A document expansion framework for tag-based image retrieval

Wei Lu and Heng Ding
School of Information Management, Wuhan University, Wuhan, China, and
Jiepu Jiang
College of Information and Computer Sciences, University of Massachusetts, Amherst, Massachusetts, USA

Received 28 May 2017
Revised 27 November 2017
Accepted 11 December 2017

Abstract
Purpose – The purpose of this paper is to utilize document expansion techniques for improving image representation and retrieval. This paper proposes a concise framework for tag-based image retrieval (TBIR).
Design/methodology/approach – The proposed approach includes three core components: a strategy of selecting expansion (similar) images from the whole corpus (e.g. cluster-based or nearest neighbor-based); a technique for assessing image similarity, which is adopted for selecting expansion images (text, image, or mixed); and a model for matching the expanded image representation with the search query (merging or separate).
Findings – The results show that applying the proposed method yields significant improvements in effectiveness, and the method obtains better performance on the top of the rank and makes a great improvement on some topics with zero score in baseline. Moreover, nearest neighbor-based expansion strategy outperforms the cluster-based expansion strategy, and using image features for selecting expansion images is better than using text features in most cases, and the separate method for calculating the augmented probability $P(q|R_D)$ is able to erase the negative influences of error images in R_D.
Research limitations/implications – Despite these methods only outperform on the top of the rank instead of the entire rank list, TBIR on mobile platforms still can benefit from this approach.
Originality/value – Unlike former studies addressing the sparsity, vocabulary mismatch, and tag relatedness in TBIR individually, the approach proposed by this paper addresses all these issues with a single document expansion framework. It is a comprehensive investigation of document expansion techniques in TBIR.

Keywords Information retrieval, Document expansion, Retrieval model, Social image representation, Social image retrieval, Tag-based image retrieval

Paper type Research paper

Introduction
The development of digital photography and social media-sharing platforms (e.g. Flickr and Instagram) has led to a rapid increase in the number of social images produced. Social bookmarks (tags) provide noisy, yet useful descriptive information to enhance traditional image retrieval technology (Firan *et al.*, 2007; Nov *et al.*, 2008; Sun, Bhowmick, Nam Nguyen, and Bai, 2011). Techniques leveraging social bookmarks for image search are called tag-based image retrieval (TBIR), which have attracted wide attention (Chen *et al.*, 2010; Gao *et al.*, 2013; Li *et al.*, 2015; Li and Snoek, 2010). These approaches are general methods for assessing the similarity between a search query and an image's tags.

Previous studies showed that social tags are usually helpful for image retrieval (Chen *et al.*, 2010; Gu *et al.*, 2011; Liu *et al.*, 2009; Sun and Bhowmick, 2008; Tang *et al.*, 2009). However, a social image usually only has a limited number of tags. For example, in the NUS-WIDE data set (an open data set for TBIR), each image has only 18 tags on average, and almost 15 percent of images own less than 8 tags. Such a tag-based image representation often suffers from serious sparsity and vocabulary mismatch issues. In addition, most image-sharing platforms do not allow users to assign the same tag multiple times, which makes it difficult to distinguish informative tags from less important

Aslib Journal of Information Management
Vol. 70 No. 1, 2018
pp. 47-65
© Emerald Publishing Limited
2050-3806
DOI 10.1108/AJIM-05-2017-0133

ones by their frequencies. Therefore, measuring the degree of effectiveness of a tag describing the tagged image also becomes a crucial issue (this we refer to as tag-relatedness issue in this paper). Many studies have been conducted to address these issues. More concretely, neighbor voting schemes (Truong *et al.*, 2012) are widely adopted to measure the degree of effectiveness of a tag describing the tagged image. Tag recommendation (Sun, Bhowmick and Chong, 2011) and tag completion (Wu *et al.*, 2013) are both put forward in addressing issues of sparsity and vocabulary mismatch. However, it is still very hard to combine these methods into a uniform framework. In this paper, we propose a concise framework based on document expansion techniques widely adopted in document retrieval to address all these issues at once. In our approach, we consider the set of tags for an image as a "document" for that image. Specifically, our approach has three core components:

(1) a strategy of selecting expansion (similar) images from the whole corpus;

(2) a technique for assessing image similarity, which is adopted for selecting expansion images; and

(3) a model for matching the expanded image representation with the search query.

We describe and evaluate our approach in this paper. We compare it with previous approaches and experiment using different implementations of the three core components. The rest of this paper is organized as follows. In the second section, we review the related work on TBIR from social tags research, related efforts on image retrieval, and research on tag-based retrieval. The third section provides a detailed description of the proposed approach. The fourth section introduces the experimental setup and analyzes the results in detail. The fifth section concludes this study.

It is worth noting that TBIR is quite different from concept-based (i.e. text-based) image retrieval, because of some characteristics of social tags. For example, in concept-based image retrieval, an image is often represented by a textual document that typically has much redundancy of words to convey its semantics. However, in TBIR, an image is represented with many fewer tags with no or minimal redundancy. Moreover, text used in concept-based image retrieval is usually provided by professional indexers, but social tags are assigned by different users having different motivations, different interpretations of the meaning of tags. Thus, traditional techniques of concept-based image retrieval, such as term frequency weighting and document length normalization, do not work well on TBIR.

Related work

Research on social tags in the search environment

Much research has examined social tags from the perspective of organization and retrieval. For example, Nov *et al.* (2008) divided tagging motivation into three categories based on target audience and tagging function into two dimensions based on a tag's intended use. They pointed out that the organization function of tags is intended to facilitate future search and retrieval by the user. Carman *et al.* (2008) found that social tags (bookmarks) are useful for approximating actual user queries from the perspective of personalized information retrieval. Gu *et al.* (2011) concluded that social tags reveal confidence issues caused by ambiguity and synonymy. They proposed a statistic model to measure the confidence of social tags and applied it to filter noisy tags with low tag confidence. The results of their experiment revealed that confidence of social tags highly influenced the performance of tag-based search methods. Wu *et al.* (2013) stated that "since many users tend to choose general and ambiguous tags in order to minimize their efforts in choosing appropriate words, tags that are specific to the visual content of images tend to be missing or noisy."

Additionally, Koutrika *et al.* (2008) asserted that misleading tags confuse users instead of increasing the visibility of some resource. Therefore, they proposed a method for ranking documents matching a tag based on taggers' reliability. Li *et al.* (2009) stated that various tagging motivations naturally lead to the personalization characteristic of social tags and create an unreliable interpretation of the relevance of a tag with respect to the visual content that it is describing. Hence, the fundamental problem in TBIR is how to reliably estimate the relevance of a tag with respect to the visual content that it is describing. Note that the above characteristics of social tags make TBIR more challenging than concept-based image retrieval, and demand a revisit rather than directly employing techniques of concept-based image retrieval (Sun, Bhowmick, Nam Nguyen, and Bai, 2011).

Related efforts on image retrieval

Inspired by research on text-information retrieval, many methods (sometimes called text-based image retrieval) have been developed to improve image search in cases in which textual descriptions of visual content are vague (La Cascia *et al.*, 1998; Sclaroff *et al.*, 1999). However, considering the problem of subjectivity in contextual information, text-based image retrieval also possesses some limitations (Inoue, 2004). To overcome the problem in text-based image retrieval, many efforts (often termed content-based image retrieval (CBIR)) attempt to utilize visual content for estimating image visual similarity (Gudivada and Raghavan, 1995; Smeulders *et al.*, 2000). One primary goal of these studies is to measure the similarity between two images based on their level features (color, texture, and shape) and semantic content (object, scene, and emotion).

More recently, advancements in image understanding, such as image classification and object recognition, have made it possible to learn and understand visual concepts from images (Krizhevsky *et al.*, 2012; Torralba *et al.*, 2008). For example, automatic image annotation focuses on assigning a few relevant and controlled keywords to an unannotated image, and then these keywords can be indexed and utilized for image retrieval. Jeon *et al.* (2003) proposed a cross-media retrieval model for annotating images with keywords from a small vocabulary of blobs[1], and their experiment demonstrates the usefulness of these keywords for the task of image retrieval.

Research on tag-based retrieval

On the other hand, with the popularity of folksonomy, uncontrolled, and personalized social tags as metadata, we see new opportunities to enhance current retrieval technology. Bao *et al.* (2007) proposed two algorithms that use social tags for web search, and they found that social tags are usually good summaries of corresponding web pages and the count of tags indicates the popularity of web pages. Xu *et al.* (2008) proposed a personalized search framework to utilize folksonomy (social tags) for personalized web search. Melenhorst *et al.* (2008) reported a study on tag-based video retrieval, and their experiment suggested that uncontrolled social tags are valuable for supporting video retrieval processes. Hsieh and Hsu (2010) proposed a method to annotate images with social tags; their method is able to solve the sparsity of user-contributed tags. Sevil *et al.* (2010) presented an automatic tag expansion approach, which is valuable for image retrieval. Efron (2010) proposed a language modeling (LM) approach to retrieve useful hashtags from posts in a microblogging environment. Inspired by document expansion and inverted index in traditional IR, Min *et al.* (2010) revisited document expansion in the context of retrieval of images annotated with brief textual labels. Zhu *et al.* (2010) and Sang *et al.* (2012) introduced the task of tag refinement which aims to solve the imprecise and incomplete issues of social tags. Lee *et al.* (2012) proposed a social inverted index for social-tagging-based IR. Li and Snoek (2013) developed a system that has the ability to select the most relevant positive and negative examples for a given tag. Recently, Li *et al.* (2016) presented a comprehensive treatise about image tag assignment, refinement, and TBIR. Lu *et al.* (2016) proposed a re-ranking approach depending on three steps: the first step, "Keyword matching," returns all images that contain the query terms and the images uploaded by the same user, grouped into user image set; then the second step, "Inter-user re-ranking," ranks user image sets by considering users' contributions to the query; and finally, the third step "Intra-user re-ranking," selects the image which has the highest score among each user image sets. A state-of-the-art research in TBIR has been performed by Sun, Bhowmick, Nam Nguyen, and Bai (2011), which quantifies the relevance score between a tagged image and a tag query by five orthogonal dimensions:

(1) tag discrimination: analogous to the idea of tf-idf in traditional IR;

(2) tag length: used to reflect the impact of the number of tags assigned to social images;

(3) tag query matching score: quantifying the matching score between a tag and the query tag t_q;

(4) query model: used for rewriting a given query, analogous to query expansion technology in traditional IR; and

(5) tag relatedness: used to measure the degree of effectiveness of a tag describing the tagged image.

Document expansion approach

Task definition

Our approach addresses the TBIR problem. TBIR refers to an image retrieval task where the images often have user-generated short text descriptions (tags). A tag is usually a single word, but can be more complex, e.g., "xmas2015." However, our approach does not consider the latter case because the majority of the tags in our data set belong to the first case – a tag is considered as one word token in this paper.

TBIR has been widely applied on image-sharing platforms in different scenarios to help users look for images. For example, the famous image-sharing website flickr.com provides two kinds of TBIR. First, users can type a text query (one or a few words) in the search bar to find relevant images annotated by other persons using similar words (social tags). Second, while browsing images, users can click on the social tags associated with the images, and the system will retrieve relevant images using the clicked tag as a query. In this paper, we only consider the second scenario.

We formally define the TBIR problem as follows. The corpus is a collection of N social images $C = \{D_1, \ldots, D_N\}$, where each image D_k is associated with a set of m tags $\{w_{k1}, w_{k2}, \ldots, w_{km}\}$. Given a query q containing s words (tags) $\{w_1, \ldots, w_S\}$, the task is to rank images by their relevance to the query.

TBIR techniques focus on improving retrieval accuracy using tag information. We can consider the set of tags for an image as a complementary representation for that image, in addition to other representations such as its content.

Specifically, we adapt document expansion techniques for the TBIR problem to address the vocabulary mismatch issue in TBIR. Here, we consider the set of tags for an image as a "document" for that image. As Sun, Bhowmick, Nam Nguyen, and Bai (2011) pointed out, the tag-based image representation (document) suffers from sparsity issues, i.e., an image is usually associated with only a few tags. This makes a query difficult to match a relevant image if they used similar but different words. In addition, most image-sharing platforms do not allow users to assign the same tag multiple times, which makes it difficult to distinguish informative tags from less important ones by their frequencies. This is also similar to the issue of lacking term frequency information in short-text retrieval (Efron *et al.*, 2012). Both issues increase the risk of vocabulary mismatch when using tag-based representation for image retrieval.

The rest of this section describes our approach.

Framework

Our system ranks a target image by the following steps:

(1) we find images similar to the target image;

(2) we compute a relevance score for the target image based on the similarity of the query to the target image itself as well as its similar images; and

(3) we rank target images by the computed relevance scores.

Figure 1 shows an example. The target image has two tags, "sky" and "helicopter," while the query includes a word "blue" that does not exist in the target image's tag-based representation. Apparently, the target image is relevant, but it has a relatively low relevance score if we directly match the query with the target image's tags. Instead, our approach expands the target image's tag-based representation using similar images. For example, if a similar image has the tags blue, "sky," and "cloud," it improves the target image's representation by helping it to match the word blue in the query (and hopefully enhances retrieval performance).

Formally, we use D for the original (unexpanded) tag-based representations for an image. We use R_D for D's augmented representation based on similar images' tags. We rank images using the following equation. The parameter $\alpha_{\exp}$ controls the weight of the expanded representation:

$$\text{Score}(q, D) = \left(1 - \alpha_{\exp}\right) \cdot P(q|D) + \alpha_{\exp} \cdot P(q|R_D) \tag{1}$$

The rest of this section introduces:

(1) the strategies of selecting similar images – we compare a cluster-based strategy and a nearest neighbor-based one;

(2) the techniques used for assessing image similarity (in order to select similar images in Step 1) – we compare using image content, tags, and a combination of the two to assess image similarity; and

(3) matching the query q and the expanded image representation R_D, i.e., estimating $P(q|R_D)$ – we compare two approaches.

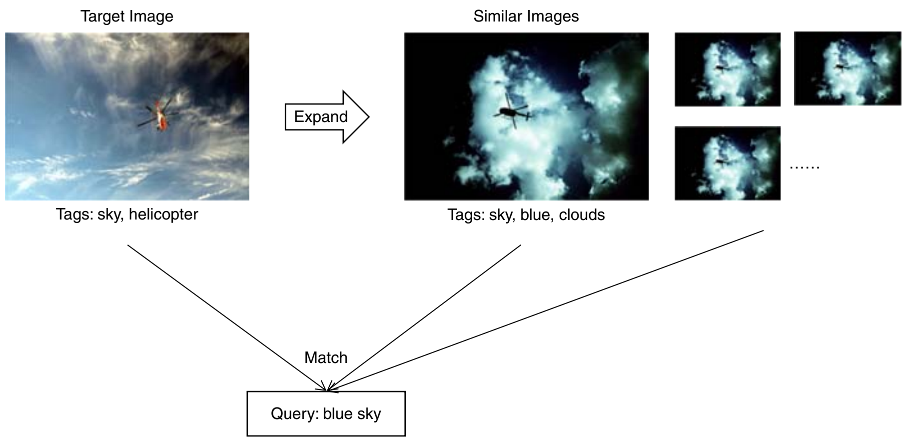


Figure 1.
An example of document expansion for tag-based image retrieval

Document expansion strategy

Inspired by previous studies on document expansion (Liu and Croft, 2004; Tao *et al.*, 2006; Wei and Croft, 2006), we compare two document expansion strategies for selecting similar images:

(1) Cluster-based strategy: we group images into clusters and select the closest cluster to the target image for expansion. We use K-means algorithm for clustering. For each image D, its expansion image set $R_D = \{D_1, D_2, \ldots, D_M\}$ consists of all images sharing the same cluster with D.

(2) Nearest neighbor strategy: the nearest neighbor strategy selects the most similar k images to the target image for expansion. We construct a pseudo query Q_D based on the target image's representation and retrieve the top k similar images. The expansion image set $R_D = \{D_1, D_2, \ldots, D_k\}$ consists of the top k similar (relevant) images retrieved for Q_D, where each image is associated with a similarity score.

The intuition behind these two expansion strategies is quite different. The cluster-based method assumes that different images in the same cluster belong to the same topic (Liu and Croft, 2004). It does not differentiate images in the same cluster while performing expansion – each expansion image has an equal weight in R_D. In contrast, the nearest neighbor strategy specifically retrieves the most similar k images for the target image for expansion. Different images are assigned different weights – their relevance scores. It assumes that more similar images provide better complementary representations for the original image (Tao *et al.*, 2006).

Image similarity

Both the cluster-based and the nearest neighbor-based expansion strategies require techniques for assessing image similarity. We compare three different ways of assessing image similarity in this paper:

(1) Text/tag-based approach – assessing image similarity only based on tag-based representation:

- While using the cluster-based strategy, we represent each image as a vector of tags, and cluster images using K-means algorithm with cosine distance.
- While using the nearest neighbor strategy, we construct a text query for the target image using the combination of its tags. We submit the query to a text retrieval IR system (such as Indri or Lucene) to obtain a ranked list of relevant (similar) images. We set the weights of the images to their relevance scores returned by the text IR system while performing expansion.

(2) Image feature-based approach – assessing image similarity only based on image content:

- While using the cluster-based strategy, we represent each image's visual features using a 500-dimension bag of "words" feature based on scale-invariant feature transform (SIFT) descriptions (Kulis and Grauman, 2009). We also use K-means algorithm with cosine distance for clustering. Although there are various visual features for representing image content, it is beyond the scope of this paper.
- While using the nearest neighbor strategy, we use the image's visual feature as a query to retrieve similar images in a CBIR system. We retrieve images using locality sensitive hashing (Jégou *et al.*, 2010) and cosine distance. Although there are various approximate nearest neighbor methods for searching k nearest neighbor images and measuring image similarity based on visual features, it is beyond the scope of this paper.

(3) Mixed feature-based approach – assessing image similarity based on the combination of tag-based representation and image content:

- While using the cluster-based strategy, we construct a mixed representation for each image by concatenating its text feature vector and visual feature vector. Then, we use K-means algorithm with cosine distance for clustering.
- While using the nearest neighbor strategy, we combine the lists of similar images retrieved using the tag-based approach and the image feature-based approach using CombMNZ, a popular method for fusing ranked lists (Fox and Shaw, 1994). The following equation explains the score of an image D_i by CombMNZ, where $F(D_i)$ refers to the number of times the image D_i appeared in the two ranked lists; $S_{\text{text}}(D_i)$ and $S_{\text{image}}(D_i)$ are the scores returned by the text IR system and CBIR system, respectively:

$$S_{\text{combMNZ}}(D_i) = F(D_i) \times \left(S_{\text{text}}(D_i) + S_{\text{image}}(D_i)\right) \tag{2}$$

Matching queries and expanded image representation

Let $R_D = \{D_1, D_2, \ldots, D_M\}$ be the set of expansion images selected using the approaches described in the previous sections. Each D_i in R_D is associated with a weight – the importance of D_i in expansion. While using the cluster-based strategy, the weight of each image is set to 1. The weight of an image is set to its relevance score if we adopt the nearest neighbor strategy for image expansion.

We compare two methods for computing the probability of a query q given R_D:

- "Merging": in this approach, we merge all "documents" (images' tag sets) in R_D as a big "document" D'. Thus, the probability $P(q|R_D)$ can be estimated as in the following equation, where $D' = \{t_1, t_2, \ldots, t_L\}$ is the bag of tags associated with the image set $R_D = \{D_1, D_2, \ldots, D_M\}$. L is the total number of unique tags in the image set R_D. sim (D, D_j) stands for the weight of the image D_j in the expansion set. S is the number of words (tags) in query q:

$$P(q|R_D) = P(q|D') = P(q|t_1, t_2, \ldots, t_L) = \sum_{i=1}^{S} P(w_i|t_1, t_2, \ldots, t_L)$$

$$= \sum_{i=1}^{S} \frac{\sum_{j=1}^{M} \text{freq}(w_i, D_j)\,\text{sim}(D, D_j)}{\sum_{l=1}^{L}\sum_{j=1}^{M} \text{freq}(t_l, D_j)\cdot \text{sim}(D, D_j)} \tag{3}$$

- "Separate": in this approach, we compute the probability $P(q|R_D)$ by marginalizing overall documents in R_D. We sum over the probability of q from each individual document D_i, weighted by D_i's weight in the expansion set. The following equation describes this approach:

$$P(q|R_D) = \sum_{D_i \in R_D} P(q|D_i)\,\text{sim}(D, D_i) \tag{4}$$

Experiment

The NUS-WIDE data set is an open and accessible benchmark for evaluating TBIR techniques released by the National University of Singapore. The data set incorporates

269,648 images acquired from Flickr[2] and 81 queries. We also test our methods on the Flickr51 data set, a smaller data set containing 81,541 images and 51 queries. All these queries are simple concepts such as "airport," "valley," etc. We refer to Chua *et al.* (2009) and Wang *et al.* (2010) for further details.

We compare with two baselines: an LM approach and Sun, Bhowmick, Nam Nguyen, and Bai's (2011) approach (the $Q_S R_V D_F L_S M_C$ model with the 500-dimension bag of "words" feature based on SIFT descriptions). The LM baseline simply treats the set of tags for an image as a document and ranks images by the query likelihood score. We stem words using the Krovetz stemmer. We report results using Jelinek-Mercer smoothing with $\lambda = 0.4$. Based on our observation, smoothing has very little impact on the search results in this data set. Sun, Bhowmick, Nam Nguyen, and Bai's approach ($Q_S R_V D_F L_S M_C$) quantifies the relevance score between a tagged image D and a query q as in the following equation, where $N_k(D)$ is the 100 similar images for image D based on visual similarity (500-dimension bag of "words" feature); t_j is one of the tags that belong to image D; $P(t_j|N_k(D))$ and $P(t_j)$ are the probabilities of observing tag t_j among images in $N_k(D)$ and collection C, respectively; N is the number of images in collection C; $f(t_j)$ is the number of images annotated by tag t_j in collection C; $|D|$ is the number of tags of image D; and $P(t_j|w_i)$ is the conditional probability of being tagged by t_j among the images tagged by w_i in collection C:

$$\mathrm{Score}(q, D) = \sum_{w_i \in q,\ t_j \in D} \left(0.5 + 0.5 \times \max\left(P\left(t_j \middle| N_k(D)\right) - P\left(t_j\right), 0\right)\right) \times \left(1.0 + \log \frac{N}{1 + f\left(t_j\right)}\right) \times \frac{1}{\sqrt{|D|}} \times P\left(t_j \middle| w_i\right) \quad (5)$$

For the cluster-based strategy on NUS-WIDE, we use cluster size 500, 1,000, 2,000, and 3,000. Because the scale of Flickr51 is much smaller than (one-third of) NUS-WIDE, we choose cluster size 160, 330, 660, and 1,000 while we conduct experiments on the Flickr51 data set. For the nearest neighbor strategy on two data sets, we compare using the top 10, 20, 50, and 100 similar images for expansion. We evaluate results using the following four metrics: first, mean average precision (MAP), the mean of average precision for a sample of queries, where average precision is a measure that combines recall and precision for ranked retrieval results; second, mean reciprocal rank (MRR), the average of reciprocal ranks for a sample of queries, where the reciprocal rank is the multiplicative inverse of the rank of the first correct answer; third, precision at 10 (P@10), a statistic measure that counts the number of relevant results on the top ten results; and fourth, normalized discounted cumulative gain at 10 (nDCG@10), a measure of ranking quality that considers cumulative gain at each position of ranking list; we refer to Sanderson (2010) for further details. We evaluate all methods using five-fold cross-validation. We train the best parameters (smooth parameter α, cluster size L, and the number of similar images k) by performing a grid search. We compare the two approaches by the mean values of the evaluation measures on their ranked list. We test statistical significance using paired t-test (Table I).

Results and discussion

Table II reports the evaluation results. Comparing the two baselines, we found that $Q_S R_V D_F L_S M_C$ outperforms LM on both NUS-WIDE and Flickr51 data sets. The limited performance of LM in TBIR is not surprising. LM ranks result mainly based on term frequency and document length. In the case of TBIR, tag (term) frequency is always 1, because many current systems do not allow assigning the same tag to the same image multiple times. Thus, the score is very sensitive to the number of tags associated with the image (document length). This means that LM, in the case of TBIR, will find images which

	Abbr.	Expansion strategy	Information modality	$P(q\|R_D)$
Baseline	B1: LM	–	–	
	B2: $Q_SR_VD_FL_SM_C$	–	–	
Experiment	M1: cluster + text + merging	Cluster	Text	Merging
	M2: cluster + text + separate	Cluster	Text	Separate
	M3: cluster + image + merging	Cluster	Image	Merging
	M4: cluster + image + separate	Cluster	Image	Separate
	M5: cluster + mixed + merging	Cluster	Mixed	Merging
	M6: cluster + mixed + separate	Cluster	Mixed	Separate
	M7: NN + text + merging	NN	Text	Merging
	M8: NN + text + separate	NN	Text	Separate
	M9: NN + image + merging	NN	Image	Merging
	M10: NN + image + separate	NN	Image	Separate
	M11: NN + mixed + merging	NN	Mixed	Merging
	M12: NN + mixed + separate	NN	Mixed	Separate

Table I. Baseline and experimental retrieval names and descriptions

	NUS-WIDE				Flickr51			
Method	MAP	MRR	P@10	nDCG@10	MAP	MRR	P@10	nDCG@10
B1	0.3124	0.7778	0.6272	0.6226	0.6166	0.7209	0.6686	0.5077
B2	0.3588	0.7644	0.6864	0.6866	0.7133	0.8326	0.7980	0.6837
M1	0.3383	0.6931	0.5975	0.5925	0.6261	0.6998	0.6275	0.4881
M2	0.3412	*0.8327*	0.6605	0.6347	0.6167	0.8062	0.6804	0.5550
M3	0.3382	0.7726	0.6432	0.6299	0.7582	0.8067	0.7471	0.6548
M4	0.3461	0.7802	0.6889	0.6962	0.7430	0.8655	0.8627	0.7211
M5	0.3393	0.7536	0.6037	0.6128	0.6193	0.6830	0.6510	0.5476
M6	0.3441	0.7526	0.6494	0.6751	0.6087	0.7361	0.7118	0.5515
M7	*0.3600*	0.7558	0.6970	0.6933	0.6080	0.6916	0.6275	0.4832
M8	0.3599	0.8244	0.6926	0.6826	0.5326	0.7265	0.6000	0.4962
M9	0.3557	0.8241	0.7012	0.6692	*0.7866*	0.8981	0.8784	0.7848
M10	0.3561	0.7712	*0.7358*	*0.7266*	0.7679	*0.9186*	*0.8941*	*0.8118*
M11	0.3308	0.7778	0.6486	0.6062	0.6981	0.7277	0.6294	0.5169
M12	0.3452	0.7737	0.6815	0.6609	0.5066	0.4517	0.4019	0.3261

Table II. All results on MAP, MRR, P@10, and nDCG@10

contain the query term (tag) and rank the images by document length (the number of tags associated with that image). In contrast, the $Q_SR_VD_FL_SM_C$ model, estimating the relevance between the tag and visual content of the image by nearest neighbor voting, seems to overcome the problem in the LM method.

Our approach uses "document" expansion techniques to improve image representation. It achieves better results in terms of all evaluation measures compared to the two baselines. We think that document expansion enhances the presentation of images in two ways: assigning the right and unannotated tags to the image; and increasing the weight of the right tags of the image. Figure 2 shows an instance of the situation. The left one is an original document D and the right one is one of the images in R_D. Under the framework of document expansion, the weight of "sky," "clouds," and "helicopter" will increase and an unannotated right tag "blue" will be assigned to the left image, then the left image can be retrieved by the query "blue."

Next, we will focus on discussing the differences in document expansion strategies, information modality for selection of R_D and computation methods of the augmented probability $P(q|R_D)$.

Sea sky norway clouds canon 350d king helicopter trondheim seaking am bulanse onlyyourbestshots redningshelikopter

Blue sky film clouds lomo lca crossprocess helicopter adamscoot

Figure 2.
An illustration about how document expansion benefits TBIR

Comparing document expansion strategies

We compare the cluster-based document expansion strategy and the nearest neighbor strategy in this section. Table III reports the evaluation results for the best cluster-based strategy run (M4) and the best nearest neighbor strategy run (M10) on two data sets. Results show that both document expansion strategies are at least as good as the baseline. As Table III shows, the cluster-based strategy run performs as good as $Q_SR_VD_FL_SM_C$ in terms of MRR (+2.1 percent), P@10 (+0.4 percent), and nDCG@10 (+1.4 percent) on NUS-WIDE, and brings significant[3] improvements in P@10 (+8.1 percent, $p < 0.05$) and nDCG@10 (+5.4 percent, $p < 0.05$) on Flickr51. In contrast, the nearest neighbor-based strategy significantly outperforms the $Q_SR_VD_FL_SM_C$ baseline in all terms of both P@10 (+7.2 percent, $p < 0.05$) and nDCG@10 (+5.8 percent, $p < 0.05$) on NUS-WIDE, and brings significantly improvements in MAP (+7.7 percent, $p < 0.05$), MRR (+10.3 percent, $p < 0.05$), P@10 (+12.0 percent, $p < 0.05$), and nDCG@10 (+18.7 percent, $p < 0.05$). Note that we did not see stable improvements in terms of MAP and MRR on the two data sets but in P@10 and nDCG@10.

In addition, results also suggest that the nearest neighbor-based strategy is better than the cluster-based one. This is probably because the cluster-based strategy introduces too much noise – it is over-optimistic to assume that images in the same cluster contribute the same to the expanded representation. In contrast, the nearest neighbor-based strategy overcomes this issue by weighting images differently during expansion. Figure 3 shows an example. Although both strategies expanded the wrong image D_j, in nearest neighbor-based strategy the sim(D, D_j) is 0.35, much less than 1 in cluster-based strategy.

Comparing approaches for assessing image similarity

We compare the three approaches for assessing image similarity (text, image, or mixed) in this section.

Table III.
Comparison of best results in three approaches

	NUS-WIDE				Flickr51			
Method	MAP	MRR	P@10	nDCG@10	MAP	MRR	P@10	nDCG@10
B2	*0.359*	0.764	0.686	0.686	0.713	0.833	0.798	0.684
M4	0.346	*0.780*	0.689	0.696	0.743	0.866	0.863*	0.721*
M10	0.356	0.771	*0.735**	*0.726**	*0.768**	*0.919**	*0.894**	*0.812**

Note: *$p < 0.05$ on a two pair-wise tests against the baseline

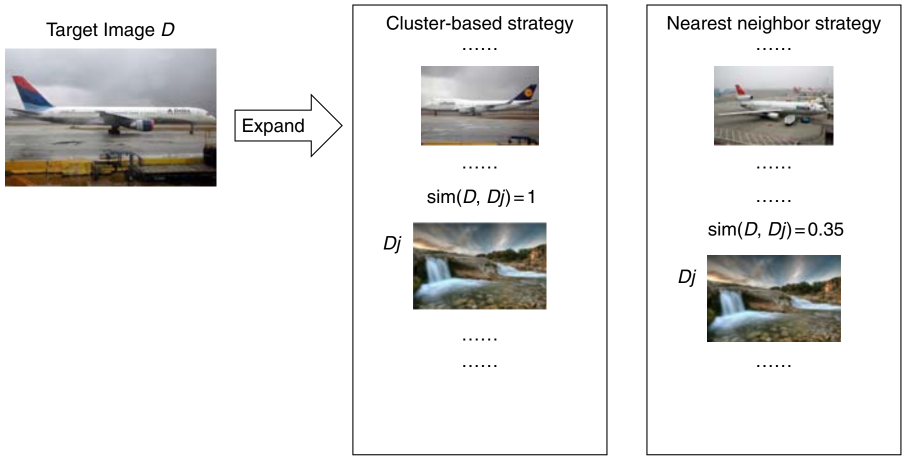


Figure 3. An illustration of why nearest neighbor-based strategy is better than cluster-based strategy

Table IV lists the average scores of different measures for runs using the text-based approach, the image-based approach, and the mixed approach. Figure 4 shows detailed results using different combinations of document expansion strategy and matching model. Results suggest that using image features for selecting expansion images is better than using text features in most cases. In addition, using mixed features does not outperform using text or image features alone.

We suspect that the sparsity issue of text/tag-based representation is a key reason for its limited performance in terms of assessing image similarity. In contrast, in such a case, using image features is usually a better option. Figure 4 shows an example. Of course, the text/tag-based representation makes it much easier to connect images and search queries. Thus, our approach is also an effective way of combining text/tag-based and content-based image retrieval – we leverage low-level image features to connect similar images and improve images' text/tag-based representation. It is worthy to note that mixed features do not demonstrate the advantage that had been expected. In the cluster-based approach, the mixed feature suffers from the curse of dimensionality. In the nearest neighbor-based strategy, the set of images expanded with image representation is quite different from the set of images expanded with tag-based representation. Therefore, the CombMNZ method fails to utilize images appeared in both two sets.

Matching queries and expanded image representation

We compare the two different approaches for computing $P(q|R_D)$ in this section – "merging" and "separate." Table V reports the differences between the two computation methods using different

Information modality	MAP	MRR	P@10	nDCG@10
NUS-WIDE				
Text (mean of M1 + M2 + M7 + M8)	*0.3499*	0.7765	0.6619	0.6508
Image (mean of M3 + M4 + M9 + M10)	0.3490	*0.7870*	*0.6923*	*0.6805*
Mixed (mean of M5 + M6 + M11 + M12)	0.3400	0.7644	0.6458	0.6387
Flickr51				
Text (mean of M1 + M2 + M7 + M8)	0.5959	0.7310	0.6339	0.5056
Image (mean of M3 + M4 + M9 + M10)	*0.7639*	*0.8722*	*0.8456*	*0.7431*
Mixed (mean of M5 + M6 + M11 + M12)	0.6082	0.6496	0.5985	0.4855

Table IV. Average scores of three information modalities

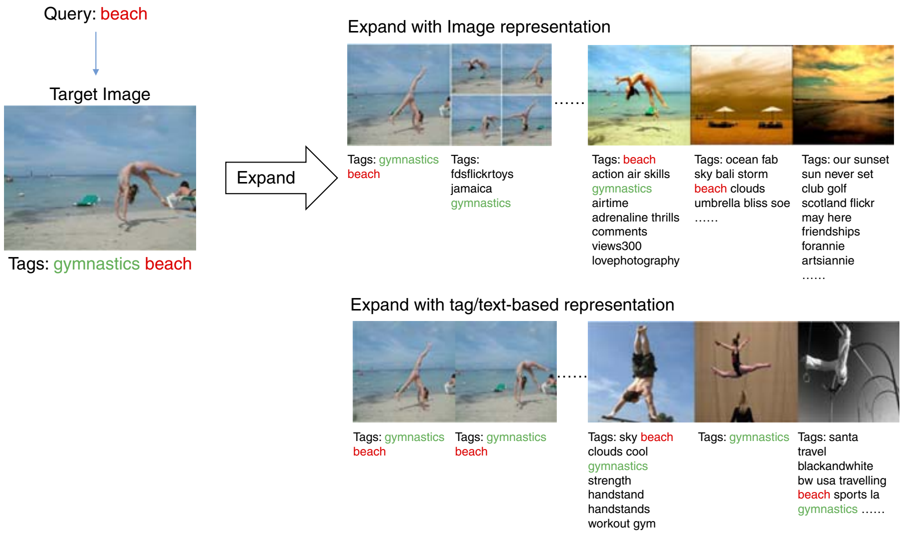


Figure 4.
Similar images found by text and image representation

Table V.
Pair-wise comparison of two computation methods

Method	MAP	MRR	P@10	nDCG@10
NUS-WIDE				
M1/M2	0.3383/*0.3412*	0.6931/*0.7558*	0.5975/*0.6970*	0.5925/*0.6933*
M3/M4	0.3382/*0.3461*	*0.8327*/0.8244	0.6605/*0.6926*	0.6347/*0.6826*
M5/M6	0.3393/*0.3441*	0.7726/*0.8241*	0.6432/*0.7012*	0.6299/*0.6692*
M7/M8	*0.3600*/0.3599	*0.7802*/0.7712	0.6889/*0.7358*	0.6962/*0.7266*
M9/M10	0.3557/*0.3561*	0.7536/*0.7778*	0.6037/*0.6486*	*0.6128*/0.6062
M11/M12	0.3308/*0.3452*	0.7526/*0.7737*	0.6494/*0.6815*	*0.6751*/0.6609
Flickr51				
M1/M2	*0.6261*/0.6167	0.6998/*0.8062*	0.6275/*0.6804*	0.4881/*0.5550*
M3/M4	*0.7582*/0.7430	0.8067/*0.8655*	0.7471/*0.8627*	0.6548/*0.7211*
M5/M6	*0.6193*/0.6087	0.6830/*0.7361*	0.6510/*0.7118*	0.5476/*0.5515*
M7/M8	*0.6080*/0.5326	0.6916/*0.7265*	*0.6275*/0.6000	0.4832/*0.4962*
M9/M10	*0.7866*/0.7679	0.8981/*0.9186*	0.8784/*0.8941*	0.7848/*0.8118*
M11/M12	*0.6981*/0.5066	*0.7277*/0.4517	*0.6294*/0.4019	*0.5169*/0.3261

combinations of expansion strategy and image similarity measures. Overall, results suggest that the "separate" approach is better than the "merging" approach when computing $P(q|R_D)$.

We suspect that the "separate" approach works better than the "merging" approach because the former is less likely affected by wrong expansion images with a lot of tags. While using the "merging" approach, an expansion image with a lot of tags will increase both tags' frequencies and document length of the merged "big document" representation by a greater extent. This may be useful if the expansion image is truly relevant to the original image, but it also introduces more noise when the expansion image is not relevant. In contrast, the "separate" approach does not have the issue.

Influence of cluster size and number of top k nearest neighbors

In cluster-based methods, cluster size has an impact on retrieval performance. In neighborhood-based methods, the parameter k (the number of top similar images) also has an important influence on retrieval performance. Thus, in this section, we discuss how these two parameters influence retrieval performance. In comparison to baselines, our methods do not bring consistent improvements in terms of MAP but in terms of P@10 and nDCG@10. This suggests that our methods have the ability to obtain better performance on the top of the rank. Therefore, in this section, we focus on how cluster size L and number of similar images k influence on P@10 and nDCG@10.

Figure 5 shows retrieval performance of the cluster-based methods using different cluster sizes. Although we did not observe any clear trends, most methods achieved excellent scores when setting cluster size to 3,000 (1,000 on Flickr51). Figure 6 shows retrieval performance of the nearest neighbor method with a different number of top similar images k. It seems that in general the nearest neighbor method prefers using more results for expansion, except in the case of NN + mixed + merging (M11) on NUS-WIDE and NN + text + separate (M8) on Flickr51.

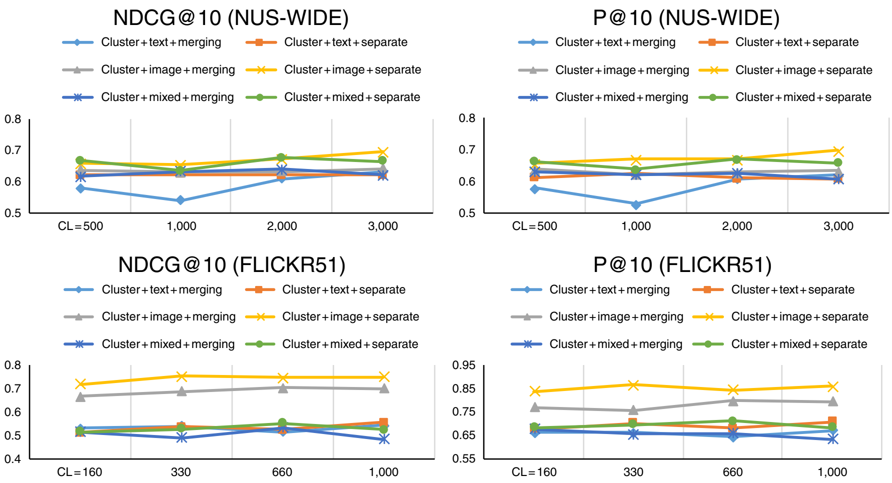


Figure 5.
Influences of the cluster size CL

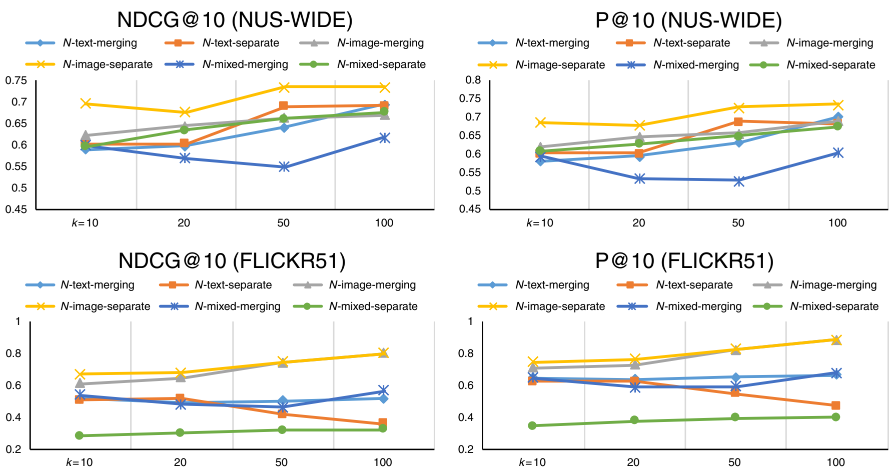


Figure 6.
Influences of the number of similar images k

This suggests that for both expansion strategies, the parameter settings are not trivial and will influence the system's performance. We also suggest to fully train these parameters before deploying our techniques to a practical scenario.

The weight of the expanded representation

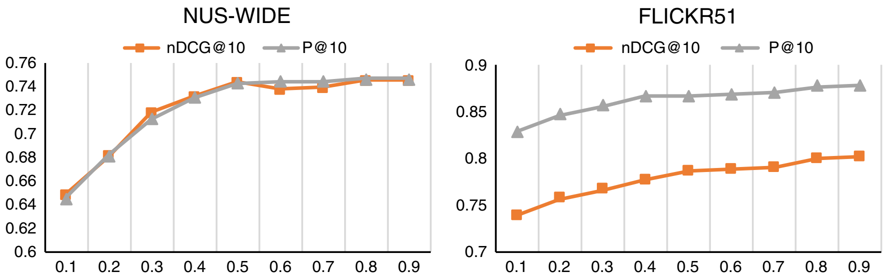
The weight of the expanded representation also affects the retrieval performance. Figure 7 shows the three measures' values for NN + image + separate (M10, the best performing run) with different values of α_{exp}. It is clear that all measures' values increase substantially when α increases from 0.1 to 0.5. The trend of increasing becomes smooth when $\alpha > 0.5$.

Results show that the optimal performance is usually achieved when using a higher weight on the expanded representation compared with the original image representation ($\alpha > 0.5$). This indicates the important role of the expanded representation in helping the original tag-based representation to achieve high retrieval performance.

Per-topic difference

In addition, it is noteworthy that the retrieval performance gains a significant and comprehensive improvement comparing with $Q_SR_VD_FL_SM_C$ (B2) when we use NN-image-separate method (M10) and set k at 100 for document expansion (Table VI). We also examine per-topic performance.

Figure 8 shows the difference of nDCG@10 between $Q_SR_VD_FL_SM_C$ model and the NN-image-separate method (k at 100) on a query-by-query basis. We found that more than 40 percent of query topics benefit from our methods on both data sets. On the NUS-WIDE data set, half of the query topics in the right-most panel increase over 0.25, and almost all query topics in the left-most panel decrease within 0.25. On the Flickr51 data set, lots of query topics in the right-most panel increase over 0.2. It is interesting to note that our method makes a great improvement on some topics with zero scores in the baseline



Notes: The x-axis indicates α_{exp} and the y-axis indicates the number of P@10 or nDCG@10

Figure 7. Performance of smoothing parameter α_{exp}

Method	MRR	±%	P@10	±%	nDCG@10	±%
NUS-WIDE						
B2	0.7644	–	0.6864	–	0.6866	–
M10 ($k = 100$)	0.8325	+8.91%*	0.7358	+7.20%*	0.7345	+6.99%*
Flickr51						
B2	0.8326	–	0.7980	–	0.6837	–
M10 ($k = 100$)	0.9003	+8.13%*	0.8839	+10.8%*	0.7971	+16.6%*

Note: $^*p < 0.05$ on a two pair-wise tests against the baseline

Table VI. Comparison between $Q_SR_VD_FL_SM_C$ and NN-image-separate (M10) method ($k = 100$)

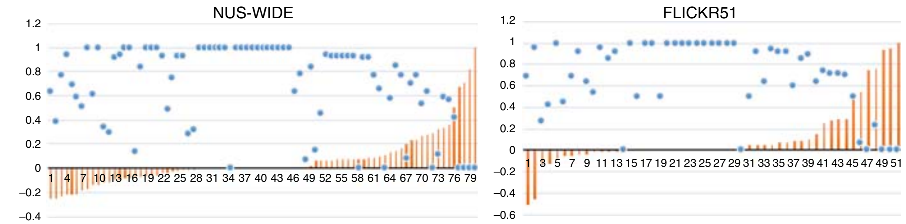
(the circle-shaped markers at the x-axis). Overall, the improvement on nDCG@10 means that our method succeeds in bringing relevant images to the top 10. Although our methods don not stably outperform in terms of MAP, they bring improvements on top ranks. Thus, the method will benefit a special scenario where users may pay more attention to top results, e.g., enjoying a TBIR service in the mobile platform.

Overall, we presented 12 methods to estimate the effectiveness of document expansion technology on TBIR. The comparison between our experiments and LM model indicate that document expansion can provide a better estimation than traditional IR in TBIR. Comparing the cluster-based and the neighborhood-based strategy, we found that the neighborhood-based strategy is the best choice for document expansion in TBIR. The investigation on two computation methods of the augmented probability $P(q|R_D)$ indicated that the individual method is able to erase the negative influences of error images in R_D. In addition, we also found that mixed features do not demonstrate the advantage that had been expected.

Computation cost

Despite its high performance, a practical concern for the nearest neighbor strategy is the cost of finding and computing expansion images. For example, if we simply re-rank the top N images retrieved by an initial approach (such as a baseline), we need to expand each of the top N images – while using the nearest neighbor strategy, this means N additional k nearest neighbor search. In contrast, the cluster-based strategy is usually cheaper at running time as long as we pre-compute and store the clustering result. However, it cannot handle dynamic data set – when the collection changes, we need to re-cluster the whole collection.

We report time cost for the bigger data set (NUS-WIDE) used in our experiments. Table VII shows the time cost of the two expansion strategies on a computer with Intel(R) Xeon(R) E5-2640 v2 @ 2.00 GHz CPU. On average, a k nearest neighbor search using image



Note: Circle-shaped markers indicate the baseline performance (y-axis in this case indicates absolute nDCG@10 scores)

Figure 8. Per-topic difference in nDCG@10 against baseline

Table VII. The time cost of clustering and neighborhood searching

Abbr.	CL	Time cost
Image clustering (using scikit-learn mini batch K-means algorithm[a])	500	115.89 s
	1,000	160.01 s
	2,000	420.85 s
	3,000	800.08 s
Text clustering (using scikit-learn mini batch K-means algorithm)	500	337.60 s
	1,000	384.71 s
	2,000	659.62 s
	3,000	1,072.94 s
Image neighborhood searching (using FLANN[b] to find similar images for per document in corpus)		5.5 ms
Text neighborhood searching (using Indri to find similar images for per document in corpus)		0.065 s

Notes: [a]http://scikit-learn.org/stable/; [b]www.cs.ubc.ca/research/flann/

features only takes 5.5 milliseconds – this means that it takes less than 1 s in the data set to expand and re-rank a list of 100 images, and it takes about 5 s to re-rank 1,000 images. Despite the increased computation cost, we believe that the technique is still reasonably fast, which makes it useful for many occasions requiring high accuracy image search results.

Conclusion

In this paper, we propose a concise framework based on document expansion techniques to address the sparsity, vocabulary mismatch, and tag-relatedness issues in TBIR. We experimented and compared different strategies, similarity measures, and models for constructing expanded image representation.

Unlike the former best performing work based on neighbor voting for pre-computation of tag relatedness, we used document expansion to measure the relation between tags and images. Our method is simple to understand and takes full advantage of the established technology of traditional IR and CBIR. With respect to the established baseline, the results of our experiments show that applying our NN-image-separate method yields significant improvements in effectiveness. Specifically, our method obtains better performance on the top of the rank and makes a great improvement on some topics with zero scores in the baseline. We also find that the neighborhood-based document expansion strategy outperforms the cluster-based document expansion strategy, and mixed features for the selection of R_D does not demonstrate the advantage that had been expected, and the separate method for calculating the augmented probability $P(q|R_D)$ is able to erase the negative influences of error images in R_D.

More recently, the development of deep learning has greatly increased the quality of automatic image annotation and makes it possible to predict multiple textual labels or generate natural language descriptions for an unseen image (Murthy *et al.*, 2015). It seems that these textual labels or descriptions can be indexed and used for image retrieval directly (Karpathy and Li, 2015). Although these methods create new opportunities to improve the performance of image retrieval, some disadvantages still exist regarding their application to image retrieval. Compared with textual labels and descriptions generated by these costly and time-consuming methods, social tags are easy-to-use and ready-to-use without requiring any additional training data. Moreover, social tags associated with images contain much abstract and personalized information, whereas general automatic image annotation focuses on assigning controlled keywords and limited concepts. So, we think that using social tags for image retrieval still constitutes a good choice in the short term.

Moreover, it is noteworthy that our approach is independent of any characteristic of social tags. If we use text labels or descriptions generated by automatic image annotation instead of social tags, the proposed method can also be applied to annotation-based image retrieval. We believe that users searching images with the TBIR system will benefit from our method. In future work, we plan to train an effective and suitable model for annotating the images in our data set and to test our approach on text generated by automatic image annotation methods.

This study has some limitations as well. First, we only consider the scenario that users search for relevant images by clicking on social tags associated with images (single concept query topics). To overcome this limitation, future research may consider multiple concepts query topics. Second, our methods only obtain better performance on the top of the rank instead of the entire ranked list. Therefore, this approach may benefit a special scenario where users may only pay attention to top rank results, e.g., searching on a mobile platform.

Notes

1. "Blobs" is a kind of image feature, we refer to the literature for further details.
2. www.flickr.com/
3. In the next, when we use "significantly," it means the differences are statistically significant.

References

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B. and Su, Z. (2007), "Optimizing web search using social annotations", *Proceedings of the 16th International Conference on World Wide Web*, p. 501.

Carman, M.J., Baillie, M. and Crestani, F. (2008), "Tag data and personalized information retrieval", *Proceeding of the 2008 ACM Workshop on Search in Social Media*, pp. 27-34.

Chen, L., Xu, D., Tsang, I.W. and Luo, J. (2010), "Tag-based web photo retrieval improved by batch mode re-tagging", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3440-3446.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y. (2009), "NUS-WIDE: a real-world web image database from National University of Singapore", *Proceedings of the ACM International Conference on Image and Video Retrieval, Santorini*, pp. 48:1-48:9.

Efron, M. (2010), "Hashtag retrieval in a microblogging environment", *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 787.

Efron, M., Organisciak, P. and Fenlon, K. (2012), "Improving retrieval of short texts through document expansion", *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 911-920.

Firan, C.S., Nejdl, W. and Paiu, R. (2007), "The benefit of using tag-based profiles", *Proceedings – 2007 Latin American Web Conference*, pp. 32-41.

Fox, E. a and Shaw, J. a. (1994), "Combination of multiple searches", *The 2nd Text Retrieval Conference, NIST SP 500215*, Vols 500-215, pp. 243-252.

Gao, Y., Wang, M., Zha, Z.J., Shen, J., Li, X. and Wu, X. (2013), "Visual-textual joint relevance learning for tag-based social image search", *IEEE Transactions on Image Processing,* Vol. 22 No. 1, pp. 363-376.

Gu, X., Wang, X., Li, R., Wen, K., Yang, Y. and Xiao, W. (2011), "Measuring social tag confidence: is it a good or bad tag?", *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6897 No. 1, pp. 94-105.

Gudivada, V.N. and Raghavan, V.V. (1995), "Content based image retrieval systems", *Computer*, Vol. 28 No. 9, pp. 18-22.

Hsieh, L. and Hsu, W. (2010), "Search-based automatic image annotation via Flickr photos using tag expansion", *ICASSP*, pp. 1-4.

Inoue, M. (2004), "On the need for annotation-based image retrieval", *Proceedings of the Workshop on Information Retrieval in Context, January*, pp. 2-4.

Jégou, H., Douze, M. and Schmid, C. (2010), "Improving bag-of-features for large scale image search", *International Journal of Computer Vision*, Vol. 87 No. 3, pp. 316-336.

Jeon, J., Lavrenko, V. and Manmatha, R. (2003), "Automatic image annotation and retrieval using cross-media relevance models", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 119-126.

Karpathy, A. and Li, F.F. (2015), "Deep visual-semantic alignments for generating image descriptions", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 7, June 12*, pp. 3128-3137.

Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P. and Garcia-Molina, H. (2008), "Combating spam in tagging systems", *ACM Transactions on the Web*, Vol. 2 No. 4, pp. 1-34.

Krizhevsky, A., Sulskever, I. and Hinton, G.E. (2012), "ImageNet classification with deep convolutional neural networks", *Proceedings of the 25th International Conference on Neural Information Processing Systems*, *Nevada, CA*, pp. 1097-1105.

Kulis, B. and Grauman, K. (2009), "Kernelized locality-sensitive hashing for scalable image search", *Proceedings of 2009 IEEE 12th International Conference on Computer Vision*, pp. 2130-2137.

La Cascia, M., Sethi, S. and Sclaroff, S. (1998), "Combining textual and visual cues for content-based image retrieval on the world wide web", *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 24-28.

Lee, K.-P., Kim, H.-G. and Kim, H.-J. (2012), “A social inverted index for social-tagging-based information retrieval”, *Journal of Information Science*, Vol. 38 No. 4, pp. 313-332.

Li, X. and Snoek, C. (2010), “Unsupervised multi-feature tag relevance learning for social image retrieval”, *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 10-17.

Li, X. and Snoek, C.G.M. (2013), “Classifying tag relevance with relevant positive and negative examples”, *Proceedings of the 21st ACM International Conference on Multimedia, ACM, New York, NY*, pp. 485-488.

Li, X., Snoek, C.G.M. and Worring, M. (2009), “Learning social tag relevance by neighbor voting”, *IEEE Transactions on Multimedia*, Vol. 11 No. 7, pp. 1310-1322.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C. and Bimbo, A.D. (2015), “Image tag assignment, refinement and retrieval”, *Proceedings of the 23rd ACM International Conference on Multimedia, ACM, Brisbane*, pp. 1325-1326.

Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C.G.M. and Bimbo, A.D. (2016), “Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval”, *ACM Computing Surveys*, Vol. 49 No. 1, pp. 14:1-14:39.

Liu, D., Hua, X.S., Wang, M. and Zhang, H. (2009), “Boost search relevance for tag-based social image retrieval”, *Proceedings – 2009 IEEE International Conference on Multimedia and Expo*, pp. 1636-1639.

Liu, X. and Croft, W.B. (2004), “Cluster-based retrieval using language models”, *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, pp. 186-193.

Lu, D., Liu, X. and Qian, X. (2016), “Tag-based image search by social re-ranking”, *IEEE Transactions on Multimedia*, Vol. 18 No. 8, pp. 1628-1639.

Melenhorst, M., Grootveld, M., van Setten, M. and Veenstra, M. (2008), “Tag-based information retrieval of video content”, *Proceeding of the 1st International Conference on Designing Interactive User Experiences for TV and Video*, p. 31.

Min, J., Leveling, J., Zhou, D. and Jones, G.J.F. (2010), “Document expansion for image retrieval”, *Proceedings of Adaptivity, Personalization and Fusion of Heterogeneous Information*, *Paris*, pp. 65-71.

Murthy, V.N., Maji, S. and Manmatha, R. (2015), “Automatic image annotation using deep learning representations”, *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, *Shanghai*, pp. 603-606.

Nov, O., Naaman, M. and Ye, C. (2008), “What drives content tagging: the case of photos on Flickr”, *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems*, pp. 1097-1100.

Sanderson, M. (2010), “Test collection based evaluation of information retrieval systems”, *Foundations and Trends® in Information Retrieval*, Vol. 4 No. 4, pp. 247-375.

Sang, J., Xu, C. and Liu, J. (2012), “User-aware image tag refinement via ternary semantic analysis”, *IEEE Transactions on Multimedia*, Vol. 14 No. 3, pp. 883-895.

Sclaroff, S., La Cascia, M., Sethi, S. and Taycher, L. (1999), “Unifying textual and visual cues for content-based image retrieval on the world wide web”, *Computer Vision and Image Understanding*, Vol. 75 Nos 1/2, pp. 86-98.

Sevil, S.G., Kucuktunc, O., Duygulu, P. and Can, F. (2010), “Automatic tag expansion using visual similarity for photo sharing websites”, *Multimedia Tools and Applications*, Vol. 49 No. 1, pp. 81-99.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A. and Jain, R. (2000), “Content-based image retrieval at the end of the early years”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22 No. 12, pp. 1349-1380.

Sun, A. and Bhowmick, S.S. (2008), “Image tag clarity: in search of visual-representative tags for social images”, *Proceedings of the 1st SIGMM Workshop on Social Media*, *Beijing*, pp. 19-26.

Sun, A., Bhowmick, S.S. and Chong, J.-A. (2011), "Social image tag recommendation by concept matching", *Proceedings of the 19th ACM International Conference on Multimedia*, p. 1181.

Sun, A., Bhowmick, S.S., Nam Nguyen, K.T. and Bai, G. (2011), "Tag-based social image retrieval: an empirical evaluation", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 12, pp. 2364-2381.

Tang, J., Yan, S., Hong, R., Qi, G.-J. and Chua, T.-S. (2009), "Inferring semantic concepts from community-contributed images and noisy tags", *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 223-232.

Tao, T., Wang, X., Mei, Q. and Zhai, C. (2006), "Language model information retrieval with document expansion", *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, *June*, pp. 407-414.

Torralba, A., Fergus, R. and Freeman, W.T. (2008), "80 million tiny images: a large data set for nonparametric object and scene recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30 No. 11, pp. 1958-1970.

Truong, B.Q., Sun, A. and Bhowmick, S.S. (2012), "Content is still king: the effect of neighbor voting schemes on tag relevance for social image retrieval", *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ACM, New York, NY*, pp. 9:1-9:8.

Wang, M., Yang, K., Hua, X.S. and Zhang, H.J. (2010), "Towards a relevant and diverse search of social images", *IEEE Transactions on Multimedia*, Vol. 12 No. 8, pp. 829-842.

Wei, X. and Croft, W.B. (2006), "LDA-based document models for ad-hoc retrieval", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 178.

Wu, L., Jin, R. and Jain, A.K. (2013), "Tag completion for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35 No. 3, pp. 716-727.

Xu, S., Bao, S., Fei, B., Su, Z. and Yu, Y. (2008), "Exploring folksonomy for personalized search", *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Vol. 31, pp. 155-162.

Zhu, G., Yan, S. and Ma, Y. (2010), "Image tag refinement towards low-rank, content-tag prior and error sparsity", *Proceedings of the 18th ACM International Conference on Multimedia, ACM, New York, NY*, pp. 461-470.

Corresponding author
Heng Ding can be contacted at: hengding@whu.edu.cn