

HOKKAIDO UNIVERSITY

Title	Extracting Hierarchical Structure of Web Video Groups Based on Sentiment-Aware Signed Network Analysis
Author(s)	Harakawa, Ryosuke; Ogawa, Takahiro; Haseyama, Miki
Citation	IEEE Access, 5, 16963-16973 https://doi.org/10.1109/ACCESS.2017.2741098
Issue Date	2017-08-17
Doc URL	https://hdl.handle.net/2115/67506
Rights	© 2017 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Type	journal article
File Information	08012375.pdf



Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 14, 2017, accepted August 13, 2017, date of publication August 17, 2017, date of current version September 19, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2741098

Extracting Hierarchical Structure of Web Video Groups Based on Sentiment-Aware Signed Network Analysis

RYOSUKE HARAKAWA, (Member, IEEE), TAKAHIRO OGAWA, (Member, IEEE), AND MIKI HASEYAMA, (Senior Member, IEEE)

Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Ryosuke Harakawa (harakawa@lmd.ist.hokudai.ac.jp)

This work was partly supported by JSPS KAKENHI Grant Numbers JP16J02042, JP17H01744.

ABSTRACT Sentiment in multimedia contents has an influence on their topics, since multimedia contents are tools for social media users to convey their sentiment. Performance of applications such as retrieval and recommendation will be improved if sentiment in multimedia contents can be estimated; however, there have been few works in which such applications were realized by utilizing sentiment analysis. In this paper, a novel method for extracting the hierarchical structure of Web video groups based on sentiment-aware signed network analysis is presented to realize Web video retrieval. First, the proposed method estimates latent links between Web videos by using multimodal features of contents and sentiment features obtained from texts attached to Web videos. Thus, our method enables construction of a signed network that reflects not only similarities but also positive and negative relations between topics of Web videos. Moreover, an algorithm to optimize a modularity-based measure, which can adaptively adjust the balance between positive and negative edges, was newly developed. This algorithm detects Web video groups with similar topics at multiple abstraction levels; thus, successful extraction of the hierarchical structure becomes feasible. By providing the hierarchical structure, users can obtain an overview of many Web videos and it becomes feasible to successfully retrieve the desired Web videos. Results of experiments using a new benchmark dataset, YouTube-8M, validate the contributions of this paper, *i.e.*, 1) the first attempt to utilize sentiment analysis for Web video grouping and 2) a novel algorithm for analyzing a weighted signed network derived from sentiment and multimodal features.

INDEX TERMS Video retrieval, video clustering, network analysis, signed network, sentiment analysis.

I. INTRODUCTION

Due to the widespread use of video hosting services such as YouTube,[1] more and more users are retrieving desired Web videos that include topics in which they are interested [1], [2]. Usually, users input queries into search engines and then Web videos associated with the input queries are returned as ranked lists to users. Remarkable progress has been made recently in semantic understanding including object recognition and event detection [3]–[6]; however, it is still difficult to obtain desired Web videos if users cannot input suitable queries [7], [8].

To overcome this difficulty, retrieval methods that provide Web video groups with similar topics have been proposed [9]–[12]. It has been reported that methods for hierarchically providing Web video groups at multiple abstraction levels are especially useful for easy access to desired Web videos since users can obtain an overview of retrieval results [13]–[19]. These methods enable accurate retrieval by analysis of multimodal features, *e.g.*, visual and textual features and features of metadata attached to contents through statistical schemes such as canonical correlation analysis (CCA) [20] and hierarchical latent Dirichlet allocation (hLDA) [21].

Sentiment analysis for multimedia contents such as images, audio and videos has attracted much attention recently [22]–[27]. Multimedia contents are tools for social media users to convey their sentiment, emotion and opinions [24]; conversely, the sentiment has an influence on topics

[1] https://www.youtube.com/

2169-3536 © 2017 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

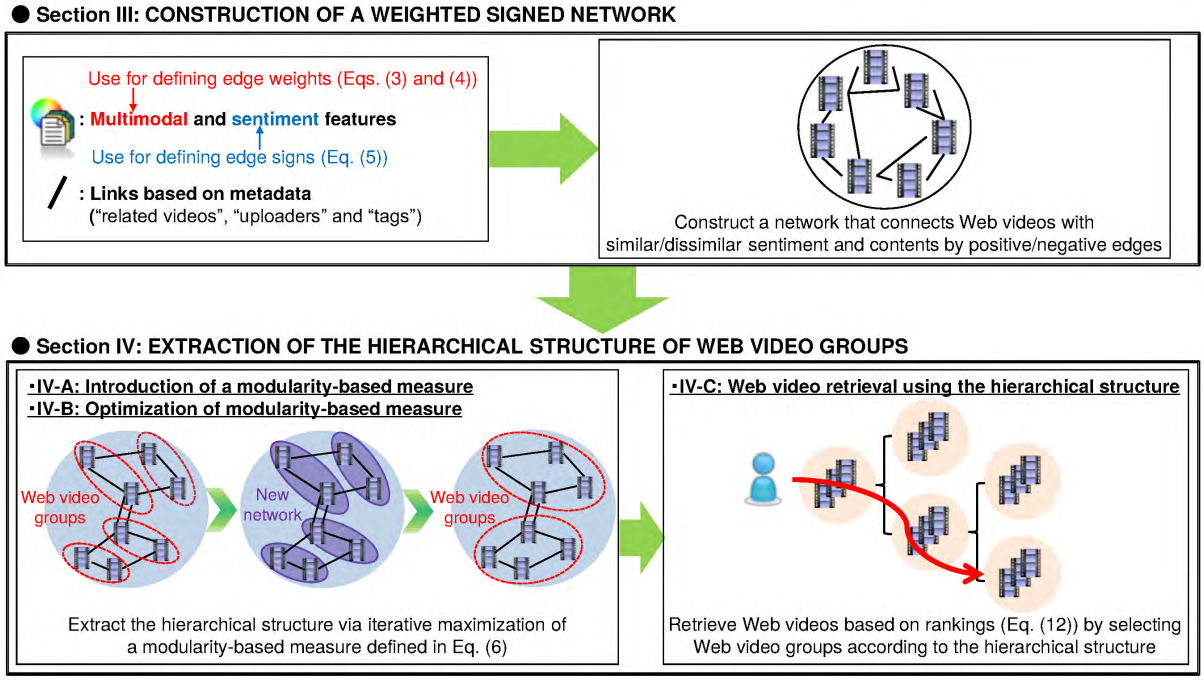


FIGURE 1. Overview of the proposed method for extracting the hierarchical structure of Web video groups based on sentiment-aware signed network analysis.

of multimedia contents. For example, we assume two users who support and oppose the government policy. The user who supports the policy is likely to upload contents with positive sentiment (*e.g.* a support speech). On the other hand, the user who opposes the policy is likely to upload contents with negative sentiment (*e.g.* a negative campaign). Thus, their uploaded contents will have topics that are different. In this way, since topics of multimedia contents are determined by sentiment features as well as multimodal features, performance of applications such as retrieval and recommendation will be improved if the sentiment in contents can be estimated. Although fundamental studies to predict sentiment in multimedia contents have been carried out [22]–[27], there have been few works that were carried out to realize such applications by utilizing sentiment analysis.

In this paper, we present a novel method to extract the hierarchical structure of Web video groups based on sentiment-aware signed network analysis. The proposed method enables users to obtain an overview of many Web videos and retrieve the desired ones through the hierarchical structure even if users cannot input suitable queries. To the best of our knowledge, this work is the first work on collaborative use of sentiment and multimodal features in order to realize Web video grouping for accurate retrieval. The technical contributions of this paper are the proposal of a novel data structure, *i.e.*, a weighted signed network that can consider positive/negative sentiment in Web videos as well as multimodal features and the development of a new algorithm to analyze the network. Although a signed network has been used for human interaction analysis [28]–[30], our work is the first work in which a signed network was introduced into multimedia content analysis (particularly Web video analysis).

The rest of this paper is organized as follows. In Section II, an overview of the proposed method is presented. In Section III, a method for constructing a weighted signed network with consideration of sentiment and multimodal features of Web videos is explained. In Section IV, a new algorithm to extract the hierarchical structure through the signed network is presented. In Section V, results of experiments for a new benchmark dataset, YouTube-8M [31], are presented to validate the contributions of this paper, *i.e.*, (1) the first attempt to utilize sentiment analysis for Web video grouping and (2) a novel algorithm for analyzing a weighted signed network derived from sentiment and multimodal features. Conclusions are given in Section VI.

II. OVERVIEW OF THE PROPOSED METHOD

In this section, we present an overview of the proposed method for extracting the hierarchical structure of Web video groups based on sentiment-aware signed network analysis (see Fig. 1).

First, we construct a weighted signed network whose nodes and edges are Web videos and latent links, respectively (see Section III). The proposed method judges whether Web videos have positive or negative sentiment by applying a state-of-the-art scheme, VADER [27], to titles and description. Since VADER is suitable for social media microblog-like contexts, we adopt VADER. From this result and content similarities obtained by selecting the statistically most sim-

ilar features from visual and textual features via empirical distribution functions [32], a weighted signed network can be constructed. In this network, Web videos with similar and opposite sentiments are linked by positive and negative edges, respectively, and edge weights represent content similarities. Since the signed network is useful for separating Web videos with opposite sentiments, *i.e.*, different topics, we introduce the signed network into the proposed method.

Moreover, we propose an algorithm to analyze the obtained network for extracting the hierarchical structure of Web video groups (see Section IV). Specifically, an algorithm for optimizing a modularity [33]-based measure, which can adaptively adjust the balance between positive and negative edges, was newly developed. Since our network can distinguish sentiment as well as multimodal features in Web videos unlike conventional methods, successful extraction of the hierarchical structure becomes feasible. By providing the hierarchical structure, users can obtain an overview of many Web videos and it becomes feasible to successfully retrieve the desired Web videos.

III. CONSTRUCTION OF A WEIGHTED SIGNED NETWORK

In this section, a method to construct a weighted signed network by estimating latent links between Web videos is explained. We denote Web videos by f_i ($i = 1, 2, \cdots, N$; N being the number of Web videos). Also, we represent a weighted signed network as $G_v = (V_v, E_v)$, where V_v and E_v are respectively sets of Web videos f_i and latent links w_{ij} ($i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, N$), which is constructed as follows.

First, we apply a sentiment analysis method called VADER [27] to titles and description of each Web video. Thus, valence scores $c(i)$ ($-1 \leq c(i) \leq 1$) for the texts of f_i can be obtained, that is, we can understand the degree of positiveness/negativeness of topics of Web videos. We define a sign of an edge between f_i and f_j as positive if $c(i) \cdot c(j) \geq 0$ and negative otherwise. The paper [24] shows that multimedia contents are tools for social media users to convey their sentiment, emotion and opinions; conversely, the sentiment has an influence on topics of multimedia contents. Therefore, by utilizing the obtained signs, the performance of Web video grouping for retrieval will be improved from conventional schemes without consideration of sentiment.

Since topics of Web videos are determined by both sentiment and multimodal features, we then calculate edge weights through multimodal features. To group Web videos with similar topics, we should judge what features are particularly similar to each other. For example, if Web videos of the same person with different viewpoints are given, we can find topics of these videos that are the same by focusing not on visual features but on textual features. Thus, we adopt the similarity calculation scheme [15] based on this intuition and estimate similarities by selecting particularly similar features in a statistical way. First, we denote features of video contents by $\boldsymbol{v}_i^m$ ($m = 1, \cdots, J$; m representing modalities and J being the number of modalities). In the experiment shown later, we adopt convolutional neural network (CNN) features at the video level [31] and Doc2Vec features [34] calculated from titles and description of each Web video. We calculate distances between the feature vectors for each modality and denote them by $d^m(i,j) = \|\boldsymbol{v}_i^m - \boldsymbol{v}_j^m\|$ where $m = 1, \cdots, J$, $i = 1, \cdots, N$, $j = 1, \cdots, N$. Here, we sort each element of $d^m(i,j)$ in ascending order and denote them by $d^m(l)$ ($l = 1, 2, \cdots, N_{comb}$; N_{comb} being the number of the combination of different Web videos). Next, we construct the empirical distribution function $F^m(x)$ of $d^m(l)$ ($m = 1, \cdots, J$) as follows [32]:

$$F^m(x) = \frac{1}{N_{comb}} \sum_{l=1}^{N_{comb}} X_l^m(x), \tag{1}$$

$$X_l^m(x) = \begin{cases} 1 & \text{if } d^m(l) \leq x \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Similarities s_{ij} and distances d_{ij} between f_i and f_j are defined as follows:

$$s_{ij} = \max_{m=1,\cdots,J} [1 - F^m\{d^m(i,j)\}], \tag{3}$$

$$d_{ij} = \max_{m=1,\cdots,J} F^m\{d^m(i,j)\}, \tag{4}$$

where $i = 1, \cdots, N, j = 1, \cdots, N$. These equations enable comparison between different kinds of features since we equalize occurrence probability of similarities in a constant interval of the similarity-axis. Thus, selection of the statistically most similar and dissimilar features becomes feasible to adaptively define the similarities and distances.

Furthermore, we employ multiple metadata. The papers [35], [36] showed that similar Web videos can be associated with each other by using "related videos," "uploaders" and "tags". According to those reports, we use information of "related videos," "uploaders" and "tags" simultaneously to estimate latent links. We consider f_i and f_j are linked to each other if "related videos" of f_i include f_j or vice versa, "uploaders" of f_i and f_j are the same, or the same tags are attached to f_i and f_j. If f_i and f_j are linked to each other, edge weights w_{ij} between f_i and f_j are defined as follows.

$$w_{ij} = \begin{cases} s_{ij} & \text{if } c(i) \cdot c(j) \geq 0 \\ -d_{ij} & \text{otherwise.} \end{cases} \tag{5}$$

If f_i and f_j are not linked to each other, their latent link is not built. In this way, we design weights so that Web videos with similar sentiment and contents are connected by positive edges and those with dissimilar sentiment and contents are linked by negative edges.

One of the contributions of this paper is construction of this sentiment-aware signed network, which will be useful for Web video grouping for retrieval since sentiment has an influence on topics of Web videos. It should be noted that this work is the first work in which a signed network was introduced into multimedia content analysis (particularly Web video analysis), though a signed network has been used for

human interaction analysis [28]–[30]. The other contribution is the proposal of a novel algorithm for analyzing the obtained network, which is explained in the next section.

IV. EXTRACTION OF THE HIERARCHICAL STRUCTURE OF WEB VIDEO GROUPS

In this section, we present a new algorithm for extracting the hierarchical structure through the signed network. The paper [37] presented an idea for detecting community structure in the signed network by a modularity-based measure, which can adjust the balance between positive and negative edges. Although detailed explanation and experimental results were not presented, that paper claimed that the use of a combination of a well-known scheme [38] and the modularity-based measure would be useful. The scheme [38] enables fast and accurate community detection; therefore, the above combination is useful for our purpose since Web video analysis needs both adequate scalability and accuracy. For this reason, we developed a new algorithm based on the scheme [38] and the modularity-based measure [37]. Note that we enhanced the algorithm presented in the paper [37] by introducing edge weights obtained via latent features. Our new algorithm is described in detail below.

A. INTRODUCTION OF A MODULARITY-BASED MEASURE

First, we explain a modularity-based measure used as an optimization measure for extracting the hierarchical structure. Modularity for a signed network was first proposed in a paper [39] to evaluate the quality of detecting communities. Maximizing the modularity results in good community detection results, in other words, we can obtain the structure in which positive edges are dense within the same Web video groups and the different Web video groups are connected by negative edges. The paper [37] improves the original modularity by adding a parameter for adjusting the balance between positive and negative edges. In the proposed method, according to the paper [37], we adopt an improved measure defined as follows.

$$Q = \alpha Q^+ - (1-\alpha)Q^-, \tag{6}$$

where

$$Q^+ = \frac{1}{2w^+}\sum_{i,j}(w_{ij}^+ - \frac{w_i^+ w_j^+}{2w^+})\delta(c_i, c_j), \tag{7}$$

$$Q^- = \frac{1}{2w^-}\sum_{i,j}(w_{ij}^- - \frac{w_i^- w_j^-}{2w^-})\delta(c_i, c_j). \tag{8}$$

Here,

$$\begin{aligned} w_{ij}^+ &= \max\{0, w_{ij}\}, \quad w_{ij}^- = \max\{0, -w_{ij}\}, \\ 2w^+ &= \sum_{i,j} w_{ij}^+, \quad 2w^- = \sum_{i,j} w_{ij}^-, \\ w_i^+ &= \sum_j w_{ij}^+, \quad w_i^- = \sum_j w_{ij}^-. \end{aligned} \tag{9}$$

In the above equations, $\delta(c_i, c_j)$ is equal to 1 if f_i and f_j belong to the same Web video group and is 0 otherwise. Also, α $(0 < \alpha < 1)$ is a parameter for adjusting the balance between positive and negative edges. If α is set to a large value, negative edges within the same Web video groups tend to be accepted since the positive edges are highly regarded. If α is set to a small value, on the other hand, positive edges between the different Web video groups tend to be accepted since negative edges are highly regarded.

B. OPTIMIZATION OF A MODULARITY-BASED MEASURE

Next, we present a new algorithm to extract the hierarchical structure that was developed by improving the scheme [38] via the signed network. The hierarchical structure of Web video groups can be extracted by maximizing Q defined in Eq. (6). Let us denote the network used in this subsection by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For the initialization, we assign each Web video f_i $(i = 1, 2, \cdots, N)$ to each node $v \in \mathcal{V}$, and each edge between nodes, $e_{ij} \in \mathcal{E}$, is defined by the obtained links w_{ij}. Also, e_{ij}^+, e_{ij}^-, e^+ and e^- are respectively defined in the same manner as Eq. (9). The maximization consists of the following two phases.

Phase 1 (Local Maximization of the Modularity-Based Measure): We assign each node of the network $\mathcal{G}$ to each Web video group. For each target node, we evaluate the gain of the proposed measure Q when the node is assigned to a Web video group including its neighborhood node. Then the target node is re-assigned to a Web video group for which the positive gain is maximum. Here, the gain ΔQ when the node $v_i \in \mathcal{V}$ is set to a Web video group including the neighborhood node $v_j \in \mathcal{V}$ can be computed as follows.

$$\Delta Q = \alpha \Delta Q^+ - (1-\alpha)\Delta Q^-. \tag{10}$$

Here, ΔQ^+ represents the gain corresponding to the positive edges, which is defined by the following equation.

$$\begin{aligned} \Delta Q^+ = \frac{1}{e^+}\Big(\sum_{k\in c_j} e_{ik}^+ - \sum_{k\in c_i\setminus\{v_i\}} e_{ik}^+\Big) \\ - \frac{1}{(2e^+)^2}\Big[\Big(\sum_{k\in c_i\setminus\{v_i\}} e_k^+\Big)^2 - \Big(\sum_{k\in c_i} e_k^+\Big)^2 \\ + \Big(\sum_{k\in c_j\cup\{v_i\}} e_k^+\Big)^2 - \Big(\sum_{k\in c_j} e_k^+\Big)^2\Big], \end{aligned} \tag{11}$$

where $k \in c_i$ denotes a node contained in a Web video group c_i. Moreover, we can derive ΔQ^- in the same manner. For more details of the derivation, refer to Appendix. The local maximization of the gain ΔQ is applied to all nodes iteratively and sequentially until no more improvement of the modularity can be obtained.

Phase 2: Update of a new network We construct a new network whose nodes are the Web video groups obtained in the first phase. There are two types of edges between pairs of nodes: positive and negative edges whose weights are the sum of positive and negative weights in the original network, respectively. Also, each new node has positive and

negative self-loops derived from the positive and negative weights of the corresponding original nodes in the first phase, respectively. Thus, we can obtain a new network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the updated values of e_{ij}, e_{ij}^{+}, e_{ij}^{-}, e^{+} and e^{-}. In this paper, a pair of the first and second phases is represented as ''a pass'' and this iteration number is denoted by $q\,(= 1, 2, \cdots, Q_h$; Q_h being the number of all passes). Furthermore, the passes (the first phase for detecting Web video groups from the network and the second phase for building the new network) are iterated until no more improvement of Q can be obtained.

As a consequence of obtaining Web video groups with different levels of resolution by the iteration of passes, it becomes feasible to extract the hierarchical structure of Web video groups. The number of new nodes recursively decreases in the second phase according to the increase of q; thus, efficient extraction becomes feasible even when a large number of Web videos are targeted. In this paper, we denote the obtained Web video groups by $Grp_{n_q}^{q}(n_q = 1, 2, \cdots, D_q$; D_q being the number of Web video groups where the iteration count is q). Finally, we note that the novelty of our method can be found in the point that we enable application of the derived algorithm based on the measure in Eq. (6) to Web video analysis through the sentiment-aware signed network (see Eq. (5)). By newly considering the sentiment as well as multimodal features, the hierarchical structure can be successfully extracted since the sentiment has an influence on topics of Web videos.

C. WEB VIDEO RETRIEVAL USING THE HIERARCHICAL STRUCTURE

A Web video retrieval method using the hierarchical structure is explained. First, we rank each Web video f_i $(i \in \{1, 2, \cdots, N\})$ that belong to the Web video group $Grp_{n_q}^{q}$ in descending order of the following measure.

$$R_{n_q}^{q}(i) = \sum_{j=1}^{N} e_{ji}\delta(c_i, c_j), \tag{12}$$

where $\delta(c_i, c_j)$ is 1 if f_i and f_j belong to the same Web video group and 0 otherwise. Thus, Web videos are ranked on the basis of the distribution of edge weights in the network of each Web video group. Moreover, we show the Web video groups in the order of $Grp_{n_q}^{Q_h}$, $Grp_{n_q}^{Q_h-1}$, $\cdots$, $Grp_{n_q}^{1}$, that is, from the larger Web video groups to the smaller Web video groups. The larger Web video groups include Web videos with various topics and the smaller Web video groups contain Web videos with similar topics. Then users select the Web video groups associated with the desired Web videos according to the exhibited hierarchical structure and retrieve Web videos based on the ranking $R_{n_q}^{q}(i)$ of each Web video f_i $(i = 1, 2, \cdots, N)$ (see Fig. 1 IV-C). Hence, the proposed method enables users to easily obtain an overview of many Web videos via the hierarchical structure of Web video groups. Consequently, users can retrieve the desired Web videos even if they cannot input suitable queries corresponding to the desired Web videos.

V. EXPERIMENTAL RESULTS

In Section V-A, experimental settings are described. In Section V-B, quantitative evaluations are performed by comparing clustering results for retrieval, which were obtained by our method and recently published reference methods. In Section V-C, we show several visualization results to qualitatively discuss the effectiveness of the proposed method.

A. SETTINGS

In the experiment, we used a newly published public dataset, YouTube-8M [31], containing YouTube videos. Multiple labels called ''entities'' are annotated to each Web video for the ground truth for tasks such as clustering and classification [31]. To construct datasets for the experiments, we collected Web videos with specific entities, which were selected as queries.[2] Note that it is not suitable to use frequent entities attached to almost all Web videos for performing accurate clustering evaluation. In particular, a query entity or entities that are high level concepts of the query entity are attached to almost all Web videos and those entities should be removed to accurately evaluate the clustering results. In this experiment, we defined entities attached to more than 50% of the Web videos in each dataset as frequent entities. We removed Web videos that had only those entities and constructed datasets shown in Table 1.

As described in Section III, we employed three kinds of features. First, we adopted CNN features at the video level, which were available on YouTube-8M [31]. In a work [31], the features were computed on the basis of a state-of-the-art deep model, *i.e.*, Inception network[3] trained on ImageNet [40], [41]. Second, we adopted Doc2Vec features [34] calculated from titles and description of each Web video. The feature representation is given by dense vectors, and good discriminative power was reported [34]. In this experiment, we lemmatized each word and removed stop words by using Natural Language Toolkit (NLTK) [42]. Third, we employed sentiment features obtained by applying VADER [27] to titles and description. In addition to these features, metadata ''related videos,'' ''uploaders'' and ''tags'' were utilized to build links defined in Eq. (5). Here, we removed tags for which document frequencies were less than five and more than 90 percentile to remove noisy tags.

B. QUANTITATIVE EVALUATIONS

The effectiveness of our method was quantitatively verified by evaluating clustering results for retrieval. Since a contribution of this paper is the proposal of a novel network analysis algorithm with sentiment and multimodal features, we compare our method with recently presented network analysis methods. Concretely, we denote the proposed method by (P)

[2]Since the sentiment analysis scheme VADER [27] targets English documents, Web videos with English titles were collected.

[3]Tensorflow: Image recognition. (https://www.tensorflow.org/tutorials/image_recognition/)

TABLE 1. Datasets constructed by using YouTube-8M [31]. "Query entity" denotes the entity used to collect Web videos. Web videos that had only entities shown in "Excluded entities" were removed from each dataset for accurately evaluating clustering results. Values in parentheses are percentages of Web videos with "Excluded entities" to all Web videos in each dataset before removing such Web videos. "Num. of Web videos" shows the number of remaining Web videos after removing Web videos that had only "Excluded entities."

Dataset	Query entity	Excluded entities	Num. of Web videos
1	News program	News Program (100%)	523
2	Zoo	Zoo (100%)	824
3	Athlete	Athlete (100%)	869
4	Money	Money (100%)	1225
5	Manga	Manga (100%), Animation (71.7%)	1433
6	Plant	Plant (100%), Gardening (74.2%)	1457
7	Robot	Robot (100%)	1643
8	British Broadcasting Corporation	British Broadcasting Corporation (100%)	1843
9	Aquarium	Aquarium (100%), Animal (78.8%), Fish (78.1%)	1958
10	Pet	Pet (100%), Animal (77.4%)	2058
11	Eating	Eating (100%), Food (69.1%)	2148
12	Tennis	Tennis (100%)	2448
13	Computer	Computer (100%)	3524
14	Fish	Fish (100%), Animal (99.9%)	4329
15	Cooking show	Cooking show (100%), Food (79.4%), Cooking (66.1%), Recipe (50.7%)	5232

and compare (P) with the following network analysis algorithms, (R1), (R2) and (R3).

- **(R1):** This is a method that replaces α for calculating Q in Eq. (6) with $w^+/(w^+ + w^-)$. Since this is a weight presented in the paper [39] in which modularity for a signed network was first proposed, we adopt this method as a baseline to verify the effectiveness of the proposed measure Q in Eq. (6).
- **(R2):** This is a method based on the recently published paper [15] that extracts the hierarchical structure of Web video groups through unsigned network analysis. Although this method uses CNN and Doc2Vec features and metadata as in (P), sentiment features obtained by VADER are not utilized. Since this method cannot handle negative edge weights, w_{ij} in Eq. (5) has to be replaced with $w_{ij} = s_{ij}$ $(i = 1, 2, \cdots, N,\ j = 1, 2, \cdots, N)$.
- **(R3):** This is a method based on a flat community detection scheme that has recently been presented [43]. This method uses both video features and metadata as in (P); however, the hierarchical structure is not provided. We calculate w_{ij} in Eq. (5) as in (R2) since this method cannot handle negative edges.

To evaluate the accuracy of extracting the hierarchical structure for Web video retrieval, we employed the following average precision.

$$AP@k = \frac{1}{R_k} \sum_{i=1}^{k} x_i \, prec_i, \tag{13}$$

where k is the number of Web videos provided as the retrieval results, R_k is the number of "relevant Web videos" within k Web videos of the retrieval results, x_i is 1 if the i-th retrieved Web videos are "relevant Web videos" and 0 otherwise, and $prec_i$ is the precision[4] when i Web videos are retrieved. We define "relevant Web videos" as Web videos with the most frequent entity for each Web video group.

Figure 2 shows the detailed results. In this figure, we also show simulation results to verify the robustness to noisy edges in addition to the results for real data. Specifically, for each network, we randomly selected $N_n(\in \{0.002, 0.004, 0.006, 0.008, 0.01\})\%$ node pairs from ones without edges and constructed edges with weights. Note that for (P), we varied α around the weight originally presented in the paper [39], *i.e.*, $w^+/(w^+ + w^-)$, to define the optimal parameter. Concretely, we varied α from $w^+/(w^+ + w^-) - 0.2$ to $w^+/(w^+ + w^-) + 0.2$ by a constant interval of 0.05, and then we adopted α when the average precision for the hierarchies was maximum. From (P) and (R1) in Fig. 2, it can be seen that the introduction of α into Q in Eq. (6) worked well for obtaining the suitable hierarchical structure. When comparing (P) with (R2), the hypothesis that sentiment in multimedia contents has an influence on their topics can be quantitatively proven. Even if irrelevant Web videos are linked by noisy edges, our method can group only Web videos with positive edges, *i.e.*, Web videos with the same valence, and separate Web videos with negative edges, *i.e.*, Web videos with the opposite valance; thus, better results can be obtained. From (P) and (R3), we can confirm the effectiveness of using the hierarchical structure rather

[4]Precision is defined by the following equation: $Precision = \frac{Num.\ of\ correctly\ retrieved\ Web\ videos}{Num.\ of\ retrieved\ Web\ videos}$.

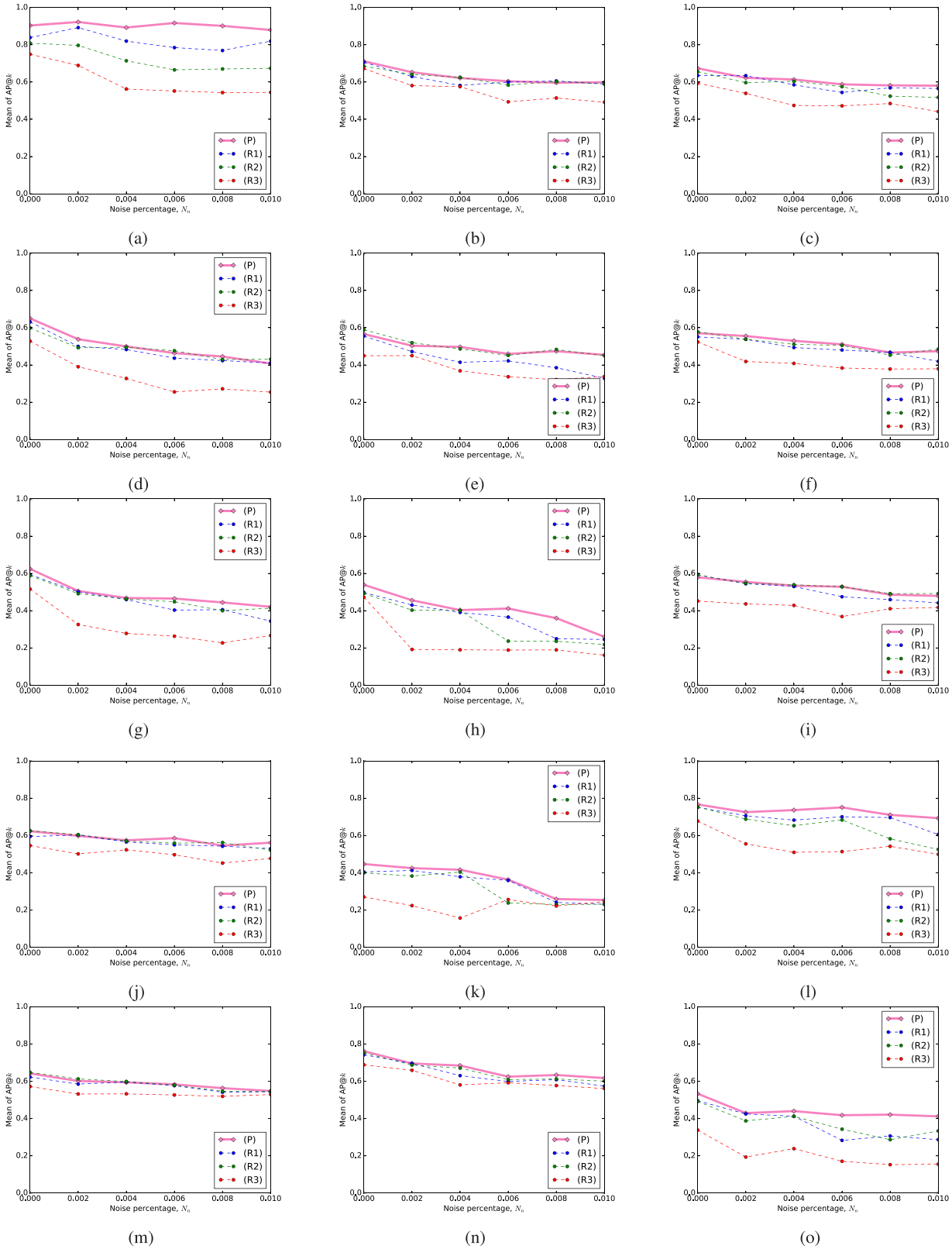
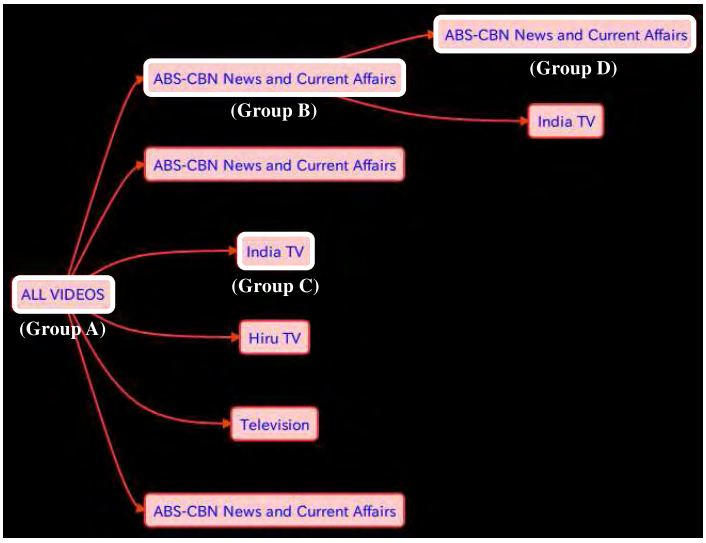
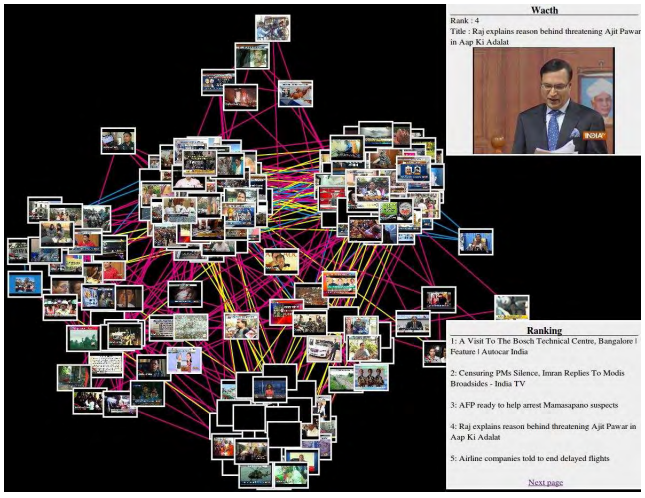


FIGURE 2. Mean of AP@k for all Web video groups in all hierarchies, which is weighted by the numbers of Web videos in Web video groups. k is defined as the number of Web videos in each Web video group. Vertical and horizontal axes denote the mean of AP@k and noise percentages N_n added to edges of each network, respectively. Results where $N_n = 0$ correspond to real data with no noise added.

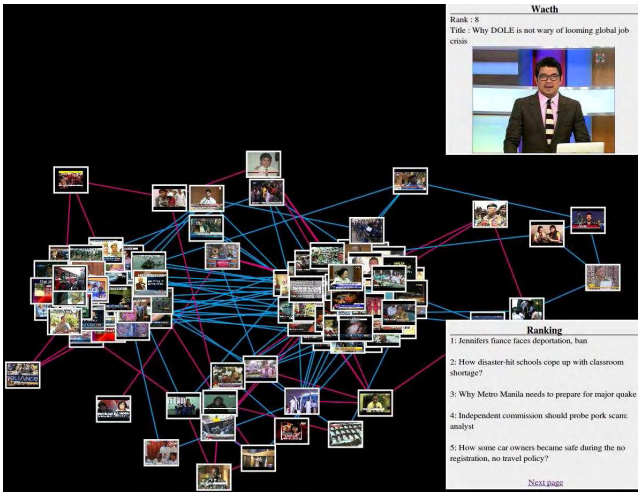
than the flat cluster structure. Since the desired abstraction levels of topics are different depending on each user, this performance improvement is significant for satisfying users' needs. Finally, it is notable that the robustness to noisy edges



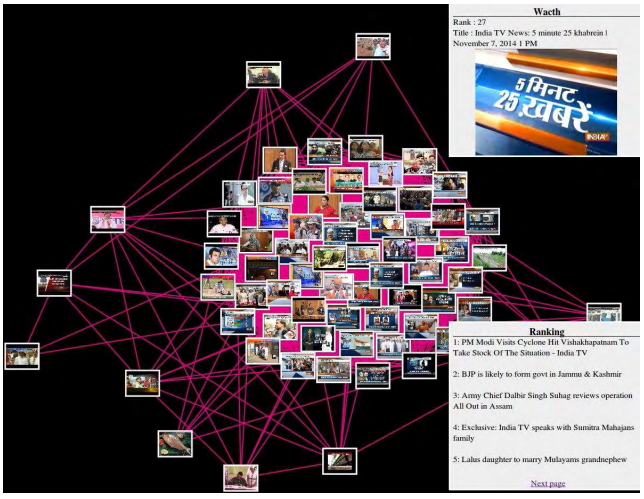
(a)



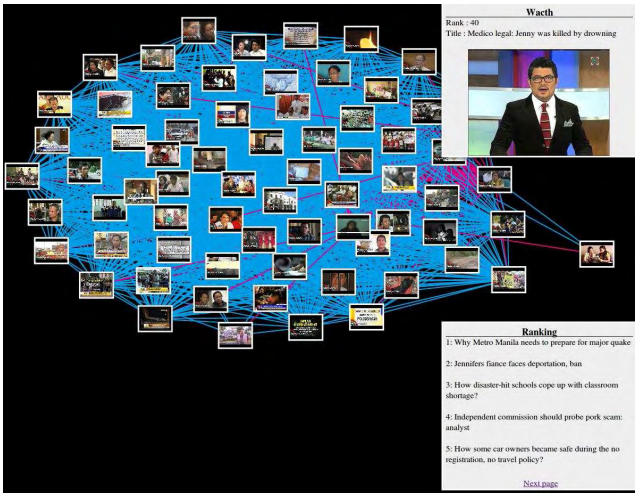
(b)



(c)

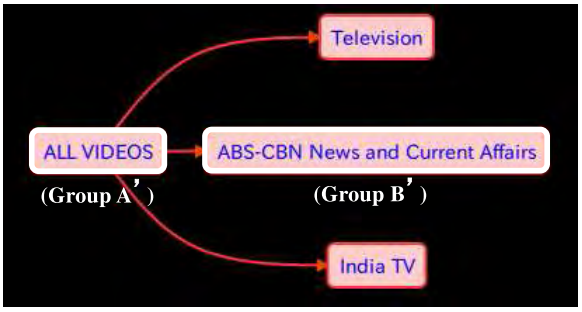


(d)

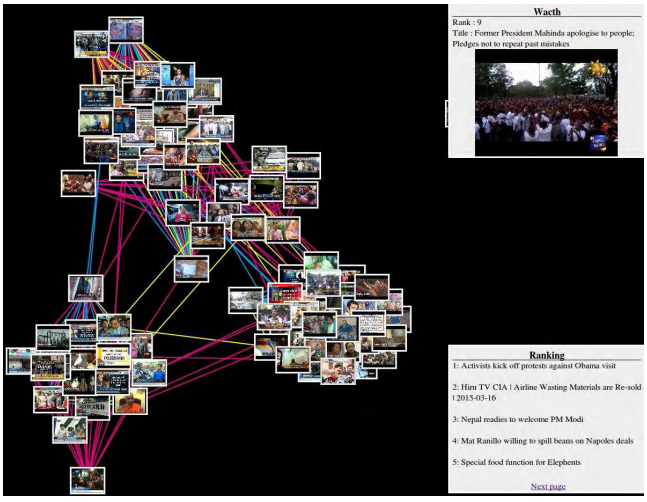


(e)

FIGURE 3. Visualization results for the hierarchical structure of Web video groups for dataset 1, which was obtained by (P) where $N_n = 0.01$. (a): Hierarchical structure of Web video groups. Each square and words in the square denote Web video groups and the most frequent entity in the group, respectively. (b), (c), (d) and (e): Groups A, B, C and D shown in (a), respectively. Red, blue and yellow lines denote edges that link nodes with positive/positive valence, negative/negative valence and positive/negative valence, respectively.



(a)



(b)



(c)

FIGURE 4. Visualization results for the hierarchical structure of Web video groups for dataset 1, which was obtained by (R2) where $N_n = 0.01$. (a): Hierarchical structure of Web video groups. (b) and (c): Groups A′ and B′ shown in (a), respectively. Notation of figures such as squares and edge colors is the same as in Fig. 3.

can be improved by our newly introducing sentiment-aware signed network (see the results for larger N_n).

C. DISCUSSION

Several visualization results are shown to qualitatively verify the effectiveness of our method. Results for (P) and (R2) where $N_n = 0.01$ are shown in Figs. 3 and 4, respectively, to further discuss the robustness to noisy edges. To improve the visibility, we highlighted edges between nodes with positive/positive valence, negative/negative valence and positive/negative valence in red, blue and yellow, respectively. From Fig. 3, it can be seen that users can obtain an overview of many Web videos through the hierarchical structure to retrieve the desired Web videos. It is notable that our method can successfully separate Web video groups with positive valence (see Fig. 3 (d)) and those with negative valence (see Fig. 3 (e)). Indeed, the Web video group shown in Fig. 3 (d) contained topics such as "marriage of influential people in India" and "cricket world cup opening ceremony;" whereas the group depicted in Fig. 3 (e) included topics such as "child abuse" and "murder case." Since there coexisted red, blue and yellow edges in the Web video group shown in Fig. 4, we can see that (R2) cannot distinguish Web videos with positive/negative valence. Therefore, (R2) cannot successfully extract the hierarchical structure (see Fig. 4 (a)) and caused performance degradation as shown in Fig. 2. Hence, these visualization results can confirm that signed edges derived from sentiment analysis are useful for separating Web videos with different topics even if noisy edges exist.

As a consequence of the above experiments, we can confirm that the proposed method in which the sentiment-aware signed network is newly adopted enables more accurate extraction of the hierarchical structure for Web video retrieval.

VI. CONCLUSIONS

In this paper, a novel method for extracting the hierarchical structure of Web video groups based on sentiment-aware signed network analysis has been presented to realize Web video retrieval. The contributions of this paper are two-fold: (1) the first attempt to utilize sentiment analysis for Web video grouping and (2) the derivation of a novel algorithm for analyzing a weighted signed network derived from sentiment and multimodal features. Since our method can separate Web videos with different topics by signed edges, our method can successfully extract the hierarchical structure in order to retrieve the desired Web videos even if noisy edges exist. Results of experiments using a new benchmark dataset,

YouTube-8M, have proven the hypothesis that sentiment in multimedia contents has an influence on their topics and have confirmed the improvement in performance by extracting the hierarchical structure for Web video retrieval.

APPENDIX

In this appendix, we describe the details of derivation of ΔQ^+ in Eq. (11). From Eq. (7), we can rewrite Q^+ as follows.

$$\begin{aligned} Q^+ &= \frac{1}{2e^+}\sum_{i,j}\Big(e_{ij}^+ - \frac{e_i^+ e_j^+}{2e^+}\Big)\delta(c_i, c_j) \\ &= \frac{1}{2e^+}\sum_{l}\sum_{i\in c_l,\, j\in c_l}\Big(e_{ij}^+ - \frac{e_i^+ e_j^+}{2e^+}\Big) \\ &= \frac{1}{2e^+}\sum_{l}\Big(\sum_{i\in c_l,\, j\in c_l} e_{ij}^+ - \frac{\big(\sum_{i\in c_l} e_i^+\big)^2}{2e^+}\Big). \end{aligned} \tag{14}$$

If we focus on the first and second terms in parentheses in Eq. (14), then we can obviously derive the gain of Q^+, *i.e.*, ΔQ^+, when a node v_i is assigned to a Web video group including the neighborhood node v_j. Furthermore, we can easily obtain ΔQ^- in the same manner.

REFERENCES

[1] V. Turner, J. Gantz, D. Reinsel, and S. Minton. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. Accessed on 2014. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

[2] J. Gantz and D. Reinsel. *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. Accessed on 2012. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[3] I.-H. Jhuo and D. T. Lee, "Video event detection via multi-modality deep learning," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 666–671.

[4] L. Yu, Y. Yang, Z. Huang, P. Wang, J. Song, and H. T. Shen, "Web video event recognition by semantic analysis from ubiquitous documents," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5689–5701, Dec. 2016.

[5] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann, "Complex event detection via multi-source video attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2627–2633.

[6] X.-J. Chen, Y.-Z. Zhan, J. Ke, and X.-B. Chen, "Complex video event detection via pairwise fusion of trajectory and multi-label hypergraphs," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 15079–15100, 2016.

[7] M. Haseyama, T. Ogawa, and N. Yagi, "A review of video retrieval based on image and video semantic understanding," *ITE Trans. Media Technol. Appl.*, vol. 1, no. 1, pp. 2–9, 2013.

[8] A. Hanjalic, "Multimedia retrieval that matters," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 1s, pp. 44:1–44:5, 2013.

[9] J. Wang, X. Zhu, and S. Gong, "Video semantic clustering with sparse and incomplete tags," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3618–3624.

[10] Y. Wang, M. Belkhatir, and B. Tahayna, "Near-duplicate video retrieval based on clustering by multiple sequence alignment," in *Proc. ACM Multimedia Conf.*, 2012, pp. 941–944.

[11] U. Gargi, W. Lu, V. Mirrokni, and S. Yoon, "Large-scale community detection on YouTube for topic discovery and exploration," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 486–489.

[12] A. Hindle, J. Shao, D. Lin, J. Lu, and R. Zhang, "Clustering Web video search results based on integration of multiple features," *World Wide Web*, vol. 14, no. 1, pp. 53–73, 2011.

[13] R. Harakawa, T. Ogawa, and M. Haseyama, "Accurate and efficient extraction of hierarchical structure of Web communities for Web video retrieval," *ITE Trans. Media Technol. Appl.*, vol. 4, no. 1, pp. 49–59, 2016.

[14] R. Harakawa, T. Ogawa, and M. Haseyama, "A Web video retrieval method using hierarchical structure of Web video groups," *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 17059–17079, 2016.

[15] R. Harakawa, T. Ogawa, and M. Haseyama, "Extraction of hierarchical structure of Web communities including salient keyword estimation for Web video retrieval," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1021–1025.

[16] J. Sang and C. Xu, "Browse by chunks: Topic mining and organizing on Web-scale social media," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7S, no. 1, pp. 30:1–30:18, 2011.

[17] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "ViBE: A compressed video database structured for active browsing and search," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 103–118, Feb. 2004.

[18] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "On clustering and retrieval of video shots through temporal slices analysis," *IEEE Trans. Multimedia*, vol. 4, no. 4, pp. 446–458, Dec. 2002.

[19] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "On clustering and retrieval of video shots," in *Proc. ACM Multimedia Conf.*, 2001, pp. 51–60.

[20] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 293–305, Mar. 2002.

[21] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 7:1–7:30, 2010.

[22] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. ACM Multimedia Conf.*, 2013, pp. 459–460.

[23] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *Comput. Res. Repository*, vol. abs/1410.8586, pp. 1–7, Oct. 2014.

[24] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[25] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 2837–2841.

[26] S. Sager *et al.*, "Audiosentibank: Large-scale semantic ontology of acoustic concepts for audio content analysis," *Comput. Res. Repository*, vol. abs/1607.03766, pp. 1–10, Jul. 2016.

[27] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 216–225.

[28] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proc. ACM SIGCHI*, 2010, pp. 1361–1370.

[29] Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang, "Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2013, pp. 657–666.

[30] P. Esmailian and M. Jalili, "Community detection in signed networks: The role of negative ties in different scales," *Sci. Rep.*, vol. 5, p. 14339, Sep. 2015.

[31] S. Abu-El-Haija *et al.*, "YouTube-8m: A large-scale video classification benchmark," *Comput. Res. Repository*, vol. abs/1609.08675, pp. 1–10, Sep. 2016.

[32] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[33] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity," *New J. Phys.*, vol. 9, no. 176, pp. 604–632, 2007.

[34] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *Comput. Res. Repository*, vol. abs/1405.4053, pp. 1–9, May 2014.

[35] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *Proc. IEEE Int. Workshop Quality Service*, Jun. 2008, pp. 229–238.

[36] J. Cao, Y. Zhang, R. Ji, F. Xie, and Y. Su, "Web video topics discovery and structuralization with social network," *Neurocomputing*, vol. 172, no. 8, pp. 53–63, 2016.

[37] T. Sugihara, X. Liu, and T. Murata, "Community detection from signed networks," (in japanese), *Trans. Jpn. Soc. Artif. Intell.*, vol. 28, no. 1, pp. 67–76, 2013.

[38] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech.*, vol. 2008, no. 10, p. P10008, 2008.

[39] S. Gomez, P. Jensen, and A. Arenas, "Analysis of community structure in networks of correlated data," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, p. 016114, Mar. 2009.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[42] S. Bird, "Nltk: The natural language toolkit," in *Proc. COLING/ACL Interactive Presentation Sessions*, Jul. 2006, pp. 69–72.
[43] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti, "General optimization technique for high-quality community detection in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, pp. 012811-1–012811-8, Apr. 2014.



RYOSUKE HARAKAWA (S'13–M'16) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 2013, 2015, and 2016, respectively, all in electronics and information engineering. He is currently a Post-Doctoral Fellow with the Graduate School of Information Science and Technology, Hokkaido University. His research interests include audiovisual processing and Web mining. He is a member of the IEICE and Institute of Image Information and Television Engineers.



TAKAHIRO OGAWA (S'03–M'08) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 2003, 2005, and 2007, respectively, all in electronics and information engineering. He is currently an Associate Professor with the Graduate School of Information Science and Technology, Hokkaido University. His research interests are multimedia signal processing and its applications. He has been an Associate Editor of the *ITE Transactions on Media Technology and Applications*. He is a member of the EURASIP, IEICE, and Institute of Image Information and Television Engineers.



MIKI HASEYAMA (S'88–M'91–SM'06) received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively, all in electronics. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Graduate School of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEICE, ITE, and the Information Processing Society of Japan IPSJ. She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), the Editor-in-Chief of the *ITE Transactions on Media Technology and Applications*, the Director of the International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE).