Multimed Tools Appl
DOI 10.1007/s11042-017-5404-4

Tracking Topic Evolution via Salient Keyword Matching with Consideration of Semantic Broadness for Web Video Discovery

Ryosuke Harakawa · Takahiro Ogawa · Miki Haseyama

Received: 27 February 2017 / Revised: 16 September 2017 / Accepted: 8 November 2017

Abstract A method to track topic evolution via salient keyword matching with consideration of semantic broadness for Web video discovery is presented in this paper. The proposed method enables users to understand the evolution of topics over time for discovering Web videos in which they are interested. A framework that enables extraction and tracking of the hierarchical structure, which contains Web video groups with various degrees of semantic broadness, is newly derived as follows: Based on network analysis using multimodal features, *i.e.*, features of video contents and metadata, our method extracts the hierarchical structure and salient keywords that represent contents of each Web video group. Moreover, salient keyword matching, which is newly developed by considering salient keyword distribution, semantic broadness of each Web video group and initial topic relevance, is applied to each hierarchical structure obtained in different time stamps. Unlike methods in previous works, by considering the semantic broadness as well as the salient keyword distribution, our method can overcome the problem of the desired semantic broadness of topics being different depending on each user. Also, the initial topic relevance enables correction of the gap from an initial topic at the start of tracking. Consequently, it becomes feasible to track the evolution of topics over time for finding Web videos in which the users are interested. Experimental results for real-world datasets containing YouTube videos verify the effectiveness of the proposed method.

Keywords Web video · video retrieval · topic evolution · tracking algorithm · network analysis

R. Harakawa · T. Ogawa · M. Haseyama
Graduate School of Information Science and Technology,
Hokkaido University, Sapporo, Japan
Tel.: +81-11-706-6078
Fax: +81-11-706-7369
E-mail: {harakawa, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

1 Introduction

With the proliferation of Web videos [16, 52], there is an increasing number of Internet users who are retrieving Web videos that include topics in which they are interested by using video hosting services such as YouTube[1]. Current retrieval methods usually target a static database containing Web videos collected at one time stamp [1, 31, 34, 41].When performing retrieval, users input queries into the retrieval systems and Web videos associated with the input queries are returned as ranked lists to users. Recent progress of semantic understanding including object recognition and event detection [11, 30, 39, 44] has enabled a continuous increase in retrieval performance for a static database [18, 19, 25, 28].

In previous papers [6, 8, 10, 47, 55], retrieval methods that focus on a dynamic database containing Web videos collected at multiple time stamps have been proposed. These methods extract Web video groups, *i.e.*, Web video sets with similar topics at each time stamp, and track changes in contents of Web video groups over time. Thus, it becomes possible to understand trends and users can successfully find their desired Web videos.

However, the approaches mentioned above have the following problems:

Problem (i)
If users do not input suitable queries, the ranked list-based retrieval methods [1, 31, 34, 41] provide many irrelevant Web videos in high ranked results [25].

Problem (ii)
Since these methods [1,31,34,41] for a static database do not target a dynamic database, users cannot find the desired Web videos by understanding trends over time. For example, for Web videos with the topic "American president" as the retrieval targets, users may want information about "changing history of presidents" or "popularity change of a certain president"; however, the above methods cannot meet such information needs.

Problem (iii)
Although Problem (ii) can be solved by approaches using Web video groups for a dynamic database [6, 8, 10, 47, 55], these approaches still have the following limitation: Generally, the desired degrees of semantic broadness of topics differ depending on each user [23]. However, these methods are based on flat clustering and thus may provide Web video groups including topics with semantic broadness that do not correspond to the user's desired videos. Therefore, these methods work well only if the user can input accurate queries that identify the user's desired semantic broadness.

To solve these problems, this paper presents a novel method that enables tracking of topic evolution considering users' desired semantic broadness via extraction of the hierarchical structure of Web video groups. In this paper, the hierarchical structure denotes the property of Web video groups being divided into sub-groups. Several methods have been proposed for extracting the hierarchical structure [20, 21, 23, 24, 45, 46, 48, 51] and for tracking topic evolution of Web video groups [6, 8, 10, 47, 55]; however, our novel method enables simultaneous realization

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

of extraction and tracking of topic evolution in the hierarchical way. For each time stamp, our novel method extracts the hierarchical structure and salient keywords that represent contents of each Web video group on the basis of network analysis [4] using multimodal features, *i.e.*, features of video contents and metadata. Here, we adopt a scheme that enables direct comparison between heterogeneous features via empirical distribution functions [53] to successfully extract the hierarchical structure and salient keywords. By providing the hierarchical structure rather than ranked lists to users, Problem (i) can be solved. Moreover, for solving Problem (ii), salient keyword matching has been newly developed by considering the salient keyword distribution; semantic broadness; and initial topic relevance to track evolution of topics over time. Here, we can solve Problem (iii) by considering the semantic broadness [14] of each Web video group as well as the salient keyword distribution. Also, for realizing accurate tracking, the initial topic relevance enables correction of the gap from an initial topic at the start of tracking. Consequently, our method enables users to understand the evolution of topics over time for finding Web videos in which they are interested.

The preliminary version of this work can be found in a conference paper [22]. This work newly introduces the initial topic relevance in the algorithm to correct the deviation from the initial topic associated with the user's selected Web video group when performing tracking over time. Also, the effectiveness of our new algorithm is shown by performing more comprehensive experiments for real-world datasets and confirming the superiority to the previous work [22].

The rest of this paper is organized as follows: In Section 3, a method for extracting the hierarchical structure and salient keywords of Web video groups at each time stamp is described. In Section 4, we present a novel method for tracking topic evolution via the obtained hierarchical structure of Web video groups with salient keywords over time. Section 5 shows experimental results for real-world datasets containing YouTube videos to confirm that our method enables tracking of topic evolution over time for finding Web videos that the users are interested in.

2 Related Work

Previously reported methods related to our work are shown in this section. We first explain clustering-based Web video retrieval methods and then describe researches on topic detection and tracking (TDT) for Web videos.

2.1 Related Work of Clustering-based Web Video Retrieval

To retrieve desired Web videos even if users cannot input suitable queries, clustering-based methods have been proposed. Flat clustering-based methods using textual features and users' viewing behavior [17, 32]; visual features [27]; or audiovisual features as well as metadata attached to Web videos [26] have been developed. These methods can assist users to find desired Web videos by providing an outline of retrieval results including varied topics. However, the obtained clusters are presented to users without focusing on the hierarchical relationships between topics. Therefore, as topics contained in the presented clusters become

varied, it becomes difficult to find desired Web videos since users need to search for clusters containing topics with desired semantic broadness from many clusters.

To solve this problem, methods based on hierarchical clustering have been proposed. Sang *et al.* [48] proposed a method to retrieve desired Web videos by presenting the hierarchical topic structure obtained via textual features and WordNet [42]. Ngo *et al.* [45,46] targeted sports videos and proposed methods to present clusters with the hierarchical structure whose upper layers contain similar color shots and lower layers contain shots with similar motion features. Taskiran *et al.* [51] proposed "a similarity pyramid" to browse video shots at various levels of detail by using visual features. To overcome the performance limitation of these methods using a single modality, we proposed methods utilizing multimodal features [20,21,23,24]. These methods enable users to retrieve Web videos containing topics with desired semantic broadness by hierarchically providing Web video groups. However, these methods do not target a dynamic database, users cannot find the desired Web videos by understanding trends over time.

2.2 Related Work of TDT for Web Videos

To overcome the problem mentioned above, methods of TDT for Web videos have been developed. As a pioneer work, a method [37] for topic discovery and tracking through the hierarchical steps: coarse topic filtering, fine topic re-ranking and topic tracking on the basis of the bipartite graph model was proposed. Reed *et al.* [47] proposed a text analysis scheme to detect emerging trends on YouTube videos by applying strongly connected component (SCC) decomposition [5] to a graph that represents co-occurrence of emerging tags. Xie *et al.* [55] proposed a method to extract visual memes, *i.e.*, frequently reposted short video segments via a network model considering color correlogram [29] for tracking and monitoring of events on YouTube. Shao *et al.* [49,50] proposed a star-structured K-partite graph to represent multimodal features for Web video topic discovery. A method [10] using auxiliary information obtained from Google Trends as well as multimodal features was presented to detect Web video topics. Moreover, Cao *et al.* [6, 8] proposed methods to discover hot topics in Web videos and understand their evolution using social metadata as well as multimodal features. Also, Li *et al.* [36] implemented a topic tracker on the basis of a semi-supervised multi-class multi-feature scheme with historical data. With the development of social media, we can expand the definition of Web videos; thus, Lu *et al.* [38] proposed a joint clustering algorithm using a K-partite graph for automatic discovering, tracking and summarizing of video topics from social media streams on Weibo.

The above methods can be useful for obtaining desired Web videos considering topic evolution or trends; however, these methods do not focus on the hierarchical characteristics of topics. Thus, these methods may provide Web video groups including topics with semantic broadness that do not correspond to desired Web videos and work well only if a user can input suitable queries that identify the user's desired semantic broadness. By contrast, our work is the first work to simultaneously realize extraction and tracking of topic evolution in the hierarchical way for Web video discovery. An overview of the proposed method is shown in Fig. 1. In the subsequent sections, we explain the details.

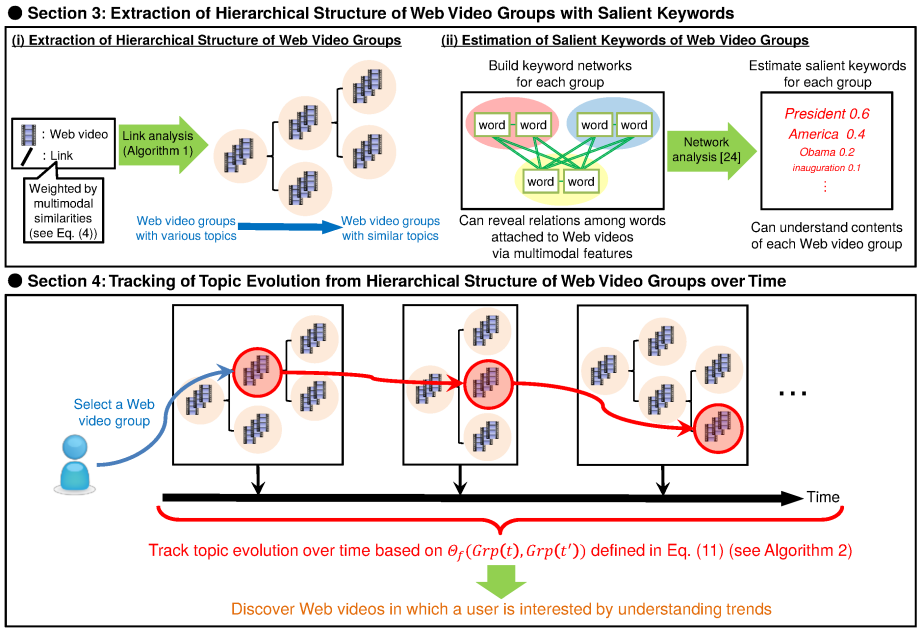


Fig. 1 Overview of the proposed method for tracking topic evolution via salient keyword matching with consideration of semantic broadness for Web video discovery.

3 Extraction of Hierarchical Structure of Web Video Groups with Salient Keywords

In this section, a method to extract the hierarchical structure of Web video groups and estimate salient keywords of them on the basis of the preliminary version [22] of this work is described.

3.1 Calculation of Similarities between Web Videos

In this subsection, we explain a scheme for calculating similarities between Web videos by the collaborative use of multimodal features in a statistical way. Let us denote Web videos by f_i ($i = 1, 2, \cdots, N$; N being the number of Web videos). Also, we represent keyframes of f_i as u_{q_i} ($q_i = 1, 2, \cdots, M_i$; M_i being the number of keyframes of f_i), and denote Web video features obtained from u_{q_i} by $\boldsymbol{x}_{q_i}^m$ ($m = 1, 2, \cdots, J$; m representing modalities and J being the number of modalities). Furthermore, for obtaining a discriminative feature vector for each Web video and for reducing computational cost of the subsequent processing, we calculate a vector of Bag of Features (BoF) [13] by using $\boldsymbol{x}_{q_i}^m$. Specifically, we train codebooks by applying k-means clustering [40] to the feature vectors and calculate histogram vectors by selecting the codebook that is most similar to each feature vector. The obtained vectors are denoted by $\boldsymbol{v}_i^m$ ($m = 1, 2, \cdots, J$).

Next, we derive similarities that enable direct comparison between heterogeneous modalities via empirical distribution functions [53]. Specifically, we calculate distances between the feature vectors for each modality, which are denoted by $d^m(i,j)$, by the following equation:

$$d^m(i,j) = \|\boldsymbol{v}_i^m - \boldsymbol{v}_j^m\|, \tag{1}$$

where $m = 1, 2, \cdots, J$, $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, N$. 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 combinations of different Web videos). Next, we construct the empirical distribution functions $F^m(x)$ of $d^m(l)$ ($m = 1, 2, \cdots, J$) as follows [53]:

$$F^m(x) = \frac{1}{N_{comb}} \sum_{l=1}^{N_{comb}} X_l^m(x), \tag{2}$$

$$X_l^m(x) = \begin{cases} 1 & \text{if } \ d^m(l) \leq x \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Furthermore, similarities s_{ij} between f_i and f_j are defined as follows:

$$s_{ij} = \max_{m=1,2,\cdots,J} [1 - F^m\{d^m(i,j)\}], \tag{4}$$

where $i = 1, 2, \cdots, N$, $j = 1, 2, \cdots, N$. Thus, comparison between heterogeneous features becomes feasible since we equalize the occurrence probability of similarities in a constant interval of the similarity-axis. Moreover, the statistically most similar feature is selected from the heterogeneous features for deriving the similarities adaptively. Note that the features that have high similarities differ depending on the video; therefore, it is reasonable to adaptively calculate similarities on the basis of the optimal feature selection via the empirical distribution functions.

3.2 Extraction of Hierarchical Structure of Web Video Groups

After obtaining multimodal similarities, we extract the hierarchical structure of Web video groups, *i.e.*, Web video sets with similar topics. First, we construct link relationships between Web videos by using metadata "related videos". We introduce the metadata into our method since the metadata is useful for associating Web videos that are similar to each other [12]. Note that most of popular video hosting services such as YouTube[1]; Dailymotion[2]; Veoh[3]; and Vimeo[4] provide the metadata "related videos". Even if "related videos" cannot be obtained, it is reported that other metadata such as "uploaders" and "tags" are also utilized to associate with similar Web videos [8]. In this paper, we consider that a Web video f_i links to a Web video f_j if "related videos" of f_i include f_j. Then we construct *a video network* G_v whose nodes and edges are Web videos and weighted

[2] http://www.dailymotion.com/
[3] http://www.veoh.com/
[4] https://vimeo.com/

links, respectively. The edge weight w_{ij} between f_i and f_j $(i = 1, 2, \cdots, N, j = 1, 2, \cdots, N)$ is defined as follows:

$$w_{ij} = \begin{cases} 2s_{ij} & \text{if } (f_i \text{ links to } f_j) \text{ AND } (f_j \text{ links to } f_i) \\ s_{ij} & \text{if } (f_i \text{ links to } f_j) \text{ XOR } (f_j \text{ links to } f_i). \end{cases} \tag{5}$$

If f_i and f_j do not link to each other, an edge between f_i and f_j is not made. Note that there is a case in which unreliable metadata causes an unsuitable network structure for extracting the hierarchical structure, and thus it becomes necessary to reconstruct G_v. Specifically, we construct the empirical distribution function $F_{\mathrm{sim}}(x)$ of Web video similarities s_{ij} in the same manner as that in Eqs. (2) and (3). Then the edge between f_i and f_j is added if $F_{\mathrm{sim}}(s_{ij})$ is more than ϵ_1 and removed if $F_{\mathrm{sim}}(s_{ij})$ is less than ϵ_2. Since the edge construction/removal have to applied to Web videos with statistically remarkable similarity/dissimilarity, we set ϵ_1 and ϵ_2 to 0.99 and 0.01, respectively, in the experiment shown later. In this way, by statistically monitoring similarities between Web videos and reconstructing link relationships, we solve the problem that some metadata is miss-labeled or some important information is missing from the metadata.

By using the obtained video network G_v, the hierarchical structure of Web video groups is extracted via a network analysis algorithm [4], which is called *Louvain method*, based on recursive modularity optimization. It should be noted that the algorithm [4] is a very fast method by which 118 million nodes can be processed in 152 minutes. Also, our proposed similarities can be easily introduced into this algorithm. For these reasons, we consider that this algorithm is optimal for accurate and efficient extraction of the hierarchical structure of Web video groups. The hierarchical structure is extracted by iterating the following two phases:

Local maximization of modularity
In the first phase, each node f_i $(i = 1, 2, \cdots, N)$ is assigned to each Web video group. For each node f_i, the gain of modularity Q when a node f_i is set to a Web video group including a neighbourhood node f_j is evaluated, and then f_i is re-assigned to a Web video group for which the positive gain is maximum. Modularity Q is an evaluation measure for detecting the group structure from a network, which is defined as follows [4]:

$$Q = \frac{1}{2\hat{m}} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_{ij} - \frac{\hat{k}_i \hat{k}_j}{2\hat{m}}) \hat{\delta}_{ij}, \tag{6}$$

where $2\hat{m} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, $\hat{k}_i = \sum_{j=1}^{N} w_{ij}$ and $\hat{\delta}_{ij}$ is 1 if f_i and f_j belong to the same Web video group and 0 otherwise. Therefore, the higher the modularity is, the better the detection results of Web video groups are. This process is applied to all nodes iteratively and sequentially until no more improvement of modularity can be obtained.

Updating a new network
In the second phase, a new network whose nodes are the Web video groups obtained by the first phase is constructed. Here, the edge weight between the two new nodes is the sum of the edge weights of the original network found during the first phase. Also, each new node has a self-loop that is derived from the weighted edges of the

corresponding original nodes obtained in the first phase. In this paper, a pair of the first and second phases is represented as "a pass" and this iteration number is denoted by q ($= 1, 2, \cdots, C_h; C_h$ being the number of all passes).

To extract the hierarchical structure, the passes (the first phase for detecting Web video groups from the network and the second phase for constructing the new network) are iterated until no more improvement in positive gain of modularity 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. In this paper, we denote the obtained Web video groups by $Grp^q_{n_q}(n_q = 1, 2, \cdots, D_q;\ D_q$ being the number of Web video groups in which the iteration count is q). Note that the number of new nodes recursively decreases in the second phase according to the increase in q; thus, efficient extraction of the hierarchical structure becomes feasible even when a large number of Web videos are targeted. The detailed algorithm for extracting the hierarchical structure is shown in Algorithm 1.

3.3 Estimation of Salient Keywords of Web Video Groups

To assist users in finding the desired Web videos easily, we need to enable users to understand the contents in each Web video group at a glance when providing the extracted hierarchical structure. To meet this necessity, we estimate salient keywords to identify contents of each Web video group. From each Web video group $Grp^q_{n_q}$ ($q = 1, 2, \cdots, C_h$, $n_q = 1, 2, \cdots, D_q$), we construct *a keyword network* $G^q_{n_q} = (V^q_{n_q}, E^q_{n_q})$ whose nodes are words obtained from texts attached to Web videos. The edge weight $(e^q_{n_q})_{ij}$ from $(v^q_{n_q})_i \in V^q_{n_q}$ to $(v^q_{n_q})_j \in V^q_{n_q}$ is defined by the following equation:

$$(e^q_{n_q})_{ij} = \sum_{m \in (K^q_{n_q})_i} \sum_{n \in (K^q_{n_q})_j} s_{mn}\ \delta_{mn}, \tag{7}$$

where $(K^q_{n_q})_i$ represents a set of Web videos that have a word $(v^q_{n_q})_i$, and δ_{mn} is 1 if m links to n or m is the same as n and 0 otherwise. In this way, we can obtain $G^q_{n_q}$ that enables association of related words on the basis of multimodal features, and thus successful extraction of salient keywords becomes feasible via a network analysis scheme [33]. Specifically, we define a matrix $\boldsymbol{L}^q_{n_q}$ whose (i, j)-th element corresponds to $(e^q_{n_q})_{ij}$, and eigenvectors of $(\boldsymbol{L}^q_{n_q})^{\mathrm{T}}\boldsymbol{L}^q_{n_q}$ are obtained. Each element of the eigenvectors of $(\boldsymbol{L}^q_{n_q})^{\mathrm{T}}\boldsymbol{L}^q_{n_q}$ represents the attribution degree of each keyword to keyword sets with similar contents [33]. By focusing on the salient keywords with a high degree of attribution, it becomes possible to understand the contents of each Web video group at a glance.

Note that, as shown in Eqs. (5) and (7), we use video features to weight edges of the video and keyword networks (see Section 5.1 for details of video features used in the experiment). Since unreliable metadata may cause performance degradation of network analysis, we use the video features as well as metadata to accurately group Web videos with similar contents or to extract salient keywords related to the same contents. Performance improvement by introducing the video features into network analysis is reported in papers [20, 21, 23].

Algorithm 1 : Extraction of the hierarchical structure of Web video groups based on *Louvain method* [4].

Input: A video network G_v whose nodes and edges are Web videos f_i ($i = 1, 2, \cdots, N$) and weighted links, respectively.
Output: The hierarchical structure of Web video groups, $Grp^q_{n_q}$ ($q = 1, 2, \cdots, C_h,\ n_q = 1, 2, \cdots, D_q$).

1: Assign each node f_i ($i = 1, 2, \cdots, N$) to each Web video group.
2: Set the index of pass as $q \leftarrow 1$.
3: **while** True **do**
4: **while** Improvement of modularity Q of the video network G_v is obtained **do**
5: **/* Local maximization of modularity */**
6: **for** each node **do**
7: Evaluate the gain of Q when a node is set to each Web video group including neighbourhood nodes.
8: Re-assign a node to the Web video group for which the positive gain of Q is maximum.
9: **end for**
10: Calculate Q of G_v.
11: **end while**
12: Denote the obtained Web video group by $Grp^q_{n_q}$ ($n_q = 1, 2, \cdots, D_q$).
13: **if** Improvement of Q of G_v is not obtained **then**
14: Break the while loop.
15: **end if**
16: **/* Updating a new network */**
17: Build the new network G_v with a self-loops whose nodes are the obtained Web video groups, where each edge weight is defined by the sum of the edge weights of the original network.
18: $q \leftarrow q + 1$
19: **end while**
20: Return the hierarchical structure of Web video groups, $Grp^q_{n_q}$ ($q = 1, 2, \cdots, C_h,\ n_q = 1, 2, \cdots, D_q$).

4 Tracking of Topic Evolution from Hierarchical Structure of Web Video Groups over Time

In this section, we describe the development of a new method to track the evolution of topics over time for finding Web videos in which the users are interested.

4.1 Overview of the Tracking Scheme

Figure 2 shows an overview of our proposed tracking scheme. First, a user selects a Web video group associated with the desired topics from the hierarchical structure. Next, our scheme sequentially searches the adjacent hierarchical structure for a Web video group that is similar to the selected one over time; thus, it becomes feasible to track evolution of topics. As a result, it becomes possible to understand

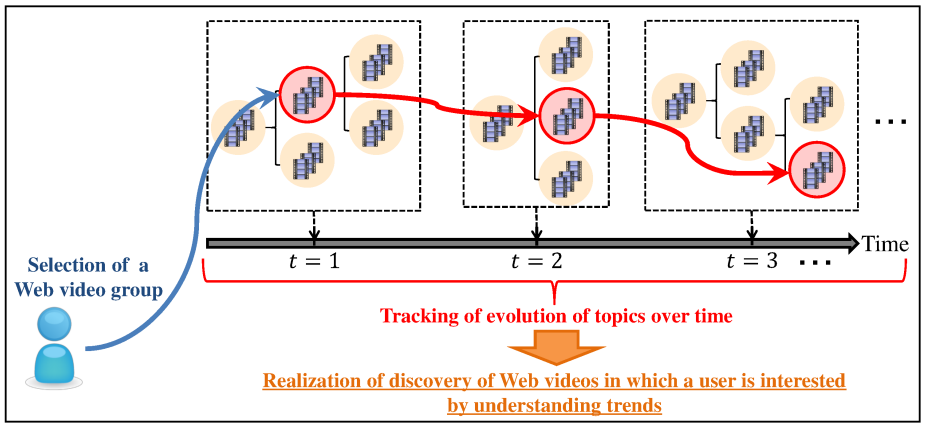


Fig. 2 Outline of the tracking of topic evolution from the hierarchical structure of Web video groups over time.

evolution of topics over time and find Web videos in which the user is interested. The proposed scheme is explained in detail below.

4.2 Tracking of the Evolution of Topics over Time

First, let us denote a set of time stamps by $\Omega = \{1, 2, \cdots, T\}$, and $t \in \Omega$ represents each time stamp. In the proposed method, a user selects a Web video group including the desired topic by browsing the hierarchical structure at $t = 1$. Next, we track the evolution of topics including the selected Web video group. This scheme consists of the two processes: (i) salient keyword matching using semantic broadness and (ii) tracking of topic evolution over time. The details of these two processes are shown below.

4.2.1 Salient Keyword Matching Using Semantic Broadness

To perform tracking of topic evolution over time, we should accurately calculate similarities between Web video groups in each hierarchical structure obtained in different time stamps. We introduce three measures, *i.e.*, salient keyword distribution; semantic broadness; and initial topic relevance into the proposed method. Each of them is explained below.

Salient keyword distribution
Let $Grp(t)$ $(t \in \Omega)$ be a Web video group within the hierarchical structure obtained in a time stamp $t \in \Omega$, and $\mathcal{K}(Grp(t))$ represents a set of keywords attached to Web videos in $Grp(t)$. Also, $u \in \mathcal{K}(Grp(t))$ denotes a keyword included in $Grp(t)$. In our method, we define the similarity based on the salient keyword distribution, which is denoted by $S_k(Grp(t), Grp(t'))$, by the following equation:

$$S_k(Grp(t), Grp(t')) = \frac{\sum_u \sum_{u'} a(u, Grp(t))\ a(u', Grp(t'))\ \tilde{\delta}_{uu'}}{\sqrt{\sum_u a(u, Grp(t))^2} \sqrt{\sum_{u'} a(u', Grp(t'))^2}}, \tag{8}$$

where $a(u, Grp(t))$ and $a(u', Grp(t'))$ are the attribution degrees of u and u' to $Grp(t)$ and $Grp(t')$, respectively, and $\tilde{\delta}_{uu'}$ is 1 if u and u' are the same and 0 otherwise. Note that this similarity reflects not only just co-occurrences of the salient keywords but also multimodal features; therefore, the similarity enables comparison between two Web video groups by considering video contents. Even if this similarity can reflect the video contents, however, there is a case in which only the use of $S_k(Grp(t), Grp(t'))$ unsuitably associates Web video groups in the upper hierarchies since such groups have too many common salient keywords.

Semantic broadness

To solve this problem, we have to consider semantic broadness of each Web video group. As the paper [14] showed that entropy of tag collections can represent semantic broadness, we introduce entropy into the proposed method. We define the similarity based on semantic broadness of Web video groups, which is represented as $S_b(Grp(t), Grp(t'))$, as follows:

$$S_b(Grp(t), Grp(t')) = \exp(-\frac{\|H(Grp(t)) - H(Grp(t'))\|^2}{2\sigma^2}), \tag{9}$$

where $H(Grp(t))$ and $H(Grp(t'))$ are entropy of keywords that appear in $Grp(t)$ and $Grp(t')$, and σ is a pre-defined constant. Furthermore, by considering both the salient keyword distribution and semantic broadness of each Web video group, the similarity $\Theta(Grp(t), Grp(t'))$ is calculated as follows:

$$\Theta(Grp(t), Grp(t')) = S_k(Grp(t), Grp(t')) \times S_b(Grp(t), Grp(t')). \tag{10}$$

Thus, it becomes feasible to suitably compare two Web video groups since we can consider the hierarchical levels of each Web video group as well as the salient keyword distribution.

Initial topic relevance

Since topic evolution is tracked over time, there may be a gradual accumulation of errors that occur when calculating similarities between adjacent Web video groups. In such a case, the topics including final tracking results will be greatly different from the initial topics. To solve this problem, we newly introduce a term to correct the above errors, namely "initial topic relevance". The introduction of initial topic relevance is one of the novelties of our method compared to our previous one [22]. We define initial topic relevance as $\Theta(Grp(1), Grp(t'))$, where $Grp(1)$ denotes the user's selected Web video group including the initial topic. Also, $Grp(t')$ represents a Web video group of a tracking target. Finally, similarities between two Web video groups $\Theta_f(Grp(t), Grp(t'))$ are defined by the following equation:

$$\Theta_f(Grp(t), Grp(t')) = \alpha \cdot \Theta(Grp(t), Grp(t')) \times (1 - \alpha) \cdot \Theta(Grp(1), Grp(t')), \tag{11}$$

where α $(0 < \alpha < 1)$ is a parameter that determines the influence on the initial topic relevance. If this parameter is set to a small value, irrelevant tracking results tend to increase, although those results may include some discoveries related to the initial topic. On the other hand, if this parameter is set to a large value, new discoveries related to the initial topic may decrease, but accurate tracking results can be obtained. In the experiments shown later, we tested several settings of this parameter. As a result of introduction of initial topic relevance, successful topic evolution tracking, which is explained in the next section, becomes feasible.

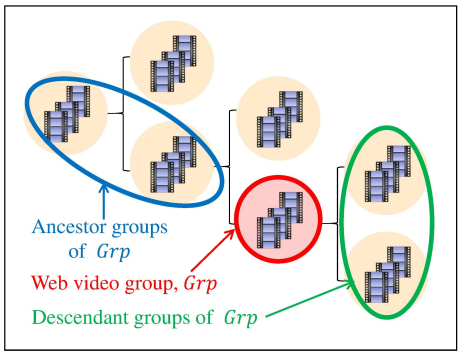


Fig. 3 Example of ancestor and descendant groups. Given a Web video group Grp with a red circle, Web video groups highlighted by blue and green become ancestor and descendant groups, respectively. The set of ancestor and descendant groups is denoted by $A(Grp)$ in this paper.

4.2.2 Tracking of Topic Evolution Over Time

Next, we propose an algorithm for tracking the evolution of topics included in a Web video group selected in the first phase by utilizing the newly derived similarities. For a Web video group selected as a tracking target, the proposed algorithm sequentially searches the hierarchical structure in the adjacent time stamp for relevant Web video groups. Here, by considering the hierarchical structure and semantic broadness of each Web video group (see Eq. (9)), we can obtain successful tracking results even if the desired semantic broadness of topics is different depending on each user. This is the novelty and contribution of this paper in comparison with conventional studies [6, 8, 10, 47, 55] on Web video tracking, which do not consider semantic broadness.

Furthermore, the use of initial topic relevance (second term on the right side of Eq. (11)) can reduce mistracking results. In particular, this effect will become large when tracking is performed over a long time since the initial topic relevance can correct the gap between topics including a tracking target and the user's selected Web video group at $t = 1$. The introduction of initial topic relevance is the novelty and contribution in comparison with the preliminary work [22] of this paper.

Consequently, by providing the tracking results to a user, it becomes possible to understand the evolution of topics associated with the user's selected Web video group over time in order to find Web videos in which the user is interested. The detailed processing of the tracking is shown in Algorithm 2. Note that, in Algorithm 2, Th is a threshold to decide whether a Web video group is tracked or not. Since the suitable numbers of tracking results, which are exhibited to users, are different depending on each user, we assume that the best Th should be selected by each user. It should be also noted that, for each Web video group, the groups in the upper and lower hierarchies are denoted by ancestor and descendant groups, respectively. We denote a set of ancestor and descendant groups of a Web video group Grp' by $A(Grp')$ (see Fig. 3).

Algorithm 2 : Tracking of the hierarchical structure of Web video groups over time.

Input: Web video group sets $\mathcal{S}_\tau$ in the hierarchical structure obtained in time stamps $\tau = 1, 2, \cdots, T$, an original tracking target, *i.e.*, a Web video group $Gr \in \mathcal{S}_1$, and a threshold Th. Here, the larger τ becomes, the newer the time stamp becomes.

Output: Tracking results, *i.e.*, Web video group sets $\mathcal{R}_\tau$ ($\tau = 2, 3, \cdots, T$).

Initialize $Src \leftarrow \{Gr\}$.
for $\tau = 2, 3, \cdots, T$ **do**
  Initialize $\mathcal{R}_\tau \leftarrow \phi$.
  for $Grp \in Src$ **do**
    for $Grp' \in \mathcal{S}_\tau$ **do**
      Calculate $\Theta_f(Grp, Grp')$ defined in Eq. (11).
      if $\Theta_f(Grp, Grp') > Th$ **then**
        if $A(Grp') \cap R_\tau \neq \phi$ **then**
          Add the element of $A(Grp') \cup \{Grp'\}$ with the maximum similarity Θ_f to R_τ and remove the other elements from R_τ.
        else
          Add Grp' to $\mathcal{R}_\tau$.
        end if
      end if
    end for
  end for
  if not $\mathcal{R}_\tau = \phi$ **then**
    $Src \leftarrow \mathcal{R}_\tau$
  end if
end for
return $\mathcal{R}_\tau$ ($\tau = 2, 3, \cdots, T$)

5 Experimental Results

This section shows experimental results for real-world datasets to verify the effectiveness of the proposed method for tracking topic evolution over time.

5.1 Datasets

We performed experiments for the following two datasets containing YouTube videos:

Dataset 1

We used a public dataset containing YouTube videos, namely, MCG-WEBV 1.0 [7, 9]. The dataset is widely utilized for the evaluation of Web video research such as classification [54]; tag refinement [56]; retrieval [15]; and topic detection [6, 8]. As shown in Table 1, the dataset contains "most viewed videos" on each month. In this experiment, we used the following two kinds of feature vectors: First, we used high-level features (HLF346 features) calculated for keyframes of Web

videos. We utilized keyframes and HLF346 features calculated from them, which were available on MCG-WEBV 1.0 [7, 9]. This feature is a high-level feature with 346 dimensions obtained per keyframe. Each dimension indicates the prediction score of a detector for 346 concepts[5]. During the training phase of concept detectors, development data of TRECVID 2011 SIN Task[6] was used. Furthermore, we extracted SIFT spatial bag of words features, and employed the Sequential Boosting SVM model proposed by CMU-informedia team[7] to obtain the concept detectors. Second, we used textual feature vectors whose dimensions were 300 obtained by applying the Doc2Vec approach [35] to titles, tags and description of each Web video. Doc2Vec is an algorithm that learns fixed-length feature representations from variable-length pieces of texts. This algorithm is realized on the basis of a neural network that predicts words in the documents. It is reported that Doc2Vec features can overcome the weakness in Bag-of-Words (BoW) features, which ignore the context. Here, model training was performed for all texts obtained from the whole dataset.

It should be noted that HLF346 features were calculated for multiple keyframes per Web video. Thus, on the basis of a BoF approach (see Section 2.1), we summarized multiple feature vectors into one vector for each Web video. Here, we constructed the codebook with 1000 bins for BoF by applying k-means clustering [40] to the HLF 346 feature vectors obtained from the whole dataset and defining cluster centers as the codebook. On the other hand, we did not apply a BoF approach to textual feature vectors since one textual feature vector was obtained per Web video.

Dataset 2

We prepared another dataset by crawling YouTube videos on our own. Concretely, we input a keyword "Tokyo Olympic (in Japanese)" as a query and obtained the top 50 Web videos. By repeatedly obtaining 10 Web videos contained in "related videos' of the selected Web videos, the dataset was constructed (see Table 2)[8].

For each Web video in this dataset, we calculated visual and textual features. To obtain visual features, we first divided Web videos into shots through a shot segmentation method [43]. Then, for each shot, we calculated HSV color histogram with 48 bins every second and obtained the vector medians [2]. Furthermore, by summarizing multiple vectors into one vector whose dimension was 1000 for each Web video via a BoF approach, we obtained visual feature vectors. On the other hand, we calculated textual feature vectors on the basis of Doc2Vec approach [35] in the same manner as textual feature vectors for dataset 1.

[5] http://mcg.ict.ac.cn/mcg-webv.files/downloads/MCG-WEBV-2012Update-HLF346-readme.pdf

[6] ttp://www-nlpir.nist.gov/projects/tv2011/tv2011.html#sin

[7] L. Bao, SI. Yu, and A. Hauptmann, CMU-informedia@TRECVID 2011 Semantic Indexing.Available: http://wwwnlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.cmu.sin.slides.pdf

[8] We collected only Web videos with lengths of less than 1800 seconds.

Table 1 Dataset 1 obtained from MCG-WEBV 1.0 [9].

	Crawling time stamp	Num. of Web videos
Dataset 1-1	December 2008	1315
Dataset 1-2	January 2009	1204
Dataset 1-3	February 2009	1141
Dataset 1-4	March 2009	1333
Dataset 1-5	April 2009	1439
Dataset 1-6	May 2009	1382
Dataset 1-7	June 2009	1087
Dataset 1-8	July 2009	1351
Dataset 1-9	September 2009	1369
Dataset 1-10	October 2009	1405
Dataset 1-11	November 2009	1446

Table 2 Dataset 2 obtained by crawling YouTube videos through a keyword "Tokyo Olympic (in Japanese)".

	Crawling time stamp	Num. of Web videos
Dataset 2-1	September 2015	3003
Dataset 2-2	November 2015	3001
Dataset 2-3	January 2016	3000
Dataset 2-4	March 2016	3005
Dataset 2-5	May 2016	3001

5.2 Evaluations

We verify the effectiveness of our method for tracking topic evolution over time. First, we extracted the hierarchical structure of Web video groups with salient keywords[9] by setting ϵ_1 and ϵ_2 to 0.99 and 0.01, respectively. Furthermore, we tracked evolution of topics associated with Web video groups in datasets 1-1 and 2-1.

In this experiment, we denote our method by (P) and set α in Eq. (11) to 0.2, 0.4, 0.6 and 0.8 to verify the parameter sensitivity. Here, (P) with $\alpha \in \{0.2,\ 0.4,\ 0.6,\ 0.8\}$ are denoted by (P-1), (P-2), (P-3) and (P-4), respectively. We compare them with the following reference methods:

(R1) Salient keyword distribution + Semantic broadness
This is a conventional method [22] that ignores the initial topic relevance $\Theta(Grp(1), Grp(t'))$, *i.e.*, the second term on the right side of Eq. (11). This method utilizes the salient keyword distribution $S_k(Grp(t), Grp(t'))$ (see Eq. (8)) and semantic broadness $S_b(Grp(t), Grp(t'))$ (see Eq. (9)).

(R2) Salient keyword distribution + Initial topic relevance
This is a method that ignores the semantic broadness of each Web video group, *i.e.*, $S_b(Grp(t), Grp(t'))$ defined in Eq. (9). This method uses the salient keyword distribution $S_k(Grp(t), Grp(t'))$ defined in Eq. (8) and initial topic rel-

[9] To estimate salient keywords, we lemmatized each word and removed stop words by using Natural Language Toolkit (NLTK) [3].

evance $\Theta(1, Grp(t'))$, *i.e.*, the second term on the right side of Eq. (11). We denote (R2) with $\alpha \in \{0.2,\ 0.4,\ 0.6,\ 0.8\}$ by (R2-1), (R2-2), (R2-3) and (R2-4), respectively.

(R3) Text + Visual
This is a method based on the earlier study [8]. Since the problem settings are not exactly same as ours, we adopt only the similarity proposed in the paper [8]. This method focuses on visually near-duplicate Web videos as well as textual similarities. Specifically, the similarity in Eq. (11) is computed as

$$\gamma \cdot Tex(Grp(t), Grp(t')) \times (1-\gamma) \cdot Vis(Grp(t), Grp(t')),$$

where $Tex(Grp(t), Grp(t'))$ and $Vis(Grp(t), Grp(t'))$ are textual and visual similarities between $Grp(t)$ and $Grp(t')$, respectively. Here, $Vis(Grp(t), Grp(t'))$ is defined as the maximum of similarities between Web videos included in $Grp(t)$ and $Grp(t')$. In this experiment, we used visual feature vectors that were the same as our method. Also, $Tex(Grp(t), Grp(t'))$ was computed by cosine similarities between Doc2Vec [35] features obtained from $Grp(t)$ and $Grp(t')$. According to the paper [8], γ was set to 0.8.

For (P) and (R1), we defined σ in Eq. (9) as the median of Euclidean distances between entropy of all different Web video groups in each dataset. For all of the methods explained above, we obtained tracking results by setting $Th = \eta + \lambda\hat{\sigma}$ ($\lambda \in \{2,\ 4,\ 6,\ 8\}$), where η and $\hat{\sigma}$ were the mean and standard deviation of all similarities between different Web video groups obtained from all datasets, respectively. Since it was trivial to provide too many tracking results for users to check their contents, we exhibited two groups at most as the tracking results for each Web video group. Note that it is difficult to directly implement other methods since there are no existing methods that completely correspond to the purpose of our work. Thus, we implement (R3) as motivated by the state-of-the-art tracking method [8] whose purpose is similar to that of our work. (R3) follows the idea of directly using visual and textual features, whose effectiveness is proven in the paper [8]; therefore, the results of (R3) can be regarded as comparable experimental results. The results of evaluation are shown below.

Qualitative Evaluations
Figure 4 shows the hierarchical structure of Web video groups with salient keywords for dataset 1-1. We show some tracking results for topics of politics since it is easy to discuss the trend changes. As an example, we show the tracking result for (Group A), which is shown in Fig. 4, by (P) in Fig. 5. It can be seen that our method accurately grasped topic changes from "American president, Bush" to "American president, Obama". On the other hand, there was a case in which (R1) and (R3) caused 77 and 357 results including unsuitable topics (*e.g.*, "Lady GaGa" and "UFO"). This is because these reference methods tend to accumulate gaps from the initial topic over time due to the lack of initial topic relevance (second term on the right side of Eq. (11)). Therefore, we can confirm the effectiveness of initial topic relevance, *i.e.*, the main novelty of this work.

Figures 6 and 7 show the tracking results for (Group B), which is shown in Fig. 4, by (P) and (R2), respectively. Figure 6 shows that our method reflected

people's attention of topics about "the Israel and Palestine situation" and "Relationship between Islamic countries and America". In Fig. 7, the results by (R2) included Web video groups with topics such as "Air France accidents" and "US military programs", which were irrelevant to the original tracking target. This is because this reference method tends to mistrack Web video groups including common salient keywords in any hierarchies due to the lack of semantic broadness defined in Eq. (9). Thus, the necessity of semantic broadness can be confirmed.

Subjective Evaluations

Although qualitative evaluations were performed as in previous works, *e.g.*, [8, 22], we show results of quantitative evaluations through a subjective experiment to further clarify the effectiveness of our method. In the experiments for datasets 1 and 2, we invited 16 and 17 evaluators with an educational background in information science, respectively. All of them were males and their average age was 24.3 years for dataset 1; meanwhile, two females were included and the average age was 24.5 for dataset 2. Subjects that do not have a basic computer literacy and are not familiar with information retrieval or social media websites may bias evaluations; thus, we invited such evaluators to fairly verify the effectiveness of the proposed method. Subjective evaluations for datasets 1 and 2 under this condition are shown below.

First, for each hierarchy at the initial time, we instructed the evaluators to select Web video groups in which they were interested one by one. In the example of dataset 1, the evaluators selected one Web video group per each hierarchy $q = 1$, 2 and 3 and obtained three Web video groups in total (see Fig. 4). Then the evaluators were asked to browse the overviews of tracking results for each parameter, $Th \in \{\eta + \lambda\hat{\sigma} \mid \lambda = 2, 4, 6, 8\}$, and select the parameter that provided the results that each evaluator considered most suitable. Next, we instructed the evaluators to check whether each Web video group was relevant to its parent groups (*i.e.*, Web video groups to be the tracking sources at the last time stamp) and the original tracking target. Here, the evaluators were asked to answer the following questions:

[**Question**] Perform evaluations for an overview of the tracking results. For [Q-1], answer with numbers from 1 to 4: (1) "Very irrelevant", (2) "Irrelevant", (3) "Relevant" and (4) "Very relevant". For [Q-2], [Q-3] and [Q-4], answer with numbers from 1 to 4: (1) "Very bad", (2) "Bad", (3) "Good" and (4) "Very good".

- [**Q-1**] Were the tracking results reasonable?
- [**Q-2**] Could you obtain the expected Web videos through the tracking results?
- [**Q-3**] Could you discover Web videos with topics that you did not know but were interested in?
- [**Q-4**] Could you get more information through the tracking results than from systems that do not consider trend changes?

The results for datasets 1 and 2 are shown in Figs. 8 and 9, respectively. For accurate analysis, we also show the results for datasets 1 and 2 in which the deviation of evaluations was reduced for each user in Figs. 10 and 11, respectively. Specifically, for each user, we standardized N_{eval} evaluation values given to each question, where N_{eval} = "Num. of methods" × "Num. of evaluation targets".

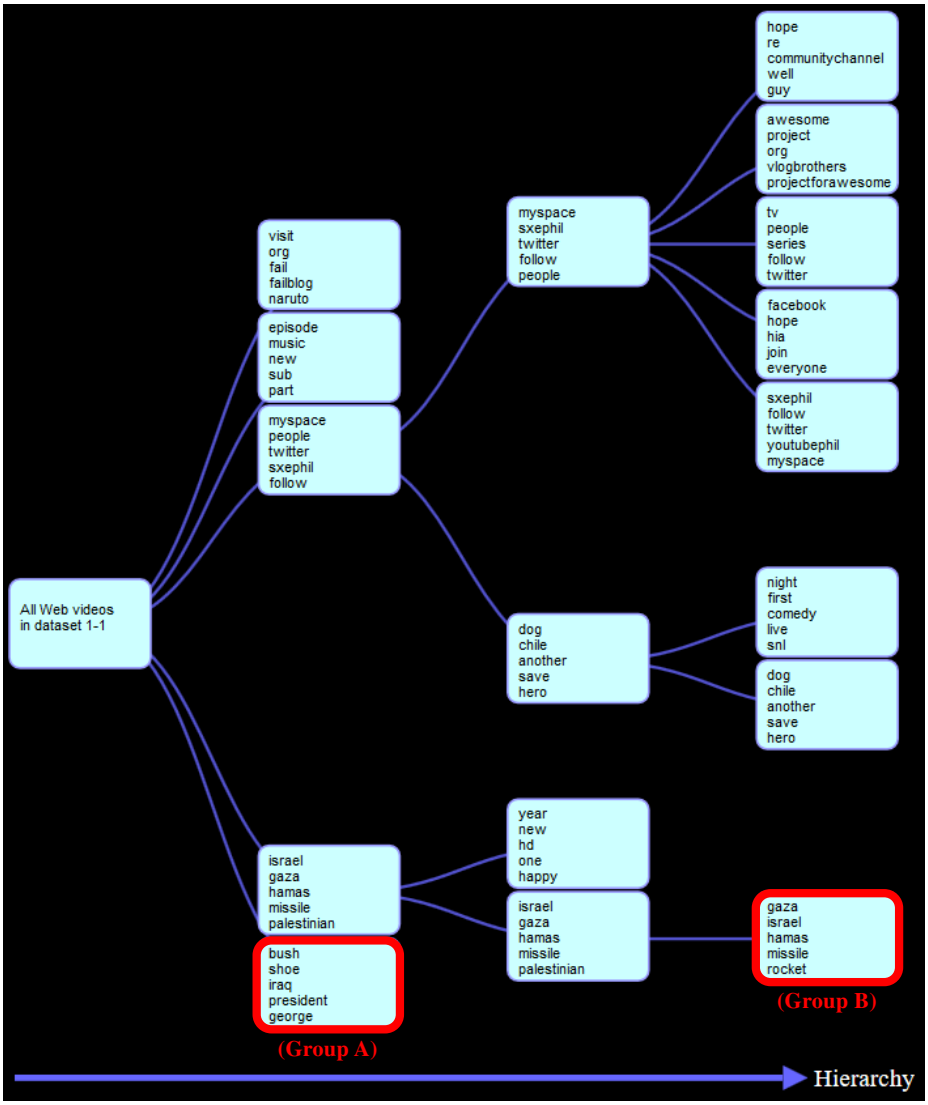
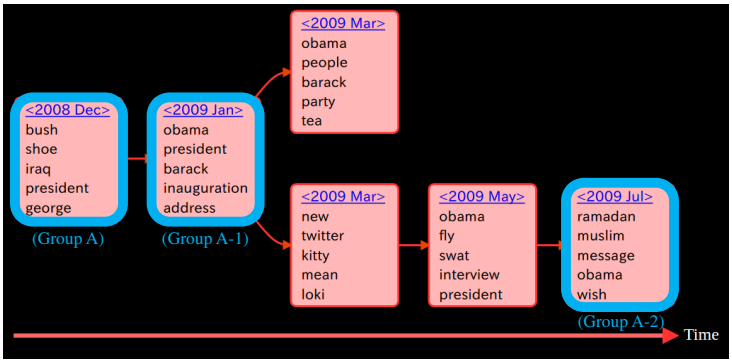


Fig. 4 Hierarchical structure of the top five largest Web video groups obtained by our method. Web video groups shown in the second, third and fourth columns correspond to ones in which $q = 1,\ 2$ and 3, respectively. The words in each square are the top five salient keywords of each Web video group obtained by our method. We show Web video groups when more than 10 Web videos were separated.

Moreover, we show details of the evaluation values shown in Figs. 8, 9, 10 and 11 in Tables 3, 4, 5 and 6, respectively. From (P-1), (P-2), (P-3) and (P-4) in these figures and tables, we can see that good results can be obtained when α, which decided the importance of initial topic relevance, was set to a value of 0.6 or less. By comparing (P-1) and (P-2) with (R1) and (R3), we can quantitatively confirm the effectiveness of initial topic relevance, *i.e.*, the main novelty of this work. In other words, it can



(a)

(b)

(c)

(d)

Fig. 5 Tracking result for (Group A) shown in Fig. 4 by (P-2) where $Th = 0.6$. (a): Overview of the results. Each square denotes Web video groups. The first lines in the squares represent the crawling time stamps and the later lines show the top five salient keywords. (b), (c) and (d): Examples of thumbnails of Web videos included in the original tracking target and some tracking results, *i.e.*, "Groups A, A-1 and A-2" shown in (a).

be found that a lack of initial topic relevance causes accumulation of unsuitable results over time even if multimodal analysis is conducted. From (P-1), (P-2), (R2-1), (R2-2), (R2-3) and (R2-4), we can numerically confirm the necessity of semantic broadness. The use of semantic broadness can lead to accurate results even if the semantic broadness of desired Web video groups is different depending on each user. If semantic broadness is not used, however, unsuitable results including common salient keywords are obtained.

Furthermore, we perform objective discussion. We verify how consistent tracking results were obtained over time by defining "consistency" for each result. For each tracking result, we first calculated mean vectors of visual and textual features of Web videos for each Web video group. Next, for each tracking result, we calculated cosine similarities between the mean vectors for all combinations between Web video groups. Finally, we obtained mean of the cosine similarities to define "consistency" of each tracking result. Table 7 shows the obtained consistency for datasets 1 and 2. We can see that methods that produce high users evaluation

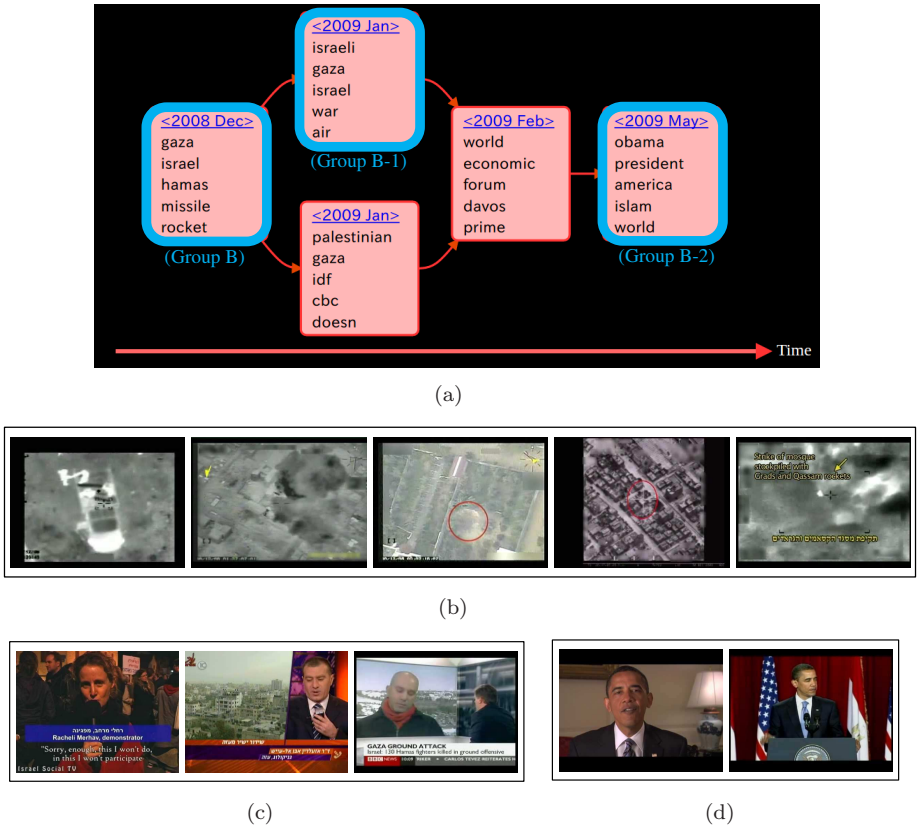
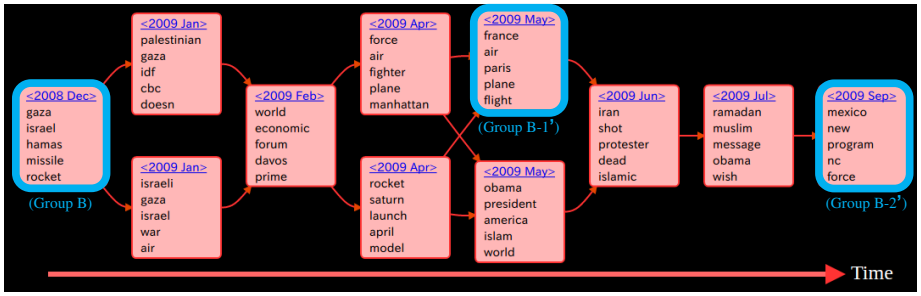


Fig. 6 Tracking result for (Group B) shown in Fig. 4 by (P-2) where $Th = 0.6$. (a): Overview of the results. Each square denotes Web video groups. The first lines in the squares represent the crawling time stamps and the later lines show the top five salient keywords. (b), (c) and (d): Examples of thumbnails of Web videos included in the original tracking target and some tracking results, *i.e.*, "Groups B, B-1 and B-2" shown in (a).

Table 3 Details of the evaluation values shown in Fig. 8. Bold emphases denote the highest evaluation values in each row.

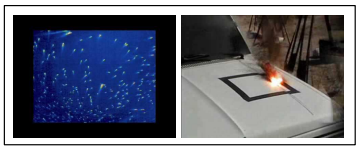
	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Q-1	2.60	**2.63**	2.19	2.00	1.81	2.31	2.15	2.08	2.08	2.08
Q-2	2.46	**2.56**	2.15	2.04	1.92	2.27	2.04	2.00	2.13	1.96
Q-3	**2.52**	2.50	2.15	2.21	1.79	2.29	1.96	2.10	2.00	1.88
Q-4	**2.56**	2.45	2.21	2.02	1.73	2.25	2.08	2.00	1.94	1.96
Average	**2.54**	**2.54**	2.17	2.07	1.81	2.28	2.06	2.05	2.04	1.97

values tend to produce high consistency, while methods that do not work well tend to produce low consistency. Thus, we consider that it is necessary for users satisfaction on Web video discovery to track Web video groups, which are similar to the original tracking target, over time. From this consideration, it can be seen that the proposed method enables successful tracking of Web video groups since



(a)

(b)



(c)

Fig. 7 Tracking result for (Group B) shown in Fig. 4 by (R2-1) where $Th = 0.4$. (a): Overview of the results. Each square denotes Web video groups. The first lines in the squares represent the crawling time stamps and the later lines show the top five salient keywords. (b) and (c): Examples of thumbnails of Web videos included in some tracking results, *i.e.*, "Groups B, B-1′" and B-2′" shown in (a). Note that thumbnails of the original tracking target, (Group B), are shown in (b) of Fig. 6.

Table 4 Details of the evaluation values shown in Fig. 9. Bold emphases denote the highest evaluation values in each row.

	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Q-1	2.85	2.88	**2.91**	2.88	2.74	**2.91**	2.71	2.76	2.74	1.50
Q-2	2.71	2.79	2.76	2.76	2.65	**2.88**	2.68	2.76	2.68	1.56
Q-3	2.50	2.59	**2.68**	2.59	**2.68**	2.62	2.53	2.44	2.56	1.85
Q-4	2.41	2.56	**2.71**	2.68	2.59	2.50	2.41	2.38	2.53	1.71
Average	2.62	2.71	**2.76**	2.73	2.65	2.73	2.58	2.59	2.63	1.65

Table 5 Details of the evaluation values shown in Fig. 10. Bold emphases denote the highest evaluation values in each row.

	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Q-1	0.45	**0.51**	0.02	-0.22	-0.44	0.10	-0.04	-0.10	-0.13	-0.16
Q-2	0.34	**0.46**	0.06	-0.08	-0.24	0.12	-0.15	-0.17	-0.05	-0.29
Q-3	**0.45**	**0.45**	-0.02	0.05	-0.41	0.19	-0.22	-0.04	-0.14	-0.31
Q-4	**0.47**	0.44	0.11	-0.10	-0.38	0.13	-0.05	-0.15	-0.25	-0.22
Average	0.43	**0.47**	0.04	-0.09	-0.37	0.13	-0.12	-0.12	-0.14	-0.24

our method can obtain more consistent tracking results over time than reference methods.

As a summary of the evaluations, we explain how "Problems (i), (ii) and (iii)" described in Section 1 were solved. We solved "Problem (i)" by providing the hierarchical structure of Web video groups. In previous works [20, 21, 23, 24], it was proven that the hierarchical structure can navigate users to the desired Web videos even if users cannot input suitable queries. We solved "Problem (ii)" by the

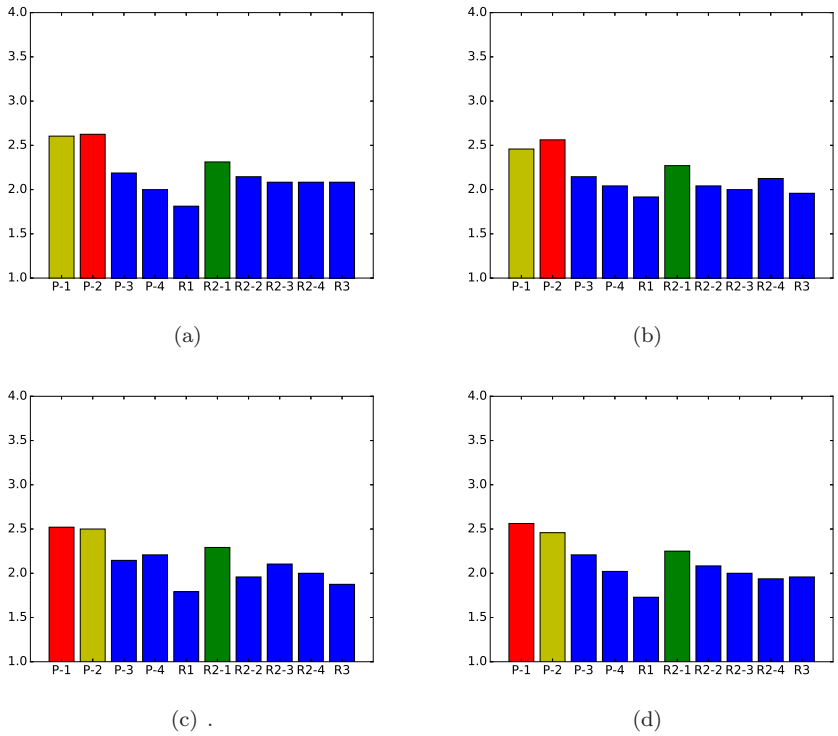
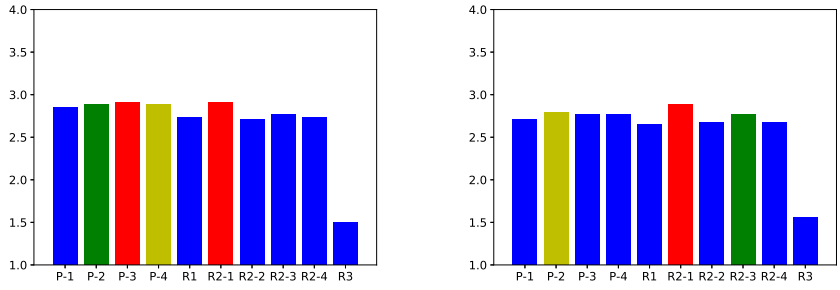


Fig. 8 Results of the subjective experiment for dataset 1. The averages of evaluation values by 16 evaluators are shown. (a), (b), (c) and (d) are the results for [Q-1], [Q-2], [Q-3] and [Q-4], respectively. The top three values are highlighted by red, yellow and green, respectively.

Table 6 Details of the evaluation values shown in Fig. 11. Bold emphases denote the highest evaluation values in each row.

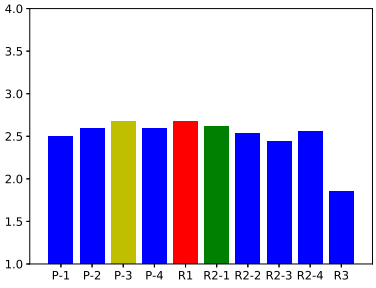
	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Q-1	0.20	0.25	**0.28**	0.25	0.02	0.23	0.08	0.15	0.07	-1.54
Q-2	0.06	0.20	0.21	0.23	0.07	**0.27**	0.09	0.21	0.06	-1.40
Q-3	-0.04	0.06	**0.31**	0.11	0.27	0.13	-0.01	-0.12	0.05	-0.77
Q-4	-0.12	0.10	**0.42**	0.32	0.22	0.02	-0.04	-0.03	0.08	-0.96
Average	0.03	0.15	**0.31**	0.23	0.15	0.16	0.03	0.05	0.06	-1.17

proposed tracking method that seeks similar Web video groups over time. From the results for [Q-4], we can see the superiority of our method compared with reference methods. We solved "Problem (iii)" by introducing semantic broadness into the proposed method. In this experiment, we showed the superiority of our method compared with reference methods that do not consider semantic broadness (see (R2) and (R3)).

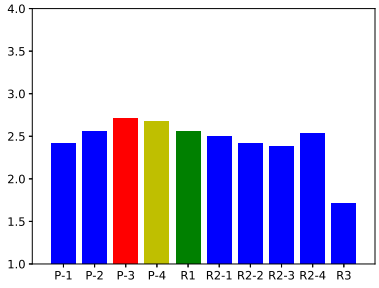


(a)

(b)



(c) .



(d)

Fig. 9 Results of the subjective experiment. The averages of evaluation values by 17 evaluators are shown. (a), (b), (c) and (d) are the results for [Q-1], [Q-2], [Q-3] and [Q-4], respectively. The top three values are highlighted by red, yellow and green, respectively.

Table 7 "Consistency" of tracking results over time. "Vis." and "Tex." show consistency for visual and textual features. Mean of consistency for all tracking results are described.

(a) Dataset 1

	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Vis.	0.327	0.283	0.206	0.181	0.156	0.253	0.226	0.191	0.164	0.095
Tex.	0.198	0.183	0.118	0.113	0.102	0.163	0.146	0.129	0.129	0.052

(b) Dataset 2

	P-1	P-2	P-3	P-4	R1	R2-1	R2-2	R2-3	R2-4	R3
Vis.	0.312	0.289	0.251	0.215	0.191	0.224	0.207	0.200	0.170	0.016
Tex.	0.725	0.729	0.725	0.700	0.654	0.622	0.623	0.612	0.584	0.142

5.3 Discussion on Computational Cost

Finally, we discuss computational cost of the proposed method. The proposed method consists of four phases, *i.e.*, (A) calculation of similarities between Web videos (Section 3.1), (B) extraction of hierarchical structure of Web video groups (Section 3.2), (C) estimation of salient keywords of Web video groups (Section 3.3) and (D) tracking of the evolution of topics over time (Section 4.2). Computational cost of each phase is explained below.

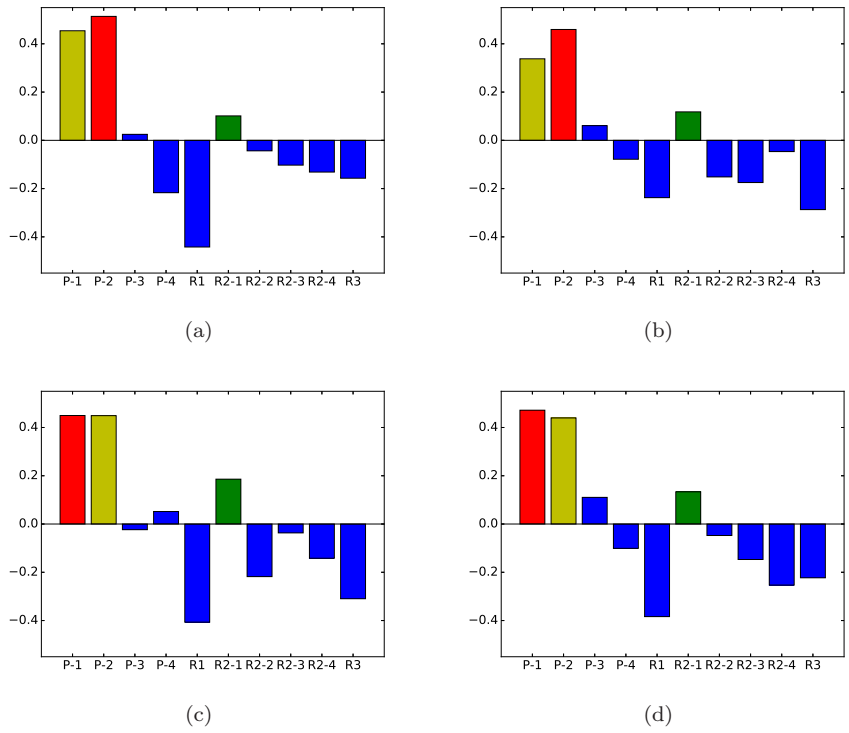


Fig. 10 Results of the subjective experiment for dataset 1. For each user, we standardized N_{eval} evaluation values given to each question, where N_{eval} = "Num. of methods" × "Num. of evaluation targets". (a), (b), (c) and (d) are the results for [Q-1], [Q-2], [Q-3] and [Q-4], respectively. The top three values are highlighted by red, yellow and green, respectively.

First, we denote a set of time stamps by $\Omega = \{1, 2, \cdots, T\}$. Then we assume that the hierarchical structure of Web video groups are extracted for each time stamp $t \in \Omega$ and evolution of topics of the hierarchical structure at $t = 1$ are tracked over time.

(A) Calculation of similarities between Web videos (Section 3.1)
We need to calculate distances for all pairs of Web videos for the similarity calculation in Eq. (4). Thus, given N videos for each time stamp $t \in \Omega$, then computational cost of this phase is $O(N^2)$.

(B) Extraction of hierarchical structure of Web video groups (Section 3.2)
For each time stamp $t \in \Omega$, we extract the hierarchical structure of Web video groups. Our method shown in Algorithm 1 is based on an efficient scheme [4] with computational cost $O(N \log N)$.

(C) Estimation of salient keywords of Web video groups (Section 3.3)
For estimating salient keywords of Web video groups, we need to calculate eigenvectors of $(\boldsymbol{L}^q_{n_q})^{\mathrm{T}} \boldsymbol{L}^q_{n_q}$ where $\boldsymbol{L}^q_{n_q}$ is a keyword network. Thus, the compu-

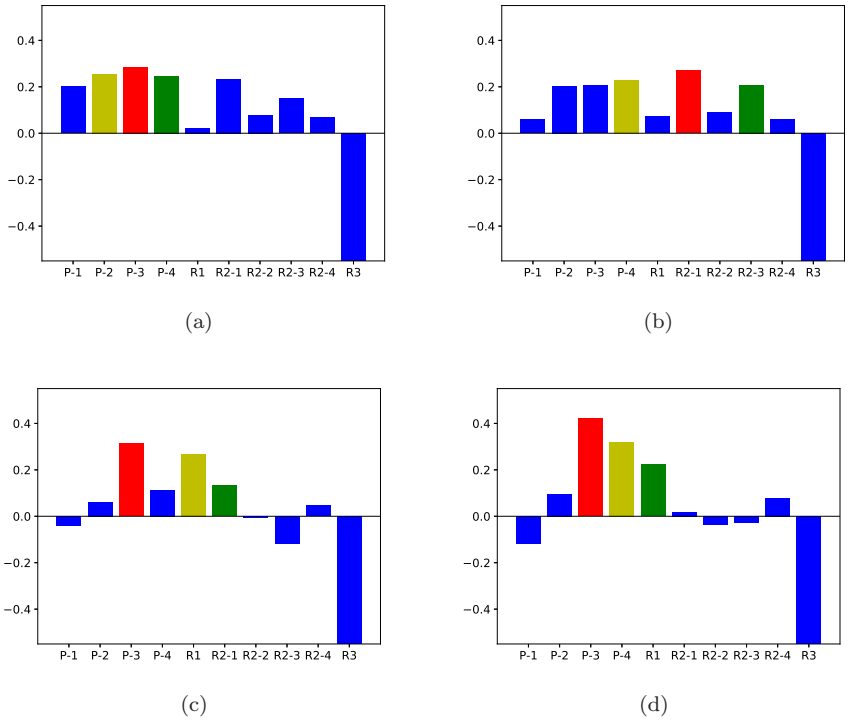


Fig. 11 Results of the subjective experiment for dataset 2. For each user, we standardized N_{eval} evaluation values given to each question, where N_{eval} = "Num. of methods" × "Num. of evaluation targets". (a), (b), (c) and (d) are the results for [Q-1], [Q-2], [Q-3] and [Q-4], respectively. The top three values are highlighted by red, yellow and green, respectively.

tational cost is the polynomial order for the number of nodes of the network.

(D) Tracking of the evolution of topics over time (Section 4.2)
As shown in Algorithm 2, for N_{tr} tracking targets, computational cost of our method is $O(N_{tr} \times (T-1) \times |Src| \times |S_\tau|)$ where Src is a set of tracking results for each step and S_τ is a set of Web video groups within the hierarchical structure at a time stamp $t = \tau$. Here, T is a constant and we can narrow down the numbers of Src and S_τ to exhibit to users even if Src and S_τ contains many Web video groups. Thus, the computational cost is $O(N_{tr})$.

When assuming that the proposed method is utilized as real-world application, (A), (B) and (C) can be calculated in advance as offline process. Although (D) can be regarded as online process, its computational cost is the linear order and the process can be easily parallelized for more efficient computation.

6 Conclusions

A method to track topic evolution via salient keyword matching with consideration of semantic broadness to find Web videos is presented in this paper. Our method enables users to understand the evolution of topics over time for discovering Web videos in which they are interested by deriving a framework to extract and track the hierarchical structure of Web video groups. Specifically, network analysis using multimodal features enables extraction of the hierarchical structure and salient keywords that represent the contents of each Web video group. Moreover, salient keyword matching by considering the salient keyword distribution, semantic broadness of each Web video group and initial topic relevance was newly derived. Unlike methods in previous works that do not consider semantic broadness and initial topic relevance, our method can accurately track topic evolution over time even if the desired semantic broadness is different depending on each user. Experimental results have qualitatively and quantitatively confirmed that our method enables discovery of Web videos that users did not know but were interested in, unlike conventional retrieval methods that do not consider trend changes over time.

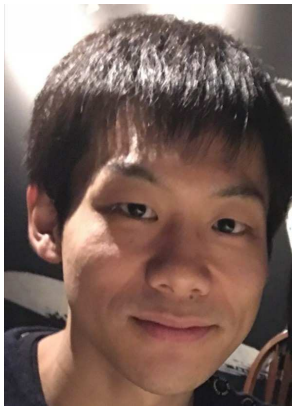
Acknowledgements This work was partly supported by JSPS KAKENHI Grant Numbers JP16J02042 and JP17H01744. We are grateful that a publisher, Springer permits us to deposit this accepted manuscript in the open access repository. The final publication is available at "http://link.springer.com/article/10.1007/s11042-017-5404-4".

References

1. Araujo, A., Chaves, J., Chen, D., Angst, R., Girod, B.: Stanford i2v: A news video dataset for query-by-image experiments. In: Proc. ACM Multimedia Systems Conf., pp. 237–242 (2015)
2. Astola, J., Haavisto, P., Neuvo, Y.: Vector median filters. In: Proc. IEEE, pp. 678–689 (1990)
3. Bird, S.: Nltk: The natural language toolkit. In: Proc. COLING/ACL Interactive Presentation Sessions, pp. 69–72 (2006)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. p. P10008 (2008)
5. Butafogo, R.A., Schneiderman, B.: Identifying aggregates in hypertext structures. In: Proc. 3rd ACM Conf. Hypertext, pp. 63–74 (1991)
6. Cao, J., Ngo, C.W., Zhang, Y.D., Li, J.T.: Tracking web video topics: Discovery, visualization, and monitoring. IEEE Trans. Circuits and Systems for Video Technology **21**(12), 1835–1846 (2011)
7. Cao, J., Zhang, Y., Ji, R., Li, X.: On application-unbiased benchmarking of web videos from a social network perspective. Multimedia Tools and Applications **75**(3), 1543–1556 (2016)
8. Cao, J., Zhang, Y., Ji, R., Xie, F., Su, Y.: Web video topics discovery and structuralization with social network. Neurocomputing **172**(C), 53–63 (2016)
9. Cao, J., Zhang, Y.D., Song, Y.C., Chen, Z.N., , X.Z., Li, J.: Mcg-webv: A benchmark dataset for web video analysis. In: Technical Report, ICT-MCG-09-001, Institute of Computing Technology (2009)
10. Chen, T., Liu, C., Huang, Q.: An effective multi-clue fusion approach for web video topic detection. In: Proc. ACM Multimedia Conf., pp. 781–784 (2012)
11. Chen, X.J., Zhan, Y.Z., Ke, J., Chen, X.B.: Complex video event detection via pairwise fusion of trajectory and multi-label hypergraphs. Multimedia Tools and Applications pp. 1–22 (2015). DOI 10.1007/s11042-015-2514-8

12. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: Proc. IEEE Int. Workshop on Quality of Service, pp. 229–238 (2008)
13. Csurka, G., Dance, C.R., L. Fan, J.W., Bray, C.: Visual categorization with bags of keypoints. In: Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–22 (2004)
14. Fang, Q., Xu, C., Sang, J., Hossain, M.S., Ghoneim, A.: Folksonomy-based visual ontology construction and its applications. IEEE Trans. Multimedia **18**(4), 702–713 (2016)
15. Feng, B., Cao, J., Chen, Z., Zhang, Y., Lin, S.: Multi-modal query expansion for web video search. In: Proc. ACM SIGIR, pp. 721–722 (2010)
16. Gantz, J., Reinsel, D.: The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. In: IDC iView (2012)
17. Gargi, U., Lu, W., Mirrokni, V., Yoon, S.: Large-scale community detection on youtube for topic discovery and exploration. In: Proc. Int. AAAI Conf. Weblogs and Social Media, pp. 486–489 (2011)
18. Greetha, P., Narayanan, V.: A survey of content-based video retrieval. Journal of Computer Science **4**(6), 474–486 (2008)
19. Hanjalic, A.: Multimedia retrieval that matters. ACM Trans. Multimedia Comput. Commun. Appl. **9**(1s), 44:1–44:5 (2013)
20. Harakawa, R., Ogawa, T., Haseyama, M.: Extraction of hierarchical structure of web communities including salient keyword estimation for web video retrieval. In: Proc. IEEE Int. Conf. Image Processing, pp. 1021–1025 (2015)
21. Harakawa, R., Ogawa, T., Haseyama, M.: Accurate and efficient extraction of hierarchical structure of web communities for web video retrieval. ITE Trans. Media Technology and Applications **4**(1), 49–59 (2016)
22. Harakawa, R., Ogawa, T., Haseyama, M.: Tracking hierarchical structure of web video groups based on salient keyword matching including semantic broadness estimation. In: Proc. IEEE Global Conf. Signal and Information Processing, pp. 1238–1242 (2016)
23. Harakawa, R., Ogawa, T., Haseyama, M.: A web video retrieval method using hierarchical structure of web video groups. Multimedia Tools and Applications **75**(24), 17,059–17,079 (2016). DOI 10.1007/s11042-015-2976-8
24. Harakawa, R., Ogawa, T., Haseyama, M.: Extracting hierarchical structure of web video groups based on sentiment-aware signed network analysis. IEEE Access **5**, 16,963–16,973 (2017)
25. Haseyama, M., Ogawa, T., Yagi, N.: A review of video retrieval based on image and video semantic understanding. ITE Trans. Media Technology and Applications **1**(1), 2–9 (2013)
26. Hatakeyama, Y., Ogawa, T., Asamizu, S., Haseyama, M.: A novel video retrieval method based on web community extraction using features of video materials. IEICE Trans. Fundamentals **E92-A**(8), 1961–1969 (2009)
27. Hindle, A., Shao, J., Lin, D., Lu, J., Zhang, R.: Clustering web video search results based on integration of multiple features. World Wide Web **14**(1), 53–73 (2011)
28. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews **41**(6), 797–819 (2011)
29. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J.: Spatial color indexing and applications. In: Int. Conf. Computer Vision (IEEE Cat. No.98CH36271), pp. 602–607 (1998)
30. Jhuo, I.H., Lee, D.T.: Video event detection via multi-modality deep learning. In: Proc. IEEE Int. Conf. Pattern Recognition, pp. 666–671 (2014)
31. Jiang, L.: Web-scale multimedia search for internet video content. In: Proc. Int. Conf. Companion on World Wide Web, pp. 311–316 (2016)
32. Kamie, M., Hashimoto, T., Kitagawa, H.: Effective web video clustering using playlist information. In: Proc. ACM Symp. Applied Computing, pp. 949–956 (2012)
33. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Jounal of ACM **46**(5), 604–632 (1999)
34. Kofler, C., Larson, M., Hanjalic, A.: Intent-aware video search result optimization. IEEE Trans. Multimedia **16**(5), 1421–1433 (2014)
35. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. Computing Research Repository **abs/1405.4053**, 1–9 (2014 (http://arxiv.org/abs/1405.4053))
36. Li, G., Jiang, S., Zhang, W., Pang, J., Huang, Q.: Online web video topic detection and tracking with semi-supervised learning. Multimedia Systems **22**(1), 115–125 (2016)
37. Liu, L., Sun, L., Rui, Y., Shi, Y., Yang, S.: Web video topic discovery and tracking via bipartite graph reinforcement model. In: Proc. ACM Int. Conf. World Wide Web, pp. 1009–1018 (2008)

38. Lu, Z., Lin, Y.R., Huang, X., Xiong, N., Fang, Z.: Visual topic discovering, tracking and summarization from social media streams. Multimedia Tools and Applications **76**(8), 10,855–10,879 (2017)
39. Ma, Z., Yang, Y., Xu, Z., Yan, S., Sebe, N., Hauptmann, A.G.: Complex event detection via multi-source video attributes. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2627–2633 (2013)
40. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, pp. 281–297 (1967)
41. Mazloom, M., Habibian, A., Snoek, C.G.M.: Querying for video events by semantic signatures from few examples. In: Proc. ACM Multimedia Conf., pp. 609–612 (2013)
42. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
43. Nagasaka, A., Tanaka, Y.: Automatic video indexing and fullvideo search for object appearances. In: Proc. IFIP 2nd Working Conf. Visual Database Systems, pp. 113–127 (1991)
44. Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., Natarajan, P.: Multimodal feature fusion for robust event detection in web videos. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1298–1305 (2012)
45. Ngo, C.W., Pong, T.C., Zhang, H.J.: On clustering and retrieval of video shots. In: Proc. ACM Multimedia Conf., pp. 51–60 (2001)
46. Ngo, C.W., Pong, T.C., Zhang, H.J.: On clustering and retrieval of video shots through temporal slices analysis. IEEE Trans. Multimedia **4**(4), 446–458 (2002)
47. Reed, C., Elvers, T., Srinivasan, P.: What's trending?: Mining topical trends in ugc systems with youtube as a case study. In: Proc. Int. Workshop on Multimedia Data Mining, pp. 4:1–4:9 (2011)
48. Sang, J., Xu, C.: Browse by chunks: Topic mining and organizing on web-scale social media. ACM Trans. Multimedia Comput. Commun. Appl. **7S**(1), 30:1–30:18 (2011)
49. Shao, J., Ma, S., Lu, W., Zhuang, Y.: A unified framework for web video topic discovery and visualization. Pattern Recognition Letters **33**(4), 410 – 419 (2012)
50. Shao, J., Yin, W., Ma, S., Zhuang, Y.: Topic discovery of web video using star-structured k-partite graph. In: Proc. ACM Int. Conf. Multimedia, pp. 915–918. ACM (2010)
51. Taskiran, C., Chen, J.Y., Albiol, A., Torres, L., Bouman, C.A., Delp, E.J.: Vibe: A compressed video database structured for active browsing and search. IEEE Trans. Multimedia **6**(1), 103–118 (2004)
52. Turner, V., Gantz, J.: The digital universe of opportunities: Rich data and the increasing value of the internet of things. In: IDC iView (2014)
53. van der Vaart, A.W.: Asymptotic statistics. Cambridge University Press (1998)
54. Wu, J., Worring, M.: Efficient genre-specific semantic video indexing. IEEE Trans. Multimedia **14**(2), 291–302 (2012)
55. Xie, L., Natsev, A., Kender, J.R., Hill, M., Smith, J.R.: Visual memes in social media: Tracking real-world news in youtube videos. In: Proc. ACM Multimedia Conf., pp. 53–62 (2011)
56. Zhang, X., Huang, Z., Shen, H.T., Yang, Y., Li, Z.: Automatic tagging by exploring tag information capability and correlation. World Wide Web **15**(3), 233–256 (2011)



Ryosuke Harakawa 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 IEEE, IEICE and Institute of Image Information and Television Engineers (ITE).

Takahiro Ogawa 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 IEEE, ACM, EURASIP, IEICE, and Institute of Image Information and Television Engineers (ITE).

Miki Haseyama 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 IEEE, IEICE, Institute of Image Information and Television Engineers (ITE), and Information Processing Society of Japan (IPSJ). She has been a Vice-President of the ITE, the Editor-in-Chief of the ITE Transactions on Media Technology and Applications, the Director of the International Coordination and Publicity of the IEICE.