A Graph-based Video Visual Reranking Method via Heterogenous Graph Analysis

Soh Yoshida* and Mitsuji Muneyasu*
*Faculty of Engineering Science, Kansai University
Email:*{sohy, muneyasu}@kansai-u.ac.jp,

***Abstract*—This paper addresses the problem of analyzing topics, included in social videos, for improving the performance of video retrieval. Unlike previous works which only focus on an individual video visual aspect, the proposed method leverages the "mutual reinforcement" of heterogeneous objects such as text tags and users. In order to represent multiple types of relationships between each heterogeneous object, the proposed method constructs three subgraphs: a user-tag graph, a video-video graph, and a video-tag graph. We combine the three types of graphs to obtain the heterogeneous graph. Then the extraction of latent features, i.e., topics, becomes feasible by applying graph-based soft clustering to the heterogeneous graph. By estimating the membership of each grouped cluster for each video, the proposed method defines a new video similarity measure. Since the understanding of video content is enhanced by taking advantage of latent features obtained from different types of data, that complement each other, the proposed method can improve the performance of video visual reranking. We conduct experiments on the YouTube-8M dataset, and the results show that our reranking approach is effective and efficient.**

I. Introduction

Current multimedia search technology is mainly relying on employing text annotations to provide users with accurate results for their queries. However, text-based search alone is not enough, due to the well-known semantic gap between textual description and video content. To overcome the semantic gap, visual search reranking, which adjusts the initial ranking order by mining visual patterns or leveraging some auxiliary knowledge, has been proposed [1]–[7]. According to their reranking objectives, the existing visual search reranking efforts can be mainly classified into two categories, i.e., relevance-based reranking [1]–[3], [5], [7] and diversified-based reranking [6].

Relevance-based methods maximize the relevance of the returned list through reordering. Recent research in visual reranking has focused on improving the relevance of the results. Since maximizing the relevance of each item in the list is only visually objective, the resulting ranking list tends to return the near duplicate videos that convey repetitive information. However, an efficient video retrieval system should be able to give a global view so that it surfaces results that are both relevant and that are covering different aspects of a query, e.g., providing different views of a object or scene rather than duplicates of the same perspective showing almost identical videos. Therefore, diversified reranking is proposed to allow the search results to convey more information by considering the topics of videos.

The existing relevance and diversified-based reranking methods capture relationships between videos by only using visual information [1]–[3], [5]–[7] or other pre-trained semantic models [4]. On the other hand, because of the various social activities, a video contains rich social media information such as tags and users, etc., which can provide meaningful contextual cues to understand its content. However, since the relations between videos and other information are not fully leveraged, the reinforcing dependence between them, which is beneficial for further improving reranking results, has not been considered. To our best knowledge, no previous works leverage the reinforcement of heterogeneous objects collaboratively to learn topics included in a video ranking for better video search reranking.

To overcome the above problem, this paper proposes a heterogeneous graph-based video search reranking (HGVR) method using topic relevance. The proposed method attempts to extract topics from a video group whose videos contain various social information. Specifically, the proposed method constructs three subgraphs: a user-video graph, a video-video graph, and a video-tag graph. We combine the three types of graphs and call it as the heterogeneous graph. We model each type of relationship as a cost function based on the concept of locality preservation. Then the extraction of topics becomes feasible by applying graph-based soft clustering [8] to the combined graph and grouping different types of data as clusters. By estimating the topic membership for each video, the proposed method defines a video similarity measure. Finally, we formulate the video search reranking as an optimization problem.

The main contributions of this work are summarized as follows:

1) We propose the method named heterogeneous graph-based video search reranking (HGVR) for improving the performance of video visual reranking.
2) We collaboratively fuse the video visual information, their associated tags, and relative users to leverage the mutual reinforcement between each heterogeneous object through the heterogeneous graph construciton.
3) An effective feature extraction method for combining three types of objects is realized by using a heterogeneous graph-based soft clustering approach. This approach is able to enhance the quality of video content understanding by taking advantage of different types of data, which complement each other.

II. Heterogeneous Graph Construction

We use a heterogeneous graph, which consists of multiple types of objects and multiple types of relationships, to preserve different kinds of information from different data sources. The basic idea used to construct the graph is that two objects are linked with a stronger relation if they are more likely to share a similar content. Let $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$, $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$, and $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$ denote videos, tags, and users, respectively, where $\{|\mathcal{V}|, |\mathcal{T}|, |\mathcal{U}|\}$ are the number of each node.

There are three types of relations between these objects including the user-video relation extracted from information of uploader, the video-video relation based on video link structure, and the video-tag relation provided by associated tags, leading to three subgraphs: the user-video graph $H_{\mathcal{U},\mathcal{V}}$, the video-video graph $G_{\mathcal{V}}$, and the video-user graph $J_{\mathcal{V},\mathcal{T}}$. We denote the constructed heterogeneous graph as G.

A. User-video Graph

We assume that *videos uploaded by the same user tend to share the similar topics*. We represent a semantic relationship between users and videos by forming a bipartite user-video subgraph. Specifically, the proposed method defines the edge between video $v_j \in \mathcal{V}$ and its uploader $u_i \in \mathcal{U}$. We then define the affinity matrix $\mathbf{C} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{V}|}$ by taking the connection between u_i and v_j as its (i, j)-th element as follows:

$$C_{ij} = \frac{1}{N_{\mathcal{U}}(u_i)}, \tag{1}$$

where $N_{\mathcal{U}}(u_i)$ denotes the number of videos managed by a user u_i.

B. Video Graph

The proposed method constructs the video graph $G_{\mathcal{V}}$ by using metadata named "related videos" since this is useful for associating videos on the Web that are similar to each other. Most of popular video hosting service such as YouTube provide the metadata "related videos". Using existing metadata such as related videos, we are able to efficiently construct graph, whose videos are connected with similar videos. In this paper, we consider that a video v_i links to a video v_j if "related videos" of v_i include v_j. Finally, by calculating the video similarity between v_i and v_j as follows:

$$W_{ij} = \exp\left\{-\frac{||\mathbf{f}_i - \mathbf{f}_j||^2}{2\sigma_{\mathcal{V}}^2}\right\}, \tag{2}$$

where $\mathbf{f}_\bullet$ is a feature vector extracted from a video $v_\bullet$ and $\sigma_{\mathcal{V}}$ is the scaling parameter estimated as the median value of all the Euclidean distance. We define the affinity matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, where W_{ij} is equal to the video similarity between v_i and v_j.

C. Video-tag Graph

Tags, which are supplied by users, describe the content of videos while providing additional contextual modalities about the videos. If tags can be used appropriately, video contents understanding is reinforced rather than in the case of using only video features and topic analysis becomes feasible. First, we collect $|\mathcal{T}|$ tags $\mathcal{T}$, which are associated with the crawled videos $\mathcal{V}$. As a preprocessing step, the standard lemmatization algorithm and stop words removal are applied. Then, we remove noisy tags based on frequency of tag appearance in the tag set $\mathcal{T}$. Second, we form a bipartite video-tag subgraph by connecting videos and tags with edges. Specifically, we link video v_i and its associated tag t_j with an edge, and then the relationship between them is calculated by using the Vote+ algorithm [9] as follows:

$$score(v_i, t_j) = tagRelevance(v_i, t_j, k) \cdot \\ \frac{k_d}{k_d + |k_d - \log(freq_{t_j})|} \cdot \frac{k_r}{k_r + (rank_{t_j} - 1)} \tag{3}$$

where k_d, k_r are the parameters and $freq_{t_j}$ is the frequency of tag t_j in the collection. $tagRelevance(v_i, t_j, k)$ is calculated by using the neighbor voting scheme. $tagRelevance(v_i, t_j, k)$ estimates tag relevance, whether associated tag t_j is appropriate for the video v_i or not, by counting neighbor votes on tags. The detailed of this voting algorithm in our method is shown in [9]. Then we define the edge strength $\mathbf{P} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{T}|}$, where P_{ij} is equal to the tag relevance of each video. We calculate P_{ij} as following equation:

$$P_{ij} = \frac{score(v_i, tj) - score_{min}}{score_{max} - score_{min}}, \tag{4}$$

where $score_{max}$, $score_{min}$ are the maximum and minimum "$score$", respectively.

III. Heterogeneous Graph-based Visual Reranking

In the proposed method, we introduce a Bayesian visual reranking framework to implement video retrieval based on our heterogeneous graph. This framework can model textual and visual information from a probabilistic perspective and formulate visual reranking as an optimization problem in the Bayesian framework. First, the textual information is modeled as a likelihood to reflect the disagreement between reranked results and text-based search results, which is called the ranking distance. Second, the visual information is modeled as a conditional prior to indicate the reranking score consistency among similar videos.

Suppose there are c different latent class associated with all three types of objects. We first define three indicator matrices $\mathbf{F}_{\mathcal{U}} \in [0, 1]^{|\mathcal{U}| \times c}$, $\mathbf{F}_{\mathcal{V}} \in [0, 1]^{|\mathcal{V}| \times c}$, and $\mathbf{F}_{\mathcal{T}} \in [0, 1]^{|\mathcal{T}| \times c}$, which describe the confidence of users, videos, and tags belonging to different search intents, respectively. In the proposed method, given a constructed heterogeneous graph G, its adjacency matrices $\mathbf{C}$, $\mathbf{W}$, and $\mathbf{P}$ learn $\mathbf{F}_{\mathcal{U}}$, $\mathbf{F}_{\mathcal{V}}$, and $\mathbf{F}_{\mathcal{T}}$ as soft-clustering indicators for all three types of objects simultaneously.

A. Heterogeneous graph-based soft clustering for learning latent features

In this section, a heterogeneous graph-based soft-clustering optimization problem is derived for unified latent features learning. Mathematically, we model each type of relationship as a cost function based on the concept of locality preservation, which requires two nearby objects in G to have similar indicators, and further derive the following optimization problem by using weighted summation of these cost functions as the objective function and imposing soft-clustering constraints on the indicators:

$$\begin{aligned}\min_{\mathbf{F}_{\mathcal{U}},\mathbf{F}_{\mathcal{V}},\mathbf{F}_{\mathcal{T}}} & \mathcal{L}(\mathbf{F}_{\mathcal{U}},\mathbf{F}_{\mathcal{V}},\mathbf{F}_{\mathcal{T}}) = \\ & \lambda_{\mathcal{U}}\sum_{i=1}^{|\mathcal{U}|}\sum_{j=1}^{|\mathcal{V}|} C_{ij}||\mathbf{F}_{\mathcal{U}i}-\mathbf{F}_{\mathcal{V}j}||_2^2+\lambda_{\mathcal{V}}\sum_{i,j=1}^{|\mathcal{V}|} W_{ij}||\mathbf{F}_{\mathcal{V}i}-\mathbf{F}_{\mathcal{V}j}||_2^2 \\ & + \lambda_{\mathcal{T}}\sum_{i=1}^{|\mathcal{V}|}\sum_{j=1}^{|\mathcal{T}|} P_{ij}||\mathbf{F}_{\mathcal{V}i}-\mathbf{F}_{\mathcal{T}j}||_2^2, \end{aligned} \tag{5}$$

where $||\cdot||_2$ denotes the L_2 norm, and $0 \leq \lambda_{\mathcal{U}\mathcal{V}}, \lambda_{\mathcal{V}}, \lambda_{\mathcal{T}} \leq 1$ are tuning parameters, which control the trade-off between the three types of relationships. With the definition of the augmented indicator $\mathbf{F} = [\mathbf{F}_{\mathcal{U}}^T, \mathbf{F}_{\mathcal{V}}^T, \mathbf{F}_{\mathcal{T}}^T]^T \in [0,1]^{(|\mathcal{U}|+|\mathcal{V}|+|\mathcal{T}|)\times t}$, Eq. (5) can be further rewritten in the following concise form:

$$\min_{\mathbf{F}} \mathcal{L}(\mathbf{F}) = \mathrm{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}). \tag{6}$$

Here, $\mathrm{Tr}(\cdot)$ denotes the trace and $\mathbf{L}$ is the following global graph Laplacian matrix:

$$\mathbf{L} = \begin{bmatrix} \lambda_{\mathcal{U}}\mathbf{D}^{(\mathcal{U})} & -\lambda_{\mathcal{U}}\mathbf{C} & 0 \\ -\lambda_{\mathcal{U}}\mathbf{C}^T & \lambda_{\mathcal{U}}\mathbf{D}^{(\mathcal{U}\mathcal{V})}+2\lambda_{\mathcal{V}}\mathbf{L}^{(\mathcal{V})}+2\lambda_{\mathcal{T}}\mathbf{D}^{(\mathcal{V}\mathcal{T})} & -\lambda_T\mathbf{P} \\ 0 & -\lambda_{\mathcal{T}}\mathbf{P} & \lambda_{\mathcal{T}}\mathbf{D}^{(\mathcal{T})} \end{bmatrix}, \tag{7}$$

where $\mathbf{D}^{(\mathcal{U})}$, $\mathbf{D}^{(\mathcal{U}\mathcal{V})}$, $\mathbf{D}^{(\mathcal{V}\mathcal{T})}$, $\mathbf{D}^{(\mathcal{T})}$ are the degree matrices which are defined as $D_{ii}^{(\mathcal{U})} = \sum_{j=1}^{|\mathcal{V}|} C_{ij}$, $D_{jj}^{(\mathcal{U}\mathcal{V})} = \sum_{i=1}^{|\mathcal{V}|} C_{ij}$, $D_{ii}^{(\mathcal{V}\mathcal{T})} = \sum_{j=1}^{|\mathcal{T}|} P_{ij}$, $D_{jj}^{(\mathcal{T})} = \sum_{i=1}^{|\mathcal{V}|} P_{ij}$, and $\mathbf{L}^{(\mathcal{V})}$ and $\mathbf{L}^{(\mathcal{T})}$ are the graph Laplacian matrices of $\mathbf{W}$ and $\mathbf{E}$, respectively.

Next, to solve the soft-clustering problem in Eq. (7), we adopt an efficient algorithm to approximately solve Eq. (7) that first embeds each object into a c-dimensional latent feature space, and then clusters the objects on the basis of the embedding latent features. First, by relaxing $\mathbf{F}$ to $\mathbf{U} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{V}|+|\mathcal{T}|)\times c}$ and imposing a constraint on $\mathbf{U}$, we can learn an optimal graph embedding $\mathbf{U}$ that is similar to a Laplacian eigenmaps [10]. This optimal graph encodes all types of relationships into a c-dimensional latent feature space as follows:

$$\min_{\mathbf{U}} \mathrm{Tr}(\mathbf{U}^T\mathbf{L}\mathbf{U}) \quad \text{s.t} \quad \mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{I}_c, \tag{8}$$

where $\mathbf{I}_c$ is the identity matrix and $\mathbf{M}$ is the global degree matrix defined as:

$$\mathbf{M} = \begin{bmatrix} \lambda_{\mathcal{U}}\mathbf{D}^{(\mathcal{U})} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda_{\mathcal{U}}\mathbf{D}^{(\mathcal{U}\mathcal{V})}+2\lambda_{\mathcal{V}}\mathbf{D}^{(\mathcal{V})}+\lambda_{\mathcal{T}}\mathbf{D}^{(\mathcal{V}\mathcal{T})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \lambda_{\mathcal{T}}\mathbf{D}^{(\mathcal{T})} \end{bmatrix}, \tag{9}$$

where $\mathbf{D}^{(\mathcal{V})}$ is the degree matrix defined as $D_{ii}^{(\mathcal{V})} = \sum_{j=1}^{|\mathcal{V}|} W_{ij}$ The optimal latent features $\tilde{\mathbf{U}}$ in Eq. (5) can be computed from c-generalized eigenvectors corresponding to the c-smallest eigenvalues of the generalized eigenvalue problem $\mathbf{LU} = \lambda\mathbf{MU}$.

Various soft-clustering methods can be adopted to optimize $\mathbf{U}_{\mathcal{U}}^T, \mathbf{U}_{\mathcal{V}}^T$, and $\mathbf{U}_{\mathcal{T}}^T$ simultaneously. In this work, we adopt the widely known fuzzy c-means algorithm [11], which optimizes the latent features as follows:

$$\{\tilde{\mathbf{O}}, \tilde{\mathbf{\Theta}}\} = \arg\min_{\mathbf{O},\mathbf{\Theta}} = \sum_{i=1}^{|\mathcal{U}|+|\mathcal{V}|+|\mathcal{T}|}\sum_{j=1}^{k} O_{ij}^2||\tilde{\mathbf{\Pi}}_i - \tilde{\mathbf{\Theta}}_j||_2^2, \tag{10}$$

where $\tilde{\mathbf{\Pi}} = [\tilde{\mathbf{U}}_{\mathcal{U}}^T, \tilde{\mathbf{U}}_{\mathcal{V}}^T, \tilde{\mathbf{U}}_{\mathcal{T}}^T]^T$ is comprised the augmented latent features for all objects and $\tilde{\Theta}_j$ is the center of j-th cluster. Details of the fuzzy c-means algorithm are given in [11]. Finally, we obtain the indicators for users, videos, and tags as $\tilde{\mathbf{O}} = [\tilde{\mathbf{F}}_{\mathcal{U}}^T, \tilde{\mathbf{F}}_{\mathcal{V}}^T, \tilde{\mathbf{F}}_{\mathcal{T}}^T]^T \in [0,1]^{(|\mathcal{U}|+|\mathcal{V}|+|\mathcal{T}|)\times c}$.

B. Reranking

After latent features learning using heterogenous graph, we implement a reranking to search videos including the topic with regard to the target query. The proposed method follows a graph-based reranking approach [2] to rank relevant videos higher. Let $\bar{\mathbf{r}} = [\bar{r}_i, \bar{r}_2, \ldots, \bar{r}_{|\mathcal{V}|}]^T$ and $\mathbf{r} = [r_i, r_2, \ldots, r_{|\mathcal{V}|}]^T$ denote the vectors of the initial ranking scores and the reranking scores, which correspond to the video set $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$. $\bar{r}_i$ and r_i are the initial ranking scores, which are calculated from the ranking position by keyword search, and the relevance scores with regard to the user's query.

All top ranked videos should include a same topic. To conduct a search that considers topics, we first reconstruct a video graph whose edges are weighted by topic-based similarity. When constructing the graph, each video is connected with its K nearest neighbors [2]. We define an affinity matrix $\mathbf{Y} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ in which Y_{ij} indicates the topic similarity between v_i and v_j, as follows:

$$Y_{ij} = \exp\left\{-\frac{||\tilde{\mathbf{F}}_{\mathcal{V}i}, \tilde{\mathbf{F}}_{\mathcal{V}j}||^2}{2\sigma^2}\right\}, \tag{11}$$

where σ is is the scaling parameter estimated as the median value of all the Euclidean distances.

By using the affinity matrix $\mathbf{Y}$, we formulate the reranking problem as follows:

$$Q(\mathbf{r}) = \sum_{i=1}^{|\mathcal{V}|} \left\{ \frac{1}{2} \sum_{j=1}^{|\mathcal{V}|} Y_{ij} \left(\frac{r_i}{\sqrt{d_i}} - \frac{r_j}{\sqrt{d_j}} \right)^2 \right\} + \rho \frac{1}{2} \sum_{i,j \in S_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j} \right)^2, \qquad (12)$$

where the first term is the local consistency term, the second term is the loss term, and ρ is a tuning parameter that controls the effect of the consistency term. Here, d_i is the sum of the ith row of $\mathbf{Y}$ and $S_{\bar{\mathbf{r}}}$ is the set of pairs (i, j) for which the relevance scores of all the sample pairs $(\mathbf{x}_i, \mathbf{x}_j)$ satisfy $\bar{r}_i > \bar{r}_j$.

The optimal solution $\mathbf{r}^*$ is obtained by minimizing $Q(\mathbf{r})$ in Eq. (12) as

$$\begin{aligned} \mathbf{r}^* &= \arg\min_{\mathbf{r}} \mathbf{r}^T \mathbf{L}_n \mathbf{r} + \rho \frac{1}{2} \sum_{i,j \in S_{\bar{\mathbf{r}}}} \left(1 - \frac{r_i - r_j}{\bar{r}_i - \bar{r}_j} \right)^2, \\ &= \arg\min_{\mathbf{r}} \mathbf{r}^T \mathbf{L}_n \mathbf{r} + \rho(\mathbf{r}^T \mathbf{L}^{(A)} - 2\mathbf{A}\mathbf{e})\mathbf{r}, \qquad (13) \end{aligned}$$

where $\mathbf{L}_n = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ is the normalized graph Laplacian matrix, where $\mathbf{D}^{(C)}$ is a diagonal matrix whose (i, i)th element is the sum of the ith row of $\mathbf{Y}$, and $\mathbf{I}$ is the identity matrix. $\mathbf{L}^{(A)}$ is a graph Laplacian matrix defined over the graph G_A, which has the same structure as $G_{\mathcal{V}}$ with the weight between nodes v_i and v_j is equal to $|\alpha_{ij}|$, $\mathbf{A} = [\alpha_{ij}]_{|\mathcal{V}|\times|\mathcal{V}|}$ is an antisymmetric matrix with $\alpha_{ij} = 1/(\bar{r}_i - \bar{r}_j)$, and $\mathbf{e}$ is a vector with all elements equal to 1.

Finally, the optimal solution $\mathbf{r}^*$ is derived by differentiating w.r.t $\mathbf{r}$ and equating it to zero to obtain

$$\mathbf{r}^* = \frac{1}{2}(\mathbf{L}_n + \rho \mathbf{L}^{(A)})^{-1} \tilde{\boldsymbol{\rho}}, \qquad (14)$$

where $\tilde{\boldsymbol{\rho}} = 2\rho(\mathbf{A}\mathbf{e})$. The proposed method returns videos in accordance with the ranking score $\mathbf{r}^*$ as the video search result.

IV. Experimental Results

A. Experimental Setup

1) **Dataset:** Out experiments were conducted using the YouTube-8M dataset [12]. The complete YouTube-8M dataset consists of approximately 7 million YouTube videos, each approximately 2-5 minutes in length, with at least 1000 views. There are 24 categories and 4716 possible classes, named "entity", given in a multi-label form. From the entire dataset, we selected 35 entities. Then, we obtained the metadata such as "uploader", "tag", and "related videos" of each video. The entire dataset of our experiments was summarized in Fig. 1.

2) **Queryset:** We applied labels of entities to queries and also regard them as ground-truth for evaluating video retrieval. Query labels were automatically selected from each entity. The number of videos for each label was between 200 and 3000. The details of the queries and examples are shown in the Table I. First, we conducted keyword searches giving the selected labels, and then performed experiments to rerank the obtained rankings. Specifically, the initial ranking list $\bar{\mathbf{r}}$, which

TABLE I
List of entities and examples of queries used in the experiments.

Entity	# of queries	Query examples
Airline	8	cloud, jet engine, cockpit
Airport	10	airport terminal, microsoft flight simulator, motorsport
Skateboarding	6	skateboarding trick, fisheye lens, vehicle
Basketball moves	8	athlete, highlight film, arena
Skateboard	6	skateboarding trick, fisheye lens, kickflip
Tennis	3	game, table furniture, racket sports equipment
Lego	15	animation, lego star wars, lego city
Shoe	7	association football, running, basketball
Doll	7	princess, the walt disney company, dollhouse
Plant	12	agriculture, forest, leaf
Underwater	6	underwater diving, scuba diving, nature
Room	3	home improvement, building, furniture
Dog	3	terrier, cat, outdoor recreation
Horse	7	livestock, horse racing, race track
Bird	9	pet, poultry, wildlife
House	14	condominium, architecture, resort
Gardening	11	cooking, farm, animal
Garden	3	nature, food, vegetable
Radio-controlled model	14	unmanned aerial vehicle, toy, four-wheel drive
Wedding	4	bride, music video, wedding dress
Sports game	9	video game, ball association football, highlight film
Fighting game	7	the king of fighters, street fighter iv, super street fighter iv
Vegetable	14	roasting, cookware and bakeware, indian cuisine
Dessert	16	dish food, cooking show, cake decorating
Meat	9	barbecue, cooking show, recipe
IPhone	6	telephone, tablet computer, game
Personal computer	20	mobile phone, smartphone, gadget
Video game console	19	handheld game console, playstation, xbox console
Fish	13	fishing rod, fishing lure, recreational fishing
Machine	6	manufacturing, woodturning, factory
Railroad car	7	rapid transit, new york city subway, rail freight transport
Fashion	15	model person, runway fashion, eye shadow
Weight training	4	human back, dance, biceps curl
Bicycle	13	mountain bike, road bicycle racing, bmx bike
Truck	14	off-road vehicle, four-wheel drive, heavy equipment
Boat	8	ocean, fish, recreational fishing
Disc jockey	4	concert, dance, mixing console
Snare drum	6	electronic drum, drum stick, musical ensemble

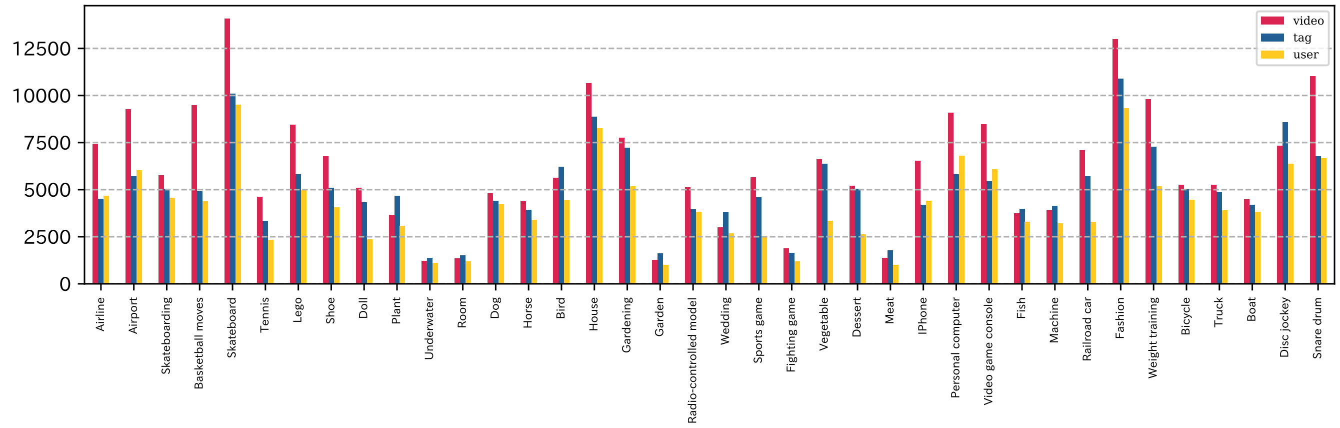


Fig. 1. The number of videos, tags, and users included in the dataset.

was substituted for Eq. (14), was calculated by using the Okapi BM-25 formula [13]. Note that we excluded the queries from the Queryset if its top 10 evaluation metric of initial scores equals 1.0 since the objective of reranking is a refining weak result.

3) **Features:** Raw visual features were extracted from Google's Inception-v3 model trained on *ImageNet 1K*. Raw audio features were extracted from a CNN-inspired architecture trained for audio classification as described in [14]. Both visual and audio features follow a PCA whitening process to further reduce the dimension to 1,024 and 128, respectively. The video-level features were mean-pooled from frame-level features. In the experiments, these video-level visual and audio features were combined by early fusion to get the feature vector in Eq. (2). The above features are opened as part of the dataset.

4) **Evaluation metrics:** The ground truth for evaluating the performance of reranking algorithms was given by using the labels of the entities in the YouTube-8M dataset. Specifically, the relevance scores related to a specific query keyword were classified automatically according to whether or not videos had a corresponding entity label. The scores were confined within the following two categories: "1=relevant" and "0=irrelevant". We denote the relevance score of video x_i as $rel_i \in \{1, 0\}$. To measure the relevance performance of retrieval results, we used the well-known normalized discounted cumulative gain under depth d (NDCG@d) and the average precision under depth d (AP@d), which are defined as

$$\mathrm{NDCG@}d = \frac{1}{W}\sum_{i=1}^{d}\frac{2^{rel_i}-1}{\log(1+i)}, \tag{15}$$

$$\mathrm{AP@}d = \frac{1}{d}\sum_{i=1}^{d}\left(\sum_{j=1}^{i}\frac{rel_j}{i}\right), \tag{16}$$

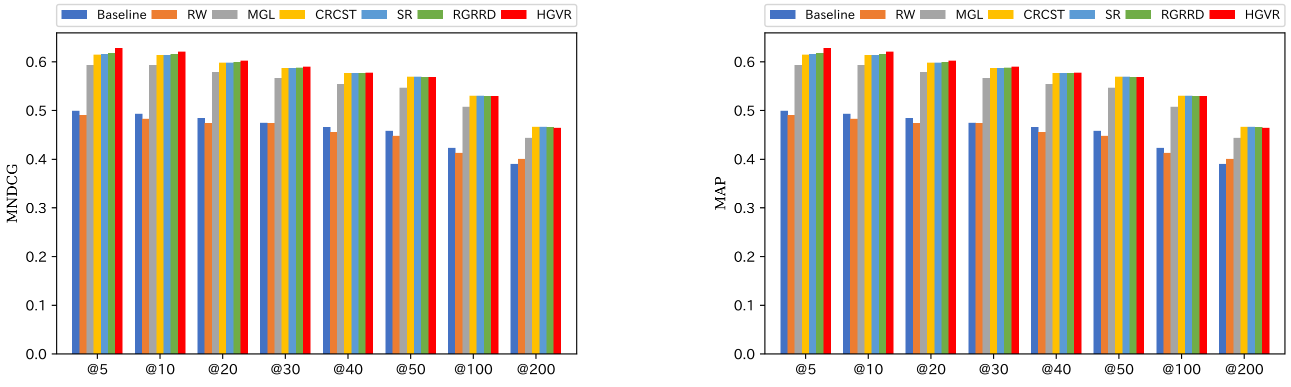
where W is a normalization constant. It makes the optimal ranking's NDCG score to be 1. AP is the mean of the precision values obtained when each relevant video occurs. We average the NDCGs and APs over all the queries to obtain the mean NDCG (MNDCG@d) and the mean AP (MAP@d) for an overall performance measurement.

5) **Methods of comparison:** To evaluate the performance of the proposed reranking algorithm, we compared the proposed method, denoted as HGVR, with the following six algorithms:

- Tag-based video retrieval using the Okapi BM-25 formula (Baseline) [13]: this method was used as the baseline TBVR to obtain the initial ranking order. The following algorithms reranked this result.
- Random walk-based reranking (RW) [1] : a representative self-reranking method which conducts random walk on a video graph where nodes are videos and edges are weighted by video visual similarities.
- Multimodal graph-based reranking (MGL) [3], which is the state-of-the-art for graph-based reranking.
- Clustering reranking with click-based similarity and typicality (CRCST) [5] : a two-step reranking method that first learns the similarity between videos to perform relevance feedback using click-through data, and then evaluates the cluster typicality to rerank videos.
- Social ranking (SR) [6]: User information is utilized to boost the retrieval performance. A regularization-based diversified framework which fuses the visual and views information is introduced.
- Robust graph reranking based on rank distance (RGRRD) [7] : a reranking method that defines a rank distance to measure the relevance of each video at the rank level and constructs a directed graph to encode the relationship between videos.

B. Comparison of Different Reranking Algorithms

Figure 2 illustrate comparisons of the MNDCG and MAP, respectively, using the above reranking methods including the state-of-the-art algorithms RGRRD, SR, CRCST, MGL, and our proposed method. We can see that our reranking method outperforms the other methods with {MNDCG, MAP}@5, 10, 20, 30, and 40. Compared with Baseline, the methods based



(a) MNDCG

(b) MAP

Fig. 2. MNDCG and MAP of all ranking methods under different depths.

on the proposed approach can achieve an improved MNDCG different depths. Specifically, the MNDCG@10 is improved by 20.4% from 0.493 for Baseline to 0.620, and MAP@10 is boosted by 33.6% from 0.461 to 0.695 over the entire dataset.

Since we can see that all methods, which employ the visual reranking approach, outperform Baseline, the visual reranking approach overcomes the semantic gap. Furthermore, we find that HGVR outperforms state-of-the art, such as CRCST, SR, and RGRRD, using relevance and diversified-based approach at different depths. These methods introduce visual, text, and other aspect individually. From this result, topic-based similarity estimated by heterogeneous graph-based learning contributes to improving the relevance performance. Therefore, the proposed feature extraction method can collaboratively fuse the video visual information, their tags, and relative users and utilize the mutual reinforcement between each heterogeneous object through the heterogeneous graph.

V. Conclusions

We have presented a method of improving the performance of graph-based Web video search reranking.The proposed method involves two procedures. We first construct a heterogeneous graph, which consists of multiple types of objects and multiple types of relationships, to preserve different kinds of information from different data sources. Secondly, we apply graph-based soft clustering to the heterogeneous graph to group the different types of data as clusters. From the clustering result, the extraction of topics becomes feasible. By estimating the topic membership for each video, the proposed method defines a video similarity measure and formulates the video search reranking as an optimization problem. As a result of these procedures, we obtain an accurate reranking score list. Consequently, the superiority of our method to the existing methods was confirmed.

Acknowledgment

This research was financially supported by JSPS KAKENHI Grant Number 17K12687.

References

[1] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang, "Video search reranking via information bottleneck principle," in *Proceedings of the ACM International Conference on Multimedia*, 2006, pp. 35–44.

[2] Xinmie Tian, Linjun Yang, Jingdong Wang, Xiuqing Wu, and Xian-Sheng Hua, "Bayesian visual reranking," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 639–652, 2011.

[3] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu, "Multimodal graph-based reranking for Web image search," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4649–4661, 2012.

[4] Junjie Cai, Zheng-Jun Zha, Meng Wang, Shiliang Zhang, and Qi Tian, "An attribute-assisted reranking model for Web image search," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 261–272, 2015.

[5] Xiaopeng Yang, Tao Mei, Yongdong Zhang, Jie Liu, and Shin'ichi Satoh, "Web Image Search Re-Ranking With Click-Based Similarity and Typicality," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4617–4630, 2016.

[6] Dan Lu, Xiaoxiao Liu, and Xueming Qian, "Tag-based image search by social re-ranking," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1628–1639, 2016.

[7] Ziqiong Liu, Shengjin Wang, Liang Zheng, and Qi Tian, "Robust ImageGraph: Rank-Level Feature Fusion for Image Search," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3128–3141, 2017.

[8] Xiang Ren, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, and Jiawei Han, "Heterogeneous graph-based intent learning with queries, Web pages and Wikipedia concepts," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, 2014, pp. 23–32.

[9] Börkur Sigurbjörnsson and Roelof van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proceedings of the International Conference on World Wide Web*, 2008, pp. 327–336.

[10] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the International Conference on Neural Information Processing Systems: Natural and Synthetic*, 2001, pp. 585–591.

[11] James C. Bezdek, Robert Ehrlich, and William Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[12] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," in *arXiv:1609.08675*, 2016, pp. 1–10.

[13] Stephen Robertson and Hugo Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2010.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 131–135.