Zhao *et al. EURASIP Journal on Image and Video Processing* (2020) 2020:53
https://doi.org/10.1186/s13640-020-00540-4

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Online social image ranking in diversified preferences

Xuezhuan Zhao[1,2†], Lishen Pei[2,3*†], Tao Li[4] and Zheng Zhang[4]

*Correspondence:
lishen198815@126.com
†Xuezhuan Zhao and Lishen Pei contributed equally to this work.
[2]Henan University of Economics and Law, Jinshui Road, 450046 Zhengzhou, China
[3]Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, 300300 Tianjin, China
Full list of author information is available at the end of the article

Abstract

Due to the prevalence of social media service, effective and efficient online image retrieval is in urgent need to satisfy diversified requirements of Web users. Previous studies are mainly focusing on bridging the semantic gap by well-established content modeling with semantic information and social tagging information, but they are not flexible in aggregating the diversified expectations of the online users. In this paper, we present OSIR, a solution framework to facilitate the diversified preference styles in online social media image searching by textual query inputs. First, we propose an efficient Online Multiple Kernel Ranking (OMKR) model which is constructed on multiple query dimensions and complimentary feature channels, and trained by minimizing the triplet loss on hard negative samples. By optimizing the ranking performance with multi-dimensional queries, the semantic consistency between the image ranking and textual query input is directly maximized without relying on the intermediate semantic annotation procedure. Second, we construct random walk-based preference modeling by domain-specific similarity calculation on heterogeneous social attributes. By re-ranking the rank output of OMKR based on each preference ranking model, we obtain a set of ranking lists encoding different potential aspects of user preference. Last, we propose an effective and efficient position-sensitive rank aggregation approach to aggregate multiple ranking results based on the user preference specification. Extensive experiment on two social media datasets demonstrates the advantages of our approach in both retrieval performance and user experience.

Keywords: Social media image retrieval, Online multiple kernel, Ranking, Preference modeling, Rank aggregation

1 Introduction

Multimedia content searching in Web space is a very challenging task. The prevalence of social media service makes this task even harder due to the diversified user preference and heterogeneous user behaviors. Online users usually present themselves by transmitting online multimedia to their social circles and contributing user-generated content ad hoc with mobile devices. For example, users share interesting photos with rich tags and comments to friends, they would like to show where they are or what they are doing at this moment with pictures and the corresponding location information, they put tags and comments to certain images to express their feelings about the content therein, and they

also categorize their favorite images into several online albums. Consequently, the online social multimedia documents, especially the huge volume of images, are associated with a lot of meta-information and social user-related attributes, e.g., location, upload time, user, and community. Despite that the huge number of social images indeed provides chance to develop models for social image retrieval, most of existing works only learn models that capture the preference towards the whole user community instead of a single user or a small group of users. As a consequence, identical results tend to be returned to online users given a specific query input, which tends to be less desirable. Therefore, effective method is required to meet the diversified preference styles among the user community.

To address this practical problem, a possible paradigm is to construct real-world image retrieval methods by content-based visual analysis [1] and semantic-based analysis [2] with the content information and co-occurred semantic information (e.g., labels and tags). For content-based analysis, retrieval models are constructed on the local (e.g., Bag-of-Visual-Word) and global (e.g., Gist and Edge histogram) visual feature representation and state-of-the-art deep convolutional neural network (CNN) features. Accordingly, the models for visual content hashing, indexing, and similarity learning are deeply investigated. For semantic-based analysis, the images and the queries (visual or textual) are projected into the multi-dimensional semantic space. The similarities between the queries and database images are calculated on the semantic space. However, for social media images with heterogeneous information beyond the visual content and semantics, the existing content-based and semantic-based approaches are not flexible to satisfy the user's true needs reflected as user preferences. For example, given a query "sunflower," some user may prefer the sunflower taken in the wild, while some others may prefer sunflower taken in the greenhouse. When we look into the retrieved result, if the images in different environments are contained in the top ranked list, it would be better that their relative positions in the rank list for different users are different. Towards this objective, the technical challenges and our proposals are as follows.

First, existing approaches bridge the semantic gap by visual modeling or semantic annotation. But their learning procedure does not maximize a criterion directly related to the final retrieval performance. Instead, they maximize alternative criteria such as the annotation performance or the descriptive power of visual features. However, in practice, user queries are highly diversified. Despite that the criteria difference can be compensated by bridging the intention gap [3] or user interaction [4–6], the query-independent approaches do not directly fit well to the user needs expressed in the queries.

In this work, we introduce a method called Online Multiple Kernel Ranking (OMKR) to directly learn the image retrieval model without relying on semantic annotation. Our model adopts a learning criterion [7] related to the final retrieval performance based on discriminative learning. It takes as input a set of training queries as well as a set of ranked online social media images, and outputs a trained model to achieve high ranking performance on new queries. By combining multiple visual features and exploring the correlation among different query words, our model achieves better model generality. OMKR is also featured with an efficient online optimization procedure which builds upon the online multiple kernel learning framework [8]. Therefore, it permits learning over large training data.

Second, when users are searching online images, the queries are words expressing what the users want to search. However, in most situations, the user preference is usually not

expressed in a query with several words. Instead, different users tend to give the same query words on a certain topic, but their intrinsic expectation on the returned images may be different from person to person. For example, with the query "car," *Alex* would like to search car listing information specifically available in London since he lives in the city, and *Tony* would expect to search for car images that receive the most positive comments or car review reports, as he will buy a car very soon, while *Anya* would expect to find specific car images shared in certain groups when she considers to join in certain online user groups with similar interests (vintage car or refitted vehicle). Such diversified preference styles are usually unavailable in practice because it is always hard to require users to be professional and precise on describing what they truly demand. Fortunately, we can retrieve user preference by exploiting from related users and the rich context information of social media Websites, e.g., the temporal adjacency, the location affinity, the gallery information, the associated user groups, and the positive/negative comments. Based on the study of McAuley et al. [9], these social network meta-data provide an informative signal for certain image categories. Therefore, promising performance has been achieved even when the visual features are not employed for image labeling and tag prediction tasks. In this paper, we call the associated meta-data as the *social attributes* of online images. We construct random walk models on each social attribute and re-rank the results of OMKR according to the potential preference expressed in each social attribute.

Moreover, as each ranking metric captures only some aspect of the consistency with respect to certain social attribute, it is beneficial to combine different ranking metrics to accurately identify what a user really needs. The problem of rank aggregation or preference aggregation has been extensively studied in social choice theory [10]. We propose an order-based technique with the weighted position-sensitive measurement. Compared with the traditional rank aggregation models, our model achieves better rank aggregation results with low computational cost. Consequently, a set of ranking results that encode both the semantic ranking and potential preference are obtained based on the user preference specification.

To summarize, in this paper, we study the problem of social image retrieval satisfying both the semantic consistency and diversified user preference styles. To solve the above challenges, we present online social image ranking (OSIR), a direct solution framework for social media image retrieval in diversified preference styles. The model is flexible in utilizing state-of-the-art visual and textual features such as word embedding and the multiple feature layers of a deep CNN.

Our approach produces the final ranking of the retrieved social media images in a way similar to the preparation of cocktail drink, a kind of alcoholic mixed drink that contains two or more ingredients to fit the diversified user preference styles. The key contributions can be summarized as follows:

(1) The *Online Multiple Kernel Ranking* maximizes the semantic consistency between the top ranked images and the multi-dimensional textual queries, by combining complimentary visual features and minimizing the hard negative-based triplet loss, similar as [11]. The learning procedure quickly converges in receiving less than 30 thousand triplets.

(2) The *Rank Aggregation* appropriately aggregates various social attribute correlations among different images. By modeling the relative importance of each top ranked image, a unified ranking that better fits user preference is obtained.

(3) *Experiments on real social image retrieval* demonstrate that OSIR outperforms state-of-the-art. Besides, the subjective study shows that the aggregate ranking satisfies the user preference beyond semantic consistency.

Roadmap. Section 2 provides a brief literature review. Section 3 gives the framework overview. Section 4 introduces the Online Multiple Kernel Ranking. Section 5 describes the preference modeling. Section 6 presents the position-sensitive rank aggregation. Section 7 provides the experimental details and discussion. Section 8 concludes the paper.

2 Related work

Great effort has been dedicated to modeling different aspects of online multimedia retrieval and different user browsing behaviors for visual content retrieval. In this paper, we provide a brief literature review from the following aspects.

2.1 Visual-semantic retrieval

For decades, image retrieval has been a core research problem in multimedia research community. Research efforts have been made to bridge the semantic gap between the user queries and the multi-dimensional content representation [1]. For example, Grangier and Bengio [7] proposed a discriminative kernel-based ranking approach for image retrieval by textual queries. Rasiwasia et al. [2] constructed a unified semantic space for cross-modal data, which was based on what documents to be retrieved by queries from other modalities. Following this idea, correlation learning from multiple modalities has been comprehensively studied [12, 13]. Zhang et al. [3] proposed an attribute-augmented semantic hierarchy for content-based image retrieval. Since then, fusing complementary information for visual content modeling has been a widely accepted paradigm to achieve better semantic consistency [14]. Li et al. [15] propose a deep collaborative embedding method for social image tagging, tag-based image retrieval, and content-based image retrieval.

Due to the success of deep neural network work, the deep model has been employed in image retrieval tasks. For example, Gordo et al. [16] proposed an end-to-end trainable deep convolutional neural network model for image retrieval, which has been treated as a standard CNN-based pipeline. To address the modality difference between visual and textual modalities, there are numerous work in recent years. Ma et al. [17] propose a multi-modal convolutional neural network which explicitly captures and aggregates the multi-level visual-textual component correlation for measuring the visual-textual correlation. Similarly, Lu et al. [18] propose a hierarchical co-attention model to adaptively learn the visual-textual component correlation for visual question answering (VQA) task. Recently, given the success of large-scale pretrained model (e.g., the BERT model [19]) and self-supervised learning paradigm, numerous joint visual-textual deep representation models have been proposed. For example, Lu et al. [18] propose a pretrained task-agnostic model for visiolinguistic representation which can be used as the backbone network for various vision-language tasks, such as image-sentence retrieval, image captioning, and VQA. However, these models are developed by assuming the textual modality to be a sentence, which is slightly different from our setting where the query is assumed to be several tags.

As an effective and efficient solution framework for image retrieval and text-to-image retrieval, online similarity learning [20–25] has been studied extensively. Specifically,

Chechik et al. [20] develop a large-scale online asymmetric similarity learning method from ranking. Xia et al. [22] develop a multiple kernel similarity learning for visual search. An online multi-modal distance learning method has also been proposed in [23]. As a recent achievement, Wu et al. propose an online asymmetric similarity learning method for text-image retrieval which aggregates the visual features of different CNN layers. An online low-rank similarity learning method is also proposed in [25] to obtain a low-rank similarity parameter matrix for measuring similarity between image and text. Despite the effectiveness of using hard negatives for retrieval model learning [11], it has not been considered in the context of online learning for text-to-image retrieval.

Based on active learning and relevance feedback, the intention gap can be effectively reduced [3, 26]. Fan et al. [5] proposed a personalized image recommendation via exploratory search modeling. Tian et al. [4] proposed an active re-ranking approach. Zhang et al. [27] propose an active learning method for image classification, which indicates if an image should be labeled by states in the generative adversarial network. On mobile platforms, Wang et al. [28] proposed an interactive mobile visual search with multi-modal queries. However, they either assume the active learning process contains less user preference, or assume that the user preference is obtained by interactions. In contrast, our approach directly fits images into the query space, and the intention gap problem is naturally avoided.

Enforcing diversity in retrieved content has become an important research issue in recent years [29–31]. Generally, from the visual content perspective, the top ranked retrieved images are expected to be as diversified as possible so as to deliver richer content information under the same semantics. For example, Ionescu et al. [29] propose to enhance the diversity of the social image dataset by multiple technical treatments, e.g., machine analysis, human-based computation, or hybrid approaches. The semantic relevance and diversity, nevertheless, are considered to be somehow contradictive in existing solutions. A supervised relevance scoring approach was proposed in [31] to re-rank the social images by optimizing the utility function that jointly considers the two issues, and finally, a better trade-off between relevance and diversity can be achieved. Wu et al. study how the diversity affects user satisfaction in image search [30]. Specifically, when users want to collect information or save images for further usage, more diversified result lists lead to higher satisfaction levels. The insights may help to design better result ranking strategies and evaluation metrics. Besides, diversity is also enforced in other applications such as image recommendation [32], movie recommendation [33], and general purpose recommendation tasks [34]. Similar as the retrieval task, it has been shown that more diversity can bring user with better experience.

2.2 Modeling social context

Online multimedia documents are believed to be correlated to each other on different aspects where such context information is delivered by their meta-data. The context and correlation usually have strong relevance to their semantics. McAuley et al. [9] showed that image labeling with mere social media meta-data performed equally or even outperformed visual content modeling method.

In existing study, the knowledge discovered from context has been employed in many recommendation tasks. For example, as a similar task with retrieval, the friend suggestion/recommendation aims to recommend friend to users according to the similarity

between friend candidates and targeted user. The user similarity can be made by joint content and context analysis [35]. The techniques can also be used for other new tasks. Based on the photographing behavior from the user crowd, Yin et al. [36] developed a socialized mobile photography model to suggest the optimal view enclosure (composition) and appropriate camera parameters by comparing the visual similarity of the query scene and the social image database with diversified photographing styles. Heterogeneous user behaviors can be modeled by the social context of online social media and effectively combine the multi-aspect behavior similarities by multiple kernel learning towards friend recommendation, advertisement, and people searching [37]. Our approach captures the potential preference styles from heterogeneous social attributes. Consequently, the user expectation on the retrieval results can be conveniently expressed by weight specification.

2.3 Ranking aggregation and refinement

Rank aggregation [38] has been recognized as a key technology for Web-based applications. The necessity to meaningfully aggregate preference ranking into a joint ranking has been deeply investigated to provide information fusion from multiple sources and diversified social choices. Rank aggregation is specially useful in crowdsourcing [39], where different users/annotators produce ranking lists with diversified results. From methodology perspective, Prati [40] proposed to combine feature ranking algorithms through rank aggregation. Ding et al. [41] propose a hierarchical ranking aggregation method. An iterative ranking aggregation method is proposed in [42] using quality improvement of subgroup ranking. Liang et al. [43] propose a manifold learning method for rank aggregation.

In multimedia research domain, Tian et al. [44] proposed a ranking SVM-based approach to identify the best ranking from a number of candidate ranking lists for image re-ranking. Yeh et al. [6] developed a personalized photograph ranking framework with various visual aspects. Zha et al. [45] constructed a probabilistic model for product ranking with hundreds of aspects. Motivated by social choice theory, a supervised Kemeny rank aggregation was proposed to aggregate multiple rankings with different credibilities [46]. Dalal et al. [47] developed a globally consistent multi-objective ranking based on Hodge decomposition. Klementiev et al. [48] proposed a probabilistic distance-based model. Our rank aggregation approach considers the relative importance of the position of a document which appears in a rank list, while existing approaches usually treat the rank of each document without discrimination.

Rank aggregation has also been used in other research topic such as person re-identification [49] and POI ranking in spatial-temporal data mining [50]. However, the time complexity of existing rank aggregation is generally prohibitive, which hinders rank aggregation to be applied to a wider range of application scenarios.

3 Method

The aim of OSIR is to provide an aggregate image ranking results given user query inputs. The ranking is expected to achieve better consistency in semantics and the preference styles of the users. The framework is illustrated in Fig. 1. OSIR is essentially composed of the following key steps:

Online Multiple Kernel Ranking. We propose an Online Multiple Kernel Ranking (OMKR) approach by minimizing the hard negative-based triplet loss to rank the images

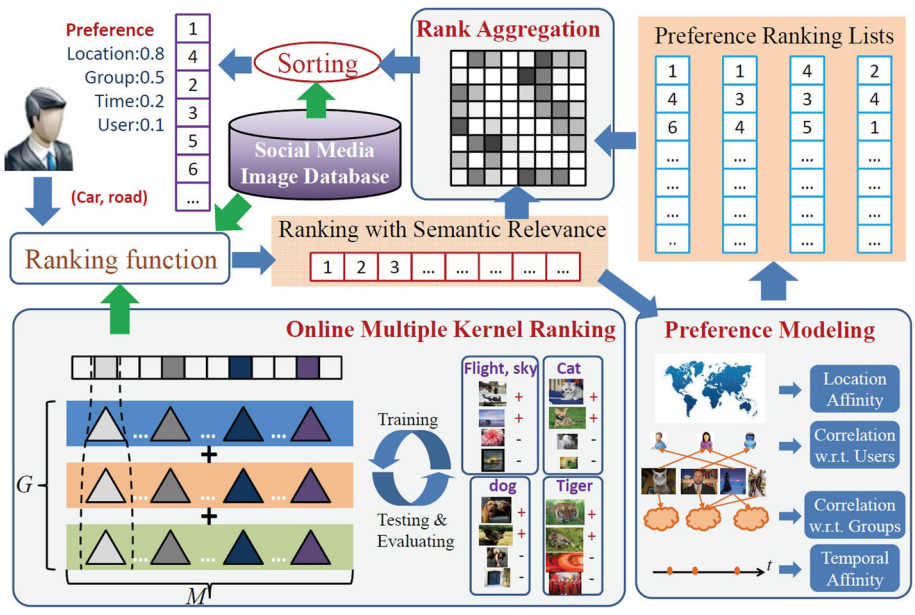

Fig. 1 The framework of OSIR. The blue arrows denote the data flow and their ranking results. The green arrows represent the support from the database and algorithm. Given a user query and preference weight specification, first, the ranking list with the semantic ranking function is learned by our OMKR model. Then, the semantic ranking is fed into the preference modeling, and we obtain a set of preference ranking lists by preference re-ranking. These ranking lists including the semantic ranking results are aggregated with our position-sensitive rank aggregation technique. The ranking results that aggregate both semantics and preference information are finally returned to the users

according to their semantic consistency with the multi-dimensional textual query input. Compared with the existing approaches, our model directly fits the images into the query space. The better semantic consistency is achieved by combining complementary visual features. We design an online learning procedure which quickly optimizes the ranking model with a large number of training triplets where the negative samples in the triplet are selected from the most similar ones to their positive counterparts, and the model quickly converges in receiving less than 30 thousand triplets. Consequently, we learn a set of semantic coherent projections which map each image into a low-dimensional semantic space where the relevance between the queries and the database images can be directly calculated by inner product.

Preference modeling. We construct random walk models on each social attribute of social media images. By using domain-specific knowledge, the social attribute correlations among different images are properly measured. We re-rank the semantic ranking respectively based on each of the preference models. Thus, a set of ranking lists encoding different potential aspects of user preference can be obtained.

Position-sensitive rank aggregation. Based on social choice theory [10], we propose a position-sensitive rank aggregation model to measure the relative importance of the top ranked results given the user preference specification. By aggregating the semantic ranking and preference ranking results, a unified ranking is obtained to achieve better consistency in both semantics and the user preference styles of the users.

4 Online Multiple Kernel Ranking

4.1 Ranking model

An image database is represented as $\mathcal{D} \in \mathbb{R}^{N\times V}$ where N denotes the number of images and V denotes the number of feature dimensions for each image. We denote each image as $d \in \mathbb{R}^V$ which represents a row of $\mathcal{D}$. We represent the textual query as an M-dimensional real value vector $q = (q_1, ..., q_M) \in \mathbb{R}^M$ where there may be multiple non-zero entries for multi-word query input. The score function $F_w(q, d)$ of an image d from $\mathcal{D}$ can be written as:

$$\begin{aligned} F_w(q,d) &= q \cdot f_w(d) = q \cdot (w_1 \cdot d, ..., w_M \cdot d) \\ &= \sum_{m=1}^{M} w_m \cdot (q_m d) = w \cdot \gamma(q,d) \end{aligned} \tag{1}$$

where $\gamma(q, d) = [q_1 d; ...; q_M d]$. For each query q, suppose we have collected the ranking information (*relevant* or *irrelevant*) of the images in the database $\mathcal{D}$. In this paper, the queries are assumed to be closed set, i.e., the number of query words is fixed. It is possible to extend the query to process those queries that are even semantically unrelated to the training set. For example, we may resort to the latent topic modeling methods which use linear/deep mapping functions to process the BOW features of the queries and derive the latent representation; then, the rank model can be constructed based on the topic level instead of the word level. To deal with unseen query words, we may also use more recent methods such as word embeddings to produce an aggregated multi-dimensional query representation.

Another important issue is the relevance/irrelevance score used in this paper. In general, one image is considered to be relevant if it contains visual content describing even one query word. Extending the relevance score to multi-level case would result in the usage of other ranking loss function such as the list-wise ranking loss.

Based on the above definition, we organize the data into a training triplet set D_{tr} where each triplet is represented as $(q, d^+, d^-) \in D_{tr}$. The ranking function learning is equivalent to minimizing the following primal ranking SVM (RSVM) objective function [7]:

$$\begin{aligned} &\min_{\mathbf{w}} \tfrac{1}{2}||\mathbf{w}||^2 + \tfrac{C}{|D_{tr}|} \sum_{(q,d^+,d^-)\in D_{tr}} \xi_{(q,d^+,d^-)} \\ &s.t. \quad \mathbf{w}\cdot\gamma\left(q,d^+\right) - \mathbf{w}\cdot\gamma\left(q,d^-\right) \geq 1 - \xi_{(q,d^+,d^-)} \\ &\qquad \xi_{(q,d^+,d^-)} \geq 0 \end{aligned} \tag{2}$$

where $\mathbf{w} = \left[w_1^T, ..., w_M^T\right]$ denotes the concatenated discriminative model parameter vector. We can introduce any kernel function $\kappa : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ for calculating the similarity among images in high-dimensional space. Consequently, the discriminative functions $f_m, m = 1, ..., M$ and the score function $F(q, d)$ can be represented as:

$$\begin{aligned} f_m(d) &= \sum_{j=1}^{|D_{tr}|} q_{mj}\alpha_j \left(\kappa\left(d_j^+, d\right) - \kappa\left(d_j^-, d\right)\right) \\ F(q', d) &= \sum_{m=1}^{M}\sum_{j=1}^{|D_{tr}|} q_m' q_{mj}\alpha_j \left(\kappa\left(d_j^+, d\right) - \kappa\left(d_j^-, d\right)\right) \end{aligned} \tag{3}$$

When the similarity among images is represented by multiple kernels $\kappa_g, g = 1, .., G$, according to the representer theorem, the discriminative function and score function are formulated by [51]:

$$
\begin{aligned}
f_m(d) &= \sum_{j=1}^{|D_{tr}|}\sum_{g=1}^{G} q_{mj}\alpha_j\beta_{mg}\left(\kappa_g\left(d_j^+,d\right)-\kappa_g\left(d_j^-,d\right)\right)\\
F(q^{'},d) &= \sum_{m=1}^{M} q_m^{'} f_m(d) = \\
&\sum_{m=1}^{M}\sum_{j=1}^{|D_{tr}|}\sum_{g=1}^{G} q_m^{'} q_{mj}\alpha_j\beta_{mg}\left(\kappa_g\left(d_j^+,d\right)-\kappa_g\left(d_j^-,d\right)\right)
\end{aligned}
\tag{4}
$$

By introducing the *Lagrangian* and Karush-Kuhn-Tucker (KKT) condition, we obtain the following dual problem:

$$
\begin{aligned}
&\min_{\beta}\max_{\alpha}\alpha^{\top}\mathbf{1}-\alpha^{\top}\left(\sum_{m=1}^{M}\sum_{g=1}^{G}\beta_{mg}\Theta_{mg}\right)\alpha\\
&s.t.\ 0\le\alpha_j\le\tfrac{C}{|D_{tr}|},\ \sum_{g=1}^{G}\sum_{m=1}^{M}\beta_{mg}=1,\forall m
\end{aligned}
\tag{5}
$$

where $\alpha\in\mathbb{R}^{M\times|D_{tr}|}$, $\beta\in\mathbb{R}^{M\times G}$, and $\Theta_{mg}\in\mathbb{R}^{|D_{tr}|\times|D_{tr}|}$ is a positive semi-definite matrix where:

$$
\begin{aligned}
&\Theta_{mg}(i,j)=q_{mi}q_{mj}\cdot\\
&\left(\kappa_g\left(d_i^+,d_j^+\right)-\kappa_g\left(d_i^+,d_j^-\right)-\kappa_g\left(d_j^+,d_i^-\right)+\kappa_g\left(d_i^-,d_j^-\right)\right)
\end{aligned}
\tag{6}
$$

To efficiently learn the ranking model, a large number of training data should be involved. The number of training triplet is approximately $O\left(|\mathcal{T}||\mathcal{D}|^2\right)$ where $|\mathcal{T}|$ denotes the number of textual queries for training and $|\mathcal{D}|$ denotes the number of images in the database. Consequently, to optimize the dual problem in Eq. 5, the prohibitive size of memory is required to load and maintain all the Θ_{mg}. To efficiently handle big data, we propose an online optimization procedure to optimize the multiple kernel ranking models, which will be introduced later.

4.2 Hard negative-based online learning

The value of hard negatives in learning machines was studied in depth in [52], where the samples in negative class that are the most similar to the single positive sample, given the classification hyperplane, are considered to be useful and informative. For many fundamental vision and multimedia tasks, for example, object detection [53] and image-text retrieval [11], mining the hard negatives during the training process will significantly boost the learning performance for both shallow model [54] and deep model [11, 53].

In this paper, we aim to improve the online learning by hard negative mining. Specifically, let us denote the hard negative as $\hat{d}^-$, and then the notion of training triplet becomes $\left(q,d^+,\hat{d}^-\right)$. One can easily replace the hard negative triplet into the original objective function from Eqs. 2 to 5. However, how to quickly select the hard negative samples for training remains a technical challenge. Specifically, the original hard negative samples are defined as those samples within a small circle area centered by the query points. However, in our model, the query sample and the database samples are in heterogeneous, which makes it hard to directly identify the hard negatives given a query input. Besides, according to study in [11], the hard negative samples need not be identified very accurately; otherwise, it will lead to prohibitive computational consumption.

To deal with this issue, considering the kernelized formulation in Eq. 5 and the triangle inequility theorem, we design a hard negative search method based on kernelized locality sensitivity hashing (KLSH) [55], which is an approximate nearest neighbor search

method in kernel space. We build one KLSH for each feature channel, where each image is encoded into an R-bit binary code. For M feature channels, each image is represented as an MR-bit code sequence with respect to M features/kernels. To guarantee the recall rate, we can also build more than one hash table for the training image data, e.g., 3 hash tables.

Based on the KLSH system, we perform hard negative mining for generating training triplets as follows. First, given a textual query q, we first identify the positive sample d^+ which is provided in the training dataset. Then, we treat d^+ as query to the KLSH system, and the samples with the same hash codes as the query are returned as the candidate hard negative samples. We remove those images from the candidate set with at least one class label that are identical to d^+, and the remaining samples that share no labels with d^+ but are highly similar to d^+ can be treated as the hard negatives. Based on this scheme, we can quickly identify a set of hard negative-based training triplets given a query and a positive sample.

4.3 Online model learning

The proposed Online Multiple Kernel Ranking (OMKR) algorithm is based on the fusion of two online learning methods: the Perceptron algorithm [56] and the Hedge algorithm [57]. Particularly, for each kernel and each textual query dimension, the Perceptron algorithm is employed to learn a kernel-based classifier with some selected kernel, and the Hedge algorithm is used to update their combination weights.

In this framework, we use θ_{mg}^t to denote the combination weight for the gth kernel classifier of mth query dimension at round t which is initially set to 1. For each learning round, we update the weight θ_{mg}^t by following the boosting style Hedge algorithm where each discriminative function can be treated as a weak learner. The weight update rule can be formulated as:

$$\theta_{mg}^{t+1} = \theta_{mg}^t \sigma^{z_{mg}^t} \tag{7}$$

where $\sigma \in (0, 1)$ is a discount weight parameter which is employed to penalize the kernel classifier that performs incorrect prediction at each learning step, and z_{mg}^t indicates that if the gth kernel classifier of the mth query dimension makes a mistake on the prediction of the training triplet $(q_j, d_j^+, \hat{d}_j^-)$, namely, $q_{mj}\left(f_{mg}\left(d^+\right) - f_{mg}\left(\hat{d}^-\right)\right) \leq 0$. When the tth training triplet is incorrectly predicted on the mth query dimension and the gth kernel, the corresponding discriminative sub-model is updated as:

$$f_{mg}^{t+1}(d) = f_{mg}^t(d) + q_{mj}\left(\kappa_g\left(d_j^+, d\right) - \kappa_g\left(\hat{d}_j^-, d\right)\right) \tag{8}$$

The main procedure of the optimization process is summarized in Algorithm 1. A support vector shrinking process is also performed in every T_b iteration to safely remove the training triplets with very high score using current model, i.e., the false positive support vectors, to enhance the efficiency of the learned model. Our model is similar to the boosting models where each of the gth kernel classifier on mth query dimension can be seen as the "weak learners." A weak learner selection procedure is performed which identifies a set of relevant discriminate weak learners with respect to the non-zero dimensions of the multi-dimensional query. The weight of each weak learner is updated according to their performance on the training triplet. The expected complexity of model update when receiving one training triplet is $O\left(2\overline{M}GC_\kappa\right)$ where $\overline{M}$ denotes the average number of non-zero dimensions in the query sets. For single query input, the complexity of per-step

model update is $O(2GC_\kappa)$. Under the non-hard negative setting, we provide Theorem 1 to estimate the error bound of our model.

Theorem 1 *After receiving a sequence of T training triplets, denoted by $\mathcal{D}_T = \left(q_t, d_t^+, d_t^-, t = 1, ..., T\right)$, the number of mistakes Ψ made by running Algorithm 1, denoted as:*

$$\begin{aligned}\Psi &= \sum_{j=1}^{T} I\left(\boldsymbol{w}_t\left(\gamma\left(q_t, d_t^+\right) - \gamma\left(q_t, d_t^-\right)\right) \leq 0\right) \\ &= \sum_{j=1}^{T} I\left(\sum_{m=1}^{M}\sum_{g=1}^{G} I\left(q_{mt} > 0\right)\beta_{mg}^t z_{mg}^t \geq 0.5\right)\end{aligned} \tag{9}$$

is bounded as follows:

$$\begin{aligned}\Psi &\leq \frac{2\ln(1/\sigma)}{1-\sigma}\min_{\substack{1\leq g\leq G\\1\leq m\leq M}}\sum_{t=1}^{T} z_{mg}^t + 2\frac{MG(\ln M + \ln G)}{1-\sigma} \\ &\leq \frac{2\ln(1/\sigma)}{1-\sigma}\min_{\substack{1\leq g\leq G\\1\leq m\leq M}} H_{mg} + 2\frac{MG(\ln M + \ln G)}{1-\sigma}\end{aligned} \tag{10}$$

By choosing $\sigma = \frac{\sqrt{T}}{\sqrt{T}+\sqrt{\ln M + \ln G}}$, we have:

$$\begin{aligned}\Psi \leq 2\left(\left(1 + \sqrt{\frac{\ln M + \ln G}{T}}\right)\right)\min_{\substack{1\leq m\leq M\\1\leq g\leq G}} H_{mg} + \\ \ln M + \ln G + \sqrt{T\left(\ln M + \ln G\right)}\end{aligned} \tag{11}$$

where H_{mg} denotes the structured loss on each individual classifier f_{mg} as:

$$H_{mg} = \min_{f_{mg}} ||f_{mg}||^2 + 2\sum_{t=1}^{T}\xi_{\left(q_j, d_j^+, d_j^-\right)} \tag{12}$$

As indicated by [8], the proof can be made by essentially combining the proof of the Perceptron [56] and the Hedge algorithm [57]. The details are omitted. Theorem 1 indicates that the error bound of the discriminative function is substantially determined by the error of the best weak learner. The error bound in the above theorem can be further improved from two aspects. First, it can be improved if we further tune the step-size or the margin. Second, it is further improved if we apply hard negative-based training scheme, since the error of the best weak learner can be further reduced by minimizing the hard negative learning objective function. For large-scale application, our proposed OMKR model needs to traverse the training data once or very limited times. Consequently, given a textual query q', we obtain a rank list τ_0 from the image database which reflects the semantic consistency between the query input and each image.

4.4 Discussion

Our model can be considered as a projection learning approach which learns an M-dimensional semantic consistent and query dependent representation for the image. The similarity between query and images can be directly compared by the simple inner product operation. We construct the projection function for each dimension by combining multiple visual features and exploring the correlation among different query dimensions. When the number of query dimensions is high, we can use latent topic models to learn

Algorithm 1 Online Multiple Kernel Ranking Algorithm

Input: training set $\left\{\left(q, d^+, \hat{d}^-\right)_j \in D_{tr}, j = 1, ..., |D_{tr}|\right\}$

$\kappa_g(\cdot,\cdot), \mathcal{X} \times \mathcal{X} \to \mathbb{R}, g = 1, ..., G$

Weights $\theta_{mg}, m = 1, ..., M, g = 1, ..., G$

Discount weight $\sigma \in (0, 1)$

Output: the kernel weight $\beta \in \mathbb{R}^{M \times G}$ and the discriminative functions $f = \sum_{m=1}^{M} \sum_{g=1}^{G} \beta_{mg} f_{mg}$

Initialization: $f_{m,g}^1 = 0, \theta_{mg}^1 = 1, m = 1, ..., M, g = 1, ..., G$

for $t = 1, 2, ..., t_{max}$ **do**

  Receive a training triplet $\left(q_t, d_t^+, \hat{d}_t^-\right)$

  for $\forall m$, where $q_{mt} > 0$ **do**

    for $g = 1, ..., G$ **do**

      Set $z_{mg}^t = I\left(q_{mt}\left(f_{mg}\left(d_t^+\right) - f_{mg}\left(\hat{d}_t^-\right)\right) \le 0\right)$

      Update $\theta_{mg}^{t+1} = \theta_{mg}^t \sigma^{z_{mg}^t}$

      Update $f_{mg}^{t+1}(d) = f_{mg}^t(d) + z_{mg}^t q_{mt}\left(\kappa_g\left(d_t^+, d\right) - \kappa_g\left(\hat{d}_t^-, d\right)\right)$

    end for

  end for

  IF $\mathrm{mod}(t, T_b) = 0$, **then** ShrinkSV(f_t); **end**

end for

$\beta_{mg}^{t+1} = \frac{\theta_{mg}^t}{\Theta^t}, g = 1, ..., G, m = 1, ..., M$, where $\Theta^t = \sum_{g=1}^{G} \sum_{m=1}^{M} \theta_{mg}^t$

M' ($M' \ll M$) dimensional compact representation on each query, and then, an M'-dimensional OMKR can be constructed on the latent topic representation of each query. Hence, the model complexity can be well controlled.

Our model can also be seen as a combination of Perceptron [56] and boosting-like learning methodologies [57]. From the perspective of Perceptron, the gth kernel of the mth query dimension from the jth training triplet (i.e., $q_{mj}\left(\kappa_g\left(d_j^+, d\right) - \kappa_g\left(\hat{d}_j^-, d\right)\right)$) can be seen as a "virtual" training sample that can be used to minimize the structure loss. Such "virtual" training sample is selected to update the model in boosting-based learning style, where the step-size is determined by the prediction of the current weak learner.

5 Preference modeling

When users are retrieving online images, their expectations with respect to the ranking results are diversified. Such potential interests can be exploited from the rich context information of social media Websites, e.g., the file upload time, the location of the image, the gallery information, the group of users that are interested in certain images, and the comments of each image. Based on the social attributes of the social media images, we construct correlation models to rank the social media images according to each social attribute.

For each image d_j, we have collected R types of social attribute features denoted by s_j. We define a set of social attribute relation matrix over each social attribute type as:

$$\begin{aligned} &\mathbf{M}_k \in \mathbb{R}^{N\times N}, k = 1, ..., R \\ &M_k(j,j) = 0, \forall j = 1, .., N \\ &M_k(j,j') = sim\left(s_j(k), s_{j'}(k)\right), \sum_j M_k(j,j') = 1 \end{aligned} \tag{13}$$

We construct R random walk models on each of the social attribute relation matrix, where the stable distribution can be calculated by iterating the following operation until r_k is converged:

$$r_k = \mu \mathbf{M}_k r_k + (1-\mu)\,\rho_k^0, r_k = r_k / \sum_{n=1}^{N} r_k(n) \tag{14}$$

where the stimulation vector ρ_k^0 can be identified by setting the dimensions of the top results of the semantic-based ranking results by OMKR or other possible user interests with larger weights, e.g., the historical retrieving records of the users or the popularity score of the images. We can also set the images with the most view counts as the stimulation of the random walk models. In fact, the non-zero weights represent the prior knowledge on the probability of those images that are likely to be both semantically relevant and also popular among the user community. Note that the social attributes of certain images may be missing. Each image may only have a fraction of social attributes, while others are vacant or unavailable during the collection stage. Therefore, $\mathbf{M}_k, k = 1, ...R$ are sparse and some rows and columns of $\mathbf{M}_k$ are zero which makes the corresponding probabilities become 0. To avoid this, we assign non-zero weights to all the images in the stimulation vector ρ_k^0. When $0 \le \mu \le 1$, each r_k is converged as:

$$r_k \to (1-\mu)\,(\mathbf{I} - \mu \mathbf{M}_k)^{-1} \rho_k^o \tag{15}$$

The rationality of constructing random walk model on social attributes can be explained by two folds. First, many studies indicate that information propagation in social media can be modeled by random walks [58]. Second, online users act similarly with other users with the same social behavior. Therefore, their inclinations can also be propagated along the social attribute correlations of the online documents. In this paper, we are primarily interested in the following social attributes:

Surrounding text. The surrounding text includes photo's title and short description carrying important indication of the semantic information. We measure the similarity of image i and j by calculating their cosine distance on the TF-IDF representation of the surrounding text. Besides, we can also use the word embedding to derive more effective surrounding text description. Specifically, we use GloVe [59] to represent each word into vector, and use a simple average pooling over the whole text to derive the final surrounding text features. Then, the similarity between image i and j with respect to this feature is calculated by Gaussian kernel.

Location. The location information indicates where an image is taken. Intuitively, if the locations of two images are close enough, their contents may deliver consistent semantics, e.g., the same objects, the same buildings, or the same scenery. The location attribute may also reflect the geo-trend that can be used to detect the local interest and location-aware topics [60]. We use RBF kernel to measure the location relevance where the similarity of geographically adjacent images is higher.

Time. The upload time of images may indicate the temporal relation of images describing hot social event. For example, the images describing the American President Election will be posted by online users frequently within a certain period. Moreover,

users would like to retrieve images in certain temporal range. We use RBF kernel to measure the temporal relevance on several temporal resolutions, e.g., year, month, and day, respectively.

Group. Similar as the category, images are associated with groups where each group is associated with uploaders' description of the semantics [9]. We collect the group information of each image and denote image i and j as relevant ($M_k(i,j) = 1$) when their associated group information is identical.

Category. On many social media Websites, semantically related images are grouped into categories by online users. Each image may be categorized into multiple categories where each category has a unique category ID. We collect the category information of each image. We denote image i and j as relevant ($M_k(i,j) = 1$) when at least one of their associated categories is identical.

User ID. The images uploaded by the same user ID may convey certain preference styles. For example, some users may be interested in the photo capturing style of specific online users. For image i and j uploaded by the same user ID, we denote them as relevant ($M_k(i,j) = 1$) when constructing the social attribute relation matrix.

Based on preference score $[r_1, ..., r_R]$, we obtain a set of ranking lists $[\tau_1, ..., \tau_R]$ by injecting semantic ranking results into correlation modeling of social attributes. We will introduce how to effectively aggregate τ_0 and $[\tau_1, ..., \tau_R]$ in a unified rank aggregation model in the subsequent section.

6 Rank aggregation

As each ranking metric captures only some aspect of the consistency with respect to certain social attribute, it is beneficial to combine them in order to more accurately identify what a user really needs. Our proposed rank aggregation model is an order-based technique with the weighted position-sensitive measurement.

For P images from the database, we have $R+1$ ranking lists $\tau_r = [\tau_{r1}, ..., \tau_{rP}], \forall r = 0, ..., R$. We define a pair-wise preference matrix $Q_{i,j}, \forall i,j = 1, ..., P$ which encodes if the ith image is preferred over the jth image by considering their ranking in all the ranking lists and the weights of individual rankers $\omega = [\omega_0, ..., \omega_R]$. When performing the retrieval given a user query, the documents are ranked according to their relevance. However, when the number P of documents is large, the user experience will be determined by the top P' ranked results where $P' \ll P$. In traditional Kemeny ranking aggregation procedure (or its weighted extensions) [46], we only need to calculate the preference score of the top P' documents as:

$$Q_{\tau_{ri},\tau_{rj}} \Leftarrow Q_{\tau_{ri},\tau_{rj}} + \omega_r, i = 1, ...P' - 1, j = i+1, ..., P' \tag{16}$$

where $\tau_{ri}, r = 1, ..., P$ indicates the ith document index in the rth ranking list. However, some documents that are ranked in the lower position in many ranking lists can be inappropriately ranked at a higher position, because the relative importance of the top ranked documents in the ranking lists is not adequately emphasized, and their relative importance over the lower ranked documents has not been sufficiently observed in the top P' results. To alleviate the disadvantages, we

revise the rank aggregation model from two aspects. First, the preference score is calculated as:

$$\begin{aligned} &Q_{\tau_{ri},\tau_{rj}} \Leftarrow Q_{\tau_{ri},\tau_{rj}} + \omega_r \cdot \big(\log(j+\epsilon) - \log(i+\epsilon)\big) \\ &i = 1, ...P' - 1, j = i+1, ..., P' \end{aligned} \tag{17}$$

where $\epsilon \in \mathbb{R}^+$ is a sensitive parameter controlling the range position importance and where a small ϵ indicates the heavier relative importance on the top ranked documents. The relative measurement ensures the top ranked documents are more carefully pondered when their positions are considered to be changed in the aggregation procedure. Second, we collect more relative importance evidence by extending the observation range P' to $\psi P'$, where $\psi > 1$ and $\psi P' \leq P$. We further consider the relative importance between the top P' documents and the $(P'+1)$-th to the $\psi P'$-th documents, and encode their relative preference into Q as:

$$\begin{aligned} &Q_{\tau_{ri},\tau_{rj}} \Leftarrow Q_{\tau_{ri},\tau_{rj}} + \tfrac{1}{2}\omega_r \cdot \big(\log(j+\epsilon) - \log(i+\epsilon)\big) \\ &i = 1, ...P', j = P'+1, ..., \psi P' \end{aligned} \tag{18}$$

Our rank aggregation model can be seen as building up a "barrier" between the top P' documents and the bottom ranked images in order to refine the top P by collecting more observations and prevent the bottom ranked documents to be ranked at a high position. A simple toy example is demonstrated in Fig. 2. We have the following theoretical analysis on our method.

Definition 1 *The Extended Condorcet Criterion* [46] *requires that if there is any partition* $\{L, R\}$ *of a ranking list* τ *for any* d_i *and* d_j*, such that a majority of rankers prefer* $d_i \in L$ *to* $d_j \in R$ *, then the aggregate ranking should prefer* d_i *to* d_j*.*

Theorem 2 *Let* τ *be the final aggregation of the positive sensitive rank aggregation procedure. Then,* τ *satisfies the Extended Condorcet Criterion with respect to the input rankings* $[\tau_0, \tau_1, ..., \tau_R]$*.*

The proof of Theorem 2 follows directly from the Theorem 4.1 in [46], and the details are omitted in this paper. Our proposed rank aggregation approach satisfies neutrality, consistency, and the Extended Condorcet Criterion. The procedure of position-sensitive

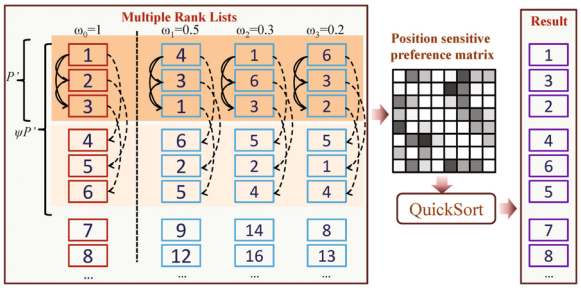

Fig. 2 Illustration of position-sensitive rank aggregation. Given a set of ranking lists ordered in columns, the list in red color denotes the semantic ranking result, and the list in blue color denotes the preference ranking lists. A preference matrix is generated according to the weights of each ranking list and the position of each document. Then, a Quicksort algorithm is employed to calculate the final ranking list, which is denoted by the list in purple color

rank aggregation is described in Algorithm 2. Similar as the Kemeny optimal aggregation, our proposed rank aggregation model also has a good maximum likelihood interpretation or even better, because we collect more observations of pair-wise preference in our framework. The rank aggregation result possesses the following properties: first, a more semantic consistent rank list compared with only using the semantic ranking results; second, it well satisfies the user's preference. The complexity of the ranking procedure is $O\left(\frac{1}{2}P(P-1)+\psi P^2\right)+O\left((R+1)P\log P\right)$ where the former is the complexity of calculating the preference matrix Q and the latter is the complexity of Quicksort on the P images. In this paper, we empirically set $\psi = 2$ for all the experiments.

Algorithm 2 Position Sensitive Rank Aggregation

Input:, $\tau_r = [\tau_{r1}, ..., \tau_{rP}], \forall r = 0, ..., R$, the rank lists of P images for $R+1$ rankers
$\omega = [\omega_0, ..., \omega_R]$, where ω_r is the weight of ranker r
$\mu = [\mu_1, ..., \mu_P]$, initial ordered arrangement of P images
P' - the number of images to consider in each τ_i, where $P' \ll P$
ψ - the range to search beyond P', where $\psi \geq 1$ and $\psi P' \leq P$
Output: τ - aggregated rank in decreasing order of importance
Initialize majority table $Q_{i,j} \Leftarrow 0, \forall i, j = 1, ..., P$
for each ranker $r = 0$ to R **do**
  for each document $i = 0$ to $P' - 1$ **do**
    for each document $j = i + 1$ to P' **do**
      $Q_{\tau_{ri},\tau_{rj}} \Leftarrow Q_{\tau_{ri},\tau_{rj}} + \omega_r \cdot (\log(j+\epsilon) - \log(i+\epsilon))$
    end for
  end for
  for each document $i = 1$ to P' **do**
    for each document $j = P' + 1$ to $\psi P'$ **do**
      $Q_{\tau_{ri},\tau_{rj}} \Leftarrow Q_{\tau_{ri},\tau_{rj}} + \frac{\omega_r}{2} \cdot (\log(j+\epsilon) - \log(i+\epsilon))$
    end for
  end for
end for
Quick Sort μ using Q_{μ_i,μ_j}:
(1) If $Q_{\mu_i,\mu_j} - Q_{\mu_j,\mu_i} > 0$, then $u_i \succ u_j$
(2) If $Q_{\mu_i,\mu_j} - Q_{\mu_j,\mu_i} = 0$, then $u_i = u_j$
(3) If $Q_{\mu_i,\mu_j} - Q_{\mu_j,\mu_i} < 0$, then $u_i \prec u_j$

7 Experimental results

In this section, we perform systematical evaluation on two real-world social media datasets on social media image retrieval task.

Datasets. The datasets we used in this paper include the following: (1) The NUS-WIDE dataset [61] consists of 269,648 images collected from Flickr. We collect their social attributes by using their URLs linked to their original pages. Six types of low-level visual features are provided by the data provider. The 81-dim tag vectors of images are treated as the ground-truth queries. (2) The Flickr dataset consists of 3.5 million images collected from Flickr covering wider visual topics than NUS-WIDE. We extract 5 types of visual

features for each image, i.e., Gist, LBP, Bag-of-Visual-Word, Color Moment, and PHOG. Similarly, we collect the same social attributes by using their URLs linked to their original pages. We select 150 common queries from the query vocabulary as the associated ground-truth queries. For both datasets, we reweigh each query dimension by TF-IDF weight to enhance the descriptive power of query inputs, and use the weighted value for each query dimension when it occurred in a query input. Besides, due to the strong representation ability of the deep convolutional neural network, we also extract deep visual features using the standard VGG-19 network pretrained using ImageNet dataset. Inspired by the feature extraction strategy in [24], we use conv2, conv4, conv5, fc6, and fc7 as the visual features that complementarily describe the visual content from low-level to high-level semantic abstraction. To deal with high dimensionality, we perform PCA on these deep features.

Data partition. For NUS-WIDE data, we randomly select 15 thousand images as the training database and another 2 thousand queries as the training queries. We select 5 thousand queries from the remaining dataset as the testing queries, and the other images excluding the training database and testing queries are used as the testing database. Note that we do not use the training/testing partition provided by the NUS-WIDE data provider. Our scheme is more suitable for evaluating the model generality of ranking model learning using small number of training data and testing database with larger size. For Flickr data, we randomly select 50 thousand images as the training database and another 6 thousand queries as the training queries. We select 10 thousand queries from the remaining dataset as the testing queries, and the other images excluding the training database and testing queries are used as the testing database. Note that despite the unified query vocabulary used on both training and testing datasets, the textual queries of the training dataset and testing dataset are still diversified and not enforced to be arranged to be the same. This setting tends to be more practical and is able to verify the generalization of the compared approaches.

Compared approaches. We compare the following approaches for the task of query-to-image retrieval and personalized social media retrieval. Note that for shallow model, we use both the shallow features and the deep features of the pretrained VGG-19 network for comparison.

(1) PAMIR-PR: PAMIR is a kernel-based discriminative text-to-image retrieval approach proposed by Grangier and Bengio [7]. We use the average kernel on the hand-crafted shallow features for training PAMIR, and re-rank the text-to-image retrieval results by PAMIR using the preference ranking proposed in this paper.

(2) MMNN-PR: MMNN is a state-of-the-art cross-modal hashing approach proposed by Masci et al. [13] using multi-layer neuro-network. We conduct dimension reduction to the concatenated visual features to reduce the feature dimension number to 300. We set the code length as 64 for the hash code learning, and re-rank the text-to-image retrieval results by MMNN using the preference ranking.

(3) SCM-PR: SCM is a semantic correlation model proposed by Rasiwasia et al. [2], which projects the text documents and image documents into a unified semantic space. In this paper, we only project the images into the semantic categories where the number of category is equal with the number of query dimension. We do not project the query text into the semantic space, since the query text is extremely sparse. We re-rank the text-to-image retrieval results by SCM using the preference ranking.

(4) CMOS$_{lg}$-PR:CMOS is an online cross-modal retrieval method [24] which learns the asymmetric bilinear similarity by aggregating deep visual features from multiple layers. Among the three layer aggregation mechanisms, we report the results derived by layer gating, and use the retrieval results to re-rank.

(5) Deep-PR: It is expected that deep models trained in an end-to-end fashion can be a strong competitor for various tasks including the text-to-image retrieval. In our study, considering that the textual queries are combination of words for the social media datasets, traditional deep learning models for text cannot be directly applied in our situation. Instead, given that the size of query vocabulary is fixed, we use an FC layer to transform the query vector into a K-dimensional representation, and use the conv1→fc6 of VGG-19 for visual feature extraction, where the parameters of VGG-19 are pretrained with ImageNet dataset. Then, the fc6 layer is connected to the K-dimensional representation as well to ensure that the similarity of image and textual queries can be measured. We train the model with the hard negative triplet loss as VSE++ and our model to guarantee good accuracy.

We implement our model using both hand-crafted visual features and the extracted deep CNN features as has been described. We test different versions of our model.

(1) OSIR$_s$: A simplified version of our proposed approach where the kernel weight for all the query dimensions is identical in the OMKR learning (OMKR-sim), i.e., we only need to learn β_g and f_g for all the query dimensions. The ranking function is:

$$\begin{aligned} F(q', d) &= \sum_{m=1}^{M} \sum_{g=1}^{G} q_m' \beta_g f_g(d) \\ &= \sum_{m=1}^{M} \sum_{g=1}^{G} \sum_{j=1}^{|D_{tr}|} q_m' q_{mj} \alpha_j \beta_g \left(\kappa_g \left(d_j^+, d \right) - \kappa_g \left(\hat{d}_j^-, d \right) \right) \end{aligned} \tag{19}$$

(2) OSIR: Our proposed approach which learns f_{mg} and β_{mg}, where $m = 1, .., M$ and $g = 1, ..., G$.

Evaluation criteria. To measure the performance of query-to-image retrieval, we adopt the mean average precision (MAP). For subjective study, we conduct evaluation on a three-level score human evaluation of the preference aggregation results by normalized discount cumulative gain (NDCG): (1) 2—*preferred and semantically relevant*; (2) 1—*semantically relevant*; and (3) 0—*irrelevant*.

Platform. Our experiments and observations are conducted on a standard server, with Intel (R) Xeon (R) Processor E7-4870 (30M Cache, 2.40 GHz, 6.40 GT/s Intel (R) QPI, 10 cores), 128 GB main memory, and 10,000 RPM server-level hard disks.

7.1 Online learning of ranking models

We conduct experiment to study the ranking model training with respect to the following aspects. We evaluate three methods, i.e., PAMIR, OMKR-sim in OSIR$_s$, and OMKR in OSIR, since they are all online ranking models. All the experiment results in this section are reported on NUS-WIDE data.

Training error curves. We record the training error curves for each method in Fig. 3. The training error of tth iteration is calculated by dividing the number of disordered training triplets (i.e., $q_t \left(F\left(d_t^+\right) - F\left(\hat{d}_t^-\right) \right) \leq 0$) with the number of total training triplets at tth iteration. From Fig. 3a, we observe that our approach achieves much lower training error after receiving the first 5 thousand training triplets, and the training error continues to decrease more quickly than the other two approaches when receiving more triplets.

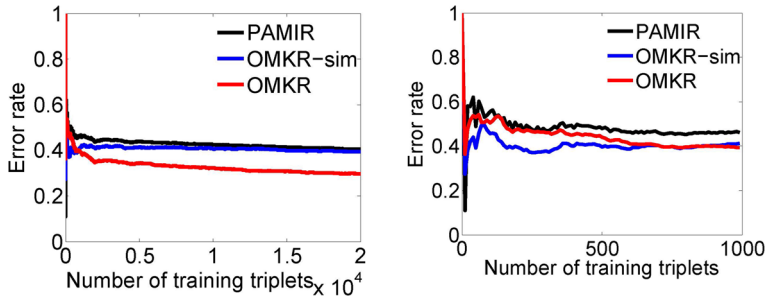

Fig. 3 The training error rate curves

The passive-aggressive learning procedure used by PAMIR possesses similar convergence property as our online optimization procedure. At the first 1 thousand training iterations, the training error of PAMIR tends to be more unstable than the other two, as shown in Fig. 3b. The lower training errors can be explained by the fact that our OMKR has more weak learners with respect to different query dimensions and different kernels. Therefore, the results indicate that OMKR possesses lower model bias and is more likely to converge to the "ideal model."

Number of support vectors. After receiving T training triplets, the number of support vectors (i.e., triplets with non-zero weights) of the online ranking methods determines both the model complexity and generality. We record the ratios of support vectors after receiving T triplets in Table 1. OMKR has the most compact support vector sets among all the approaches, which means OMKR ranks the training triplets more correctly. Therefore, less training triplets are incorporated as support vectors. Another reason is the shrinking operation in Algorithm 1 which can generally filter about $0.15T_b$ support vectors each time.

Number of training data. We evaluate the potentials of the three models when increasing the number of the training queries and the number training triplets. The results are shown in Fig. 4. In Fig. 4a, we randomly select 512 training queries to generate different numbers of training triplets, and randomly select 3 thousand test queries to measure the MAP with respect to different numbers of triplets. In Fig. 4b, we fix the number of training triplets generated on each training query as 300, and select different numbers of training queries to train the three online learning models. We evaluate MAP on the same test queries as in Fig. 4a. From the result curves, we observe that the ranking performance can be enhanced by increasing both the number of triplets and the number of different queries. Comparatively, increasing the number of queries tends to produce higher performance gain, since the ranking models benefit from capturing more patterns in the query

Table 1 Number of support vectors when receiving T triplets

SV(%)@T	10K	20K	40K	80K	160K
PAMIR [7]	32.3	30.4	29.1	28.5	26.9
OMKR-sim (shallow)	27.6	25.8	23.2	21.7	20.4
OMKR-sim (deep)	24.3	23.4	21.1	19.5	18.9
OMKR (shallow)	21	19.2	16.1	14.3	10.2
OMKR (deep)	19.8	18.6	16.0	13.9	9.7

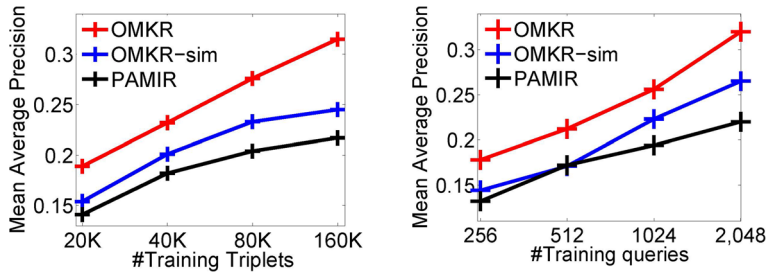

Fig. 4 The performance with different number of training triplets and training queries

inputs. The performance gain of OMKR by increasing the training data is higher than the other two models.

Training time. We record the training time of the three methods by using 50 thousand training triplets in Table 2. The training time consumptions of the three methods are mainly determined by the numbers of support vectors and complexity of kernel calculation. Although OMKR has more weak learners with respect to each kernel and each query dimension, its time consumption does not grow very significantly since the model has lower ratio of support vectors. The time efficiency of OMKR can be attributed to its model generality and the support vector shrinking in Algorithm 1. Moreover, it can be observed that the model using deep features consumes more time than using shallow features, because of the fact that it takes more time to calculate kernels using deep features with higher dimension.

The kernel weight learning mechanism. One of the most enjoyable properties of OMKR can be attributed to its query word-specific weighted kernel combination. To evaluate how the kernel weight learning scheme works, we calculate the accumulated kernel weights with respect to each feature channel and each query dimension, respectively, in Fig. 5a, b. The results are reported on the NUS-WIDE data, while similar observations can be found on the Flickr data. In Fig. 5a, the accumulated weight of BOW feature is larger than other global features. Color histogram (CH) performs the worst. Its accumulated weight is smaller than others. The accumulated weight of edge histogram (EDH) is the second larger because the texture statistics delivered in each image is informative in identifying visual objects. The result is consistent with the empirical judgment on the feature effectiveness.

In contrast, the accumulated weight distribution with respect to different query dimensions is much more imbalanced, as shown in Fig. 5b. Some query dimensions possess much larger accumulated weights, e.g., *swimmers*, *computer*, *whales*, *elk*, and *earthquake*. The reason may be three folds: (1) the query dimensions with higher weights are easy to be distinguished, (2) there are many images having the query words which produce a larger number of training triplets, and (3) these queries usually co-occurred with other queries which borrow more discriminative information from other rankers. Results in Fig. 5b can

Table 2 The training time statistics on NUS-WIDE using 20K training triplets

Method	PAMIR [7]	OMKR-sim (shallow)	OMKR-sim (deep)	OMKR	OMKL (deep)
Time (s)	2594	5638	7437	6209	8563

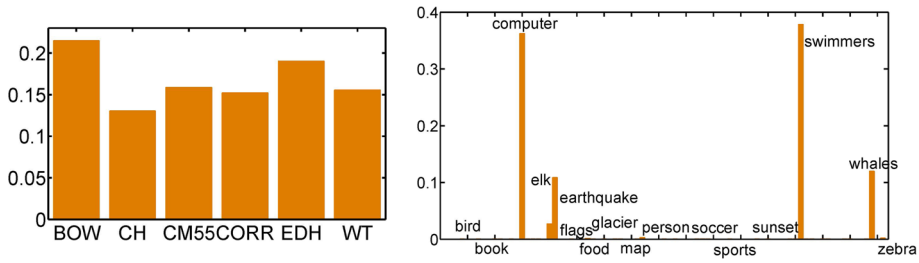

Fig. 5 The accumulated weights with respect to different features and query dimensions

also be considered as an informative query word selection procedure which identifies the most important query words on a large-scale social media dataset.

7.2 Retrieval performance

We perform extensive experiment to evaluate the retrieval performance of all the compared methods. The MAP measurements on top 500 retrieved results of different methods are shown in Table 3. We denote the retrieval results by semantic ranking of the original retrieval method (PAMIR, MMNN, CMOS_{lg}, Deep, OSIR, etc.) as SR, and rank aggregation with both SR and the re-ranking results of the surrounding text as SR+BOW and SR+GlV, which represents using Bag-of-Word or GloVe embedding for extracting the feature of surrounding text, respectively. Similarly, SR+lc is used for location, SR+tm for time, SR+gp for group, SR+ctg for category, SR+id for user ID, and SR+all for the weighted aggregation using SR and all the preference ranking lists. By appropriately aggregating the semantic ranking and preference ranking results, our approach achieves much better retrieving performance than other approaches on both datasets.

First, the results indicate that different social attributes carry different implications on the true semantics of the social media images. For example, by aggregating SR and location (SR+lc), the retrieval performance of all the compared approaches is improved over SR. By aggregating SR and user ID (SR+id), our approaches consistently obtain improved results, while other approaches may perform worse on either NUS-WIDE or Flickr. The upload time information is less relevant to the semantics, since the results by aggregating SR and user ID (SR+id) usually underperform results on SR.

In general, from the results, we observe that, among all the social attributes considered in this paper, the attributes with higher semantics, despite being noisy in some case, tend to produce better results in refining the results. For example, among all the social attributes, the social tag (ctg) and surrounding texts (GlV with GloVe feature) tend to be the most effective attributes for re-ranking performance enhancement. The reason is straightforward, i.e., incorporating the affinity structure into the re-ranking model can be seen as introducing more semantic information expressed by different users, so that the re-ranked results can be significanlty improved and better reflect the user preference.

On the other hand, introducing social atrributes that are less semantically relevant would introduce inappropriate relation information among images. For example, the uploading time of different images is not a semantic-related feature, which only encodes the temporal co-occurrence pattern of different images. For some case, these adjacencies do reflect certain cues on popularity. For instance, an image tends to be more popular if it is uploaded next to a very popular image, and a group of images uploaded

Table 3 The retrieval performance of the top 500 ranked images in MAP (%) for difference approaches on NUS-WIDE and Flickr

Datasets	NUS-WIDE								
Methods	SR	SR+BOW	SR+GIV	SR+lc	SR+tm	SR+gp	SR+ctg	SR+id	SR+all
PAMIR-PR (shallow)	27.6	28.3	30.6	27.9	26.9	28.0	28.2	27.9	31.4
PAMIR-PR (deep)	29.1	30.2	31.3	28.9	28.2	29.3	29.4	29.4	32.6
MMNN-PR (shallow)	26.3	26.9	27.8	26.3	25.5	26.9	27.3	25.9	29.5
MMNN-PR (deep)	28.2	29.0	30.4	29.4	28.3	29.1	30.1	28.1	31.3
SCM-PR (shallow)	21.5	21.9	23.7	21.5	20.1	21.9	22.1	21.2	24.4
SCM-PR (deep)	23.4	23.8	25.6	23.9	23.2	23.7	24.2	23.0	25.5
$CMOS_{lg}$-PR	36.4	36.8	38.4	37.0	36.7	37.1	37.6	36.0	38.8
Deep-PR	38.7	38.9	40.9	39.8	39.3	39.8	39.9	38.5	41.6
$OSIR_s$ (shallow)	31.7	32.2	32.9	32.0	31.8	32.1	32.5	31.9	32.6
$OSIR_s$ (deep)	35.9	36.7	38.3	36.4	36.1	36.4	36.9	36.3	38.9
OSIR (shallow)	38.5	39.4	40.3	39.1	38.4	38.7	40.4	39.1	43.8
OSIR (deep)	40.3	41.2	41.8	41.1	40.1	40.6	41.3	40.7	**44.4**
Datasets	Flickr								
Methods	SR	SR+BOW	SR+GIV	SR+lc	SR+tm	SR+gp	SR+ctg	SR+id	SR+all
PAMIR-PR (shallow)	12.4	13.2	13.9	12.8	12.0	13.1	14.0	12.3	14.5
PAMIR-PR (deep)	15.1	15.9	16.7	15.5	15.0	16.1	17.2	15.2	17.9
MMNN-PR (shallow)	12.1	12.9	13.4	12.6	11.8	13.2	13.6	12.0	14.0
MMNN-PR (deep)	14.8	15.3	15.9	15.5	14.2	16.1	16.4	14.9	16.9
SCM-PR (shallow)	5.8	6.2	6.7	5.9	5.5	6.5	6.9	6.3	7.6
SCM-PR (deep)	7.2	7.8	8.6	7.4	7.8	8.1	8.9	7.1	9.4
$CMOS_{lg}$-PR	19.4	20.1	21.6	19.9	19.7	20.3	21.4	19.2	23.7
Deep-PR	19.9	20.3	21.8	20.2	20.0	20.9	22.0	19.5	24.1
$OSIR_s$ (shallow)	14.3	15.1	15.8	14.9	14.4	15.8	16.2	15.1	18.1
$OSIR_s$ (deep)	16.5	17.1	18.3	16.9	16.7	17.2	17.8	17.1	20.3
OSIR (shallow)	18.1	19.3	19.8	19.2	18.3	20.2	20.3	18.9	23.4
OSIR (deep)	20.6	21.2	21.9	21.4	20.7	21.7	22.7	21.4	**25.3**

by the same popular user on the social image Website tends to be more popular than images from other users. However, in other case, these adjacencies reflect nothing, mainly due to the low correlation with the true semantic meaning. Therefore, the performance enhancement brought by using these social attributes tends to be less statistically significant.

Second, the results imply that by preference ranking and aggregating, the performance of all the semantic-based models can be enhanced by incorporating multiple heterogeneous social attributes. For example, the performance of all the approaches on SR+all outperforms SR on both datasets. Generally, by aggregating more social attributes, the retrieval performance of all the methods on SR+all outperforms rank aggregation with single social attribute, e.g., SR+ctr.

Third, we observe that different social attributes contribute differently to different types of queries. Specifically, we observe that if the query contain words indicating location information, the re-ranked results may be better refined. In contrast, the frequently occurring query words generally do not contain words from time, user, and group attributes, so that these attributes tend to perform equally for most queries.

Last, the rank aggregation results on Flickr dataset shows that, when processing large-scale social media with weak semantic information such as the noisy tags, fusing the

semantic relevance delivered in different social attributes will boost the retrieval performance in a more promising manner. Such a claim is made by observing that all of the compared approaches perform at least 15% better on SR+all vs. others on Flickr dataset. Despite that some attributes may even lead to a performance degradation compared to the original semantic-relevance ranking results, but the average aggregated results still tend to be better, due to the robustness of our rank aggregation technique.

Failure case. We provide some discussion from the failure case. We observe that the semantic ranking accuracy imposes direct influence on the final re-ranked results. If the truly relevant images are not ranked at top 10 positions, then the re-ranking would also fail or even push the truly relevant images backward for a small number of queries. Further study is required to address this issue to ensure better robustness of the rank aggregation.

7.3 Subjective study of preference fitting

For subjective study, given a specific query, the users are served with the same set of results without any post-processing. In offline evaluation situation, it is unable for us to provide any tailored results for subjective study because we do not have any user preference information for a specific subject. In fact, the key idea that we conduct the subjective study is to provide as many ranking choices as possible to users, and see which ranking result they would prefer. This may be a little different from the traditional view of recommendation, where the user preference has to be obtained for measuring the user-item similarity for recommendation. In our study for retrieval, if the provided top ranked results are more diversified but staying as semantically relevant, the results may be more appreciated by as many users as possible.

To this end, we randomly select 100 queries from both datasets, and ask ten normal users to provide weight specification on social attributes, and judge whether the returned aggregate ranking results can better reflect what they really like. The evaluation results are recorded in terms of NDCG@50 in Table 4, where SG denotes rank aggregation with single social attribute, and MP denotes rank aggregation with multiple social attributes. Experiments show that our approach better facilitates the diversified preference styles of online users, as it outperforms all the other approaches under different settings. The promising performance can be attributed to the good semantic retrieval performance and the position-sensitive rank aggregation that protects the top ranked results to be appropriately located in the final ranking.

7.4 Parameter sensitivity

The weight ω in rank aggregation. The setting of weight ω determines the rank aggregation performance. Existing approaches estimate the weights of multiple ranking

Table 4 The NDCG@50 (%) of the subjective study

Datasets	**NUS-WIDE**			**Flickr**		
Methods	SR	Sg	Mp	SR	Sg	Mp
PAMIR-PR	24.2	24.3	26.2	12.9	16.0	18.2
MMNN-PR	13.3	12.9	15.6	13.2	15.8	17.4
SCM-PR	12.1	11.7	14.5	4.1	5.7	8.5
$OSIR_s$	32.2	34.5	38.0	17.8	22.5	25.3
OSIR	39.6	40.6	46.1	22.6	26.7	33.4

results according to their retrieval performance with respect to certain criterion such as MAP [46]. We adopt similar tuning procedure by a validation process. Consequently, we set $\omega_0 = 1$ in any type of rank aggregation. The results in Table 3 are based on the following setting: On NUS-WIDE data, $\omega_1 = 0.8$ for SR+txt, $\omega_2 = 0.5$ for SR+lc, $\omega_3 = 0.2$ for SR+tm, $\omega_4 = 0.3$ for SR+gp, $\omega_5 = 0.9$ for SR+ctg, $\omega_6 = 0.3$ for SR+id, and $[\omega_1, ..., \omega_6] = [0.45, 0.3, 0.1, 0.2, 0.5, 0.15]$ for SR+all. On Flickr, $\omega_1 = 0.7$ for SR+txt, $\omega_2 = 0.6$ for SR+lc, $\omega_3 = 0.1$ for SR+tm, $\omega_4 = 0.25$ for SR+gp, $\omega_5 = 0.85$ for SR+ctg, $\omega_6 = 0.4$ for SR+id, and $[\omega_1, ..., \omega_6] = [0.4, 0.3, 0.1, 0.25, 0.45, 0.2]$ for SR+all.

The penalty C of online learning. We empirically set $\frac{C}{|D_{tr}|} = 1$ for PAMIR, OMKR-sim, and OMKR, since the setting guarantees good model generality.

The kernel coefficients of OMKR. According to Theorem 1, the performance of OMKR mainly depends on the performance of the best learner. We conduct a cross-validation process to tune the kernel coefficients. Details are omitted due to space limit.

The weight μ of preference modeling. This parameter determines how well the semantic ranking results is re-ranked towards the preference consistency. When μ is small, the preference re-ranking is similar to the semantic ranking which means that rank aggregation is unnecessary. When μ is large, the preference re-ranking tends to be cluttered. A reasonable setting of μ is [0.4,0.6]. In all the experiments, we set $\mu = 0.5$ for better trade-off between semantic divergence and consistency.

7.5 Findings and discussions

Retrieval examples. We provide some examples on NUS-WIDE data in Fig. 6. Given each textual query, we show the top 10 retrieved images with respect to semantic ranking and different preference styles, where each row indicates the corresponding ranking results. The semantically relevant images are marked with red dots. Although all the top ranked results are semantically relevant, their preference ranking tends to be diversified. For example, in the first example with query "animal, flowers," all the ranking lists appreciate the panda image as the top ranked image. But the results from the 4th to the 10th tend to be diversified. The aggregated ranking results are most appreciated since the top ranked images are more semantically consistent than other ranking strategies.

More efficiency on retrieving large-scale data. When retrieving large-scale social media data, it is time-consuming to conduct preference re-ranking and rank aggregation. To address this concern, a simple scheme can be used to reduce the retrieving complexity. Specifically, when processing single word queries, we quickly select P images whose semantic projection values are larger than a predefined threshold of the non-zero query dimension where P is much smaller than the database size N. When processing multi-word queries, we quickly select P_1 images based on the scores of the first non-zero query dimension, and select P_2 similarly from the P_1 selected images where $P_2 \ll P_1 \ll N$. The top ranked images with high semantic relevance can be quickly identified by the much more efficient "find" operations instead of the inner product and sorting operations.

The query word patterns. We observe from the retrieval results that the retrieval performance of multi-word queries is generally higher than single-word queries. When the user queries are multi-word, the retrieval results tend to be boosted by involving more weak learners from different query dimensions. This phenomenon can be attributed to the tag co-occurrence existing on real social media image data. Our approach is capable of capturing such correlation in the complicated patterns in user queries.

Fig. 6 Retrieved examples on NUS-WIDE

8 Conclusion

In this paper, we proposed OSIR as a solution framework to facilitate the diversified preference styles in social media image searching by combining heterogeneous information sources. First, we proposed an efficient Online Multiple Kernel Ranking model constructed on multiple query dimensions and complimentary feature channels. By optimizing the ranking performance, the semantic consistency between the image ranking and textual query input is directly maximized without relying on intermediate semantic annotation procedure. Second, we constructed random walk-based preference modeling by domain-specific similarity calculation on heterogeneous social attributes. By re-ranking the rank output of OMKR based on each of the preference models, we obtained a set of ranking lists encoding different potential aspects of user preference. Last, we proposed an effective and efficient position-sensitive rank aggregation approach to aggregate multiple ranking results based on the user's preference specification. Extensive experiments on two social media datasets have demonstrated the advantages of our approach in both retrieval performance and user experiences. In future work, we will investigate how to model the online user behaviors in a more comprehensive way to better facilitate the user preference.

Abbreviations

OSIR: Online social image ranking; OMKR: Online Multiple Kernel Ranking; CNN: Convolutional neural network; SVM: Support vector machine; POI: Point of interest; KLSH: Kernelized locality sensitivity hashing; TF-IDF: Term frequency-inverse document frequency; RBF: Radial basis function; LBP: Local binary pattern; PHOG: Pyramid histogram of oriented gradients; MAP: Mean average precision; BOW: Bag-of-features; CH: Color histogram; EDH: Edge histogram

Acknowledgements

Thanks to all those who have suggested and given guidance for this article.

Authors' contributions

All authors participated in the preliminary research and discussion of this study. XZZ designed the framework and details of the whole algorithm. LSP realized the experimental effect of the algorithm and the writing of the manuscript. LT put forward suggestions for revision of manuscript and experiments. ZZ participated in the survey and gave suggestions on the experimental analysis. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (61806073, U1904119, 31700858), the Research Programs of Henan Science and Technology Department (192102210097, 192102210126, 192102210269, 182102210210, 172102210171), the Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (NO. CAAC-ITRB-201607), and the key scientific research projects of colleges and universities in Henan Province (18A520050).

Availability of data and materials

The datasets supporting the results of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Zhengzhou University of Aeronautics, Wenyuan Road, 450046 Zhengzhou, China. [2]Henan University of Economics and Law, Jinshui Road, 450046 Zhengzhou, China. [3]Information Technology Research Base of Civil Aviation Administration of China, Civil Aviation University of China, 300300 Tianjin, China. [4]HeNan Radio Television University, Wenyuan Road, 450046 Zhengzhou, China.

Received: 13 February 2020 Accepted: 21 October 2020
Published online: 23 November 2020

References

1. Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics. Pattern Recog. **40**(1), 262–282 (2007)
2. N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, N. Vasconcelos, in *ACM Multimedia*, A new approach to cross-modal multimedia retrieval (ACM, Firenze, 2010), pp. 251–260
3. H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, T.-S. Chua, in *ACM Multimedia*, Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval (ACM, Barcelona, 2013), pp. 33–42

4. X. Tian, D. Tao, X.-S. Hua, X. Wu, Active reranking for web image search. IEEE Trans. Image Process. **19**(3), 805–820 (2010)
5. J. Fan, D. A. Keim, Y. Gao, H. Luo, Z. Li, Justclick: personalized image recommendation via exploratory search from large-scale flickr images. IEEE TCSVT. **19**(2), 273–288 (2009)
6. C.-H. Yeh, Y.-C. Ho, B. A. Barsky, M. Ouhyoung, in *ACM Multimedia*, Personalized photograph ranking and selection system (ACM, Firenze, 2010), pp. 211–220
7. D. Grangier, S. Bengio, A discriminative kernel-based approach to rank images from text queries. IEEE Trans. Pattern Anal. Mach. Intell. **30**(8), 1371–1384 (2008)
8. S. C. Hoi, R. Jin, P. Zhao, T. Yang, Online multiple kernel classification. Mach. Learn. **90**(2), 289–316 (2013)
9. J. McAuley, J. Leskovec, in *ECCV*, Image labeling on a network: using social-network metadata for image classification (Springer Berlin Heidelberg, 2012), pp. 828–841
10. J. G. Kemeny, Mathematics without numbers. Daedalus. **88**(4), 577–591 (1959)
11. F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, VSE++: improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612. **2**(7), 8 (2017)
12. J. Song, Y. Yang, Y. Yang, Z. Huang, H. T. Shen, in *SIGMOD*, Inter-media hashing for large-scale retrieval from heterogeneous data sources (SIGMOD conference, New York, 2013)
13. J. Masci, M. M. Bronstein, A. A. Bronstein, J. Schmidhuber, Multimodal similarity-preserving hashing. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4), 824–830 (2013)
14. J. Song, Y. Yang, Z. Huang, H. Shen, R. Hong, in *ACM Multimedia*, Multiple feature hashing for real-time large scale near-duplicate video retrieval (ACM MM, Scottsdale, 2011)
15. Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding. IEEE Trans. Pattern Anal. Mach. Intell. **41**(9), 2070–2083 (2018)
16. A. Gordo, J. Almazán, J. Revaud, D. Larlus, End-to-end learning of deep visual representations for image retrieval. Int. J. Comput. Vis. **124**, 237–254 (2017)
17. L. Ma, Z. Lu, L. Shang, H. Li, in *Proceedings of the IEEE International Conference on Computer Vision*, Multimodal convolutional neural networks for matching image and sentence (ICCV, Santiago, 2015), pp. 2623–2631
18. J. Lu, D. Batra, D. Parikh, S. Lee, in *Advances in Neural Information Processing Systems*, Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, (Vancouver, 2019), pp. 13–23
19. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, in *Advances in Neural Information Processing Systems*, Attention is all you need, (2017), pp. 5998–6008
20. G. Chechik, V. Sharma, U. Shalit, S. Bengio, Large scale online learning of image similarity through ranking. J. Mach. Learn. Res. **11**, 1109–1135 (2010)
21. J. Wan, P. Wu, S. C. Hoi, P. Zhao, X. Gao, D. Wang, Y. Zhang, J. Li, in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, Online learning to rank for content-based image retrieval, (Buenos Aires, 2015)
22. H. Xia, S. C. Hoi, R. Jin, P. Zhao, Online multiple kernel similarity learning for visual search. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 536–549 (2013)
23. H. Xia, P. Wu, S. C. Hoi, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, Online multi-modal distance learning for scalable multimedia retrieval (ACM, 2013), pp. 455–464
24. Y. Wu, S. Wang, G. Song, Q. Huang, Online asymmetric metric learning with multi-layer similarity aggregation for cross-modal retrieval. IEEE Trans. Image Process. **28**(9), 4299–4312 (2019)
25. Y. Wu, S. Wang, Q. Huang, Online fast adaptive low-rank similarity learning for cross-modal retrieval. IEEE Trans. Multimedia. **22**, 1310–1322 (2019)
26. Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, X.-S. Hua, Visual query suggestion: towards capturing user intent in internet image search. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM). **6**(3), 1–19 (2010)
27. B. Zhang, L. Li, S. Yang, S. Wang, Z. Zha, Q. Huang, in *CVPR*, State-relabeling adversarial active learning (CVPR, 2020)
28. Y. Wang, T. Mei, J. Wang, H. Li, S. Li, in *ACM Multimedia*, Jigsaw: interactive mobile visual search with multimodal queries (ACM, 2011), pp. 73–82
29. B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, in *Proceedings of the 5th ACM Multimedia Systems Conference*, Div400: a social image retrieval result diversification dataset (ACM, 2014), pp. 29–34
30. Z. Wu, K. Zhou, Y. Liu, M. Zhang, S. Ma, Does diversity affect user satisfaction in image search. ACM Trans. Inf. Syst. (TOIS). **37**(3), 1–30 (2019)
31. E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, I. Vlahavas, in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Improving diversity in image search via supervised relevance scoring (ACM, 2015), pp. 323–330
32. W. Niu, J. Caverlee, H. Lu, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Neural personalized ranking for image recommendation (ACM, 2018), pp. 423–431
33. L. Gu, P. Yang, Y. Dong, Diversity optimization for recommendation using improved cover tree. Knowl.-Based Syst. **135**, 1–8 (2017)
34. Z. Zhang, X. Zheng, D. D. Zeng, A framework for diversifying recommendation lists by user interest expansion. Knowl.-Based Syst. **105**, 83–95 (2016)
35. T. Yao, C.-W. Ngo, T. Mei, in *ACM Multimedia*, Context-based friend suggestion in online photo-sharing community (ACM, 2011), pp. 945–948
36. W. Yin, T. Mei, C. Chen, S. Li, Socialized mobile photography: learning to photograph with social context via mobile devices. IEEE Trans. Multimedia. **16**, 184–200 (2013)
37. J. Zhuang, T. Mei, S. C. Hoi, X.-S. Hua, S. Li, in *ACM Multimedia*, Modeling social strength in social media community via kernel-based learning (ACM, 2011), pp. 113–122
38. C. Dwork, R. Kumar, M. Naor, D. Sivakumar, in *Proceedings of the 10th International Conference on World Wide Web*, Rank aggregation methods for the web (ACM, New York, 2001), pp. 613–622
39. X. Chen, P. N. Bennett, K. Collins-Thompson, E. Horvitz, in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, Pairwise ranking aggregation in a crowdsourced setting, (Rome, 2013), pp. 193–202
40. R. C. Prati, in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Combining feature ranking algorithms through rank aggregation (IEEE, Brisbane, 2012), pp. 1–8

41. J. Ding, D. Han, J. Dezert, Y. Yang, A new hierarchical ranking aggregation method. Inf. Sci. **453**, 168–185 (2018)
42. J. Ding, D. Han, Y. Yang, Iterative ranking aggregation using quality improvement of subgroup ranking. Eur. J. Oper. Res. **268**(2), 596–612 (2018)
43. S. Liang, I. Markov, Z. Ren, M. de Rijke, in *Proceedings of the 2018 World Wide Web Conference*, Manifold learning for rank aggregation, (Lyon, 2018), pp. 1735–1744
44. X. Tian, Y. Lu, L. Yang, Q. Tian, in *ACM Multimedia*, Learning to judge image search results (ACM, Scottsdale, 2011), pp. 363–372
45. Z. Zha, J. Yu, M. Wang, T. Chua, Product aspect ranking and its applications. IEEE Trans. Knowl. Data Eng. **26**, 1211–1224 (2013)
46. K. Subbian, P. Melville, Supervised rank aggregation for predicting influence in networks (2011). arXiv preprint arXiv:1108.4801
47. O. Dalal, S. H. Sengemedu, S. Sanyal, in *Proceedings of the 21st International Conference on World Wide Web*, Multi-objective ranking of comments on web (ACM, Lyon, 2012), pp. 419–428
48. A. Klementiev, D. Roth, K. Small, in *Proceedings of the 25th International Conference on Machine Learning*, Unsupervised rank aggregation with distance-based models (ACM, Helsinki, 2008), pp. 472–479
49. M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, R. Hu, Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Trans. Multimedia. **18**(12), 2553–2566 (2016)
50. Y. Cui, L. Deng, Y. Zhao, B. Yao, V. W. Zheng, K. Zheng, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Hidden poi ranking with spatial crowdsourcing (ACM, Anchorage, 2019), pp. 814–824
51. F. R. Bach, G. R. Lanckriet, M. I. Jordan, in *Proceedings of the 21st International Conference on Machine Learning*, Multiple kernel learning, conic duality, and the SMO algorithm (ACM, Banff, 2004), p. 6
52. T. Malisiewicz, A. Gupta, A. A. Efros, in *International Conference on Computer Vision (ICCV)*, Ensemble of exemplar-svms for object detection and beyond, (Barcelona, 2011), pp. 89–96
53. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, in *Proceedings of the IEEE International Conference on Computer Vision*, Focal loss for dense object detection, (Venice, 2017), pp. 2980–2988
54. K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
55. B. Kulis, K. Grauman, Kernelized locality-sensitive hashing. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1092–1104 (2011)
56. F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**(6), 386 (1958)
57. Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
58. P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, R. Zadeh, in *Proceedings of the 22nd International Conference on World Wide Web*, Wtf: the who to follow service at twitter, (New York, 2013), pp. 505–514
59. J. Pennington, R. Socher, C. Manning, in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Glove: global vectors for word representation, (Doha, 2014), pp. 1532–1543
60. C. Budak, T. Georgiou, D. A. A. El Abbadi, Geoscope: online detection of geo-correlated information trends in social networks. Proc. VLDB Endowment. **7**(4), 229–240 (2013)
61. T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, in *Proceedings of the ACM International Conference on Image and Video Retrieval*, NUS-WIDE: a real-world web image database from National University of Singapore (ACM, Santorini Island, 2009), pp. 1–9

Publisher's Note