Int J Multimed Info Retr (2015) 4:233–245
DOI 10.1007/s13735-015-0088-x

REGULAR PAPER

aMM: Towards adaptive ranking of multi-modal documents

Mohammad Akbari[1] · **Liqiang Nie**[2] · **Tat-Seng Chua**[2]

Received: 4 August 2015 / Revised: 9 September 2015 / Accepted: 16 September 2015 / Published online: 28 September 2015

Abstract Information reranking aims to recover the true order of the initial search results. Traditional reranking approaches have achieved great success in uni-modal document retrieval. They, however, suffer from the following limitations when reranking multi-modal documents: (1) they are unable to capture and model the relations among multiple modalities within the same document; (2) they usually concatenate diverse features extracted from different modalities into one single vector, rather than adaptively fusing them by considering their discriminative capabilities with respect to the given query; and (3) most of them consider the pairwise relations among documents but discard their higher-order grouping relations, which leads to information loss. Towards this end, we propose an adaptive multi-modal multi-view (**aMM**) reranking model. This model is able to jointly regularize the relatedness among modalities, the effects of feature views extracted from different modalities, as well as the complex relations among multi-modal documents. Extensive experiments on three datasets well validated the effectiveness and robustness of our proposed model.

Keywords Information reranking · Adaptive ranking · Multi-modal multi-view search

✉ Mohammad Akbari
akbari@u.nus.edu; akbari.ma@gmail.com

Liqiang Nie
nieliqiang@gmail.com

Tat-Seng Chua
chuats@comp.nus.edu.sg

[1] NUS Graduate School for Integrative Sciences and Engineering, School of Computing, National University of Singapore, Singapore, Singapore

[2] School of Computing, National University of Singapore, Singapore, Singapore

1 Introduction

Multimedia documents are widely used for information sharing in the Web 2.0 era. Besides the textual modality, these documents are often accompanied with other heterogeneous modalities, such as videos, images and sometimes audios. Take the example of a news article talking about ebola outbreak, it may contain the textual descriptions of the news stories, images of the suffering areas and perhaps the video clips of social assistance activities. Indeed, a multi-modal document is the intrinsic nature of information available in the real world. It exhibits three advantages: (1) **Intuition**. For some visual themes and dynamic procedures, pure textual documents cannot vividly convey their contents. Comparatively, a picture is worth a thousand words and a video is worth a million. (2) **Comprehension**. Multiple modalities are able to capture the content from different aspects and enhance a comprehensive understanding by the audiences. (3) **Coherence**. Although multiple modalities within the same document express the content from different angles, they are often consistently describing the same topics. In fact, one multi-modal document can be regarded as a bi-level composition: the document consists of multiple medium types such as image and text, and each medium type can be represented by a rich set of features. In this work, we use **modality** to refer to the medium type and **feature view** to represent the feature type extracted from each modality. For example, in the known-item search (KIS) problem [6], each document consists of a video and its textual ASR (automatic speech recognition). Meanwhile, each of these two modalities can be represented by a rich set of feature views.

Understanding and retrieval of multi-modal documents are, however, non-trivial due to the following reasons: (1) **Modality Agreement**. Instead of isolation, heterogeneous modalities collectively highlight the same semantics of

the given documents with different medium types. How to effectively model the consistency relations is a tough challenge. (2) **Adaptive Confidences**. The discriminative capabilities of modalities are query specific [3,33,38]. For instance, image modality may dominate the reranking performance when searching for a visual concept such as a dog. On the other hand, textual modality may contribute more when searching for research articles. Besides modality, the representativeness of feature views varies significantly in accordance to specific queries [14,25]. Take the color and edge features under image modality as an example. Color feature is effective in differentiating the day from the night, while shape-like feature can signal the appearance of the involved objects [5,36]. We thus have to adaptively fuse and modulate the effects of different modalities and feature views. (3) **Complex Relations**. A multi-document usually has a group of similar neighbours. Traditional ranking methods consider only the pairwise relation between two samples, and they ignore the relation in a higher grouping order. Essentially, modeling the higher-order relation among samples will significantly improve the ranking performance [26,44].

It is worth mentioning that several research efforts have been dedicated to multi-modal document retrieval [37,39,43]. Roughly speaking, they can be classified into two major categories: **early fusion** [18,32,41] and **late fusion** [7,8,35,42]. Early fusion methods, such as the work in [32], construct a joint feature space by merging all the extracted features from different modalities into a single concatenated feature vector. However, they suffer from the following limitations: (1) they are unable to differentiate and leverage the discriminative capabilities of distinct views and modalities, because they usually treat them equally; (2) the features extracted from various modalities may not fall into the same semantic space and simply merging all feature views actually bring in a certain extent of noise and ambiguity [12]; and (3) they may lead to the curse of dimensionality [23], since the final feature vector would be of very high dimension. On the other hand, late fusion [32,42] analyzes each modality and feature view separately and then integrates their results. The fused result however might not be reasonably accurate due to two reasons: (1) individual feature space might not be sufficiently descriptive to represent the complex semantics of the multi-modal documents. Therefore, separate results would be suboptimal and the integration may not result in the desired outcome; and (2) it is labor-intensive to tune the fusion weights for different modalities and feature views. Even worse, the optimal parameters for one query cannot be directly applied to another query.

To address the aforementioned challenges, we propose an adaptive multi-modal multi-view (**aMM**) reranking approach to refine the initial search result of multi-modal documents. Figure 1 illustrates the framework of our model. In particular, given a collection of multi-modal documents, we first extract various feature views from each modality to comprehensively represent its content. For each feature view, we then construct a hypergraph to characterize the higher-order relations among documents and ensure that the document relevances tend to be close if they share a similar feature view. We next adaptively unify all these hypergraphs into one model and co-regularize the initial admissibility and modality agreement. Initial admissibility is to partially reserve the information of initial search results; while modality agreement leverages the inter-relations among modalities to reinforce the ranking performance. In addition, we have theoretically proven that our proposed model is a linear model, which makes it feasible for use in the large-scale applications. Extensive experiments on three real-world datasets show its superiority over other state-of-the-art approaches.

The contributions of this work are threefold:

– To boost the search performance of multi-modal documents, we propose a novel hypergraph-based model. Beyond the smoothness and initial admissibility considered by traditional reranking approaches, it also explores the higher-order inter-relations among multi-modal documents and semantic intra-agreements within multi-modal documents.

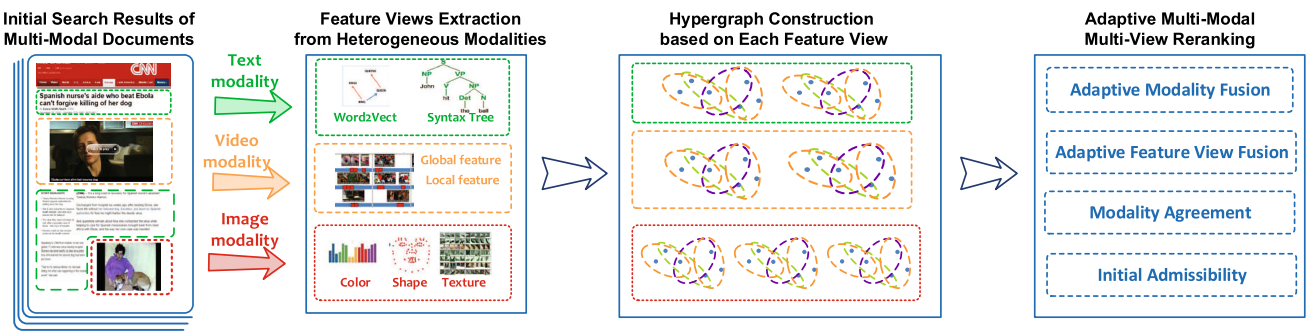

Fig. 1 The conceptual view of a multi-modal multi-view search. Each document may comprise of multiple modalities. Various feature views are extracted to represent each modality. Based upon each feature view, we construct a hypergraph to link all the multi-modal documents. We finally propose a unified model to rerank multi-modal documents by jointly regularizing various prior assumptions

– Our proposed model is able to adaptively learn the fusion weights of modalities and feature views simultaneously. Meanwhile, we theoretically demonstrate that our proposed model is a linear model and practically analyze its computational complexity.
– Besides evaluation on our manually constructed dataset, we verified the robustness of our model on other well-known datasets.

The remainder of the paper is organized as follows. Section 2 reviews the traditional reranking approaches based upon simple graph and hypergraph. Section 3 details our proposed reranking approach. Experimental results and analysis are presented in Sect. 4, followed by the conclusions and future work in Sect. 5.

2 Graph-based reranking

In the past decade, graph-based reranking has attracted a lot of research interests [3,4,10,28,29,44], which can be roughly categorized into two classes: simple graph-based reranking [13,14,37,41,47] and hypergraph-based reranking [3,11,19,29,43–45]. In fact, most of them are designed based on the following two assumptions:

– *Smoothness* The relevance probability function is continuous and smooth in the semantic space. In other words, the relevance probabilities of semantically similar documents should be close.
– *Initial admissibility* The initial ranking result partially reflects correct information, and hence reranking result should not deviate too much from the initial list.

The optimization framework which seamlessly integrates these two assumptions into a learning framework, is defined as the following objective function, $\Phi(.)$,

$$\arg\min_{\mathbf{f}} \Phi(\mathbf{f}, \mathbf{y}, \mathbf{X}) = \arg\min_{\mathbf{f}} \{\Omega(\mathbf{f}, \mathbf{X}) + \gamma R(\mathbf{f}, \mathbf{y})\}, \quad (1)$$

where $\mathbf{y}$ is the initial search result based on the given query; $\mathbf{f}$ is a ranking function we aim to learn; and $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}\}$ denotes the sample set with n documents. In addition, $\Omega(.)$ is a regularization term that enforces the smoothness assumption; and $R(.)$ is a loss term that computes the difference between $\mathbf{f}$ and $\mathbf{y}$. $R(.)$ is used to guarantee the initial admissibility and is frequently defined as a least squares function,

$$R(\mathbf{f}, \mathbf{y}) = \|\mathbf{f} - \mathbf{y}\|^2. \quad (2)$$

Note that the parameter γ is a regularization parameter to balance the regularizers on smoothness assumption and the empirical loss.

2.1 Simple graph-based reranking

In a simple graph $G_s = (V_s, E_s)$, vertices V_s are used to represent data items and each edge in E_s connects two related vertices. The constructed graph can either be undirected or directed, depending on the symmetric or asymmetric nature of the relations between the data items [47,48]. The learning process is then conducted on the constructed graph to propagate the initial ranking information until convergence. In particular, relevant items at the bottom positions in the initial ranking list are pulled up based on their similarities with items at the top of the list. In simple graph-based learning, regularizer term, $\Omega(.)$ is formulated as,

$$\begin{aligned}\Omega(\mathbf{f}, \mathbf{X}) &= \frac{1}{2} \sum_{v_i, v_j \in V_s} W_{i,j} \left(\frac{f_i}{\sqrt{D_{ii}}} - \frac{f_j}{\sqrt{D_{jj}}} \right)^2 \\ &= \mathbf{f}^T (\mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}) \mathbf{f} = \mathbf{f}^T \mathbf{\Delta_s} \mathbf{f}, \end{aligned} \quad (3)$$

where $\mathbf{\Delta_s} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}$ is the so-called graph Laplacian [29], and $\mathbf{D}$ is a diagonal matrix with its (i, i)th element equal to the sum of the ith row of the affinity matrix $\mathbf{W}$. We use $\mathbf{W}$ to denote the similarity matrix and W_{ij}, its (i, j)th element, indicates the similarity between two documents v_i and v_j. Typically, it is estimated as

$$W_{ij} = \begin{cases} K(v_i, v_j) & \text{if } v_j \in N_K(v_i) \text{ or } v_i \in N_K(v_j) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $N_K(v_i)$ denotes the index set for the K nearest neighbours of document v_i computed by Euclidean distance and noting that W_{ii} is set as 1, so that self-loop is included.

Even though great success has been achieved, simple graph-based reranking approaches encounter a tough problem, especially for multi-modal documents. Instead of grouping relations among documents, it only captures the pairwise relations due to their intrinsic nature. For example, from a simple graph, we can easily find two close samples according to the pairwise similarities, but it is not easy to predict whether there are three or more close samples. On the other hand, previous efforts have pointed out that the relations among multi-modal documents are much more complex and beyond the pairwise relations [11,19,44,49].

2.2 Hypergraph-based reranking

Hypergraph-based reranking well addresses the problems faced by simple graph-based reranking [3,29,44,45]. In particular, hypergraph is an extension of simple graph and is able

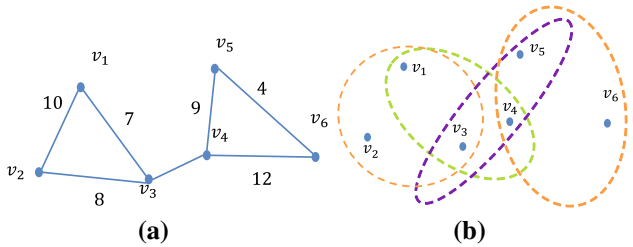

Fig. 2 Comparison illustration between simple graph and hypergraph: **a** a simple graph with six data items, where each edge captures the pairwise relations among data items; and **b** a hypergraph with six data items and four hyperedges, where each vertex and its two nearest neighbors form a hyperedge. It can capture the grouping relations

to capture the local grouping relations among more than two vertices [49]. In hypergraph, a set of vertices is connected by a hyperedge. The weight of a hyperedge indicates to what extent the vertices in a hyperedge belong to the same group. Figure 2 shows an example to contrast the difference between simple graph and hypergraph.

A hypergraph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{W})$ is composed of the vertex set $\mathcal{V}$, the hyperedge set $\mathcal{E}$, and the diagonal matrix of hyperedge weight $\mathbf{W}$. Unlike the edge in a simple graph, each hyperedge $e_i \in \mathcal{E}$ connects an arbitrary subset of $\mathcal{V}$, and is assigned a specific weight $W(e_i)$. A probabilistic hypergraph $\mathcal{G}$ can be represented by a $|\mathcal{V}| \times |\mathcal{E}|$ incidence matrix $\mathbf{H}$ with the following entries,

$$H(v_i, e_j) = \begin{cases} P(v_i, e_j), & \text{if } v_i \in e_j, \\ 0, & \text{if } v_i \notin e_j, \end{cases} \tag{5}$$

where $P(v_i, e_j)$ expresses the probability that vertex v_i associated with the hyperedge e_j. Considering $\mathbf{H}$ as the incidence matrix, the vertex degree of $v_i \in \mathcal{V}$ is defined as,

$$d(v_i) = \sum_{e_j \in \mathcal{E}} W(e_j) H(v_i, e_j). \tag{6}$$

For each hyperedge $e_j \in \mathcal{E}$, the hyperedge degree $\delta(e)$ is defined as the number of nodes it contains. The hyperedge weight $W(e_j)$ is computed as,

$$W(e_j) = \sum_{v_i \in e_j} H(v_i, e_j). \tag{7}$$

Different approaches have been proposed for learning based on a hypergraph [1,49,50]. A majority of them defined the normalized regularizer on the hypergraph as,

$$\Omega(\mathbf{f}, \mathbf{X}) = \frac{1}{2} \sum_{e \in \mathcal{E}} \sum_{u,v \in \mathcal{V}} \frac{W(e) H(u,e) H(v,e)}{\delta(e)} \times \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2, \tag{8}$$

where $\mathbf{f}$ contains the relevance probabilities of all the documents that we aim to learn. Let $\mathbf{D}_v$ and $\mathbf{D}_e$ denote diagonal matrices of vertex degrees and hyperedge degrees, respectively. By defining $\boldsymbol{\Theta} = \mathbf{D}_v^{-1/2} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^T \mathbf{D}_v^{-1/2}$, we can further derive Eq. (9) as

$$\Omega(\mathbf{f}, \mathbf{X}) = \mathbf{f}^T (\mathbf{I} - \boldsymbol{\Theta}) \mathbf{f}, \tag{9}$$

where $\mathbf{I}$ is an identity matrix. Let $\boldsymbol{\Delta} = \mathbf{I} - \boldsymbol{\Theta}$, which is a positive semidefinite matrix, the so-called graph Laplacian. Then $\Omega(\mathbf{f}, \mathbf{X})$ can be simply rewritten as

$$\Omega(\mathbf{f}, \mathbf{X}) = \mathbf{f}^T \boldsymbol{\Delta} \mathbf{f}. \tag{10}$$

The above equations can be solved using the well-known graph-based random-walk framework [2,15,21]. It can be seen that the simple graph Laplacian is a special case of hypergraph Laplacian where all hyperedges have degree two and each of them links only two vertices [29].

3 Adaptive multi-modal multi-view reranking

We have introduced the basic concept of reranking based on simple graph and hypergraph. In this section, we will detail the formulation and optimization of our proposed **aMM** methods for the reranking of multi-modal documents.

3.1 Formulation

Adaptive modality fusion Without loss of generality, we assume that each multi-modal document comprises of K distinct modalities, and they collectively describe the document from various perspectives. It is thus reasonable to take all these modalities into consideration to enhance the reranking performance. It is natural to first estimate the ranking score of each document utilizing individual modality and then merge all of them together equally,

$$\mathbf{f} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{f}^i, \tag{11}$$

where $\mathbf{f}^i$ denotes the ranking function of the ith modality and $\mathbf{f}$ is the final ranking function which aggregates the ranking results of different modalities. However, it may lead to suboptimal results, because different modalities convey different information and consequently have different contribution confidences towards the final result. In addition, the confidence values of distinct modalities are query specific. Towards this end, we adaptively fuse them by,

$$\mathbf{f} = \sum_{i=1}^{K} \lambda^i \mathbf{f}^i, \quad \text{s.t.} \quad \sum_{i=1}^{K} \lambda^i = 1, \tag{12}$$

where λ^i is the adaptive modality weight that modulates the effect of the ith modality. Instead of manually tuning the λ^i, they are automatically learned, based upon the given query. We intentionally do not constrain λ^i to be larger than zero. In this way, we can infer which modality is negatively correlated with the reranking task.

Adaptive feature view fusion As aforementioned, various feature views can be extracted from each individual modality. For each feature view, we construct one hypergraph to capture the grouping relations among multi-modal documents. The Laplacian regularizer of the ith modality is a linear combination of various Laplacians of its feature views. Mathematically, it is stated as,

$$\Omega(\mathbf{f}, \mathbf{X}) = \sum_{i=1}^{K} \lambda^i \sum_{j=1}^{N_i} \alpha_j^i \mathbf{f}^{i^T} \mathbf{\Delta}_j^i \mathbf{f}^i,$$
$$\text{s.t.} \quad \sum_{j=1}^{N_i} \alpha_j^i = 1, \quad 0 \le \alpha_j^i \le 1, \tag{13}$$

where $\mathbf{\Delta}_j^i$ denotes graph laplacian constructed based on the jth feature view extracted from the ith modality; α_j^i is the weight for the jth feature view, and N_i is the number of feature views extracted from the ith modality. For instance, if the ith modality of one multi-modal document is text and we extract topic-level, part-of-speech, and uni-gram features to represent this modality. In the context of this example, N_i is set as 3. In fact, $\mathbf{\Delta}_j^i$ implicitly ensures the **Smoothness** as discussed in the Sect. 2.

Modality disagreement penalty For a given multi-modal document, its ranking score estimated by exploring different modalities should be the same or sufficiently close. This is because heterogeneous modalities consistently convey the same underlying semantics via different medium types. Therefore, large deviations in ranking results based on different modalities should be penalized. Considering the consistency among distinct modalities, we can extend Eq. (13) to,

$$\Omega(\mathbf{f}, \mathbf{X}) = \sum_{i=1}^{K} \lambda^i \sum_{j=1}^{N_i} \alpha_j^i \mathbf{f}^{i^T} \mathbf{\Delta}_j^i \mathbf{f}^i$$
$$+\beta \sum_{\bar{i} \neq i} \|\mathbf{f}^i - \mathbf{f}^{\bar{i}}\|^2 + \mu \sum_{i=1}^{K} \|\boldsymbol{\alpha}^i\|^2,$$
$$\text{s.t.} \quad \sum_{i=1}^{K} \lambda^i = 1, \quad \sum_{j=1}^{N_i} \alpha_j^i = 1. \tag{14}$$

The first term of the above equation is to adaptively fuse the feature views and modalities. The second term reflects the agreement among different ranking functions computed from distinct modalities. It indeed enforces the final ranking scores computed from different modalities to be close to each other. The last term controls the sparsity of the model to avoid overfitting.

Initial admissibility Since we take the discriminative capabilities of each modality into consideration, we have to restate the loss penalty in Eq. (2) as

$$R(\mathbf{f}, \mathbf{y}) = \|\mathbf{f} - \mathbf{y}\|^2 = \|\sum_{i=1}^{K} \lambda^i \mathbf{f}^i - \mathbf{y}\|^2. \tag{15}$$

Unifying all the aforementioned regularizers, we arrive at our proposed adaptive multi-modal multi-view reranking model for the multi-modal documents.

$$\Phi(\mathbf{f}, \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{K} \lambda^i \sum_{j=1}^{N_i} \alpha_j^i \mathbf{f}^{i^T} \mathbf{\Delta}_j^i \mathbf{f}^i + \beta \sum_{\bar{i} \neq i} \|\mathbf{f}^i - \mathbf{f}^{\bar{i}}\|^2$$
$$+ \gamma \|\sum_{i=1}^{K} \lambda^i \mathbf{f}^i - \mathbf{y}\|^2 + \mu \sum_{i=1}^{K} \|\boldsymbol{\alpha}^i\|^2. \tag{16}$$

We can see that we need to learn the following variables: (1) $\mathbf{f^i}$: the ranking function for the ith modality; (2) λ^i: the contribution confidence for the ith modality; and (3) α_j^i: the optimal weight for the jth feature view extracted from the ith modality. We derive the solutions via alternative optimization.

3.2 Optimization

We adopt an alternative optimization method to optimize Eq. (16). Specifically, we update $\mathbf{f}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$ alternatively with others fixed to minimize the objective function.

3.2.1 Optimization of $\mathbf{f}$

We first fix $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. We then take the derivative of $\Phi(.)$ with respect to $\mathbf{f}^i$. We have

$$\frac{\partial \Phi}{\partial \mathbf{f}^i} = 2\Big\{\lambda^i \sum_{j=1}^{N_i} \alpha_j^i \mathbf{\Delta}_j^i + \Big[\beta(K-1) + \gamma {\lambda^i}^2\Big]\mathbf{I}\Big\}\mathbf{f}^i$$
$$+ 2\sum_{\bar{i} \neq i} (-\beta + \gamma \lambda^i \lambda^{\bar{i}}) \mathbf{f}^{\bar{i}} - 2\gamma \lambda^i \mathbf{y}. \tag{17}$$

Setting Eq. (17) to be zero, we obtain the following set of equations,

$$\begin{cases} \mathbf{t}^i & = \gamma \lambda^i \mathbf{y}, \\ \mathbf{L}^{ii} & = \lambda^i \sum_{j=1}^{N_j} \alpha_j^i \mathbf{\Delta}_j^i + \Big[\beta(K-1) + \gamma {\lambda^i}^2\Big]\mathbf{I}, \\ \mathbf{L}^{i\bar{i}} & = \Big(\gamma \lambda^i \lambda^{\bar{i}} - \beta\Big)\mathbf{I}. \end{cases}$$

We can rewrite Eq. (17) as,

$$\mathbf{t}^i = \mathbf{L}^{ii}\mathbf{f}^i + \sum_{\bar{i}\neq i} \mathbf{L}^{i\bar{i}}\mathbf{f}^{\bar{i}}. \tag{18}$$

The above equation systems explicitly suggest that we have to jointly learn all $\mathbf{f}^i$ and $\mathbf{f}^{\bar{i}}$, where $i \neq \bar{i}$. Meanwhile, we can re-write Eq. (18) as the following linear system,

$$\begin{bmatrix} \mathbf{L}^{11} & \mathbf{L}^{12} & \dots & \mathbf{L}^{1K} \\ \mathbf{L}^{21} & \mathbf{L}^{22} & \dots & \mathbf{L}^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}^{K1} & \mathbf{L}^{K2} & \dots & \mathbf{L}^{KK} \end{bmatrix} \begin{bmatrix} \mathbf{f}^1 \\ \mathbf{f}^2 \\ \vdots \\ \mathbf{f}^K \end{bmatrix} = \begin{bmatrix} \mathbf{t}^1 \\ \mathbf{t}^2 \\ \vdots \\ \mathbf{t}^K \end{bmatrix}. \tag{19}$$

It is worth noting that we can equivalently further simplify this linear system as,

$$\hat{\mathbf{L}}\hat{\mathbf{f}} = \hat{\mathbf{t}}, \tag{20}$$

where $\hat{\mathbf{L}}$ is a block symmetric matrix with $K \times K$ blocks. We can prove that $\hat{\mathbf{L}}$ is positive definite, and thus invertible; $\hat{\mathbf{f}}$ and $\hat{\mathbf{t}}$ are both block vectors with K elements. Noticeably, $\hat{\mathbf{t}}$ and $\hat{\mathbf{L}}$ are constant since λ^i and $\mathbf{y}$ are fixed. The analytical solution of $\hat{\mathbf{f}}$ can be derived as

$$\hat{\mathbf{f}} = \hat{\mathbf{L}}^{-1}\hat{\mathbf{t}}. \tag{21}$$

3.2.2 Optimization of λ^i

We then apply Lagrangian method to incorporate the constraints on λ^i as follows,

$$\begin{aligned} \Phi = &\sum_{i=1}^{K} \lambda_i \sum_{j=1}^{N_i} \alpha_j^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i \\ &+ \gamma \left\| \sum_{i=1}^{K} \lambda^i \mathbf{f}^i - \mathbf{y} \right\|^2 + \xi \left(1 - \sum_{i=1}^{K} \lambda^i \right) + C, \end{aligned} \tag{22}$$

where ξ is a Lagrangian multiplier and $C = \beta \sum_{\bar{i}\neq i} \|\mathbf{f}^i - \mathbf{f}^{\bar{i}}\|^2 + \mu\|\boldsymbol{\alpha}\|^2$ is constant with respect to λ^i. We now fix $\mathbf{f}^i$, α_j^i and $\lambda^1, \dots, \lambda^{i-1}, \lambda^{i+1}, \dots, \lambda^K$. We then take the derivative of Φ with respect to λ^i,

$$\begin{aligned} \frac{\partial \Phi}{\partial \lambda^i} = &2\gamma {\mathbf{f}^i}^T \mathbf{f}^i \lambda^i + \sum_{\bar{i}\neq i}^{K} 2\gamma {\mathbf{f}^i}^T \mathbf{f}^{\bar{i}} \lambda^{\bar{i}} \\ &+ \sum_{j=1}^{N_i} \alpha_j^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i - 2\gamma {\mathbf{f}^i}^T \mathbf{y} - \xi. \end{aligned} \tag{23}$$

We set Eq. (23) to be zero and we can derive

$$\lambda^i = \frac{2\gamma {\mathbf{f}^i}^T \left(\mathbf{y} - \sum_{\bar{i}\neq i}^{K} \lambda^{\bar{i}} \mathbf{f}^{\bar{i}} \right) - \sum_{j=1}^{N_i} \alpha_j^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i + \xi}{2\gamma {\mathbf{f}^i}^T \mathbf{f}^i}. \tag{24}$$

To compute the value of Lagrangian multiplier, we rearrange the terms in Eq. (24) as follows,

$$\begin{cases} \lambda^i &= \frac{\Pi^i}{\Psi^i} + \frac{\xi}{\Psi^i}, \\ \Pi^i &= 2\gamma {\mathbf{f}^i}^T \left(\mathbf{y} - \sum_{\bar{i}\neq i}^{K} \lambda^{\bar{i}} \mathbf{f}^{\bar{i}} \right) - \sum_{j=1}^{N_i} \alpha_j^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i, \\ \Psi^i &= 2\gamma {\mathbf{f}^i}^T \mathbf{f}^i. \end{cases} \tag{25}$$

With the prior knowledge, we know that the sum of λ^i equals to one. We can compute the Lagrangian multiplier as

$$\sum_{i=1}^{K} \lambda^i = \sum_{i=1}^{K} \left(\frac{\Pi^i}{\Psi^i} + \frac{\xi}{\Psi^i} \right) = 1. \tag{26}$$

We can then obtain the following results,

$$\xi = \frac{1 - \sum_{i=1}^{K} \frac{\Pi^i}{\Psi^i}}{\sum_{i=1}^{K} \frac{1}{\Psi^i}}. \tag{27}$$

3.2.3 Optimization of α_j^i

To minimize the cost function with respect to α_j^i, we fix $\mathbf{f}^i$s and λ^is first. We then compute the value of α_j^i. To consider the constraint on α_j^i, we also turn to the Lagrangian multiplier,

$$\begin{aligned} \Phi = &\sum_{i=1}^{K} \lambda^i \sum_{j=1}^{N_i} \alpha_j^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i + \mu \sum_{i=1}^{K} \|\boldsymbol{\alpha}^i\|^2 \\ &+ \eta \left(1 - \sum_{j=1}^{n_i} \alpha_j^i \right) + C, \end{aligned} \tag{28}$$

where C is constant with respect to α_j^i. Now we take derivation with respect to α_j^i as follows

$$\frac{\partial \Phi}{\partial \alpha_j^i} = \lambda^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i + 2\mu\alpha_j^i - \eta, \tag{29}$$

where η is the Lagrangian multiplier. Setting Eq. (29) to zero, we can find the solution of α_j^i as

$$\alpha_j^i = -\frac{\left(\lambda^i {\mathbf{f}^i}^T \mathbf{\Delta}_j^i \mathbf{f}^i - \eta \right)}{2\mu}, \tag{30}$$

where Lagrangian multiplier can be computed from the constraint $\sum_{j=1}^{n_i} \alpha_j^i = 1$ as follows,

$$\sum_{j=1}^{n_i} \alpha_j^i = -\sum_{j=1}^{n_i} \frac{\left(\lambda^i \mathbf{f}^{i\,T} \boldsymbol{\Delta}_j^i \mathbf{f}^i - \eta\right)}{2\mu} = 1. \tag{31}$$

Therefore, we have

$$\eta = \frac{\left(\sum_{j=1}^{n_i} \lambda^i \mathbf{f}^{i\,T} \boldsymbol{\Delta}_j^i \mathbf{f}^i\right) + 2\mu}{n_i}. \tag{32}$$

3.3 Computational complexity

The computational cost of the proposed approach mainly contains three parts, which are the cost of updating of $\mathbf{f}$, λ^i, and α_j^i:

1. The computational complexity of updating $\mathbf{f}$ using the Eq. (21) is $O(K^3N^3)$.
2. From Eq. (24), we can infer that the computational complexity of computing λ^i is $O(N^2)$. We have K modalities, and hence the complexity of updating all λ^is is $O(KN^2)$.
3. For updating α_j^i, the complexity is $O(N^2)$, and it is $O((K+D)N^2)$ for all feature views extracted from all the modalities.

In summary, the complexity of the whole optimization process is $O(T(K^3N^3 + KN^2 + (K+D)N^2))$, where N is the number of samples, D is the dimension of the largest feature view, K is the number of modalities and T is the number of iterations of alternating optimization, respectively. It is possible to use an iterative method to solve Eq. (21) instead of using matrix inversion, which is analogous to the method in [47]. In this way the computational complexity becomes $O(T(KN^2 + KN^2 + (K+D)N^2)$.

4 Experiments

In this section, we aim to answer the following questions:

1. How well does the proposed **aMM** reranking model work as compared to other state-of-the-art methods?
2. How effective is each of the components within the proposed model?
3. How sensitive is our proposed model to the involved parameters?
4. Besides our manually constructed dataset, how robust is the proposed model with respect to other publicly accessible datasets?

Table 1 The representative complex queries collected from Google Image Search Engine

Query ID	Complex query string
1	Baby eating apple while lying
2	Boy swimming in the river
3	Lady wear sunglasses on the sea beach
4	Lion hunting zebra in the grassland
5	Two kids playing with a ball in the park

Here we do not list all of the queries due to limited space

In the rest of the section, we first introduce the experimental settings. We then respectively explore the answers to the aforementioned four experimental questions. We finally summarize the key findings from the experiments.

4.1 Experimental setting

4.1.1 Data collection

To verify our proposed model, we build a multi-modal dataset with complex queries. The complex queries are constructed based on the suggestion of Google Image Search Engine. To be more specific, inspired by [27], we first selected a set of visual concepts from community question answering forums, such as Yahoo! Answer.[1] We then issued them to Google Image search Engine. We manually selected a set of 20 verbose queries from the suggestion lists. Some representative queries are listed in Table 1. For each complex query, we collected its top ranked images returned by the Google Image Search Engine together with their corresponding textual web pages. Consequently, documents in our constructed dataset contain both images and texts. In total, it comprises of 4341 multi-modal documents and on average there are approximately 217 documents for each query. The reasons that we constructed the dataset in this way are as follows: (1) the suggested complex queries are hot and real queries, which are frequently issued by information seekers; (2) current commercial web search engines do not, in general, perform well with verbose queries, especially for image retrieval;[2] and (3) this dataset contains original text-based ranking information and we thus can easily evaluate whether our approach can outperform the search engine.

To obtain the ground truth of each multi-modal document, we conducted a manual labelling process. Three human annotators were involved in the process. Each document was labelled to be very relevant (score 2), relevant (score 1) or

[1] https://answers.yahoo.com/.

[2] A study in [30] shows that a failed image query tends to be longer than the average successful query, which indicates longer queries' higher specificity of contents and also reveals the limitations of current web image search engines for complex queries.

irrelevant (score 0) with respect to each given query. We performed a voting to establish the final relevance level of each document. For cases when each of the three relevance classes for the given document has only one voting ballot, a discussion was carried out among the annotators to decide the final ground truths.

4.1.2 Evaluation metrics

In order to evaluate the reranking performance for each query, we apply the normalized Discounted Cumulative Gain (nDCG) [26], which is a standard measurement in information retrieval [25,31]. For a given query, the nDCG at position n can be computed as

$$nDCG@n = \frac{\mathrm{rel}_1 + \sum_{i=2}^{n} \frac{\mathrm{rel}_i}{\log_2 i}}{\mathrm{DCG}}, \tag{33}$$

where rel_i is the relevance score of the ith multi-modal document in the ranked list, and IDCG is the normalizing factor that makes $nDCG@n$ equals to 1 for perfect ranking. To compute the overall performance, we report the average $nDCG@n$ over all queries.

4.1.3 Feature views extraction

There are two different modalities in our dataset: image and text. For each modality, we extracted a rich set of features to comprehensively describe the multi-modal documents from various views. The extracted features are summarized in Table 2. In total, we have extracted 7 feature views from image modality with 929 dimensional visual features; and 3 feature views from text modality with 2020 dimensional textual features.[3]

4.2 On model performance comparison

To examine the efficacy of our proposed model, we conducted experiments to compare the performance of our model with other state-of-the-art competitors:

- *GISE* It is the original search result of Google Image Search Engine without any further processing and reranking.
- *SGRW* It is the simple graph based random walk reranking method [41]. It concatenates various feature views from all modalities in to a vector. We empirically set the involved tradeoff parameter γ as 0.9.
- *sMM* It is a simple graph-based multi-modal reranking method introduced in [37], which integrates different features but does not differentiate modalities. We set the involved parameters λ and ξ to be 0.1 and 10, respectively.
- *aMM* It is our proposed model. We set the parameters β, γ, and μ to be 20, 10, and 0.9, respectively.

Table 2 Illustration of different feature views in image and text modalities

ID	Modality	Feature view	Dimensions
1	Image	Block-wise color moments (5×5 partition)	255
2		Color histogram in HSV space	64
3		Color histogram in RGB space	256
4		Wavelet texture features	128
5		Color autocorrelogram in HSV space	144
6		Edge direction histogram	75
7		Face features	7
8	Text	Term frequency	1000
9		Tf-Idf weight feature	1000
10		Topic-level feature (20 extracted LDA topics)	20

For each method mentioned above, the respective parameters are carefully tuned, and the parameters with the best performance are used to report the final comparison results.

The comparison performance is shown in Fig. 3a with respect to different depth of nDCG metrics. We have the following observations: (1) our proposed method consistently outperforms other baselines. In particular, the average improvements at depth 10 are 7.22, 6.02, and 2.54 %, as compared to **GISE**, **SGRW** and **sMM**, respectively. The consistent superiority demonstrates the advantages of our approach since it has modulated the effects of different modalities and views adaptively; (2) **SGRW** method only achieves a slight improvement over the results returned by the search engine. This clearly shows that simply concatenating all features with simple graph-based reranking approach cannot significantly boost the ranking results. Indeed, for several queries, it degrades the quality of the original ranking results, especially for complex queries. This phenomenon is also reported in some previous reranking literatures [15,22,34]; and (3) **sMM** approach is the second best approach since it considers the effects of different extracted features in a similar way. However, it only uses the pairwise information and does not consider the discriminative capabilities of different modalities, which are essential in multi-modal information retrieval.

The possible reasons that our proposed **aMM** method outperforms the other two reranking baselines are as follows: (1) due to the intrinsic limitation of the simple graph, learning on simple graph naturally cannot leverage higher-order relations among documents, which can only capture pairwise relations

[3] http://nlp.stanford.edu/software/tmt/.

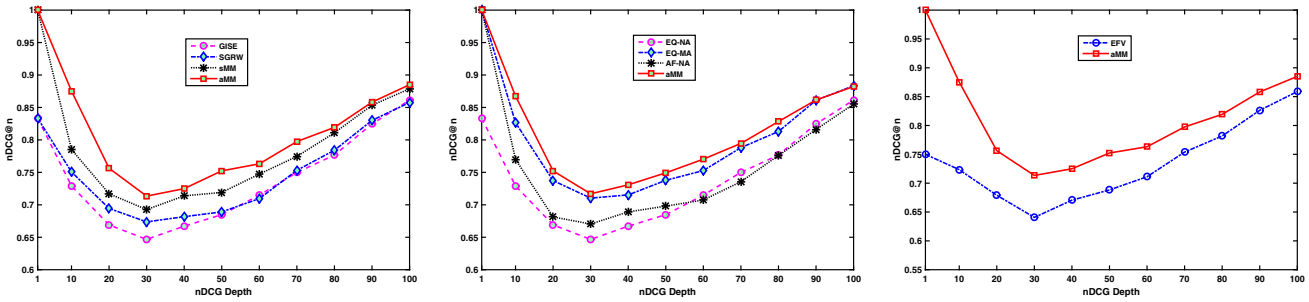

Fig. 3 These three subfigures respectively show the results of: **a** overall model comparison, **b** component-wise analysis, and **c** adaptive fusion of distinct feature views in terms of nDCG

between documents [16,19,44,49]; (2) the baselines consider all modalities and views to have equal strength in representing the content of multi-modal documents, while in fact, they have query-specific contribution confidences [37,44]; (3) simply merging all the feature vectors suffers from the curse of dimensionality [17]; and (4) they do not capture and model the relatedness among modalities.

We also conducted the analysis of variance (known as ANOVA). In particular, we performed a paired t test between our proposed **aMM** model and each of the benchmarks based on various nDCG depth. We found that all the p values are much smaller than 0.05, which demonstrates that the improvement of our model is statistically significant.

4.3 On component-wise analysis

We are now interested to find out the effectiveness of different components in our proposed model. In particular, we consider the bi-level evaluation of: (1) modality level; and (2) feature view level.

4.3.1 On multi-modality analysis

From the perspective of modality level, we conducted experiments with the following settings. It is worth noting that all of the settings below considered the adaptive fusion of feature views.

- *EQ-NA* We neither considered the adaptive weights of various modalities, nor the modality agreement. In fact, this is a special case of our model when all λ are equal and β is equal to zero.
- *AF-NA* We took the adaptive fusion of modalities into consideration, but did not consider the modality agreement. This was achieved by setting β to be zero.
- *EQ-MA* We considered modality agreement, but we did not differentiate the weights of modalities. We implemented this by setting all λ^i to be equal.
- *aMM* This is our proposed model, which considered all the components.

The performance of aforementioned settings is illustrated in Fig. 3b. From this Figure, we can observe the following points: (1) approaches with adaptive fusion of distinct modalities perform consistently better than those treating all modalities with equal weight. In particular, **aMM** consistently outperforms its counterpart **EQ-MA**, and **AF-NA** often outperforms **EQ-NA** approach too. The reason is probably that they can assign higher weights to more discriminative modalities based on the given query. This result is consistent with several past research efforts in multi-feature fusion [37,43]. Moreover, it is worth noting that **EQ-NA**, indeed, outperforms **AF-NA** at some depths. However, this happens at higher depths, i.e. higher that 60, which are less important that top-ranked results since many users might not check these depths. (2) **EQ-MA** and **aMM** significantly improve the performance as compared to **EQ-NA** and **AF-NA** respectively. The results reveal that modality agreement is more important in multi-modal information reranking than adaptive weights learning. It verifies our initial expectation that modality agreement is an important indicator for reranking process. It is worth noting that perviously proposed models, like **sMM** [37] (see Sect. 4.2), cannot incorporate the agreement of different modalities into the reranking framework.

4.3.2 On multiple feature view analysis

At the feature view level, we conducted experiments to comparatively validate the following experimental settings:

- *EFV* We treated different feature views during fusion equally. But we considered the modality agreement and adaptive fusion of modalities. It can be derived that this is a special case of our method when the α_j^i are equal for all views.

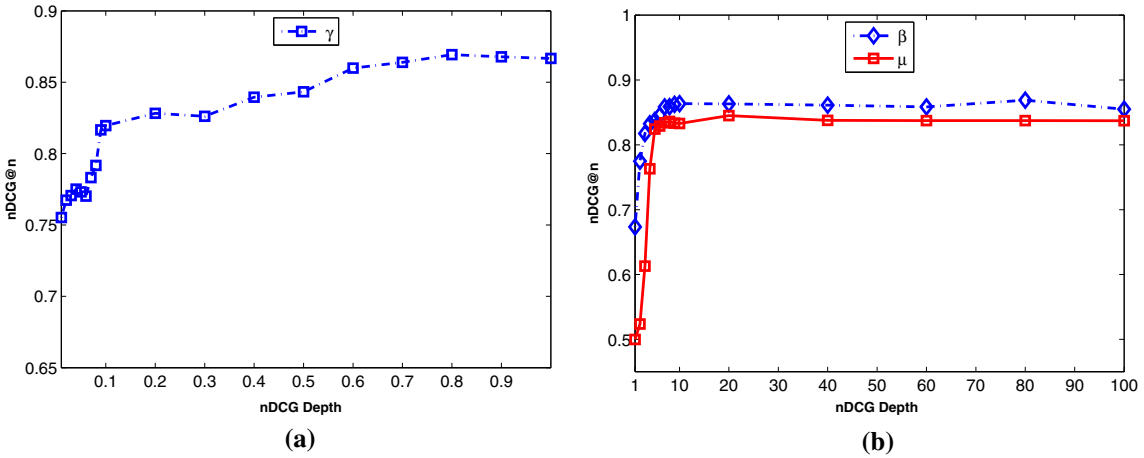

Fig. 4 The studies of parameter sensitivities. Each of the three key parameters was tuned with the other two fixed

– *aMM* This is our proposed model, which considered all the components.

Figure 3c shows the comparison results of the above two methods. From this Figure, we have two observations: (1) our proposed **aMM** model, which adaptively fuses the feature views, is significantly and stably better than **EFV**. Recent research efforts on multi-view retrieval and reranking [9,40] well support this experimental results. It shows that distinct feature views can contribute differently in learning approach. (2) The performance gap is widened as the nDCG depth becomes smaller. This implies that in the multi-modal document retrieval, it is useful to keep the top documents only.

4.4 On the parameter tuning

We also studied the parameter sensitivity of our proposed model. Our model holds two sets of parameters: (1) query-specific parameters, i.e. λ^i and α^i_j, are adaptively learned by the optimization procedure; and (2) regularization parameters include β, γ, and μ. These parameters respectively control the effects of modality agreement, initial admissibility, and sparsity of the model. In the past experiments, we have demonstrated the importance of the adaptive parameters. It is also highly desireable to evaluate the sensitivity of our method with respect to the later set of parameters.

We first fixed β and μ, and varied γ from 0.01 to 1 with flexible step size. We selected smaller step size, 0.01, between 0.01 and 0.1 to analyze the effect of very small value of γ. We then selected a larger step size of 0.1 for the rest of the range. Figure 4a demonstrates the sensitivity curve of **aMM** approach with respect to various γ values. Following that, we fixed γ and β and varied μ from 0.01 to 100 with flexible step size. The red curve in Fig. 4b shows the stability of our model with respect to the various μ values. Finally, we fixed γ and μ, and changed the modality agreement parameter to evaluate its effects on the performance of the approach. We selected different values in the range of 0.01 to 100. Figure 4b shows the performance curve with respect to the various β values. From the results, it can be seen that the performance of our model will not significantly degrade when the two parameters β and μ changes after 10. Comparatively, the model is more sensitive to the parameter γ which modulates the effects of initial reranking but it still varies in a fairy narrow range and it is thus acceptable.

4.5 Evaluation of public datasets

We are also interested to find out the robustness and applicability of our model on different multimodal information retrieval tasks. In particular, we consider two well-known multimedia tasks: (1) web Image reranking on the MSRA-MM Version 2.0 dataset [20]; and (2) KIS on TRECVid 2012.[4]

We selected these two datasets to further verify our approach due to the following reasons: (1) the queries in MSRA-MM 2.0 are relatively short, while the queries for KIS are very verbose, hence we can validate the effectiveness of our model on simple and complex queries, respectively; (2) they are publicly accessible and widely used with well-labelled ground-truth; and (3) they respectively comprise of multimodal documents in terms of image + text and video + text, which are the dominant media form to date.

We first evaluated our model on the MSRA-MM 2.0 dataset [20]. It contains 1, 165 frequent queries of a commercial search engine. This dataset is manually categorized into nine categories containing of 1,011,738 images with their surrounding texts. The same set of features summarized in Table 2 were extracted for each multimodal document. Table 3 comparatively summarizes the reranking perfor-

[4] http://www-nlpir.nist.gov/projects/tv2012/tv2012.html.

mance of various models over MSRA-MM 2.0 in terms of nDCG@100. It can be observed that our proposed model is superior to the baselines. This reflects that our model is robust for the multimodal reranking task with simple queries. Figure 5 visually illustrates the reranking results of different models. We can see that the reranking results of our proposed model are more visually consistent and coherent as compared to those obtained by baselines.

We also applied our proposed model on the KIS task. KIS is a specialized task of the general multimedia search problem where information seekers have already seen a video before and they want to find that specific video again based on their memory recall. The text queries are hence very long to depict their pieces of memories. The KIS task takes only a text query and returns a ranked list of videos which are most likely to match the known item. Table 4 illustrates the examples of the KIS task. Each video and its ASR naturally form a multimodal document. We conducted experiments on the repository of TRECVid 2012. This repository approximately includes eight thousands videos in MPEG-4/H.264 format. In total, it contains about 200 h of videos with duration between 10 s to 3.5 min. Inspired by [6,46], we employed term frequency and inverse document frequency weighting scheme (tf-idf) to compute the initial ranking list for each given query, which has been shown as an effective approach for KIS task. Based on the initial ranking list, we implemented our rerank-

Table 3 Performance comparison of various reranking models over MMRA 2.0 dataset in terms of average nDCG@100

Categories	GISE	SGRW	sMM	aMM
Animal	0.734	0.724	0.791	0.797
Cartoon	0.807	0.819	0.865	0.879
Event	0.788	0.779	0.811	0.836
Object	0.703	0.708	0.741	0.745
People	0.714	0.716	0.742	0.746
Person	0.908	0.939	0.940	0.955
Scene	0.703	0.712	0.792	0.794
Time08	0.830	0.830	0.870	0.879
Misc	0.736	0.757	0.790	0.824
Overall	0.769	0.776	0.816	0.828

Table 4 Query examples in KIS and their ground truth video keyframe from TRECVid 2012 KIS

Topic	Query example	Keyframe example
900	Find the video with people celebrating in a street yelling and running	
1152	Find the video showing a turtle on a log with trees in background	

Table 5 Different feature views extracted from text and video modalities for KIS task

ID	Modality	Feature view	Dimensions
1	Text	Tf-Idf extracted from ASR	1000
2		Topic distribution (50-D LDA topic)	50
3		Bag-of-Words	1000
4		Count of Named Entities in ASR	3
5	Video	Color histogram in HSV space	64
6		Edge direction histogram	75
7		Semantic concepts in the video key frames	1000

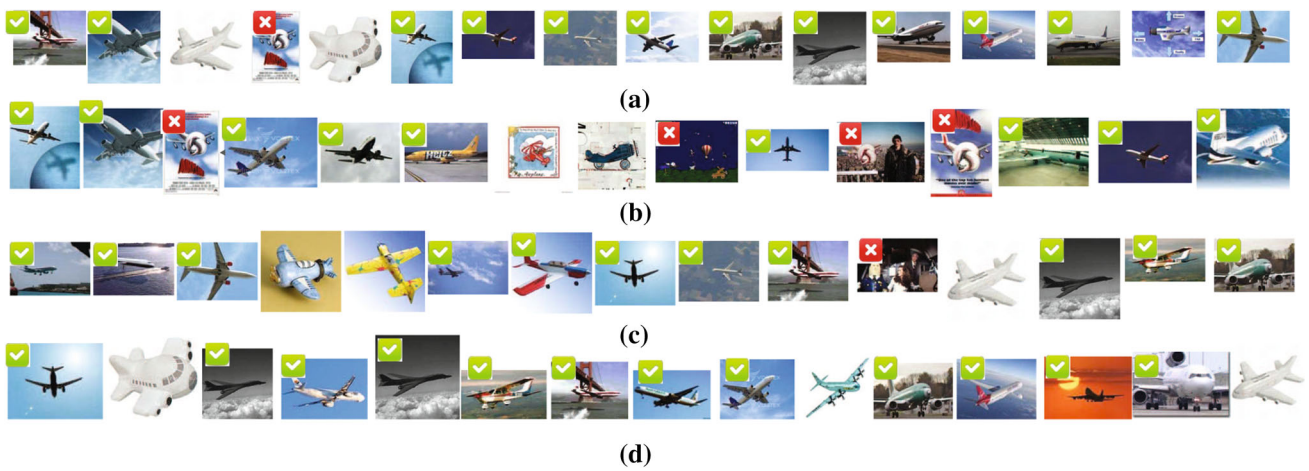

Fig. 5 Illustration of reranking results for an example query "*airplane*": **a** the original ranking list; **b SGRW**; **c sMM**; and **d** our proposed **aMM** model. The results are from position 6 to 20. Due to the limited space, we do not show the top 5 images since they are all relevant. Images with *tick symbols* are very relevant; images with *cross symbols* are irrelevant; while those *without marks* are something between

Table 6 Performance comparison of reranking models over KIS task in terms of MRR

Initial search results	SGRW	sMM	aMM
0.112	0.090	0.115	0.128

ing model and its competitors. The feature views extracted from ASR and video modalities are summarized in Table 5.

Table 6 illustrates the comparison performance in terms of mean reciprocal rank (MRR) [24]. Higher MRR values indicate better performance. From Table 6, we observed that: (1) **SGRW** makes matters worse as compared to the initial search results. The reason might be that KIS items are complex and vary in visual semantics. Therefore, simply aggregating feature vectors of different modalities would lead to a meaningless feature space. (2) The improvement in reranking performance by **sMM** is not significant. This tells us that the relations among multi-modal documents are very complex, especially for those complex queries. (3) Our proposed model significantly outperforms others. This reveals that semantics between ASR and videos are consistent, and have query-specific discriminative capabilities.

4.6 Summary

To sum up, the experiments support four main findings: (1) modality agreement is of vital importance and is more effective than the adaptive fusion in the multi-modal reranking (Sect. 4.3.1); (2) discriminative capabilities of modalities and feature views are query specific and adaptive fusion of them remarkably improves the performance (Sects. 4.3.1 and 4.3.2); (3) higher-order relations among multi-modal documents are much more informative as compared to pairwise relations (Sect. 4.2); (4) our proposed model is insensitive to the involved parameters and robust to many multi-modal reranking tasks (Sects. 4.4 and 4.5).

5 Conclusions and future work

This paper presented a hypergraph-based adaptive reranking model to boost the search performance of multi-modal documents. Besides the smoothness and initial admissibility concerned by the traditional reranking approaches, it seamlessly unifies modality agreement, adaptive fusion of feature views from distinct modalities, as well as higher-order relations among multi-modal documents. Extensive experiments on our constructed dataset well verified our proposed model and each of its components. In addition, experiments on two other widely used public datasets also demonstrated its robustness.

In the current work, we only consider relevance of multi-modal documents. On the other hand, diversity is an important aspect of multi-modal document retrieval. In the future, we plan to extend our model to jointly consider relevance and diversity.

References

1. Bolla M (1993) Spectra, euclidean representations and clusterings of hypergraphs. Discrete Math 117(1):19–39
2. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh World Wide Web Conference
3. Cai J, Zha ZJ, Wang M, Zhang S, Tian Q (2015) An attribute-assisted reranking model for web image search. Image Process IEEE Trans 24(1):261–272
4. Deng C, Ji R, Tao D, Gao X, Li X (2014) Weakly supervised multi-graph learning for robust image reranking. Multimed IEEE Trans 16(3):785–795
5. Dollár P, Tu Z, Tao H, Belongie S (2007) Feature mining for image classification. In: Computer Vision and Pattern Recognition, 2007. IEEE Conference on, pp 1–8
6. Etter D, Domeniconi C (2014) Semi-supervised rank learning for multimedia known-item search. In: Proceedings of International Conference on Multimedia Retrieval, p 257
7. Faria FF, Veloso A, Almeida HM, Valle E, Torres RdS, Gonçalves MA, Meira Jr W (2010) Learning to rank for content-based image retrieval. In: Proceedings of the international conference on Multimedia information retrieval, pp 285–294
8. Farseev A, Nie L, Akbari M, Chua TS (2015) Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp 235–242
9. Fu Y, Hospedales TM, Xiang T, Fu Z, Gong S (2014) Transductive multi-view embedding for zero-shot recognition and annotation. In: Computer Vision–ECCV, pp 584–599
10. Gao Y, Wang M, Tao D, Ji R, Dai Q (2012) 3-d object retrieval and recognition with hypergraph analysis. Image Process IEEE Trans 21(9):4290–4303
11. Gao Y, Wang M, Zha ZJ, Shen J, Li X, Wu X (2013) Visual-textual joint relevance learning for tag-based social image search. Image Process IEEE Trans 22(1):363–376
12. Gehler P, Nowozin S (2009) On feature combination for multiclass object classification. In: Computer Vision, IEEE 12th International Conference on, pp 221–228
13. He J, Li M, Zhang HJ, Tong H, Zhang C (2004) Manifold-ranking based image retrieval. In: Proceedings of the 12th annual ACM international conference on Multimedia, pp 9–16
14. He X, Ma WY, Zhang HJ (2004) Learning an image manifold for retrieval. In: Proceedings of the 12th annual ACM international conference on Multimedia, pp 17–23
15. Hsu WH, Kennedy LS, Chang SF (2007) Video search reranking through random walk over document-level context graph. In: Proceedings of the 15th international conference on Multimedia, pp 971–980
16. Huang Y, Liu Q, Zhang S, Metaxas DN (2010) Image retrieval via probabilistic hypergraph ranking. In: Computer Vision and Pattern Recognition, IEEE Conference on, pp 3376–3383
17. Indyk P, Motwani R (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp 604–613
18. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Pro-

ceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp 119–126
19. Jin T, Yu J, You J, Zeng K, Li C, Yu Z (2015) Low-rank matrix factorization with multiple hypergraph regularizer. Pattern Recognit 48(3):1011–1022
20. Li H, Wang M, Hua XS (2009) Msra-mm 2.0: A large-scale web multimedia dataset. In: Data Mining Workshops, IEEE International Conference on, pp 164–169
21. Liu J, Lai W, Hua XS, Huang Y, Li S (2007) Video search reranking via multi-graph propagation. In: Proceedings of the 15th international conference on Multimedia, pp 208–217
22. Liu Y, Mei T, Hua XS (2009) Crowdreranking: exploring multiple search engines for visual search reranking. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp 500–507
23. Marchenko Y, Chua TS, Jain R (2006) Semi-supervised annotation of brushwork in paintings domain using serial combinations of multiple experts. In: Proceedings of the ACM International Conference on Multimedia, pp 529–538
24. McFee B, Lanckriet GR (2010) Metric learning to rank. In: Proceedings of the 27th International Conference on Machine Learning, pp 775–782
25. Mei T, Rui Y, Li S, Tian Q (2014) Multimedia search reranking: a literature survey. ACM Comput Surv 46(3):38
26. Nie L, Akbari M, Li T, Chua TS (2014) A joint local-global approach for medical terminology assignment. In: Medical Information Retrieval Workshop at SIGIR, p 24
27. Nie L, Yan S, Wang M, Hong R, Chua TS (2012) Harvesting visual concepts for image search with complex queries. In: Proceedings of the 20th ACM international conference on Multimedia, pp 59–68
28. Nie L, Zhao YL, Akbari M, Shen J, Chua TS (2015) Bridging the vocabulary gap between health seekers and healthcare knowledge. Knowl Data Eng IEEE Trans 27(2):396–409
29. Nie L, Zhao YL, Wang X, Shen J, Chua TS (2014) Learning to recommend descriptive tags for questions in social forums. ACM Trans Inf Syst 32(1):5
30. Pu HT (2008) An analysis of failed queries for web image retrieval. J Inf Sci 34(3):275–289
31. Qiu S, Wang X, Tang X (2013) Anchor concept graph distance for web image re-ranking. In: Proceedings of the 21st ACM international conference on Multimedia, pp 713–716
32. Snoek CG, Worring M, Smeulders AW (2005) Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia, pp 399–402
33. Song X, Nie L, Zhang L, Akbari M, Chua TS (2015) Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 213–222
34. Tian X, Yang L, Wang J, Yang Y, Wu X, Hua XS (2008) Bayesian video search reranking. In: Proceedings of the 16th ACM international conference on Multimedia, pp 131–140
35. Wang C, Zhang L, Zhang HJ (2008) Learning to reduce the semantic gap in web image retrieval and annotation. In: Proceedings of the 31st annual international ACM conference on Research and development in information retrieval, pp 355–362
36. Wang L, Yang L, Tian X (2009) Query aware visual similarity propagation for image search reranking. In: Proceedings of the 17th ACM international conference on Multimedia, pp 725–728
37. Wang M, Li H, Tao D, Lu K, Wu X (2012) Multimodal graph-based reranking for web image search. Image Process IEEE Trans 21(11):4649–4661
38. Wang X, Qiu S, Liu K, Tang X (2014) Web image re-ranking usingquery-specific semantic signatures. Pattern Anal Mach Intell IEEE Trans 36(4):810–823
39. Wang Y, Wu F, Song J, Li X, Zhuang Y (2014) Multi-modal mutual topic reinforce modeling for cross-media retrieval. In: Proceedings of the ACM International Conference on Multimedia, pp 307–316
40. Wu J, Hong Z, Pan S, Zhu X, Cai Z, Zhang C (2014) Exploring features for complicated objects: Cross-view feature selection for multi-instance learning. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp 1699–1708
41. Xu B, Bu J, Chen C, Cai D, He X, Liu W, Luo J (2011) Efficient manifold ranking for image retrieval. In: Proceedings of the 34th international ACM conference on Research and development in Information Retrieval, pp 525–534
42. Ye G, Liu D, Jhuo IH, Chang SF (2012) Robust late fusion with rank minimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition
43. Yu J, Rui Y, Chen B (2014) Exploiting click constraints and multi-view features for image re-ranking. Multimed IEEE Trans 16(1):159–168
44. Yu J, Tao D, Wang M (2012) Adaptive hypergraph learning and its application in image classification. Image Process IEEE Trans 21(7):3262–3272
45. Yu J, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. Cybern IEEE Trans 45(4):767–779
46. Yuan J, Zhao YL, Luan H, Wang M, Chua TS (2014) Memory recall based video search: finding videos you have seen before based on your memory. ACM Trans Multimed Comput Commun Appl 10(2):21
47. Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B (2004) Learning with local and global consistency. Adv Neural Inf Process Syst 16(16):321–328
48. Zhou D, Huang J, Schölkopf B (2005) Learning from labeled and unlabeled data on a directed graph. In: Proceedings of the 22nd international conference on Machine learning, pp 1036–1043
49. Zhou D, Huang J, Schölkopf B (2006) Learning with hypergraphs: Clustering, classification, and embedding. In: Advances in neural information processing systems, pp 1601–1608
50. Zien JY, Schlag MD, Chan PK (1999) Multilevel spectral hypergraph partitioning with arbitrary vertex sizes. Comput-Aided Des Integr Circuits Syst, IEEE Trans 18(9):1389–1399