Placing Videos on a Semantic Hierarchy for Search Result Navigation

SONG TAN, City University of Hong Kong
YU-GANG JIANG, Fudan University
CHONG-WAH NGO, City University of Hong Kong

Organizing video search results in a list view is widely adopted by current commercial search engines, which cannot support efficient browsing for complex search topics that have multiple semantic facets. In this paper, we propose to organize video search results in a highly structured way. Specifically, videos are placed on a semantic hierarchy which accurately organizes various facets of a given search topic. To pick the most suitable videos for each node of the hierarchy, we define and utilize three important criteria: relevance, uniqueness and diversity. Extensive evaluations on a large YouTube video dataset demonstrate the effectiveness of our approach.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Web video browsing, video organization, visualization

ACM Reference Format:

Tan, S., Jiang, Y.-G. and Ngo, C.-W. 2013.Placing Videos on a Semantic Hierarchy for Search Result Navigation ACM Trans. Multimedia Comput. Commun. Appl. 2, 3, Article 1 (May 2010), 20 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

The amount of videos available on the Web is growing explosively, which brings a grand challenge for search engines to locate and visualize contents according to user needs, particulary for search topics of complex events. For example, a search of "Beijing Olympics 2008" on YouTube will return over 200,000 videos, organized in a list where the top ones are deemed to be more relevant. Such a

This work was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119610) and a grant from the National Natural Science Foundation of China (#61228205).
Author's address: S. Tan, Creative Media Center CMC5001, Dept. of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Town, Hong Kong; email: songtan-c@my.cityu.edu.hk; Y.-G. Jiang, School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, 825 Zhangheng Road, Shanghai 201203, China; email: ygj@fudan.edu.cn; C.-W. Ngo, Creative Media Center CMC5003, Dept. of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Town, Hong Kong; email: cscwngo@cityu.edu.hk
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2010 ACM 1551-6857/2010/05-ART1 $10.00
DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

relevance-based list without any discernable structure leaves users with fragmented and incomplete understanding of the topic, as many search topics have multiple facets which are all very relevant. For this example query, 10 out of the top 20 results of the query "Beijing Olympics 2008" are about the "Opening Ceremony". Only one is about the "Venue" and the rest are about specific games. There is no coverage of "Closing Ceremony", "Bidding Process", "Torch Relay", etc., which are all important facets for users to fully understand this search topic. Although this issue could be tackled by using a specific query for a facet, it is often difficult to come up with a suitable query when users do not know much about a topic in advance. Therefore, structured browsing of search results is a highly demanded function in modern video search engines.

Apparently, manual generation of the topic structures by a search engine company is an impossible mission simply because firstly there exist too many topics and new hot topics also emerge almost daily. Secondly, new or duplicate videos are being uploaded every second. In this paper, we propose an effective and efficient framework to automate this process. Specifically, videos are placed on a predefined hierarchy that semantically organizes multiple facets of a search topic. In this way users can browse search results in a well-organized and more user-friendly fashion. The hierarchical semantic structures used here are from the Web pages in the Wikipedia knowledge base, which are generated by the powerful crowdsourcing. Pages on Wikipedia are very informative and well-organized. Figure 2 illustrates the idea of using a semantic hierarchy for organizing search results. With a simple but very informative hierarchy, users can navigate search results very flexibly based on their interests. For instance, one may follow a top-down strategy which allows first a coarse understanding with high-level introductions, and then in-depth details of various interesting aspects of the topic.

The main technical issue to realize this idea is to find the best matches between each node in the hierarchy and the videos returned from initial search, which is nontrivial due to the variability of video contents and the semantic complexity of search topics. We therefore define three criteria to achieve high quality of video-to-node matching: relevance, uniqueness, and diversity. Relevance is obviously critical in any search tasks, which is first calculated using text matching based on surrounding context (i.e., title, tag and description), and then augmented by considering both textual and visual similarities among videos in a random walk process. Uniqueness measures how unique a video is to a specific node in the hierarchy. A unique video should be highly focused on a specific node on which it is placed. The last criterion, diversity, ensures that little content redundancy shall exist in the search results, particularly among the videos attached to the same node.

Both textual and visual similarities are exploited to reach the goal of maximizing the three criteria. In addition to proposing the idea of using semantic hierarchies for video result navigation, the technical contributions of this paper include the tailored definitions of the three criteria for this task and a simple optimization framework that integrates the three criteria in a unified fashion. We conduct experiments on a large dataset with manual labels, which—upon the acceptance of this work—will be released to stimulate further research on this interesting and challenging problem.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 first gives an overview of the proposed solution and then elaborates the detailed techniques for each module of our framework. Experimental validations and comparisons are given in Section 4 and Section 5. Finally, Section 6 concludes this paper.

2. RELATED WORKS

Several works have explored the organization of Web contents or search results through classification and clustering, mostly in the text domain. Classification based organization has been extensively studied (e.g., in [Xue et al. 2008; Huang 2004], among others), which needs a large number of training samples that do not exist in our case. In image/video search, clustering is a more popular approach

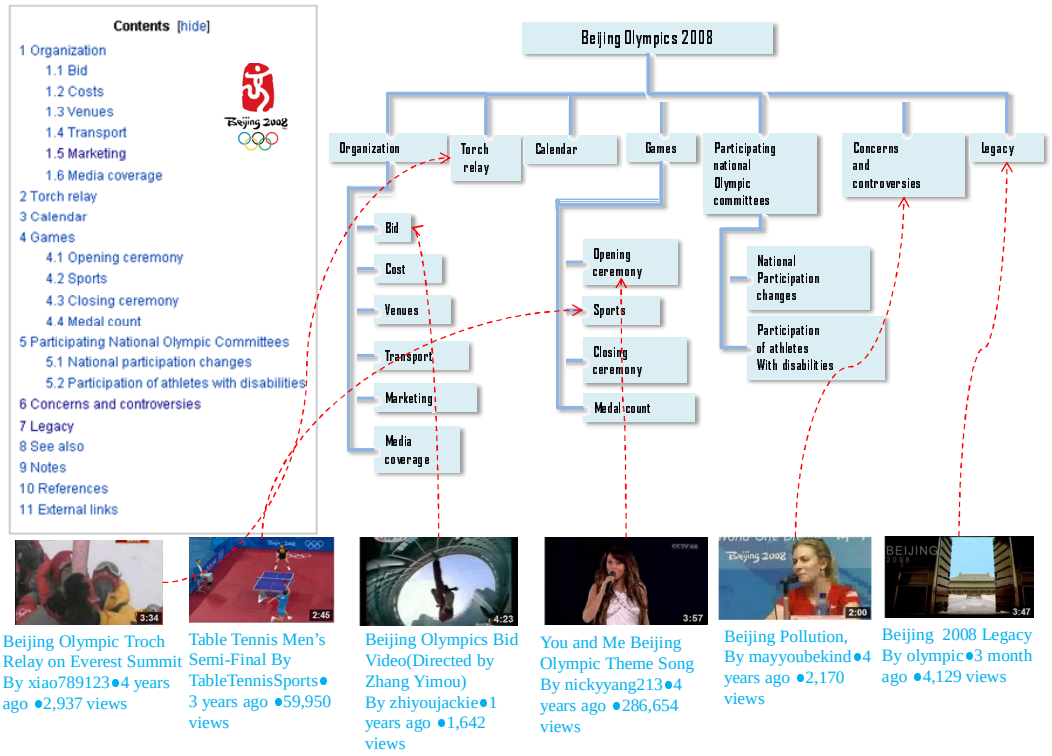

Fig. 1. Illustration of placing videos on a hierarchy from Wikipedia using an example query "Beijing Olympics 2008". This work exploits semantic hierarchies to organize video search results for better navigation.

for result organization [Cai et al. 2004; Jing et al. 2006; Slimi et al. 2013; de Rooij and Worring 2010], where visually and semantically similar images/videos were grouped together for better visualization. In [van Leuken et al. 2009], authors also emphasize on the diversification of search results during clustering. A representative image has then been selected from each cluster to form a diverse result set.

The purely data-driven approach has a clear drawback in that the resulted clusters are difficult to be interpreted semantically, which is important for user navigation of search results. In [Carmel et al. 2009], Carmel et al. proposed to use Wikipedia as background knowledge to label document clustering outputs, where candidate labels were first extracted from Wikipedia and then associated to clusters based on content similarity. This may be useful for generating good labels of the resulted clusters, but will not be helpful for user navigation if initial clusters are noisy. To alleviate this issue, Jing et al. [Jing et al. 2006] proposed to learn image cluster names (before performing clustering) by mining results of Google Web Search and PicSearch. However, the mined cluster names often contain noises, which will be propagated into image clustering, making the final results even more noisy. Our work differs from these existing works in that we directly adopt semantic hierarchies in Wikipedia to organize video search results. The Wikipedia contents are very well-organized and suitable for grouping search results of complex topics that have multiple facets. Although Wikipedia does not cover all the possible search topics, we underline that many topics like finding a movie trailer can be answered by a single video, which do not need deep result organization. The main goal of our work is to organize search results of complex topics. These topics have multiple facets and require issuing of multiple queries to

have an overview of the topics, which could be a difficult task for users who do not know exactly what the facets are.

Perhaps the work most similar to ours is from Ming et al. [Ming et al. 2010], where the Wikipedia hierarchies were used to cluster and organize web-scale collections like Yahoo! Answers. The inputs of their approach are a semantic hierarchy and a set of Query-Answer (QA) pages, and the output is the hierarchically organized collection, with each document assigned to one node in the hierarchy. The biggest difference of our work is that we focus on video search instead of organizing a QA document collection. Videos contain rich multimodal information. In addition to the textual similarities computed on noisy video contexts, we can explore video content-based features which are also important clues. Furthermore, in search result organization, only considering relevance is not sufficient, as other measures like uniqueness and diversity are all important and may affect user-experience significantly. Our approach integrates all these criteria in a unified way.

In addition, Sang and Xu [Sang and Xu 2011] also proposed to use a hierarchy to organize search results, where the topics were automatically mined from Web video collections using a probabilistic topic model. Topic labels were represented by tag clouds which are noisy and thus difficult for users to interpret. The automatically generated hierarchies are still far from satisfactory, which is the main reason that we adopt professionally edited hierarchies from Wikipedia in this work.

Research on automatic topic discovery and tracking (TDT) is also loosely related to our work, since TDT aims at discovering and associating important events of a topic in a large collection. Most of the existing TDT works were done in text domain [Chen and Chen 2008; Fung et al. 2007]. In video search, several works focused on mining important events in news videos [Wu et al. 2008; Ide et al. 2003], for which overlaid captions and text transcripts from speech recognition are available. For Web videos, Liu et al. [Liu et al. 2008] used a bipartite graph to model the relations between videos and associated texts. Graph cut was then applied to cluster videos into groups (topics), which were then further analyzed to discover interesting topics. Chen et al. [Chen et al. 2012] proposed a multi-clue fusion approach for web video topic detection. The hot tags are first extracted and then utilized as the basis for the generation of topic structures. However, these approaches again fall into the category using data-driven clustering, which often leads to noisy topics and structures. Our work avoids this drawback by directly using semantic structures on Wikipedia.

3. PLACING VIDEOS ON A SEMANTIC HIERARCHY

The proposed semantic hierarchy based search result navigation approach is depicted in Figure 2. Inputs of the framework include all the videos searchable by issuing a query topic to a search engine, and a Wikipedia article of the search topic which can be found by matching query terms to article titles in Wikipedia. The well-edited Wikipedia page gives constructive clues for organizing the search results of a corresponding topic. Notice that, although Wikipedia is adopted in this work for its suitability, the entire framework can be deployed on any semantic hierarchy that is suited for video search result organization.

A Wikipedia page is represented as a hierarchy tree with nodes capture the section information (see Figure 1 for illustration). In the key process of placing videos on each node of a hierarchy, three criteria are considered. Initial video-to-node relevance is measured through matching the contextual texts of videos and the textual descriptions in the corresponding section of the Wikipedia page. To boost the reliability of the relevance measure, the text-based estimation score is then refined by a random walk process on a graph constructed by videos, where each vertex is a video and transition probabilities are defined by both textual and visual similarities.

The relevance scores between the videos and the nodes constitute a video-node matrix, where each row vector represents how a video is related to every node in the hierarchy. The uniqueness criterion,

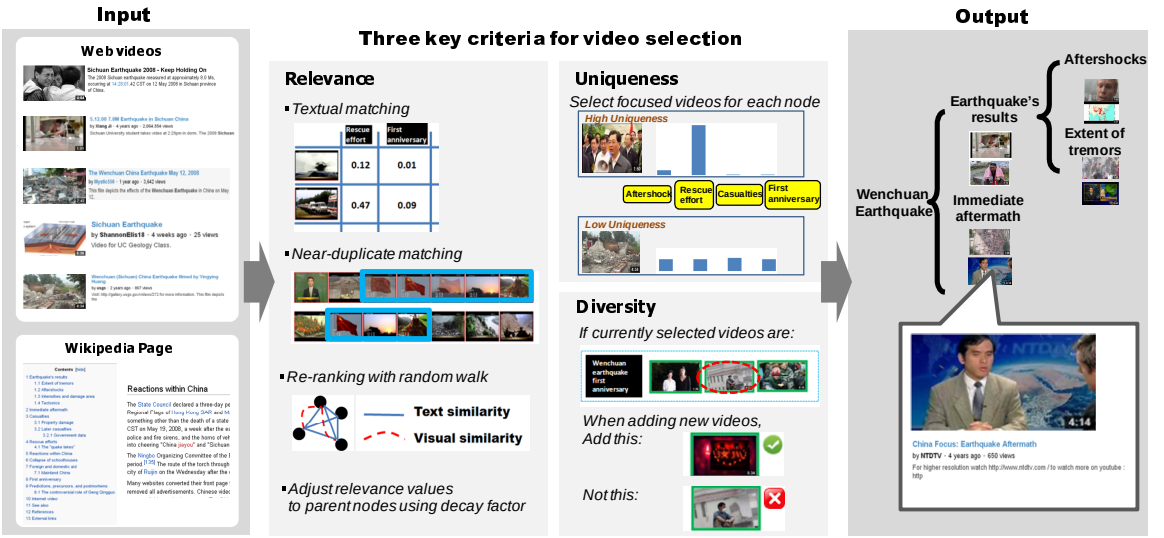

Fig. 2. The framework of our proposed approach, where candidate videos from initial search are placed on a semantic hierarchy by coherently considering three criteria: relevance, uniqueness, and diversity.

which measures exclusive content suitability, is computed based on this row vector. A video is considered less unique if its scores distribute evenly on multiple nodes. To further ensure diversity of search results, we remove duplicate videos in the final set.

Finally, these three criteria are integrated in an optimization framework to place the most suitable videos on each node of the hierarchy. In the following subsections, we describe each of the modules in this framework in detail.

3.1 Relevance

Relevance is the most important criterion in search applications. In this work, we consider both text and visual information in evaluating the relevance between videos and nodes in a semantic hierarchy.

3.1.1 *Textual Matching.* Surrounding texts of videos, such as the titles, tags and descriptions created by users, contain rich information, which are the key clue used by current commercial video search engines. As shown in Figure 2, the Wikipedia page of a topic consists of several sections/subsections (nodes). Each node corresponds to a section name, followed by several paragraphs giving highly focused details. These are used to compare with the surrounding texts of videos for textual matching.

Specifically, we first remove meaningless nodes in video search application such as “see also”, “references” and “external links”. Stop words are then removed in the main texts, after which we represent each node using the standard vector space model, using the traditional weighting scheme tf-idf. However, the quantitative digest of each node using standard tf-idf is not ideal in many cases. For instance, words in the theme song of Beijing Olympics “One World, One Dream” should not be treated independently. In addition, some words are more discriminative than others, which also cannot be reflected by typical tf-idf weighting. For example, in the topic “Beijing Olympic Torch Relay”, the document frequency[1] of “propane” and “tradition” is the same. Although subjectively “propane” could be more discriminative for the node “Torch”, the tf of “propane” is lower than “tradition” in this example. To

[1]In our application, we treat each section of a Wikipedia page as a “document”.

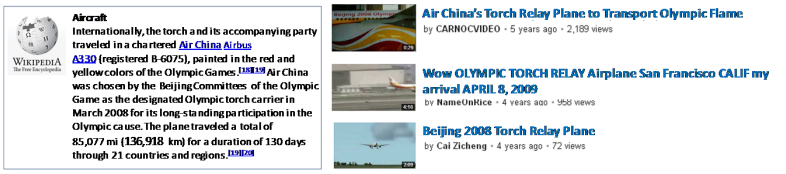

Fig. 3. Texts in a Wikipedia page and the surrounding contexts of three Youtube videos. Video contexts are sparse, noisy and subjective, creating difficulties for reliable textual matching.

tackle both issues, we make use of the hyper-links embedded on the Wikipedia pages[2], which not only can help identify salient terms like the name of the theme song, but also reflect the importance of words to some extent. We therefore analyze the html scripts to identify salient terms with hyper-links, which are then added to the vocabulary for bag-of-word representation. Since salient terms are more important, we assign them higher weights by an empirically chosen factor of five. With this node representation, video-to-node relevance can be computed by its cosine similarity with the bag-of-word representation of the video contexts.

3.1.2 *Visual Similarity by Near-duplicate (ND) Matching.* The textual contexts generated by Web users are mostly very sparse and noisy, making textual matching less reliable. Figure 3 gives one example, in which one section of the Wikipedia page introduces the "Aircraft" used for torch relay, with iconic words like "Air China" and "Aircraft". However, in the very related videos as shown in the figure, only one has the words "Air China" in its context. We mitigate this issue through using visual information.

There can be two ways of using visual matching. First, normally there exist a few images and (sometimes) very few video links on the Wikipedia pages, which can be used to match with the candidate videos. However, we found that the number of images and videos on Wikipedia is too few. As a result, the matching quality is very unreliable. The second way of using visual matching–which is adopted in this work–is to measure the similarity among the candidate videos. These similarity estimations are then used in a random walk process to select representative videos, as will be described later in next subsection. Similar idea of using random walk in visual search has been adopted in [Hsu et al. 2007; Jing and Baluja 2008].

To compute the visual similarity between two videos, we adopt a state-of-the-art solution of near-duplicate (ND) video detection, which consists of two critical stages. The first stage processes videos by uniformly sampling keyframes and extracting local interest points like SIFT [Lowe 2004]. The keyframes are represented with the widely adopted bag-of-visual-words (BoW). In BoW, a two-layer visual vocabulary tree of $20k$ visual words is adopted [Nistr and Stewnius 2006]. The BoW features of all the keyframes are then indexed by an inverted file structure for efficient retrieval. To alleviate the effect caused by quantization loss in BoW, hamming embedding [Jegou et al. 2008] is used to produce a short signature for each word.

The second stage contains ND retrieval using inverted file based BoW matching and signature verification, which returns a set of candidate ND frames for a given query frame. Purely using BoW matching is nevertheless not reliable enough because geometric information is fully ignored. We therefore employ two algorithms for geometric verification: enhanced weak geometric consistency (E-WGC) [Jegou et al. 2008; Zhao et al. 2010] and scale-rotation invariant pattern entropy (SR-PE) [Zhao and Ngo 2009]. Both E-WGC and SR-PE estimate the scale and rotation parameters of local features in two

[2]Note that only the terms associated with the hyper-links are used. The content of the pages linked by a Wikipedia page are not considered in our work, though using them may include more context to improve the performance.

keyframes, from two different views. E-WGC utilizes the by-product of local interest point detection to estimate geometry parameters, while SR-PE relies on the BoW matches for estimation. The two approaches have their own pros and cons. Given a correspondence set of two frames produced by a matching algorithm, SR-PE evaluates if the frame pair is a near-duplicate pair by evaluating the spatial regularity of the matching patterns. However, the correspondence pair can be very noisy especially if BoW retrieval is used. To alleviate this problem, EWGC is deployed to perform initial filtering by retaining only correspondences with the dominant scale, rotation and translation. This provides clean correspondence set for SR-PE to operate on. SR-PE provides a more in-depth analysis by measuring the pattern entropy of the horizontal and vertical matching pattern. EWGC, on the other hand, only selects the ones with the dominant rotation, scale and translation.

The employment of ND detection is for measuring video similarity rather than content diversity (as will be introduced in section 3.3). Based on the detected ND frame pairs, video similarity can be easily estimated based on the number of ND frames shared between a video pair, normalized by the total number of frames in the two videos. Note that the similarity measure is symmetric and we do not consider the case when one video is an excerpt of the other as in [Kender et al. 2010]. Therefore, using the measure will result in low similarity score in this case. Nevertheless, the potential bias due to symmetric measure is not important here because the uniqueness and diversity measures (sections 3.2 and 3.3) will be further employed for video selection (section 3.4).

3.1.3 *Random Walk on a Video Graph.* As mentioned above, we perform random walk on a video graph to make use of the visual similarities among videos, so that the most relevant and representative videos can be selected for each node of a given hierarchy.

Given a graph $\mathcal{G}$ with videos as vertices, the random walk process is modeled as

$$\mathcal{W} = \{\mathcal{G}, P, x_\pi\},$$

where $P = [p_{ij}]$ is the transition matrix, and x_π is a column vector containing the stationary probabilities of the videos at a state π.

The transition probability p_{ij} between two videos v_i and v_j indicates the probability of reaching v_j from v_i, which is set as:

$$p_{ij} = \frac{Sim(v_i, v_j)}{\sum_{\forall t, t \neq j} Sim(v_t, v_j)},$$

where $Sim(v_i, v_j)$ is computed by adding the textual similarity (using surrounding contexts) and visual similarity (based on ND matching) between the two videos v_i and v_j:

$$Sim(v_i, v_j) = \lambda Sim_{textual}(v_i, v_j) + (1 - \lambda) Sim_{visual}(v_i, v_j), \tag{1}$$

where $\lambda \in [0, 1]$ controls the contribution of the two modalities.

We initialize the stationary probabilities $x_{(0)}$ in $\mathcal{W}$ by the relevance values from the textual matching method described earlier (cf. Section 3.1.1), which are iteratively updated in the random walk process. At time k, the update of a video v_j is in the following form:

$$x_k(j) = \alpha \sum_{i \neq j} x_{(k-1)}(i) p_{ij} + (1 - \alpha) x_{(0)}(j) \tag{2}$$

where $\alpha \in (0, 1]$. The updates can be executed iteratively until meeting the convergence condition of $|x_{(k+1)} - x_{(k)}| \to 0$. The probabilities at the time of convergence indicate the final relevance values of the videos to a node. We run the random walk process for every node in the hierarchy.

3.1.4 *Adjusting Relevance Values to Parent Nodes.* The parent nodes in a semantic hierarchy provide a broad view of the topic. Ideally, videos mapped to a parent node should not only relevant to the parent node but also its child nodes. A straightforward way is by comparing a video to the information pooled from the parent and descendent nodes. Nevertheless, such measurement cannot truly reflect whether the video has a good coverage of all or most information from parent and descendent nodes. Therefore, the relevance of a video to a parent node is computed as the relevance score obtained by the parent node itself (if there are texts directly attached to it), added by the relevance scores to the child nodes. We then propose a decay factor DF to adjust the score of a parent node n, as well as selecting videos which are relevant to most child nodes for a parent node. Let $\mathcal{C}_n$ be the set of child nodes of a node n and p_{v,n_i} be the probability of v being relevant to a child node $n_i \in \mathcal{C}_n$, DF is defined as:

$$DF(v,n) = \frac{\sum_{n_i \in \mathcal{C}_n} p_{v,n_i} \log \frac{1}{p_{v,n_i}}}{\log |\mathcal{C}_n|}, \tag{3}$$

where p_{v,n_i} can be computed by

$$p_{v,n} = \frac{Relevance(v,n)}{\sum_{n_i \in \mathcal{C}_n} Relevance(v,n_i)},$$

where $Relevance(\cdot)$ gives the video-to-node relevance score obtained after applying random walk.

Based on the above definition, the relevance of v to the parent node n can be computed as

$$Relevance(v,n) = \sum_{n_i \in \mathcal{C}_n} Relevance(v,n_i) \times DF(v,n) + Relevance'(v,n), \tag{4}$$

where $Relevance'(v,n)$ is the relevance score computed using the parent node's own texts, which is set as 0 if there is no text directly attached to it.

Note that the decay factor is defined in an entropy form, which is an indicator for uncertainty. For a parent node having $|\mathcal{C}_n|$ child nodes, the maximum uncertainty is $\log(|\mathcal{C}_n|)$, which is used to normalize the decay factor. DF prefers videos that contain relevant contents to most of the child nodes instead of just a few, which are exactly what we want to place on the higher level parent nodes. We underline that the decay factor is useful as we want to place videos over all the nodes of a hierarchy, which enables a more comprehensive result view that allows users to navigate search results at different levels. In the case that-for instance-only leaf nodes are assigned with videos, the decay factor is not needed.

3.2 Uniqueness

Another factor to consider is how unique a video is in the search result. When selecting a video for a node in the hierarchy, we would like the video to be specifically relevant to that node, while in the meantime irrelevant to most of the other nodes. Since it is unreasonable to avoid content overlap between child nodes and parent nodes, in our work uniqueness is a measure applied particularly to ensure that no content redundancy exists among the leaf nodes of a hierarchy.

Let $\mathcal{C}_{leaf}$ be the set of all the leaf nodes, we propose the following measure to quantify the uniqueness of a video v:

$$Uniqueness(v) = \frac{\log |\mathcal{C}_{leaf}| - \sum_{n_i \in \mathcal{C}_{leaf}} p*_{v,n_i} \log \frac{1}{p*_{v,n_i}}}{\log |\mathcal{C}_{leaf}|}, \tag{5}$$

where $p*_{v,n_i}$ is the probability that v is relevant to a leaf node $n \in \mathcal{C}_{leaf}$:

$$p*_{v,n} = \frac{Relevance(v,n)}{\sum_{n_i \in \mathcal{C}_{leaf}} Relevance(v,n_i)}.$$

Apparently, the uniqueness definition in Equation 5 is in a similar form to the decay factor defined in Equation 3, but has an inverse effect of preferring videos that do not contain contents relevant to multiple nodes. In other words, if a video is relevant to every leaf node, it achieves a large entropy with a very low uniqueness value.

3.3 Diversity

The last criterion to be considered is diversity. When one browses video search results, particularly for those videos under the same node, duplicate or redundant contents are not desired. For example, under a node "Beijing Olympics Avenue", people will not like watching videos all about "the Bird's Nest", although they are highly relevant and also unique for this node.

We propose to define two levels of diversity. The first focuses on intra-node diversity of a node n, measured by the opposite definition of content redundancy in the ith video under n:

$$IntraRedundancy(v_i, n) = \max_{\{v_j \in n\}} (Sim_{visual}(v_i, v_j)), \tag{6}$$

where $j < i$ (i.e., v_j is an already-selected video for node n) and $Sim_{visual}(v_i, v_j)$ is the visual similarity computed based on ND matching as described in Section 3.1.2. This intra-node diversity measures the redundancy of a new video v_i compared with existing videos attached to the node, which is used—together with the inter-node redundancy described below—to decide if v_i could be further added.

The second level of diversity measure evaluates content redundancy across nodes along a top-down path of the hierarchy. This is useful since one may not want to see too many duplicate contents in top-down result navigation. We pool the videos in a top-down path together and quantify the inter-node redundancy as:

$$InterRedundancy(v_i, n) = \max_{\{v_j \in \mathcal{T}_n\}} (Sim_{visual}(v_i, v_j)), \tag{7}$$

where the formulation is similar to Equation 6, except $\mathcal{T}_n$ represents the set of nodes on all possible top-down paths of a node n where v_i resides.

Notice that although the decay factor defined earlier has the effect of selecting videos with full coverage of multiple semantic facets for parent nodes, and the uniqueness criterion tends to pick videos with focused content for leaf nodes, imposing both of them does not ensure low content redundancy over a top-down path. We achieve the goal by explicitly using the inter-node redundancy measure.

3.4 Integrating All the Criteria for Video Selection

Now that we have a list of candidate videos for each node, ranked by their relevance scores. We also have a uniqueness score computed across nodes for each video, and a way to compute the diversity/redundancy of a candidate video when compared with the already-selected videos for a node. This section presents a solution to optimally integrate all the three criteria for video selection.

More rigorously speaking, suppose we want to choose a total of k videos for each node, the first goal is to maximize a function $\mathcal{F}$, defined as:

$$\mathcal{F} = \sum_i \sum_j Relevance(v_i, n_j) * Uniqueness(v_i), \tag{8}$$

while at the same time minimizing

$$\mathcal{G} = \sum_i \sum_j (IntraRedundancy(v_i, n_j) + InterRedundancy(v_i, n_j). \tag{9}$$

Notice that, in the function $\mathcal{F}$, we set the uniqueness value to one if n_j is not a leaf node, since uniqueness is a criterion only applied to videos on leaf nodes.

Table I. Information of the dataset used in our experiments.

ID	Topic name	# Videos	Wikipedia URL	# Nodes (leaf nodes)
1	Economic collapse	1025	en.wikipedia.org/wiki/Late-2000s_financial_crisis	35 (30)
2	US president election 2008	738	en.wikipedia.org/wiki/US_Presidential_Election,_2008	44 (35)
3	Beijing olympics	1098	en.wikipedia.org/wiki/2008_Summer_Olympics	19 (16)
4	Somali pirates	410	en.wikipedia.org/wiki/Somali_pirates	30 (22)
5	Virginia Tech massacre	682	en.wikipedia.org/wiki/Virginia_Tech_massacre	18 (15)
6	Beijing olympics torch relay	655	en.wikipedia.org/wiki/2008_Summer_Olympics_torch_relay	12 (9)
7	Sichuan earthquake	1458	en.wikipedia.org/wiki/Wenchuan_earthquake	19 (13)
8	Iran nuclear program	1056	en.wikipedia.org/wiki/Nuclear_program_of_Iran	27 (21)
9	Swine flu	1153	en.wikipedia.org/wiki/2009_flu_pandemic	28 (22)
10	Death of Michael Jackson	1865	en.wikipedia.org/wiki/Death_of_Michael_Jackson	21 (15)

Combining both terms, we propose the following multi-criterion optimization problem for video selection ($0 \leq \beta \leq 1$):

$$\max_{\{v\}}(\beta\mathcal{F} - (1-\beta)\mathcal{G}). \tag{10}$$

While the global optimal solution can be achieved using an exhaustive search algorithm, we pursue a more efficient solution which was found satisfactory in practice. At each step of adding a new video, we pick the one that maximizes the objective function. The process continues until all nodes are inserted with k videos.

4. EXPERIMENTAL SETUP

In this section, we describe the setup of our experiments, including dataset, performance measures, evaluating several systems using different combinations of system modules, as well as time effect in our proposed approach.

4.1 Dataset

We used the dataset in [Wu et al. 2011] for experiments. Ten significant topics from the recommended hot topics from CNN, Xinhua and Times were selected. By using the name of the topics as queries and searching on YouTube, the top-ranked videos were downloaded which form a big collection of 10,140 videos. The semantic hierarchy of each topic was retrieved from Wikipedia. We used the topic names to query Wikipedia and downloaded the corresponding page for each topic. The details of the dataset are given in Table I, including topic names, the number of downloaded videos per topic, the URL of the Wikipedia pages, and the numbers of nodes in the semantic hierarchies from the Wikipedia pages.

To facilitate quantitative performance evaluation, we need manual annotations on the relevance of each video to every node of the semantic hierarchies, which is not a trivial process. Two annotators were involved, who were firstly asked to carefully read the Wikipedia pages, and then watch the videos and mark if they are relevant to each of the nodes in the semantic hierarchies. Due to the semantic relationships between leaf nodes and high-level parent nodes, videos were only annotated to the leaf nodes, and a video is assumed relevant to a parent node if it is relevant to a child node under the parent node.

4.2 Performance Evaluation

We use the following measures to evaluate the relevance, uniqueness and diversity of the structurally organized search results.

4.2.1 *Relevance.* Precision is used for this most important criterion in search applications. It measures the proportion of the selected videos that are relevant to the corresponding nodes.

4.2.2 *Uniqueness.* We evaluate uniqueness on the relevant videos (true positives) for each node using the following equation:

$$M_{uniqueness} = \sum_{n_i} \sum_{v_{ij} \in n_i} \frac{1/l_{ij}}{T},$$

where n_i represents a node in a hierarchy, v_{ij} represents the selected relevant videos under the node n_i, l_{ij} indicates the number of nodes that are relevant to v_{ij} in the hierarchy, and T is the total number of relevant videos placed on the hierarchy. Apparently, this measure prefers videos relevant to just a single node in the hierarchy (i.e., $l_{ij} = 1$).

4.2.3 *Diversity.* Diversity is evaluated based on the number of ND videos found in the selected set. Specifically, we compute the percentage of ND videos within each leaf node, and along each top-down path. Given a set of videos, we view each of them carefully and a video is marked as redundant if it has over 50% of overlap with previously viewed videos. Finally, we calculate the percentage of redundant videos in the give set. Diversity is computed as the mean of all the percentage values.

4.3 Evaluated Systems

To validate the effectiveness of our approach and understand the contribution of each component, we compare the following systems:

—Text: Directly using textual matching to select videos, as described in Section 3.1.1.
—RW: Selecting videos using random walk, where both textual matching and visual similarity based on ND matching are considered.
—RW+U: Videos are selected based on the multiplication of their relevance scores and uniqueness scores.
—RW+U+D: Integrating relevance, uniqueness and diversity in a unified fashion using the multi-criterion optimization discussed in Section 3.4.

5. EXPERIMENT

We split the experiments into three parts. The first part (Section 5.1) evaluates the contribution of three criterions towards search performances. The second part (Section 5.2) conducts subjective evaluation to study the effect of the proposed work on user search experience in comparison to two other systems. Finally, the third part (Section 5.3) performs "timeline" simulation to provide insight on how the performance varies with respect to change of topic structure and video volume.

5.1 Objective Evaluation

5.1.1 *Effect of Parameters.* There are three key parameters in our approach: λ, α, and β. λ balances the contributions of the textual and visual channels in computing video similarities (only text is used when $\lambda = 1$). α is a parameter used in random walk. No random walk is performed when it is set to 0. The last parameter β controls the effect of diversity while integrating the three criteria.

We evaluate one parameter at a time, and fix the other parameters to reduce the complexity of settings. Results are visualized in Figure 4. We see that higher precision can be achieved with a λ value between 0 and 1, indicating that combining both textual and visual information is better than using any of them alone. For α, the highest precision is obtained when it is equal to 0.6. This shows that random walk is helpful (vs. the precision when $\alpha = 0$), and also the initial scores computed by text matching is also very important. Since the last parameter β affects both relevance (measured by precision) and diversity, we use a figure with two Y-axes. Clearly there is a tradeoff between precision

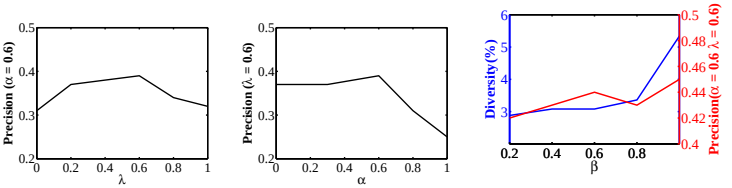

Fig. 4. Effect of parameters in our approach. See texts for more explanations. Note that diversity is evaluated by the percentage of redundant contents (lower value is better).

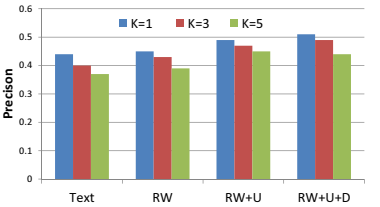

Fig. 5. Mean precision of the four systems with different k, the number of selected videos per node.

Table II. Performance evaluation on all the ten topics using the four systems, measured by relevancy, uniqueness, and diversity (grouped from left to right, see Section 4.2 for definitions of the evaluation measures). Note that diversity is evaluated by improvement in reducing the percentage of redundant contents.

ID	Relevance (Precision)				Uniqueness				Diversity Improvement (%)	
	Text	RW	RW+U	RW+U+D	Text	RW	RW+U	RW+U+D	RW+D (v.s. RW)	RW+U+D (v.s. RW+U)
1	0.40	0.37	0.37	0.26	0.36	0.38	0.37	0.36	7.2	0.0
2	0.22	0.25	0.28	0.30	0.78	0.78	0.81	0.82	39.4	53.6
3	0.40	0.39	0.55	0.55	0.78	0.76	0.79	0.81	22.3	26.3
4	0.18	0.20	0.33	0.33	0.60	0.64	0.59	0.53	78.5	100.0
5	0.26	0.28	0.39	0.28	0.69	0.66	0.74	0.78	0.0	0.0
6	0.52	0.57	0.67	0.67	0.62	0.57	0.59	0.57	62.7	53.6
7	0.48	0.51	0.63	0.63	0.78	0.78	0.83	0.93	0.0	0.0
8	0.27	0.29	0.34	0.33	0.35	0.32	0.37	0.39	0.0	0.0
9	0.36	0.36	0.43	0.46	0.81	0.83	0.82	0.84	45.6	42.3
10	0.62	0.65	0.57	0.57	0.63	0.61	0.67	0.70	0.0	0.0
Mean	0.37	0.39	0.45	0.44	0.64	0.63	0.66	0.67	25.6	42.3

and diversity. The best diversity is attained when $\beta = 0$, while the highest precision is achieved when $\beta = 1$. We see that a value around 0.6 gives a good balance between the two.

Another parameter, which is not in our core selection algorithm, is the number of videos to be placed under each node, i.e., k. Figure 5 shows the mean precisions of the four systems with different k. When k increases, the precision of all the systems drops quickly, which is quite intuitive. In the remaining experiments, we fix the parameters by setting λ, α and β all equal to 0.6. The number of videos placed under each node (i.e., k) is set to 5 for providing users more choices of videos to browse.

5.1.2 *Relevance.* Table II (the first group of columns under "Relevance") summarizes the results measured by precision. The mean precision over all the 10 topics is 0.37 for the text-only system, which is not bad since only surrounding texts of the videos are used. Random walk, which integrates both textual and visual similarities, boosts the performance to 0.39. This is not as significant as we expected. However, since random walk on the small graphs of the top returned videos is very efficient, it is still worthwhile applying it. Note that the visual similarities (based on ND matching) are also used in evaluating uniqueness and diversity, and thus are not computed solely for the random walk process. The uniqueness measure discussed in Section 3.2 significantly improves the mean precision to 0.45, which is very significant considering that uniqueness is taking into account by simply multiplying it with relevance (see Equation 8). This indicates that videos estimated to be solely relevant to a single node tend to be more precise than those determined to be relevant to multiple nodes. Furthermore, by considering diversity, the mean precision slightly drops 1%, while in the meantime the results are more diversified, which will be discussed later.

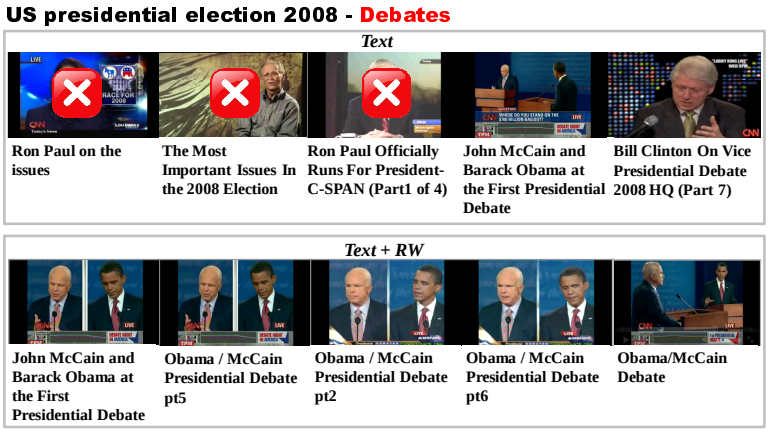

Fig. 6. Top five videos selected for node "Debates" in topic "US President Election 2008", using Text (top row) and RW (bottom row), respectively. RW effectively removes the three false positives marked by cross buttons.

For the results of individual topics, we find that the text matching system is already good for a few topics like topic 10 (T10). The surprisingly high precision for this topic is because each node in the hierarchy has its own discriminant words and phrases. While it is nice to see such cases, we also observe that the performance of the text-only system is relatively low for topic 2 (T2), T4, T5 and T8. For T2 ("US presidential election 2008"), the number of nodes contained in the Wikipedia hierarchy is the largest among all the topics. We find that sometimes frequent words cause confusion in video selection, because some nodes have significant overlap of such words. For example, a video titled "President-Elected Barack Obama in Chicago", which is supposed to be under the node *"election"*, is wrongly assigned to the node *"Party Convention"* since the word "Chicago" appears in the Wikipedia descriptions of both nodes with high tf-idf scores. In addition, a phrase "presidential debate" is obviously a keyword for node *"Debates"*, which however also appears under a node "Before the primaries". As a result, some presidential debate videos are wrongly assigned to this node. For T4, T5 and T8, the poor text matching performance is mainly because the metadata in videos is sparse. For examples, there are several videos in T4 having very short and less informative titles like "Pirates" or "Somalia Pirates", many videos in T5 are with title "Virginia Tech Shooting" or "Shooting on Virginia Tech", and "Iran nuclear" in T8. We expect that these limitations in text matching can be alleviated by visual matching utilized in both the random walk process and the computation of uniqueness and diversity.

Random walk improves 7 out of 10 topics. In Figure 6, the first 3 videos in the top row are false positives involved by text matching. Through random walk, the relevance scores are refined using visual similarities, and thus true debate videos are pulled to the top. RW also degrades the performance of a few topics like T1, where the news anchor person appears frequently and is thus connected by visual matching.

Uniqueness is found to have very positive effects on improving relevance, which is very interesting and out of our original expectation. In Figure 7, the top two rows are the selected videos by RW and RW+U for a node "Domestic Leg" under topic "Beijing Olympics torch relay". The first video in the RW result is a false positive, having the words "Sichuan" and "China", which have higher weights. This video survives from random walk because the other positive videos also have these textual words. However, these words also appear in other nodes, and thus the video is also ranked very high for nodes "Media Coverage", "Torch Security", and "Route". With the consideration of uniqueness, which discourages the matching of a video to multiple nodes, this video is filtered. From our result analysis,

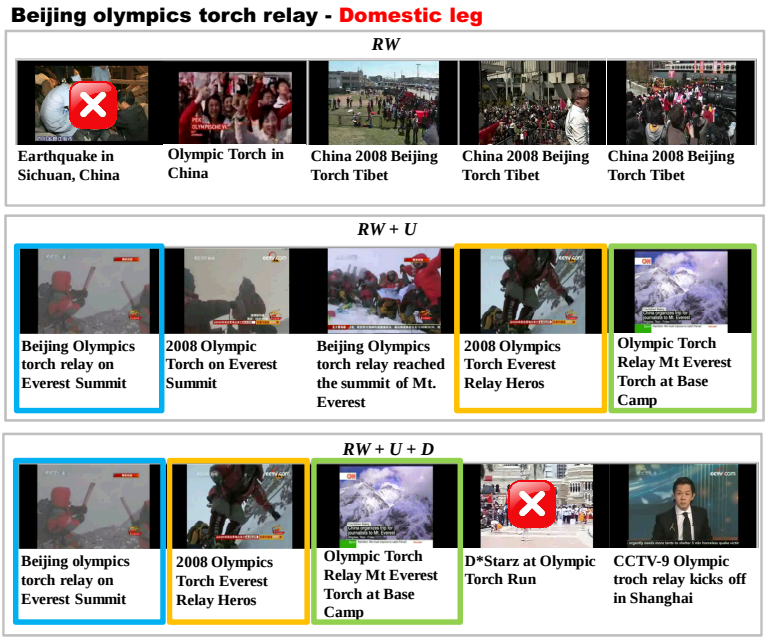

Fig. 7. Top five videos selected for node "Domestic Leg" in topic "Beijing Olympics Torch Relay", using RW (top), RW+U (middle), and RW+U+D (bottom), respectively. The same videos selected by the bottom two methods are circled by boxes of the same colors.

the uniqueness has the side effect of pruning noisy videos, especially for videos having lengthy text description irrelevant to the video content. On the other hand, uniqueness may degrade precision in the case that a video relevant to multiple nodes of a topic structure will be pruned.

By further including the diversity criteria, the relevance drops slightly for most topics. This is because of the clear tradeoff between diversity and relevance. The last row in Figure 7 shows the results after considering diversity. The videos circled with the same color as in the second row remain unchanged after using this criteria, while the two near identical videos (second and third) from RW+U are removed. However, by RW+U+D, there is a video incorrectly included among the two.

5.1.3 *Uniqueness.* The next measure to be evaluated is uniqueness, whose results are shown in the middle group of columns in Table II. We see that the results of Text and RW are close because both systems focus on relevance and do not pay special attention to this criteria. By incorporating the uniqueness score, the performance is improved to 0.66. The gain can be observed on most topics, except that for a few topics the performance drops marginally. The performance indeed varies depending on topics. For topics such as "Sichuan Earthquake", the information in different sections is more mutually exclusive and there exist videos that can uniquely match some of the sections, which leads to larger degree of improvement. Although the gain is not very significant on average, we think that the measure is also useful, particularly when considering the fact that, as a by-product of using the uniqueness criterion, relevance is significantly improved.

Figure 8 further gives an example from topic "Beijing Olympics". Three videos (marked using boxes of the same colors) are chosen by both systems. Two videos from RW are removed. While looking into one of the videos "Beijing Olympics Main Stadium, Bird's Nest (1 of 5)", as shown in the bottom of the figure, we see that there are also contents related to the node "Bid", and the different voices judging the

Fig. 8. Top five videos selected for node "Venues" in topic "Beijing Olympics", using RW (the first row) and RW+U (the second row), respectively. The same videos selected by the two methods are circled by boxes of the same colors. The callout rectangle at the bottom row includes a few key frames from the corresponding video, which has contents related to other nodes and therefore is not unique. See texts for more discussions.

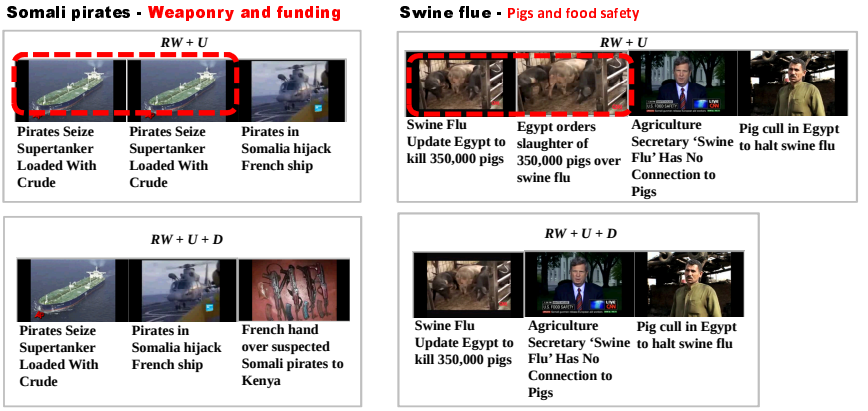

Fig. 9. Two examples about removing redundant contents by considering *Diversity*. The two videos in the same red dashed box are duplicates of each other.

design and construction of the Bird's nest, which are related to the node "Media Coverage". Therefore, this video is not unique. The newly added videos are more focused in content.

5.1.4 *Diversity*. We evaluate the effect of diversity on two baseline systems RW and RW+U. We first calculate diversity (i.e. redundancy) using metric defined in Section 4.2.3. Relative performance gains, which are calculated by improvement in reducing the percentage of redundant contents, i.e., improvements in diversity, are reported.

After considering diversity, the improvement for RW+D vs. RW is 25.6% and for RW+U+D vs. RW+U is 42.3% on average. Figure 9 gives two examples. For the first example, two duplicate videos with identical surrounding texts were chosen for node "Weaponry and funding" of the topic "Somalia pirates". One of them is removed after considering diversity. The other example has two news videos describing that many pigs are killed in Egypt because of swine flu. The two videos are from different TV channels but have similar contents, and one of them is successfully removed.

5.1.5 *Significance Test.* We conduct significance test to verify whether the performance improvement presented in Table II is by chance. The testing is based on the randomization test in [Romano 1990], where the target of iterations is set to 100,000. At the 0.05 level of significance, RW+U+D is significantly better than RW and Text in the relevance measure. There is also a significant difference between RW+U+D and RW+U in terms of diversity but not relevance. No difference is observed among the four systems for the uniqueness measure.

5.2 User Study

The aim of subjective evaluation is to investigate the presentation of search results in affecting user experience towards search effectiveness and efficiency. Specifically, the evaluation studies: 1) the degree to which the subjects using a system can complete a task more effectively and efficiently than another group of subjects using a different system; 2) the overall experience when the subjects are asked to complete similar tasks using two different systems.

We compare four video browsing systems: list-based, galaxy-based [Tan et al. 2011], the proposed hierarchical browsers with and without uniqueness criteria (namely the RW+U+D and RW+D systems). The interface of list-based browser is similar to commercial search engine, where videos are linearly ranked but the relationship among videos is not known. In contrast, galaxy browser [Tan et al. 2011] organizes similar videos as "visual snippets" through visual and textual clustering, and the relationship among the videos in a snippet is presented as a graph. Furthermore, to facilitate fast browsing, each snippet is annotated with the text snippets obtained from Wiki timeline and the sentences extracted from relevant Wikipedia pages. A total of 24 evaluators from different education backgrounds, including computer science (5), electronic engineering (6), chemistry (2), industry (3), business (1) and linguistics (2), are invited for the evaluation. There are 11 females among the 24 evaluators with average ages of 31. All the evaluators are frequent users of social media websites.

The evaluation is conducted in two rounds. In the first round, we invite 16 evaluators and they were split into four groups, each using a different system. The task is to identify as many videos as possible that are relevant to three given questions under the topic "Economic Collapse". The questions are designed such that they can be answered from different perspectives and potentially there is more than one video that can serve as the answers. The three questions are "what's the trigger of financial crisis?", "what's the effect on U.S. economy?" and "what are the international responses to the economic crisis?". The evaluators were asked to complete the task within 30 minutes. An evaluator was allowed to end the task earlier so long as no more videos can be identified as answers. The following statistics was collected in the first round of evaluation: time spent in completing the task, the number of videos (or answers) found, and the percentage of incorrect answers among the identified videos or error rate.

In the second round of evaluation, another topic "US president election 2008" is given. The 16 evaluators are asked to find videos as answers to three questions under the topic, but using a system different from the first around. The group not using RW+U+D is asked to use the system in this round, in order to draw the difference between the proposed and other systems. On the other hand, the group using RW+U+D in the first round is asked to use RW+D. We further invite 8 subjects and divide them into two groups. The first group compares the list-based browser with RW+U+D, while the second group compares the galaxy browser with RW+U+D. In this case, the number of subjects for each pair-wise comparison is identical. Furthermore, we also change the order of topics in the evaluation to alleviate the potential bias due to the topic itself. In short, the evaluation compares RW+U+D with three browsers using two topics. Each pair-wise system comparison involves 8 subjects. After completing the assigned task, the subjects are finally given a questionnaire to evaluate the systems used in the first and second rounds. The questionnaire consists of following criteria in the scale of seven (highly preferable) to one (not acceptable):

—Presentation: To what degree the organization and presentation of the videos help in understanding the given topic?
—Precision: Is the presented multimedia content relevant to topic?
—Conciseness: Is the presented content concise with minimum redundant information?
—Engagement: How useful the system is in providing guidance to complete the assigned tasks?
—Acceptance: Will the system offer better user experience, if it is used by social media websites?

The first three criteria examine whether the presented videos under a system are relevant (Precision), less redundant (Conciseness) and properly organized (Presentation) that allow the quick understanding of a topic. The last two criteria assess the usability aspect, of whether a system is useful in completing the tasks (Engagement) and can offer better user experience (Acceptance).

Figure 10 shows the results of first round evaluation for the topic "Economic collapse", averaged over the statistics collected from four subjects in a group. Overall, the group using RW+U+D system returned the most number of videos as answers, with the lowest error rate and the minimum search time. Relying on the hierarchical structure, the subjects can focus on the videos on relevant nodes for finding the potential answers. Compared to RW+D, RW+U+D has lower relevance score based on the objective evaluation in Table II. However, some of the relevant videos across nodes in RW+D are actually identical and thus are less informative. Besides, without the uniqueness criterion, videos are not unique to a node thus more time is needed to digest and locate the relevant part in the video. Because of these reasons, RW+D requires more time in identifying answers, and yet the number of videos is less than RW+U+D. In contrast to hierarchical browser, the list browser has the worst performance with the least number of returned videos and the highest error rate. Some subjects gave up the task earlier before the 30 minutes time limit for the reasons of tiring and frustrating with the system. By clustering videos as snippets, galaxy browser offers better performance, where more videos with low error rate are found. Nevertheless, as commented by the subjects, while snippets are summarized with timeline descriptions, the relationship among them is not revealed. Compared to RW+U+D, the subjects needed to spend more time in reading the text description annotated on snippets before deciding whether to investigate videos under a snippet.

The subjective evaluation in the second round is shown in Figure 11. RW+U+D shows preferable rating than list-based, galaxy browser and RW+D across all the criteria. The "Acceptance" of galaxy browser is close to that of RW+U+D. The subjects like the interface of galaxy browser, which provides a grand overview of snippets annotated with text summaries. Nevertheless, as the relationship among snippets is not explicit, and furthermore the videos in a snippet are intertwisted with partially duplicate content, the other four scores are lower. RW+D without uniqueness consideration has the lower "Conciseness" score compared with RW+U+D. Without uniqueness criterion, RW+D actually has more duplicate videos which are difficult to be removed by the diversity criterion if the videos are located in conflict paths. In this case, the same videos may be seen by the subject several times in different paths of the hierarchy, and hence results in lower "Conciseness" score of RW+D. Overall, our significance test indicates that RW+U+D is significantly better than three other systems across all the criteria at the level of 0.01.

5.3 Performance Change with Timeline

We categorize the topics in Table I into two groups depending on the update frequency of a topic structure. The examples of topic for the first group include "US presidential election 2008" (T2) and "Beijing Olympics" (T3), which have relatively comprehensive structure even before the occurrence of events. For example, the section "opening ceremony" has already been created for the topic "Beijing Olympics" in the page of year 2007, though the opening ceremony actually happened in August of 2008.

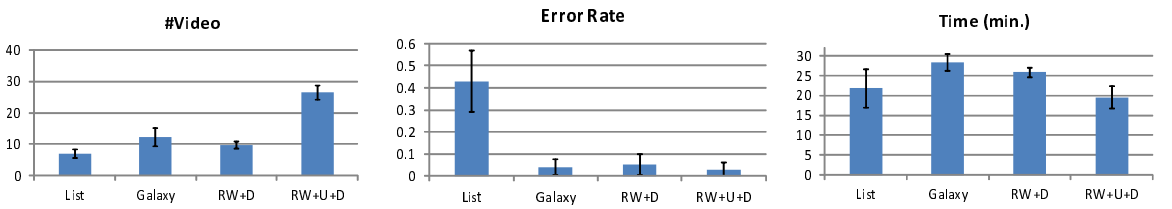

Fig. 10. Subjective evaluation of four systems, measured by the number of videos found as answers, error rate, and the average time spent for the assigned task.

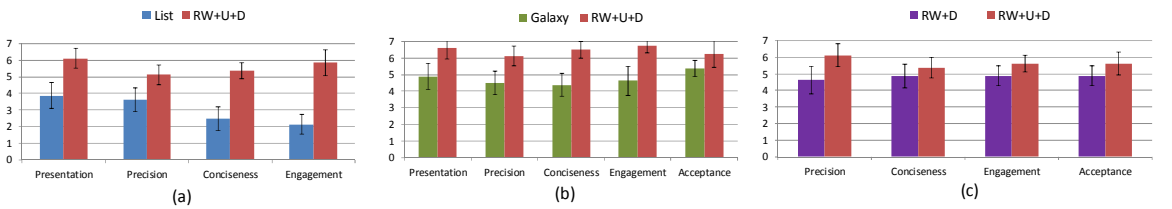

Fig. 11. The average rating of user studies for List, Galaxy and RW+D against RW+U+D. Note that in (a) the "acceptance" criteria is not applicable to list browser and thus is not shown, in (c) the "presentation" scores are the same and thus is not shown.

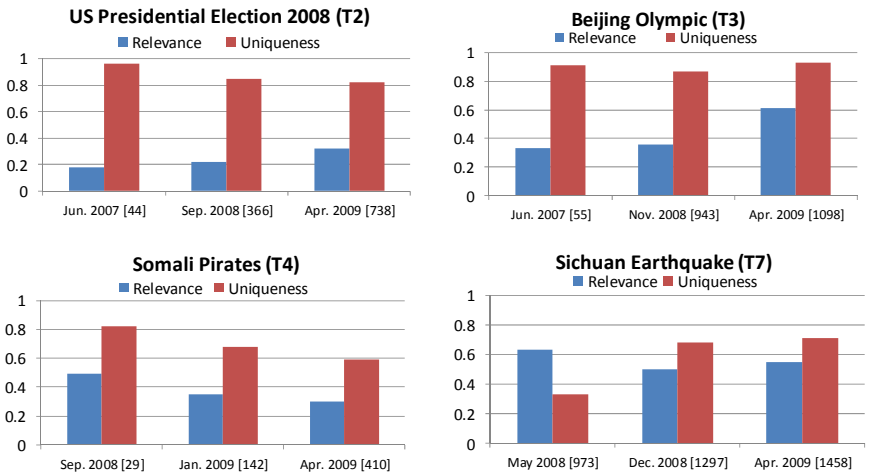

Fig. 12. Effect of time (Wikipedia page evolution) on the relevance and uniqueness measures of our hierarchical browser. The number of videos placed on the corresponding hierarchy is shown next to the dates.

The examples of topic for the second group include "Somali pirates" (T4) and "Sichuan earthquake" (T7). These topics have a brief table-of-content initially, and the content is changed dramatically with the emergence of various peripheral events, such as the sections "aftershock" and "property damage" that were created for topic T7 after the event occurrence.

Figure 12 shows the result for four topics selected for experiment. We manually select three wikipages that basically reflect the largest changes in table-of-content along the timeline of revision history for each topic. The videos are then matched to the topic structures of which the corresponding wikipages were created after the video upload time. The results show that the performance in terms of relevance measure has some correlations with the degree of changes in the topic structure, and the uniqueness

measure appears to be dependent on the clarity of section description. For T2 and T3 belonging to the first group, the relevancy improves with the progress of time for the reason that the nodes, which have no suitable videos (e.g., "Democratic Party primaries" in T2, "opening ceremony" in T3) initially, are filled with the videos which were uploaded when the actual events happened. There is no significance change observed however for uniqueness measure. For T4 and T7 in the second group, their performance trend appears quite differently. Most of the initial upload videos are relevant to the earlier created Wikipedia page that has a simpler structure. As the topic evolves with more sections included, the relevancy drops as well when more diverse content of videos are uploaded. For T7, the uniqueness improves as the newly included sections add specificity to the potentially ambiguous sections. For example, the new section "collapse of school houses" can uniquely host the videos about "tofu-dregs schoolhouses" that are partially related to the sections "immediate aftermath", "casualty" and "property damage" before the creation of this new section. However, similar trend is not observed in T4. This is because the newly added sections such as "casualty", "jurisdiction" and "trials" are somewhat related, and there is no new videos that can uniquely match each of these sections.

6. CONCLUSION AND DISCUSSION

We have presented a complete system to generate a hierarchically organized view of video search results. This is particularly suitable for organizing search results containing multiple semantic facets, which are very difficult to be demonstrated using a traditional list view. Our approach starts from a structured Web page on Wikipedia that is retrieved using a given search topic. Initial video search results (e.g., from YouTube) are then placed on each node under the Wikipedia structure based on three criteria: relevance, uniqueness, and diversity. We have provided clear definitions of the three criteria and a unified way to integrate all of them to reach our goal. Extensive evaluations on a well-designed benchmark have clearly shown the effectiveness of our approach.

There are also a few limitations of our proposed approach. First, not all the complex user queries can be matched to Wikipedia pages. This may be solved by exploring multiple knowledge bases, which is an interesting future work. Second, not all the relevant videos can be placed on a node in a hierarchy. For example, there are many opinion videos about presidential election, which are frequently viewed but cannot be associated to any node of the corresponding hierarchy. Therefore, another promising future work is the adaptive adjustments of the semantic hierarchies from Wikipedia, based on the available videos to be displayed. Nodes with no suitable videos should be removed, and new nodes may be added if sufficient evidences exist to prove their informativeness. Overall, we believe that using well-organized structures is suitable and important for dealing with complex video search queries, and systems with such a capacity can be deployed in many applications in the near future.

REFERENCES

CAI, D., HE, X., LI, Z., MA, W.-Y., AND WEN, J.-R. 2004. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th annual ACM international conference on Multimedia*. MULTIMEDIA '04. ACM, New York, NY, USA, 952–959.

CARMEL, D., ROITMAN, H., AND ZWERDLING, N. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '09. ACM, New York, NY, USA, 139–146.

CHEN, C. C. AND CHEN, M. C. 2008. Tscan: a novel method for topic summarization and content anatomy. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '08. ACM, New York, NY, USA, 579–586.

CHEN, T., LIU, C., AND HUANG, Q. 2012. An effective multi-clue fusion approach for web video topic detection. In *Proceedings of the 20th ACM international conference on Multimedia*. MM '12. ACM, New York, NY, USA, 781–784.

DE ROOIJ, O. AND WORRING, M. 2010. Browsing video along multiple threads. *IEEE Transactions on Multimedia 12,* 2, 121–130.

FUNG, G. P. C., YU, J. X., LIU, H., AND YU, P. S. 2007. Time-dependent event hierarchy construction. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '07. ACM, New York, NY, USA, 300–309.

HSU, W. H., KENNEDY, L. S., AND CHANG, S.-F. 2007. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th international conference on Multimedia*. MULTIMEDIA '07. ACM, New York, NY, USA, 971–980.

HUANG, C.-C. 2004. Liveclassifier: creating hierarchical text classifiers through web corpora. In *WWW 04: Proceedings of the 13th international conference on World Wide Web*. ACM Press, 184–192.

IDE, I., MO, H., AND KATAYAMA, N. 2003. Threading news video topics. In *Multimedia Information Retrieval* (2006-02-15), N. Sebe, M. S. Lew, and C. Djeraba, Eds. ACM, 239–246.

JEGOU, H., DOUZE, M., AND SCHMID, C. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*. ECCV '08. Springer-Verlag, Berlin, Heidelberg, 304–317.

JING, F., WANG, C., YAO, Y., DENG, K., ZHANG, L., AND MA, W.-Y. 2006. IGroup: web image search results clustering. In *Proceedings of the 14th annual ACM international conference on Multimedia*. MULTIMEDIA '06. ACM, New York, NY, USA, 377–384.

JING, Y. AND BALUJA, S. 2008. Visualrank: Applying pagerank to large-scale image search. *IEEE Trans. Pattern Anal. Mach. Intell. 30,* 11, 1877–1890.

KENDER, J. R., HILL, M. L., NATSEV, A. P., SMITH, J. R., AND XIE, L. 2010. Video genetics: A case study from youtube. In *Proceedings of the International Conference on Multimedia*. MM '10. ACM, New York, NY, USA, 1253–1258.

LIU, L., SUN, L., RUI, Y., SHI, Y., AND YANG, S. 2008. Web video topic discovery and tracking via bipartite graph reinforcement model. In *Proceedings of the 17th international conference on World Wide Web*. WWW '08. ACM, New York, NY, USA, 1009–1018.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60,* 2, 91–110.

MING, Z.-Y., WANG, K., AND CHUA, T.-S. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '10. ACM, New York, NY, USA, 2–9.

NISTR, D. AND STEWNIUS, H. 2006. Scalable recognition with a vocabulary tree. In *IN CVPR*. 2161–2168.

ROMANO, J. P. 1990. On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc. 85,* 411, 686–692.

SANG, J. AND XU, C. 2011. Browse by chunks: Topic mining and organizing on web-scale social media. *TOMCCAP 7,* Supplement, 30.

SLIMI, J., MANSOURI, S., BEN AMMAR, A., AND ALIMI, A. M. 2013. Semantic browsing in large scale videos collection. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*. OAIR '13. Paris, France, 53–56.

TAN, S., NGO, C.-W., TAN, H.-K., AND PANG, L. 2011. Cross media hyperlinking for search topic browsing. In *Proceedings of the 19th ACM international conference on Multimedia*. MM '11. ACM, New York, NY, USA, 243–252.

VAN LEUKEN, R. H., GARCIA, L., OLIVARES, X., AND VAN ZWOL, R. 2009. Visual diversification of image search results. In *Proceedings of the 18th international conference on World Wide Web*. WWW '09. ACM, New York, NY, USA, 341–350.

WU, X., LU, Y.-J., PENG, Q., AND NGO, C.-W. 2011. Mining event structures from web videos. *IEEE MultiMedia 18,* 1, 38–51.

WU, X., NGO, C.-W., AND HAUPTMANN, A. G. 2008. Multimodal news story clustering with pairwise visual near-duplicate constraint. *IEEE Transactions on Multimedia 10,* 2, 188–199.

XUE, G.-R., XING, D., YANG, Q., AND YU, Y. 2008. Deep classification in large-scale text hierarchies. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '08. ACM, New York, NY, USA, 619–626.

ZHAO, W.-L. AND NGO, C.-W. 2009. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *Trans. Img. Proc. 18,* 2, 412–423.

ZHAO, W.-L., WU, X., AND NGO, C.-W. 2010. On the annotation of web videos by efficient near-duplicate search. *Trans. Multi. 12,* 5, 448–461.