Social image retrieval based on topic diversity

Yaxiong Wang¹ · Li Zhu¹ · Xueming Qian²

Received: 21 November 2019 / Revised: 17 September 2020 / Accepted: 9 December 2020 /

Published online: 11 January 2021

© The Author(s) 2021

Abstract

Image search re-ranking is one of the most important approaches to enhance the text-based image search results. Extensive efforts have been dedicated to improve the accuracy and diversity of tag-based image retrieval. However, how to make the top-ranked results relevant and diverse is still a challenging problem. In this paper, we propose a novel method to diversify the retrieval results by latent topic analysis. We first employ *NMF* (Non-negative Matrix Factorization) Lee and Seung (Nature 401(6755):788–791, 1999) to estimate the initial relevance score to the query q . Then, the initial relevance score is fed into an adaptive multi-feature fusion model to learn the final relevance score. Next, the diversification process is conducted. We group all the images by semantic clustering and estimate the topic distribution of each cluster by topic analysis. The clusters are ranked based on the topic distribution vector and the final retrieval image list is obtained by a greedy selection mechanism based on the estimated relevances. Experimental results on the NUS-Wide dataset show the effectiveness of the proposed approach.

Keywords Tag-based image retrieval · Social media · Feature fusion · Re-ranking · Topic analysis

This work was supported in part by the NSFC under Grant 161772407, 61701391, 61732008, and 61902309 and National Key Research and Development Project with No: 2019YFB2102500; and Emergence Mechanism and Calculation Method of Group Intelligence based on Internet with No: 2018AAA0101100.

✉ Li Zhu
zhuli@mail.xjtu.edu.cn

✉ Xueming Qian
qianxm@mail.xjtu.edu.cn

Yaxiong Wang
wangyx15@stu.xjtu.edu.cn

¹ School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

² Ministry of Education Key Laboratory for Intelligent Networks and Network Security, SMILES LAB, Xi'an Jiaotong University, Xi'an, China

1 Introduction

Social websites, such as Flickr, and Instagram, are visited daily by hundreds of millions of users, and the shared various multimedia resource and amounts of images are uploaded on the Internet. For example, 4.5 million new photos are added per day on Flickr, 136,000 million new photos, and 40 million photos are posted daily to the service Instagram [17]. The continuously growth of online images requires building an efficient retrieval system.

Conventional tag-based image retrieval methods cannot achieve satisfying results due to the tag mismatch and the query ambiguity. Social networks allow users to annotate their shared images with a set of tags to describe the content of image. Unfortunately, many tags provided by users are irrelevant. As reported in [7], only about 50% of the tags provided by users are really related to the image. Numerous noisy tags make the image retrieval problem difficult. Query ambiguity is another much-vexed problem. Users can't precisely describe their query intention with a single query word. For example, when the users provide a query "apple", they may refer to fruit, computer, mobile phone, or the trademark. It is hard to identify which topic the user prefers from a single word. The frequently-used image search systems such as Bing, Google, and Yahoo! all suffer from the drawback that textual information cannot comprehensively and substantially describe the rich content of images. As a consequence, the retrieval results of the ambiguity query are always unsatisfactory.

Extensive efforts have been dedicated to overcome the two problems mentioned above. As for the tag mismatch, tag refinement [19, 22, 28, 54, 63], tag relevance ranking [12, 22, 23, 59] and image relevance ranking [4, 7, 29, 33, 39, 45, 46, 56] are proposed to solve the problem. On the subject of query ambiguity, diversifying the retrieval results is an efficient and popular approach, which aims at making the top-ranked images cover as many topics as possible. Currently, image clustering [5, 21, 41, 51] and duplicate removal [23, 35, 38] are commonly used to diversify the retrieval results.

Particularly, the diversity is defined based on the average similarity between several samples, the higher similarity leads to worse diversity. For retrieval task, a diverse result means the ranking list contains enough images with different appearances. In recent years, more and more scholars dedicate themselves to the retrieval results diversifying [3, 23, 25, 30, 41, 51]. There is no doubt that the relevance of retrieval results stands in the first place, which is regarded as the bedrock in information retrieval field. However, only relevance without diversity cannot meet users' requirements. Thus, many algorithms have been proposed to diversify the retrieval results [5, 23, 25, 30, 41, 51]. In [51], the authors first apply graph clustering to assign the images to clusters, then the random walk is used to obtain the final results. The diversity is achieved by setting the transition probability of two images in different clusters higher than those in the same cluster. Tian et al. think the topics in the initial list are of tree structure [41]. Images are first assigned to corresponding leaf topics, then a topic cover score is defined. Finally, they use a greedy algorithm to form the image list with the highest topic cover score as the retrieval results. Lin et al. [23] first rank the images according to the initial confident scores, a graph-based greedy selection algorithm is applied to get the final image list. The diversity is achieved by selecting the next image based on the dissimilarity to the images appearing before. Dang-Nguyen et al. [5] obtain a topic tree by the proposed clustering algorithm, then the topics are sorted according to the number of images. In each cluster, the image uploaded by the user who has the highest visual score is selected as the top-ranked image. The diversity is achieved by greedily selecting according to the dissimilarity between images. The works mentioned above diversify the retrieved images form the visual view by selecting the visually dissimilar images. However, the query is from the semantic field, diversifying the results from the semantic view is more reasonable and

can also alleviate the semantic gap problem. For example, some scholars expand the query utilizing the WordNet or Wikipedia to capture the possible topics or concepts of the query tag. Hoque et al. [16] expand the ambiguous query by Wikipedia and exhibit a broad range of images that represent the various possible interpretations of the query to diversify the results. In [27] and [14], the authors expand the query to the possible concept or phrase by WordNet to mine the latent topics of the query. However, it is inadequate to represent all the latent topics of the query by limited expanding words or phrases. In our previous work [25], we diversify the results from user view and regard the images uploaded by the same user as a group and pick one image from each group to achieve diversity. In [30], we first assign images to different clusters by clustering. Then inter-cluster and intra-cluster ranking are conducted, and the final image list is obtained by selecting one image from each cluster.

In this paper, we focus on semantic topic diversity. We first apply a regularization model to learn the relevances of images. Then the semantic clustering is conducted. We next rank the clusters based on the estimated topic distribution vectors. Finally, we pick one image with the highest relevance score in each cluster to achieve our semantic topic diversity. The topic of this work refers to the semantics embodied in each image cluster, since the images in different clusters share the semantics, we regard each image cluster as one semantic topic. In the following diversity step, we pick one image from each semantic topic to achieve the topic diversity.

Our previous work [30] also diversifies the results by clustering from topic views, however, there are many considerable differences with this work:

- 1) The overall algorithm design is different. Previous work [30] fuse visual and textual information to enhance relevance and diversity simultaneously. While this paper diversifies the results by a two-step mechanism and uses the visual and textual information separately to avoid the semantic gap.
- 2) The estimation of image relevance is different. In [30], we first conduct clustering and apply a regularization model for each image cluster to estimate the relevance. In this paper, we apply a regularization model on all images associated with the query to estimate the image relevance.
- 3) The cluster representation is different. We simply employ the tag histogram to represent the image cluster in the previous work [30]. While this paper analyzes the topic distribution as the cluster feature representation.
- 4) Cluster ranking algorithm is different. In [30], we apply an adaptive random walk model to rank the image clusters. While we directly compute the relevance of image cluster to query by the estimated topic distribution vector in this paper, by which we rank the image clusters.
- 5) We simply splice two features to represent the image and apply a regularization model to estimate the image relevance in work [30]. While in this paper, we use multiple visual features and fuse them by an adaptive feature fusion model to estimate the image relevance to the query.

The contributions of this paper are summarized as follows:

- We propose a semantic diversity approach based on topic analysis. We first apply clustering to assign semantically similar images into the same cluster. Then the topic distribution vector is estimated based on topic analysis. In this way, we can directly rank the image clusters.
- We learn the relevance scores of images based on the initial relevance scores from *NMF*. Then, a regularization framework is applied to learn the final relevance score from the

initial estimations. Relevance learning is conducted by visual information completely, by which we can avoid the problem of noisy tags.

- For each query tag, we generate documents to analyze the latent topic distribution. Compared with representing topics by limited expanding words or phrase [23, 33, 58], we will have enough query description that supports us to thoroughly analyze the latent topics of the query tag. Our latent topic number is also adaptively determined to build a flexible system.

The remainder of this paper is organized as follows. In section 2, we review the related works about the tag-based image retrieval. Section 3 presents the details of each step in our system. Experiments on NUS-Wide dataset are shown in section 4. In Section 5, we discuss the parameters in our approach systematically. Finally, conclusion and future work are given in Section 6.

2 Related work

Web image retrieval is usually achieved by using the tags provided by users as indexes. However, the noisy tags and duplicated images make the retrieval results lack accuracy and diversity, which can't satisfy the users well. Extensive efforts have been dedicated to improve the relevance and diversity of the retrieval results. The following parts will present the existing works related to the above two aspects respectively.

2.1 Image relevance ranking

It is an essential demand for a retrieval system to guarantee the relevances of results. For image retrieval, amounts of methods are proposed to improve the retrieval accuracy. Most of them are designed based on the visual and semantic clues. Wu et al. [50] propose a learning-based method to preserve both visual and semantic consistency in the similarity ranking. They learn the final relevance score by a self-tune manifold ranking mechanism that takes the visual and semantic similarity into consideration simultaneously. Wang et al. [43] fuse the tag information and visual feature into a graph model to re-rank images. They regard the images as vertices and construct a multigraph by investigating the textual and semantic correlation between images. Then each vertex in the multigraph is untied into two vertices, one for semantic information and another for visual content. By this way, the multigraph degrades into a complete graph. The re-ranking is conducted by the graph learning algorithm. In [10], Gao et al. propose a hypergraph based approach to utilize the visual and textual information simultaneously. A tag refinement strategy is first applied to filter the noisy tags, and each image is represented by the bag of words model. Then they construct a hypergraph based on the tags and visual words. The final relevance score is obtained by the hypergraph learning algorithm. Haruechaiyasak and Damrongrat [13] transfer a content-based image retrieval method to tag based image retrieval for relevance enhancement. Deng et al. [67] treat the top-ranked images in the initial retrieval list as the pseudo positive samples and construct multiple graphs to learn the final relevance score. In [62], Zhang et al. utilize the click information and employ the simpleMKL model to re-rank the images. The authors select training samples based on the click-through data, by which a multi-kernel SVM is trained for image ranking. Yu et al. [60] re-rank the images by a deep method. First, the initial ranked images are used to build the training set. Next, they adopt click features related to those images and apply several Auto-Encoders to initially obtain the distance metric. Then, multi-view Deep Metric Learning (DML) is adopted to

identify a set of optimal weights of multiple visual features. The final prediction model is trained by an alternating optimization algorithm. Song et al. [34] utilize the generative adversarial networks [11] and propose a unified framework to simultaneously achieve the faster image retrieval and compression. Learning to re-rank is also a popularly used strategy. Ren et al. [32] learn textual and visual mapping matrices respectively. Thus, the query tag and the images can be mapped into the common subspace by the mapping matrices. Finally, they re-rank the images by the distance between the query and the images in the sharing subspace. In [58], the authors learn a bilinear function to measure the similarity between query tag and images by canonical correlation analysis (CCA). Instead of using only one visual feature, fusing multiple visual features is a popular strategy to enhance the relevance, and many papers have validated its effectiveness. Wang et al. [42] construct graph using multiple features. They integrate the learning of relevance scores, weights of features, the distance metric, and its scaling for each modality into a unified framework for image re-ranking. Yang et al. [57] use multiple visual features and construct a graph for every feature, they get the final relevance score by graph learning. In [36], the authors focus on the efficient video retrieval, they design a hierarchical autoencoder to capture the temporal dependencies in videos. Yu et al. [59] construct five hypergraphs for five visual features and integrate the visual consistency constraints of these hypergraphs to learn a linear model for image ranking. Sharing space learning can also be applied in re-ranking task, image ranking can be conducted based on distance metric in the learned subspace. The work [9] develops a novel hierarchical LSTM with adaptive attention to simultaneously consider the visual and language context for visual caption task. Wang et al. [47] propose a position attention mechanism to estimate the relevance between visual and textual data for image ranking. Wang et al. focus on ranking the images using the sketch information only [48, 49]. In [61], Zhai et al. pursue the high-efficiency for image retrieval, they integrate the hash and deep learning techniques to achieve faster image retrieval.

2.2 Retrieval results diversification

The relevance based retrieval approaches attempt to pick up as many relevant images as possible for users. However, just exhibiting the relevant images in top retrieval results cannot satisfy the users completely. Multiple similar results will discount users' retrieval experience, even if these images are all relevant. Therefore, many researchers dedicate their extensive efforts to make the top-ranked results as diverse as possible. In [37], Spyromitros-Xioufis et al. propose a greedy selection algorithm to pick up the visually dissimilar image to form the final image list. Wang et al. [44] propose a new measure named ADP (Average Diverse Precision) to take the diversity and relevance into consideration simultaneously and the re-ranking is conducted by a greedy selection process. Calumby et al. [3] design an iteration optimization algorithm based on relevance feedback. Similar to paper [23], diversity is achieved by the greedy selection. Besides greedy selection, diversifying based on clustering is also widely used. In [2], Cai et al. propose a hierarchical clustering method and diversify the retrieval results by grouping the search results into different clusters according to visual, textual, and link information. Yan et al. [51] design a two-layer graph model to diversify the retrieval results. They first conduct clustering to assign images into different topics and construct an undirected graph. The images are re-ranked by random walk on the constructed graph, and the diversity is achieved by setting the transition probability of two images within the same topic higher than those within different topics. Lin et al. [24] first cluster the images to find the latent topics and then propose a greedy selection algorithm to fuse the relevance score, visual similarity and level of topic coverage. In [21], the authors propose

three clustering algorithms to diversify the retrieval results from three different emphases. In our previous work [30], we assign the semantically similar image into the same cluster first. Then inter-cluster and intra-cluster are conducted. The diversity is also achieved by clustering. Sun et al. [38] design a ranking method according to the relevance, typicality, and diversity criteria. They explore both semantic information and visual content of images. Yanai et al. [53] diversify the retrieval results based on the sink point of random walk. After selecting an image, they update the transition matrix by setting the selected images as a sink point. In [16], Hoque et al. expand the ambiguous query by Wikipedia and exhibit a broad range of images, which attempts to capture the various possible interpretations of the query. Ksibi et al. propose a method to balance the relevance against diversity dynamically according to the ambiguity level of the given query [18]. Qian et al. [31] propose an approach to diversify the landmark summarization from relative viewpoint perspectives. They design a 4-dimensional vector to represent the relative viewpoint of each image and select the relevant images with large viewpoint variations as top-ranked images.

3 Methodology

3.1 Method

The pipeline of our proposed approach is shown in Fig. 1. For query q , our system conducts the following steps on the images containing tag q : 1) Initial Relevance Estimation. Visual similarity is calculated first, and *NMF* is introduced to get the initial relevance score, by which we learn the final relevance score. 2) Multi-feature fusion. A regularization framework is applied to fuse the multiple visual features adaptively. We learn the weights of different visual features and final relevance scores of images to the query tag in this step. 3) Semantic clustering. We conduct clustering on all the co-occurrence tags of the query first and then assign semantic similar images into the same cluster to obtain the image clusters. 4) Cluster ranking by topic analysis. One document is generated for each image cluster as well as the query tag. Then, their topic distributions are estimated respectively, by which we rank the clusters. We sequentially select the most relevant image in each ranked cluster as our final retrieval results.

Fig. 1 The framework of our proposed method

Among the four steps above, we ensure the relevance of retrieval results in the first two steps, while in the last two steps, we focus on diversity. The final images are sequentially selected to make the retrieval results as diverse as possible on the basis of relevance. Hereinafter, the details of each component are presented.

3.2 Initial relevance estimation

In this subsection, we elaborate our initial relevance estimation method. The visual similarity matrix is calculated firstly. Then *NMF* is employed to get the initial relevance scores of images.

Let $X = \{x_1, x_2, \dots, x_n\}$ be the image set associated with query tag q , n is the number of images. The similarity of any two images is calculated by Gaussian kernel function:

$$w_{ij} = \exp\left(-\frac{\|r_i - r_j\|^2}{2\sigma^2}\right), \quad (1)$$

where r_i is the visual feature of image x_i and σ is a constant, we simply set it as the average distance between all images.

Let $W \in R^{n \times n}$ be the image similarity matrix, whose elements are defined by (1). After we obtain the similarity matrix, the initial relevance can be estimated by the Non-negative Matrix Factorization (*NMF*) [20]. The *NMF* can decompose the nonnegative matrix $W \in R^{n \times n}$ into the product of two nonnegative matrices $Z \in R^{n \times s}$ and $U \in R^{s \times n}$ approximately as follow:

$$W = Z \times U, \quad (2)$$

where s is an integer. If we set $s = 1$, we can decompose the similarity matrix W into two vectors $Z = (z_1, z_2, \dots, z_n)^T \in R^{n \times 1}$ and $U = (u_1, u_2, \dots, u_n) \in R^{1 \times n}$. This situation is the special case of general *NMF*. Thus, we can obtain:

$$\begin{pmatrix} w_{11} & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & w_{nn} \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} (u_1 \dots u_n), \quad (3)$$

From (3), the i -th column of W is the vector Z product u_i :

$$W_{:,i} = Z \times u_i, \quad (4)$$

where $W_{:,i}$ is the i -th column of W and represents the similarity of image x_i to the other images. The (4) reveals that the similarities of image x_i to the other images can be represented by the product of Z and the component u_i in U . Thus, we can consider Z as the base vector, then u_i somehow measures the average similarity of image x_i to the other images. From the truth that relevant images share high visual similarity [23], then the more relevant images will have higher u . Therefore, it is reasonable to treat the vector U as the initial estimation, which would serve as the cornerstone to learn the final relevance score.

If we apply multiple visual features, multiple visual similarity matrices can be calculated accordingly. Thus, we can obtain an initial relevance score for each feature by *NMF*. In the next subsection, we will fuse these initial relevance estimations to learn the final relevance score.

3.3 Multiple features fusion

Many works employ multiple features to boost the relevance learning, and their results validate the effectiveness of multiple features fusion [42, 57, 59, 62], since different visual

features express an image from different views, only employing a single visual feature is unilateral. Therefore, in this subsection, we fuse multiple visual features adaptively to learn the final relevance scores based on the initial estimation.

Feature fusion model The features fusion algorithm is a regularization model, it is proposed based on following considerations: 1) visually similar images should be assigned close relevance scores. 2) the final relevance score should be not far away from the initial relevance. 3) the stronger discriminative visual features should be assigned higher weights. To meet the above three constraints, we design the fusion model as (5), where the term 1 to 3 sequentially meets the considerations 1)-3). The fourth term is a regularization term, which is popular to avoid overfitting.

$$\begin{aligned} \arg \min_{\alpha, f} & \sum_{k=1}^K \sum_{i,j=1}^n \alpha_k w_{ij}^k \left(\frac{f_i}{\sqrt{D_{ii}^k}} - \frac{f_j}{\sqrt{D_{jj}^k}} \right) + \gamma \|f - \sum_{k=1}^K v_k y^k\|^2 \\ & + \beta \sum_{k=1}^K \|\alpha_k - v_k\|^2 + \xi \sum_{k=1}^K \alpha_k^2, \\ \text{s.t.} & \sum_{k=1}^K \alpha_k = 1, 0 \leq \alpha_k \leq 1, k = 1, 2, \dots, K. \end{aligned} \quad (5)$$

In (5), α_k is the weight of the k -th visual feature, f_i is the relevance score of image x_i to query q , $\alpha \in R^K$ and $f \in R^n$ are two column vectors consists of α_k and f_i . $D_{ii}^k = \sum_{j=1}^n w_{ij}^k$ is the normalization factor, where w_{ij}^k is the similarity of image x_i and x_j according to the k -th visual feature and K is the number of features. We employ y^k to represent the initial relevance score from *NMF* by the k -th visual similarity matrix W^k , v_k is the discriminative power of the k -th visual feature. For each feature, we calculate the variance over all image similarities within the set X , the similarity variance is used as the discriminative power measure. That is $v_k = \text{var}(W^k)$, where $\text{var}(\cdot)$ is the variance operator [21]. γ, β, ξ are three weight hyperparameters. In this paper, treating the variance over all image similarities as the discriminative power of visual feature is also based on the truth that the relevant images share higher similarity, while the irrelevant images are dissimilar with each other. If a feature could well distinguish these images, it should: 1) make the relevant close to each other, 2) make the relevant images be far away from the irrelevant ones, 3) make the irrelevant images far away from each other. Thus, for this feature, the variance of similarity between images will be higher accordingly, and we assign the feature higher initial weight.

Equation 5 can be rewritten as the matrix form:

$$\arg \min_{\alpha, f} \sum_{k=1}^K \alpha_k f^T L_k f + \gamma \|f - Y\|^2 + \beta \|\alpha - v\| + \xi \|\alpha\|^2, \quad (6)$$

where $L_k = I_n - (D^k)^{-\frac{1}{2}} W^k (D^k)^{-\frac{1}{2}}$, $I_n \in R^{n \times n}$, $Y = \sum_{k=1}^K v_k y^k$, D^k is the diagonal matrix whose diagonal elements are $D_{ii}^k, i = 1, 2, \dots, n$. $v \in R^K$ consist of $v_k, k = 1, 2, \dots, K$.

Optimization solution The final relevance score f and feature weight α are two parameters needed to be learned. The optimization problem defined by (6) can be solved by updating α and f alternatively.

Fix α and optimize f When we fix α , we differentiate the objective function with respect to f and follow some simple algebraic steps, we can derive that:

$$f = \left(\sum_{k=1}^K I_k + \sum_{k=1}^K \frac{\alpha_k}{\gamma} L^k \right)^{-1} Y, \quad (7)$$

where $I_k \in R^{K \times K}$ is the identity matrix.

To reduce the computational cost of matrix inversion, we can use following iteration formulation to update f :

$$f^{t+1} = \frac{1}{c} (M f^t + Y), \quad (8)$$

where $c = \frac{\gamma}{1+\gamma}$ and $M = \frac{1}{\gamma} \sum_k \alpha_k (D^k)^{-\frac{1}{2}} W (D^k)^{-\frac{1}{2}}$.

Fix f and optimize α When we fix f , the optimization problem is equivalent to:

$$\begin{aligned} \arg \min_{\alpha} \quad & \sum_{k=1}^K \alpha_k f^T L_k f + \beta \| \alpha - v \|^2 + \xi \| \alpha \|^2, \\ \text{s.t.} \quad & \sum_{k=1}^K \alpha_k = 1, 0 \leq \alpha_i, i = 1, 2, \dots, K. \end{aligned} \quad (9)$$

The optimization problem defined by (9) can be rewritten as:

$$\begin{aligned} \arg \min_{\alpha} \quad & \frac{1}{2} \alpha H \alpha + (p - 2\beta v), \\ \text{s.t.} \quad & \begin{cases} e^T \alpha = 1, \\ \mathbf{0} \leq \alpha, \end{cases} \end{aligned} \quad (10)$$

where $H = (2\beta + 2\xi)I_K$, $e \in R^K$ is an all one vector, $p = (p_1, \dots, p_k)^T \in R^K$ and $p_k = f^T L_k f$, $\mathbf{0} \in R^K$ is all-zero vector.

The problem defined by (10) is a standard quadratic programming problem and can be solve by Lagrange multiplier method. Following the *KKT* condition and some algebra steps, we can obtain:

$$\begin{pmatrix} \alpha \\ \lambda \\ \mu \end{pmatrix} = \begin{pmatrix} H & -I & -e^T \\ I & 0 & 0 \\ e & 0 & 0 \end{pmatrix}^{-1} \begin{pmatrix} -(p - 2\beta v) \\ 0 \\ e \end{pmatrix}, \quad (11)$$

where λ, μ are Lagrange multipliers.

Each of the above steps decreases the objective function which has a lower bound 0, therefore, the convergence of the alternating optimization is guaranteed.

3.4 Semantic clustering

Our diversity is achieved by clustering. To obtain the semantic clusters of images, we first conduct clustering on the co-occurrence tags of the query and then assign images to the corresponding tag set to obtain the image clusters. Images in the same cluster will share similar semantics through this mechanism. Different from many multimedia works whose semantics is embodied in the abstract deep features [9, 45], this work expresses the semantics with specific tag clusters, which is corresponding to the different perspectives of the query tag as show in Fig. 2.

Fig. 2 Tag clusters and corresponding image clusters

Tag clustering Let $T = \{t_1, t_2, \dots, t_m\}$ be co-occurrence tag set of query q , which is collected from the image tags in set X . To obtain an efficient representation of the tag, we first train the word vector using word2vec model [26] by the English Wikipedia dataset [40]. The skip-gram is also applied for better word vector generating. Based on the trained word2vec model, each tag can be represented by a vector, for example with 100-dimension. We use cosine similarity to measure the similarity between tags:

$$s_{tag}(t_i, t_j) = \frac{\langle vt_i, vt_j \rangle}{\|vt_i\| \times \|vt_j\|} \quad (12)$$

where vt_i is the word vector of tag t_i and $\|\cdot\|$ represents the Euclidean 2-norm.

The similarity matrix of tags can be obtained according to (12), which is the input of the clustering algorithm. AP-Clustering algorithm [8] is employed to assign similar tags into the same cluster. We choose AP-Clustering is mainly because it automatically determines the cluster number [50] and only needs the similarity matrix of tags, which is quite in line with our requirements. We denote the tag clustering results by $C = (c_1, c_2, \dots, c_{\bar{C}})$, where $\bar{C}$ is the number of tag cluster.

Image clustering After finishing the tag clustering, we can obtain the image clusters by assigning images to the corresponding tag cluster according to the tag overlap between image and tag cluster:

$$x_i \in c^* = \max_{c \in \bar{C}} \{ov(c, \mathcal{T}_i)\}, \quad (13)$$

where $ov(\cdot, \cdot)$ denotes the number of common tags of $\mathcal{T}_i$ and tag cluster c :

$$ov(c, \mathcal{T}_i) = |c \cap \mathcal{T}_i|, \quad (14)$$

where $|\cdot|$ denotes the element number of a set.

After the above steps, we can obtain the image clusters $IC = \{ic_1, ic_2, \dots, ic_{\hat{C}}\}$, where $\hat{C}$ is the number of image clusters, and $\hat{C} \leq \bar{C}$, since there may be some tag clusters that no image is assigned. Images in the same cluster are of similar semantics.

Figure 2 shows two exemplar tag clusters and corresponding image clusters in our experiment. Although there are some noisy tags in tag cluster, we can still find the main semantics. The tag cluster in Fig. 2a expresses a specific semantics: bird, while Fig. 2b shows general semantics: domestic animal. The semantics of images are coincident with corresponding tag cluster.

3.5 Cluster ranking by topic analysis

Many papers regard an image cluster as a topic, their algorithms are designed to find the possible topics of the query tag [2, 21, 51] and select images from different clusters to diversify the results. However, these methods don't estimate the relevance of different clusters to query and treat these topics equally. In this work, we calculate the relevance of different

clusters to the query based on the estimated topic distribution and sequentially select the images from the ranked clusters to enhance the diversity while ensuring the relevance as much as possible.

Hereinafter, we present our image clusters ranking algorithm. We first generate the query document and cluster documents respectively. Then, the topic distribution of query tag and image clusters are estimated by latent topic analysis. Finally, we rank the image clusters based on the topic distribution vector.

Query document generation To learn the semantic topic distribution of the query tag, there are two problems needed to be solved.

The first is that the textual method can't analyze the topic distribution for a single word (the tag). To overcome this problem, we expand the query tag into a document as follows: we collect all the tags F^q in the image set X and regard F^q as the document describing the query tag q .

Another problem is that it is not reasonable for only one document to analyze the topic distribution, because there is no distinction between documents without references. To address this problem, we generate the reference documents from the co-occurrence tags. We select top P tags with the highest frequency from co-occurrence tag set T , denoted by $t^i, i = 1, 2, \dots, P$. For a selected tag t^i , we pick the images with tag t^i in the whole image set and collect all the tags of these images as a reference document rf_i . Thus, we can obtain the P reference documents $\mathcal{R} \in \{rf_1, rf_2, \dots, rf_P\}$ and rf_i is the co-occurrence tag set of the i -th selected tag $t^i, i = 1, \dots, P$.

To avoid confusion, it's necessary to explain that the T is the co-occurrence set whose elements are unique, while the tags in documents generated above are repeatable.

Cluster documents generation For the sake of analyzing the topic distribution of the image cluster, we generate one document for each image cluster. The cluster documents are generated based on the image clusters in Section 3.4.

Let $ic \in IC$ be an image cluster, we traverse the images in ic and collect all the tags of the images. When the traversing is done, we can obtain a repeatable tag collection F^{ic} , which is regarded as the document describing the cluster.

Thus, for each image cluster $ic \in IC$, we can generate one document F^{ic} . Different from the single query tag, the cluster documents can serve as the references for each other.

Topic distribution estimation After we obtain the query and the cluster documents, the topic distributions of the query tag and clusters could be estimated accordingly using textual topic analysis method like PLSA [15] or LDA [1]. Hereinafter, we take PLSA as an example to estimate the topic distribution of the query tag. The topic distribution of the image cluster can be obtained analogously.

Probabilistic Latent Semantic Analysis (PLSA) is a generative model from the statistics text field. In the analysis of text, it is used to discover topics in a document using the bag-of-words document representation. PLSA thinks that there is a latent topic variable h in the generative process of document F^q and attempts to learn the distribution of different topics for F^q . Conditional probability, the topic distribution of document $P(h|F^q)$ is the parameter we need to learn.

To estimate the topic distribution $P(h|F^q)$ for query q , we treat the tags in the co-occurrence tag set T as the vocabulary and employ the BOW model to represent query documents F^q as well as the reference documents in $\mathcal{R}$. Then EM algorithm [6] can be conducted to obtain the topic distribution of documents. As for the topic number, we think

the tags in the same cluster share similar semantic, therefore we regard a tag cluster as a topic class and set the topic number as $\bar{C}$. When the topic analysis is done, we can obtain $\bar{C}$ dimensions topic distribution vector for each document. The topic distribution Sq of query document F^q is regarded as the topic estimation of the query tag q .

Following similar steps, we can obtain the topic distribution vector SC for the image cluster $ic \in IC$.

3.6 Final diversity ranking

In this subsection, we utilize the learned relevance score f and estimated topic distributions to obtain the final image list by following two steps:

- 1) Select the image with the highest relevance score as the first retrieval result, then the image cluster that the first image belongs to is removed. This step aims at pursuing the higher top 1 accuracy.
- 2) Rank the left image clusters by descending order according to rc_i , which is defined as follow:

$$rc_i = \exp\left(-\frac{\|SC_i - Sq\|^2}{2\sigma_1^2}\right) \quad (15)$$

where SC_i and Sq are the the topic distributions of cluster $ic_i = IC$ and the tag q respectively. σ_1 is a constant, we simply set it as 0.5 in our experiment.

The subsequent images are selected by picking the most relevant image from each ranked cluster according to the relevance score f .

4 Experiments

We conduct experiments on the NUS-Wide dataset by utilizing the same query tags as work [30]. We systematically make comparisons for the proposed semantic diversity ranking (SDR) and following six tag-based image retrieval approaches:

- MMS: Multimodal Stacking [37]. The authors integrate multiple features to train a multi-model classifier. The diversity is achieved by recurrently selecting images that are dissimilar with the items appearing before.
- DRCR: Diverse, Relevance, Topic Coverage, representativeness Ranking [24]. A DRCR score is defined by taking the relevance, topic coverage, and diversity into consideration simultaneously, the final results are obtained by greedy selection.
- TDR: Topic Diverse Ranking [30]. Tag graph and community detection method are utilized to boost diversity performance. A regularization framework that fuses the semantic, visual, and view information is introduced to improve the relevance performance.
- VDR: Visual Diversification Ranking [21]. With the relevance ranking list, three greedy selection algorithms are proposed to diversify the retrieval results. The diversity is achieved by clustering.
- USDM: Uncertainty Sampling with Diversity Maximization [55]. This algorithm learns the importance of training data, it iteratively selects the most helpful training samples and avoid similar samples. The images can be ranked by the classifier trained from the selected samples.

- LUPI: Learning Using Privileged Information [52]. The authors consider the privileged information and actively sample the training points according to the uncertainty and visual diversity. The final ranking can be conducted by the trained classifier.

In our baseline approach, we set $\gamma = 0.2$, $\beta = 5$ and $\xi = 0.1$, $P = 100$ respectively, and PLSA [15] is chosen for topic analysis. We will discuss above four parameters and topic analysis method in Section 5. To make fair comparisons for six methods, we use the parameters that the corresponding paper suggests for MMS [37], DRCR [24], TDR [30], VDR [21], USDM [55], and LUPI [52].

4.1 Visual features

In our experiment, we use the following six visual features (i.e. $K=6$) in our multi-feature fusion mechanism: 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moment and 500-D bag of words based on SIFT descriptions. All these six features are normalized and fed into the multi-feature fusion algorithm.

4.2 Performance evaluation

In the NUS-Wide dataset, each image is manually labeled into two relevance levels: 1-relevant and 0-irrelevant. Thus, we can evaluate the seven methods objectively.

Evaluation Criteria We use the NDCG [50] and average precision under depth n to evaluate the relevance performance, which is defined as follows:

$$NDCG@n = \frac{1}{W} \sum_{i=1}^n \frac{2^{rel_i} - 1}{\log(1 + i)} \quad (16)$$

$$AP@n = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^i \frac{rel_j}{i} \right) \quad (17)$$

where W is a normalization constant introduced to make the optimal NDCG score be 1, rel_i indicates the relevance level of image x_i to the query tag q , which is defined as:

$$rel_i = \begin{cases} 1, & \text{if } x_i \text{ is relevant to } q, \\ 0, & \text{if } x_i \text{ is irrelevant to } q. \end{cases} \quad (18)$$

Denote the top n retrieval image results by $I = \{i_1, i_2, \dots, i_n\}$ and M_k stands for the tag number of image i_k . Then diversity performance of retrieval results can be estimated as follows:

$$DS@n = \frac{1}{n} \sum_{k=1}^n DS_I(i_k) \quad (19)$$

$$DS_I(i_k) = \frac{1}{M_k} \sum_{j=1}^{M_k} \frac{1}{N_{t_j}^I} \quad (20)$$

where $DS_I(i_k)$ denotes the diversity score of image i_k in I , and $N_{t_j}^I$ denotes the number of images in the top-ranked list I that are associated with tag t_j . $DS@n$ represents the average diversity score.

Fig. 3 The performances Under four evaluation criteria. **a** The MNDCG of seven methods under different depths. **b** The MAP of seven methods under different depths. **c** The MDS of seven methods under different depths. **d** The MADP of seven methods under different depths

Moreover, we can calculate the average diverse precision under depth n (denoted as $ADP@n$ [44]) as follow:

$$ADP@n = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^i \frac{rel_j}{i} \right) * DS@i \quad (21)$$

Performance analysis Let $MNDCG@n$, $MAP@n$, $MDS@n$ and $MADP@n$ denote the average value for all test queries. The performance of all methods is shown in Figs. 3a–d, from which we can see that both the relevance and diversity of our method are better than the competing approaches. Figures 3a and b show the relevance performances. For example, from Fig. 3b, the $MAP@20$ of MMS, DRCR, TDR, VDR, USDM, and LUPI are 0.619, 0.624, 0.672, 0.698, 0.726, and 0.736. While our method SDR can reach 0.757. When the depth is 1, the MAP of SDR can reach 0.9, which is much better than the comparison methods. Figure 3a shows the consistent tendency with Fig. 3b, and similar observations can be seen.

Figures 3c and d exhibit the diversity performance. In Fig. 3d, our approach performs much better, and when the depth varies from 5 to 20, our approach all achieves the best performance. The MDS of MMS and DRCR shown in Fig. 3c are very competitive with our method, since the MMS and DRCR pay much attention to the diversity, however they sacrifice too much relevance. For example, the $MDS@5$ of MMS can reach 0.86, while the $MAP@5$ is only 0.57. The situations of USDM and LUPI are opposite with MMS and DRCR, they achieve good relevance, while the diversity is ordinary. The relevance of LUPI is very competitive, whose $MNDCG@20$ even exceeds our method SDR, however, its diversity is unsatisfactory. From Figs. 3a–d, we can find that the relevance and diversity of our method SDR are both satisfactory, which reveals the effectiveness of our method.

Figures 4a and b show the top 10 retrieval results for query “animal” and “birds”, in which the red frame indicates the irrelevance images and the same color dashed frame

Fig. 4 The exemplary results for query “animal” and “birds”

indicates the similar results. As shown in Fig. 4a, compared with our method SDR, the competing methods all lack relevance or diversity. For example, in Fig. 4a, the results of MMS for “animal” is not relevant and diverse enough: the second, the fifth and the tenth images are irrelevant, and the eighth and the ninth image show similar animals. The results determined by USDM are all relevant, however, it introduces many similar images. The results of VDR also lack diversity. While the results determined by our method are not only all relevant but diverse enough.

The semantics of “animal” is general and easy to achieve relevance and diversity. Figure 4b provides the results for a more specific query “birds”. From Fig. 4b, we can see that the methods all lack diversity or relevance. Even SDR and TDR introduce one and two irrelevant images respectively. This reveals that there are still many challenges to enhance the relevance and diversity simultaneously for the difficult query.

5 Discussion

In this section, we will discuss the parameters of our method including the hyperparameters in the multi-feature learning model (6) and the selected tag number P for reference documents generation. Besides, the topic analysis method is also discussed in this section.

5.1 Discussion about the learning parameters in (6)

In this subsection, we discuss the effects of parameter γ , β , ξ in (6). Figures 5a–d show the performances of our method with different hyperparameters.

The average NDCG of different γ is shown in Fig. 5a, the x-axes represents the depth, while the y-axes is the corresponding MNDCG value and the different bars express the SDR with different γ . When the depth is 1, the MNDCG of $\gamma = 0.2, 0.5, 1$ can reach 0.9. With the depth deepening, $\gamma = 0.2$ performs better than the other situations. Another important

Fig. 5 The performances under difference hyperparamters. **a** The MNDCG of $g = 0, 0.1, 0.2, 0.5, 1, 5, 10$ under different depths. **b** The MDS of $g = 0, 0.1, 0.2, 0.5, 1, 5, 10$ under different depths. **c** The MNDCG@20 of different β and ξ . **d** The MDS@20 of different β and ξ

truth from Fig. 5a is that the situations of $\gamma \neq 0$ all perform much better than $\gamma = 0$, which means abandoning the initial score Y (in (6)). This indicates that our initial relevance y provided by *NMF* is reliable and helpful to the retrieval accuracy.

Figure 5b shows the diversity performance with different γ . From Fig. 5b, diversities of different γ are close, and the change of MDS is relatively smooth. The case of $\gamma = 0$ performs litter better than the other cases, however, the relevance performance of $\gamma = 0$ is much poor. The diversity without relevance makes no sense. The case $\gamma = 0.2$ achieves the best relevance and acceptable diversity, therefore it is the appropriate trade-off between relevance and diversity.

Figure 5c shows the relevance performance of depth=20 under different β and ξ , where the x-axes represents the SDR with different ξ , while the different color bars indicate the SDR with different β . For example, the leftmost bar is the MNDCG@20 of SDR with $\beta = 0$ and $\xi = 0$. From Fig. 5c, we can see that the cases of $\beta \neq 0$ all perform better than $\beta = 0$. This reveals that the discriminative power described by the variance of similarity is helpful to the relevance. From Fig. 5d, we can see that the diversity performances under different β and γ are close. Another important truth from Figs. 5c–d is that either the relevance or diversity of our method under different β and ξ are better than the comparison methods.

5.2 Discusstion about the number of reference documents

The parameter P is the number of tags selected from the co-occurrence tag set of q , and each of them is utilized to generate the reference documents for the query topic analysis, more details can be found in Section 3.4. In this subsection, we investigate the effect of parameter P .

Figures 6a–b show the performances with different P , where the meaning of x-axes, y-axes and bars are analogous to Fig. 5's. The relevance performance is shown in Fig. 6a. From Fig. 6a, we can see that the MNDCG@1 can all reach 0.9, when the P varies from

Fig. 6 The performances under different numbers of reference documents. **a** The MNDCG of $P = 20, 50, 100, 200, 400$ under different depths. **b** The MDS of $P = 20, 50, 100, 200, 400$ under different depths

10 to 400, the case $P = 100$ achieves better performances than the other cases. While the MDS shown in Fig. 6b exhibits smooth change under different P .

From the discussion above, the case of $P = 100$ achieves better relevance and diversity. Too lower P will miss the meaningful reference documents and too higher P will introduce some noisy references, both situations will confuse the topic analysis and make the performance drop.

Although the case $P = 100$ performs best, the other situations still achieve acceptable performance, which validates the effectiveness of our method.

5.3 Discuss the cluster feature representation

In our method, we analyze the topic distribution of cluster as its feature representation, by which we directly compute the relevance score to the query. In our previous work [30], we use the tag frequency histogram to represent the cluster and apply an adaptively random walk model to rank the clusters. In this subsection, we compare the efficiency of these two diversity mechanisms.

Figure 7 shows the performances of three different mechanisms, where SDR-H and SDR-T represent our SDR applying tag frequency histogram and topic distribution vector for cluster representation respectively. From Fig. 7, the topic distribution is better than the frequency histogram except for MDS, which means that the topic distribution vector is a more efficient representation for clusters. Another advantage of topic distribution representation

Fig. 7 The performance of difference cluster feature representations

is that we can directly compute the relevance of the image cluster to the query tag, which is more convenient and intuitive.

SDR-H and TDR apply the same diversity mechanism, and the only difference is the relevance ranking method. From Fig. 7, we can observe that the SDR-H performs better than TDR under all criteria. This reveals that the relevance ranking algorithm proposed by this paper is more outstanding than our previous work TDR.

From Fig. 7, we can also observe that although SDR-H is not as good as SDR-T, it also achieves acceptable performance. Therefore, the topic distributions of clusters can also simply be replaced by histogram mechanism. This reveals that the topic analysis technique is not the key to our algorithm. In this paper, the reasons for using the topic distribution are two-fold: 1) Topic distribution is more intuitive. Once the topic distributions of query tag and the clusters are estimated, the relevance scores between clusters and the query can be calculated directly. 2) Employing topic distribution can sacrifice less relevance, which is clear from Fig. 7. Therefore, the topic distribution representation is more appropriate for our architecture.

5.4 Validate the importance of reference documents

In the process of query topic analysis, we not only generate one document for the query tag but generate some reference documents to analyze the topic distribution of the query tag. In this subsection, we conduct experiments to validate the necessity of the references.

Figure 8 shows the performance of two mechanisms: SDR-R is our method with references while SDR-Q only employs the query document. From Fig. 8, we can observe that the SDR with the reference documents performs much better than SDR with query document only. For example, the MAP of SDR-R can reach 0.756, while the SDR-Q is only 0.718, and the diversity of SDR-R is still better than SDR-Q.

Although topic analysis methods like PLSA or LDA can analyze the topic distribution for only one document, the analysis process is inadequate without the references, consequently, the obtained topic distribution is not reliable. Therefore, the reference documents are necessary and important for generating satisfactory topic distribution of the query.

5.5 Discuss about the topic analysis method

For our baseline approach, we choose the PLSA for topic analysis in our cluster ranking step. LDA [1] is also widely used for topic analysis. In this subsection, we employ the LDA for comparison and analyze the effects of different topic analysis methods.

Fig. 8 The performances of SDR with references and without references

Table 1 Results of our method with PLSA and LDA under depth=20, the bold indicates better results

	MNDCG	MAP	MDS	MADP
PLSA	0.738	0.757	0.748	0.621
LDA	0.726	0.738	0.764	0.612

Table 1 shows the performance of four evaluation criteria under depth=20. From Table 1, we can see that the performance of LDA is also acceptable and better than the comparison methods. Compared to the PLSA, our approach based on LDA can achieve better diversity, while the relevance of LDA is litter worse than PLSA. There is no doubt that we should ensure the relevance first, therefore PLSA is more appropriate for our approach. Easy to understand and implement is another reason for choosing the PLSA.

6 Conclusion and future work

In this paper, we propose a novel diversifying framework for tag-based image retrieval. We first employ a regularization model to refine the relevance score from *NMF* and then diversify the retrieval results by topic analysis, in which the topic number is adaptively chosen. We achieve diversity by picking one image from each image cluster. Extensive experiments on NUS-Wide dataset reveal that the initial relevance score provided by *NMF* is credible, and the topic distribution vector estimated from our generated documents is a reliable representation for query topic, and diversifying the result based on topic analysis is efficient. The discussion experiments show that our method achieves satisfactory performance with different parameters and dependency methods.

However, we don't investigate the semantic difference between clusters and implicitly think the images in different clusters are semantically dissimilar. This problem is left for our future work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Blei D, Ng A, Jordan M (2011) Latent dirichlet allocation. NIPS, pp 601–608
2. Cai D, He X, Li Z, Ma W, Wen J (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proc. ACM Multimedia
3. Calumby R, Torres R, Goncalves M (2014) Diversity-driven learning for multimodal image retrieval with relevance feedback. IEEE International Conference on Image Processing, pp 2197–2201
4. Cui H, Zhu L, Li J, Yang Y, Nie L (2019) Scalable deep hashing for large-scale social image retrieval. IEEE Trans Image Process 29:1271–1284
5. Dang-Nguyen D, Piras L, Giacinto G, Boato G, Natale F (2014) Retrieval of diverse images by pre-filtering and hierarchical clustering, MediaEval

6. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39(1):1–8
7. Du X, Liu Q, Li Z, Qin Z, Tang J (2019) Cauchy matrix factorization for Tag-Based social image retrieval. *IEEE Access*, pp 123–132
8. Frey B, Dueck D (2007) Clustering by passing messages between data points. *Science* 315(5814):972–976
9. Gao L, Li X, Song J, Shen HT (2019) Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Transactions on Pattern analysis and machine intelligence* 42(5):1112–1131
10. Gao Y, Wang M, Zha Z, Shen X, Wu X (2013) Visual-textual joint relevance learning for tag-based social image search. *IEEE Trans on Image Processing* 22(1):363–376
11. Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (2014) Generative adversarial nets. In: *NIPS*
12. Gu Y, Qian X, Li Q (2015) Image annotation by latent community detection and multi-kernel learning. *IEEE Trans Image Process* 24(11):3450–3463
13. Haruechaiyasak C, Damrongrat C (2010) Improving social tag-based image retrieval with CBIR technique. Springer, Berlin, pp 212–215
14. Haubold A, Natsev A, Naphade M (2006) Semantic multimedia retrieval using lexical query expansion and model-based reranking. *ICME*, pp 1761–1764
15. Hofmann T (1999) Learning the similarity of documents: an information-geometric approach to document retrieval and categorization. *NIPS*, pp 914–920
16. Hoque E, Hoerber O, Gong M (2013) CIDER: Concept-based Image diversification. *Exploration, and Retrieval* 49(5):1122–1138
17. Ksibi A, Ammar A, Amar C (2014) Adaptive diversification for tag-based social image retrieval. *International Journal of Multimedia Information Retrieval* 3(1):29–39
18. Ksibi A, Feki G, Ammar A, Amar C (2013) Effective diversification for ambiguous queries in social image retrieval. In: *Computer Analysis of Images and Patterns*, pp 571–578
19. Lee S, Neve W, Ro Y (2010) Tag refinement in an image folksonomy using visual similarity and tag co-occurrence statistics. *Signal Process Image Commun* 25(10):764–773
20. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791
21. Leuken R, Garcia L, Olivares X, Zwol R (2009) “Visual diversification of image search results. In: *Proc WWW Conf*, pp 341–350
22. Li X, Snoek C, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. *Proceedings of the ACM international conference on multimedia information retrieval*, pp. 180–187
23. Lin B, Wei A, Tian X (2016) Visual re-ranking through greedy selection and rank fusion. *International Conference on Multimedia Modeling*, pp 189–300
24. Lin X, Zhang T (2017) Image search reranking with relevance, diversity and topic coverage. *Proceeding of the International Conference on Internet Multimedia Computing and Service*, pp 105–109
25. Lu D, Liu X, Qian X (2016) Tag based image search by social re-ranking. *IEEE Transactions on Multimedia* 18(8):1628–1639
26. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space, *ICLR Workshop*
27. Natsev A (2007) Semantic concept-based query expansion and re-ranking for multimedia retrieval. *ACM International Conference on Multimedia*, pp 991–1000
28. Qian X, Hua X, Tang Y, Mei T (2014) Social image tagging with diverse semantics. *IEEE Trans Cybernetics* 44(12):2493–2508
29. Qian X, Liu X, Zheng C (2013) Tagging photos using users’ vocabularies. *Neurocomputing* 111(111):144–153
30. Qian X, Lu D, Wang Y, Zhu L, Tang Y, Wang M (2017) Image re-ranking based on topic diversity. *IEEE Trans on Image Processing* 26(8):3734–3747
31. Qian X, Xue Y, Tang Y, Hou X (2015) Landmark summarization with diverse viewpoints. *IEEE Trans. Circuits and Systems for Video Technology* 25(11):1857–1869
32. Ren Z, Jin H, Lin Z, Fang C, Yuille A (2016) Joint image-text representation by gaussian visual-semantic embedding. *Proceeding of the 2016 ACM on Multimedia Conference*, pp 207–211
33. Shen J, Mei T, Tian Q, Gao X (2014) Image search reranking with multi-latent topical graph. *IEEE International Symposium on Circuits & Systems*, pp 1–4
34. Song J, He T, Gao L, Xu X, Hanjalic A, Shen HT (2020) Unified binary generative adversarial network for image retrieval and compression, *IJCAI*, pp 1–22
35. Song K, Tian Y, Huang T, Gao W (2006) Diversifying the image retrieval results. In: *Proc ACM Multimedia Conf*, pp 707–710

36. Song J, Zhang H, Li X, Gao L, Wang M, Hong R (2018) Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Trans Image Process* 27(7):3210–3221
37. Spyromitros-Xioufis E, Papadopoulos S, Ginsca A, Popescu A, Kompatsiaris Y (2015) Improving diversity in image search via supervised relevance scoring. *Proceeding of the 5th ACM on international conference on multimedia retrieval*, pp 323–330
38. Sun F, Wang M, Wang D (2012) Optimizing social image search with multiple criteria: Relevance, diversity, and typicality. *Neurocomputing* 95:40–47
39. Tang J, Li Z (2017) Weakly supervised multimodal hashing for scalable social image retrieval. *IEEE TCSVT* 28(10):2730–2741
40. Tang J, Li Z (2018) Weakly supervised multimodal hashing for scalable social image retrieval. *IEEE TCSVT* 28(10):2730–2741
41. Tian X, Yang L, Lu Y (2015) Image search reranking with hierarchical topic awareness. *IEEE Trans on Cyber* 45(10):2177–2189
42. Wang M, Li H, Tao D, Lu K, Wu X (2012) Multimodal graph-based reranking for web image search. *IEEE Trans on Image Processing* 21(11):4649–4661
43. Wang G, Xu X (2012) Joint-rerank: a novel method for image search reranking. *Proceeding of the 2nd ACM International Conference on Multimedia Retrieval*, article no 37
44. Wang M, Yang K, Hua X, Zhang H (2010) Towards relevant and diverse search of social images. *IEEE T Multimedia* 12(8):829–842
45. Wang Y, Yang H, Qian X, Ma L, Lu J, Li B, Fan X (2019) Position focused attention network for image-text matching. *IJCAI*, pp 3792–3798
46. Wang Y, Zhu L, Qian X (2019) Joint hypergraph learning for tag-based image retrieval. *IEEE TIP* 27(9):4437–4451
47. Wang Y, Yang H, Bai X, Qian X, Ma L, Lu J, Li B, Fan X (2020) PFAN++: Bi-Directional Image-Text retrieval with position focused attention network. *IEEE TMM*. <https://doi.org/10.1109/TMM.2020.3024822>
48. Wang L, Qian X, Zhang Y, Shen J, Cao X (2020) Enhancing sketch-based image retrieval by cnn semantic re-ranking. *IEEE Trans Cybern* 50(7):3330–3342
49. Wang L, Qian X, Zhang X, Hou X (2020) Sketch-based image retrieval with multi-clustering re-ranking. *IEEE TCSVT* 30(12):4929–4943
50. Wu D, Wu J, Lu M (2014) A two-step similarity ranking scheme for image retrieval. In: *Parallel architectures, IEEE Algorithms and Programming*, pp 191–196
51. Yan Y, Liu G, Wang S, Zheng K (2017) Graph-based clustering and ranking for diversified image search. *Multimedia Systems* 23(1):41–52
52. Yan Y, Nie F, Li W, Gao C, Yang Y, Xu D (2016) Image classification by Cross-Media active learning with privileged information. *IEEE Trans. Multimedia* 18(12):2494–2502
53. Yanai K, Nga D (2013) UEC Tokyo at MediaEval, 2013 retrieval diverse social images tasks. *MediaEval*
54. Yang Y, Huang X, Shen H, Zhou X (2011) Mining multi-tag association for image tagging. *World Wide Web* 14(2):133–156
55. Yang Y, Ma Z, Nie F, Chang X, Hauptmann A (2015) Multi-class active learning by uncertainty sampling with diversity maximization. *IJCV* 113(2):113–127
56. Yang X, Mei T, Zhang Y, Liu J, Satoh S (2016) Web image search re-ranking with click-based similarity and typicality. *IEEE Trans on Image Processing* 25(10):4617–4630
57. Yang X, Zhang Y, Yao T, Ngo C, Mei T (2015) Click-boosting multi-modality graph-based re-ranking for image search. *Multimedia Systems* 21(2):217–227
58. Yao T, Mei T, Ngo C (2015) Learning query and image similarities with ranking canonical correlation analysis. *ICCV*, pp 28–36
59. Yu J, Tao D, Wang M (2014) Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans on Cybernetics* 45(4):767–779
60. Yu J, Yang X, Gao F, Tao D (2016) Deep multimodal distance metric learning using click constraints for image ranking. *IEEE Trans on Cyber* 47(12):1–11
61. Zhai H, Lai S, Jin H, Qian X, Mei T (2020) Deep transfer hashing for image retrieval. *IEEE TCSVT*. <https://doi.org/10.1109/TCSVT.2020.2991171>
62. Zhang Y, Yang X, Mei T (2014) Image search reranking with query-dependent click-based relevance feedback. *IEEE Trans on Image Processing* 23(10):4448–4459
63. Zhu G, Yan S, Ma Y (2010) Image tag refinement towards low-rank, content-tag prior and error sparsity. *Proceeding of the 18th ACM Multimedia*, pp 461–470