Organizing Video Search Results to Adapted Semantic Hierarchies for Topic-based Browsing

Jiajun Wang[1], Yu-Gang Jiang[1], Qiang Wang[1], Kuiyuan Yang[2], Chong-Wah Ngo[3]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[2]Microsoft Research Asia, Beijing, China
[3]Department of Computer Science, City University of Hong Kong, Hong Kong, China
{jiajunwang13, ygj, qiangwang}@fudan.edu.cn,
kuyang@microsoft.com, cscwngo@cityu.edu.hk

ABSTRACT

Organizing video search results into semantically structured hierarchies can greatly improve the efficiency of browsing complex query topics. Traditional hierarchical clustering techniques are inadequate since they lack the ability to generate semantically interpretable structures. In this paper, we introduce an approach to organize video search results to an adapted semantic hierarchy. As many hot search topics such as celebrities and famous cities have Wikipedia pages where hierarchical topic structures are available, we start from the Wikipedia hierarchies and adjust the structures according to the characteristics of the returned videos from a search engine. Ordinary clustering based on textual information of the videos is performed to discover the hidden topic structures in the video search results, which are used to adapt the hierarchy extracted from Wikipedia. After that, a simple optimization problem is formulated to assign the videos to each node of the hierarchy considering three important criteria. Experiments conducted on a Youtube video dataset verify the effectiveness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Search result organization, video search, hierarchical structure, hierarchy adaptation.

1. INTRODUCTION

Existing Web video search engines, such as the YouTube, return video ranking list given a specific query, which is acceptable if the users only want to find a single video like a movie trailer or an MTV. But when the query is a complex topic with hidden structures, finding or understanding different aspects of the topic based on a long list of videos is time-consuming, and often is impossible with a single search. For example, when a user searches for "Steven Spielberg", he/she may want to know the career of the director, his personal life, his achievements and so on. In this case, a hierarchical organization of the videos with nodes covering the aforementioned facets is a much better choice.

In the domain of *reorganizing* search results, a popular method is to cluster the results into several units or groups [14, 7, 3]. Although clustering is a good way to capture the hidden facets of a complex topic, a major drawback is that, without an overall semantic structure, even a set of well-extracted clusters may appear messy to the users, let alone clusters without proper human-readable labels.

Hierarchical clustering is slightly better than the traditional *flat* clustering, and has been adopted to build a hierarchical structure [1, 2]. However, no matter agglomerative or divisive, hierarchical clustering lacks the ability to create a semantically cohesive topic structure due to the noise in Web contents. Prototype-based hierarchical clustering (PHC) [9] is a method proposed to overcome this disadvantage. By clustering on the basis of a predefined prototype hierarchy, PHC can ensure the semantic cohesiveness of the hierarchy, but the prototype hierarchy need to be adapted to better befit the video domain.

Near-duplicate video search has also been widely applied to group visually similar videos, which is weakly related to this work. In [13, 5, 12], the authors focused on mining topic structures with near-duplicate search in news videos, since there could be a large amount of near-duplicates on the Internet when a breaking news event happens. However, when the search topic extends beyond news events, near-duplicate search becomes not so useful, but still can serve as a tool to remove content redundancy in the search results.

In this paper, we propose an approach to organize the ranking lists from video search to a semantic hierarchy for efficient browsing, which combines the advantages of both the traditional clustering techniques and the PHC. The entire process is illustrated in Figure 1. First, we cluster textual information of the retrieved videos of a query topic to discover the hidden topic structures, called salient phrases. Second, the salient phrases are assigned to the leaf nodes of a prototype hierarchy extracted from corresponding Wikipedia pages, and the nodes with no video clusters are pruned from the hierarchy. In this way, the prototype hierarchy is

MM'14, November 03–07, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2647868.2655012.

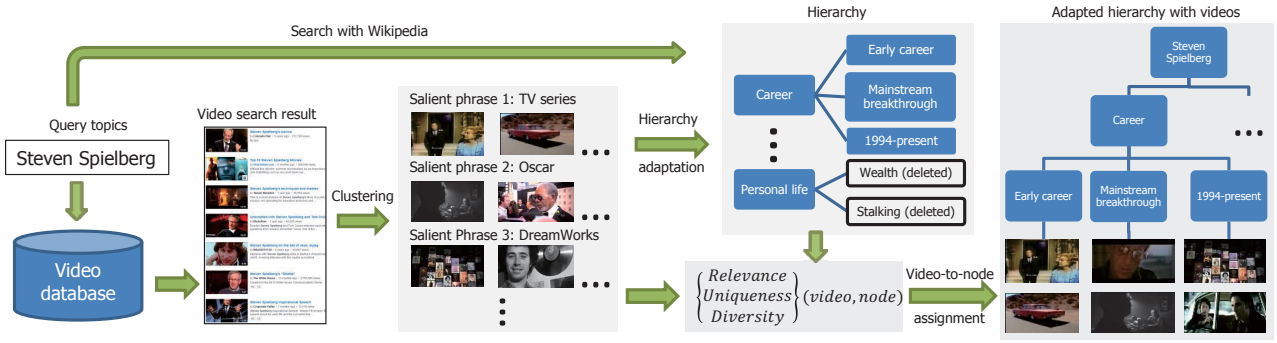

Figure 1: Illustration of our approach. Given a hot query topic, we first retrieve relevant videos using online video search engines and also download a hierarchy for the topic from Wikipedia. The hierarchy is then adapted by analyzing the retrieved videos. Finally, videos are selected and assigned to the nodes of the adapted hierarchy for efficient browsing, by considering three criteria. See texts for more explanations.

adapted to better reflect the topic structure of video search results. After that, we formulate an optimization problem in which videos are assigned to each node of the adapted hierarchy by considering multimodal information. This work is similar to [11] which also assigns videos to hierarchies for visualization. However, our approach has the capability of adapting the prototype hierarchies, which is more suitable for organizing the video search results.

2. OUR APPROACH

In this section, we will first propose the method for prototype hierarchy adaptation, and then describe a simple optimization-based video assignment process to put the videos on the modified hierarchy.

2.1 Prototype Hierarchy Adaptation

Since the prototype hierarchy extracted from the Wikipedia was constructed for textual descriptions of the topic, some nodes are not suitable for videos and should be removed. To discover the hidden facets in the search results for prototype hierarchy adaptation, we use the clustering algorithm in [14], which formulates the clustering process into a salient phrase ranking problem. Given a set of candidate phrases, it calculates several textual properties to model the probability of each phrase as a topic cluster. The calculated textual properties include tf-idf, phrase length, intra-cluster similarity, cluster entropy and phrase contextual entropy. Then a regression model is trained to combine the five properties to a final salient score. The top-ranked phrases are considered as salient phrases, and the video cluster corresponding to each salient phrase contains the videos with that phrase in textual descriptions. Finally, video cluster pairs with an overlapping rate (i.e., the percentage of shared videos between the pair) over 50% are deemed as similar salient phrases and are merged together.

The candidate phrases used above are prepared by the following procedure. We first extract all the n-grams ($n \leq 3$) from the title of each video as candidate phrases. Then, the mutual information (MI) [4] between phrases and video texts is used to quantify how informative each phrase is to all videos. The MI value of a phrase t is defined as:

$$MI(t) = \sum_{v} p(v,t) \log \frac{p(v,t)}{p(v)p(t)},$$

where v denotes a video, and the probability terms in the above equation can be derived from phrase co-occurrence statistics in the video contexts. With uninformative phrases discarded, the final set of candidate phrases is created.

The mined video clusters is used to adapt the prototype hierarchy by assigning the video clusters to the leaf nodes of the hierarchy. We use the method proposed in [11] to define the video-to-node similarity. All the texts are represented by the vector space model with tf-idf weighting, and the terms with hyper-links in the Wikipedia page are defined as important terms and assigned with empirically enlarged weights. The video-to-node similarity is then calculated as the Cosine similarity between the vector representations of the video texts and the texts under the node from the Wikipedia page. After that, the similarity between a video cluster C and a node n is defined as follows:

$$Sim_t(C,n) = \frac{1}{|C|} \sum_{v \in C} Sim_t(v,n),$$

where $Sim_t(v,n)$ is the video-to-node similarity and the subscript t indicates that the similarity is based on texts. Finally, a video cluster is assigned to the leaf node with the highest similarity value. After assigning all the video clusters to the nodes, leaf nodes with no assigned video clusters are pruned from the hierarchy.

2.2 Video-to-Node Assignment

In the browsing of a video hierarchy, there are three problems to be addressed to improve user experiences: 1) videos under each node should be semantically related to the node; 2) videos related to only one node rather than to several nodes are preferred (i.e., highly focused contents); and 3) near-duplicate videos should be avoided. The algorithm in [11] solves the three problems by using three criteria: relevance, uniqueness, diversity, which are integrated into an optimization problem. We adopt a similar method but use a different definition of uniqueness, which was found to be better. The criteria are defined as follows:

$$Rel(v,n) = Sim_t(v,n),$$

$$Uniq(v,n) = \frac{Sim_t(v,n)}{\sum_{n_j \in Leaf} Sim_t(v,n_j)},$$

$$Div(v,n) = \max_{v_j \in n} (Sim_v(v,v_j)),$$

where uniqueness is only defined on videos under the leaf nodes since videos relevant to a child node is surely related to the parent nodes. The subscript v in Sim_v indicates that similarity is calculated by near-duplicate visual search. To compute the visual similarity, we apply a state-of-the-art solution of near-duplicate search with the SIFT feature [8], Hamming embedding based fast matching [6], and temporal frame alignment [10]. Given a video pair, their visual similarity is defined as the percentage of the shared near-duplicate keyframes.

With the three criteria, the objective function for video-to-node assignment is defined as follows:

$$\mathcal{F} = \sum_{v,n} \beta \cdot Rel(v,n) * Uniq(v,n) - (1-\beta) \cdot Div(v,n).$$

By maximizing the objective function, the most suitable videos can be added to each node. For the sake of efficiency, we employ a greedy algorithm to solve the optimization, which inserts one video to a node at each step. To speed up the process, the videos are only selected from the corresponding video clusters (related to the salient phrases) assigned to the node in the hierarchy adaptation process. Note that using the salient phrase clustering outcomes also helps improve the precision of the selected videos, compared with the brute-force selection from the entire collection. The video assignment terminates after all the nodes are added with sufficient videos.

3. EXPERIMENTS

3.1 Dataset

We pick two very popular topic categories, celebrities and cities, to validate the effectiveness of our proposed approach. In collecting the dataset, for each topic, its respective Wikipedia page was first downloaded, from which the hierarchy, the texts under each node, and the terms with hyper-links (i.e., the more important terms) were extracted. We then searched the topic name in YouTube to download videos and their surrounding contexts. Table 1 summarizes the dataset.

3.2 Evaluation Plan

We split the evaluations into three parts. The first part is to evaluate the salient phrases produced from video clustering. We manually labeled the correct salient phrases and calculated the percentage of the correctly assigned phrases.

The second part is to inspect the video-to-node assignment results using the three criteria. The metrics used for evaluating the three criteria are listed below:

Relevance: Precision is used to measure the relevance criterion. Videos are manually labeled to their relevant nodes to compute precision.

Uniqueness: The average of uniqueness value of the videos, i.e., the mean value of $Uniq(v,n)$, in the leaf nodes are computed.

Diversity: Diversity is measured by the percentage of the selected videos that are redundant. If the visual similarity between two videos is over 0.5, one of them is redundant.

In the assignment procedure, at most 5 videos are selected for each parent (non-leaf) node. Since all the leaf nodes have a number of assigned salient phrases during the hierarchy adaptation process, we assigned at most $2\times\#SalientPhrases$ videos to each leaf node. The number of assigned videos is dependent on the number of salient phrases of each leaf node, which is good as a leaf node is semantically more diversified if it has more salient phrases and thus should be assigned more videos to fully capture the diversity. For the sake of comparison, we also assigned videos to the original hierarchy without adaptation.

Table 1: Statistics of the collected dataset. The first five topics are celebrities, and the last five are cities. Numbers in the parenthesis are the amount of leaf nodes in the corresponding Wikipedia hierarchies.

ID	Topic	# Videos	# Nodes (leaf)
1	Barack Obama	1887	31 (22)
2	Steven Spielberg	1618	23 (20)
3	Michael Jackson	1404	25 (22)
4	Michael Schumacher	1371	26 (19)
5	Steve Jobs	1870	31 (24)
6	London	2144	41 (33)
7	Chicago	2020	36 (28)
8	Los Angeles	2020	27 (20)
9	New York City	1834	31 (23)
10	Paris	1936	45 (35)

The third part of the evaluation is a user study. We evaluated three systems:

List: Videos are organized to a rank list (from YouTube search engine).

H: Videos are organized to the original hierarchy.

AH: Videos are organized to the adapted hierarchy by our approach.

Ten human subjects were involved in the user study, who were asked to compare the results for each topic and to rate the three systems from four aspects:

Presentation: Is the organization of the videos useful in quickly locating needed information?

Precision: Are the videos generally relevant to the topic?

Conciseness: Is there significant redundant information?

Satisfaction: Overall satisfaction of the results.

3.3 Results and Discussions

The results of first two evaluations are summarized in Table 2. We first discuss the first part which aims to evaluate the precision of the assigned salient phrases. From the results, we can see that the performance varies across topics. This measure highly depends on the relevance of the retrieved videos from YouTube to the query topics. If the retrieved videos are not very relevant, the precision of the salient phrases will be low. Therefore this number has close connections to the relevance criterion in the second part of the evaluation. Overall, with a mean precision of 63%, the procedure of salient phrase discovery is rather effective.

Then we move to the next part which evaluates the three criteria in video-to-node assignment (columns 3-8 in Table 2). Compared with the baseline ("H"), results of all the three criteria are significantly improved. The relevance value has a dramatic increase of 20% due to the hierarchy adaptation function of our approach. For instance, in the hierarchy of topic "London", unsuitable nodes with no related videos like "Toponymy", "Prehistory and antiquity" are deleted, and useful nodes like "Accent" and "Leisure and entertainment" are preserved. After the adaptation, the hierarchy originally to organize texts becomes more suitable for videos, and thus the relevance of the assigned videos can

Table 2: Results of the first two evaluations. For the salient phrase clustering (indicated by "Clustering" in the table), P stands for the precision of the salient phrases. In video-to-node assignment, R, U and D are relevance, uniqueness, and diversity respectively. See Section 3.2 for the metric definitions. Note that for diversity, lower value is better.

ID	Clustering	Video-to-Node Assignment					
	P (%)	R (%)		U (%)		D (%)	
		H	AH	H	AH	H	AH
1	78	41	70	41	47	5.8	1.0
2	55	39	59	36	42	8.7	1.7
3	50	31	51	27	31	10.4	1.6
4	46	15	42	32	44	7.7	0.7
5	61	21	45	29	41	6.5	0.0
6	67	33	58	35	56	5.4	0.0
7	78	44	53	39	45	4.4	0.0
8	82	43	61	53	65	1.5	0.0
9	70	39	51	36	40	2.0	0.6
10	43	26	35	29	34	0.9	1.1
Mean	63	33	53	36	45	5.3	0.7

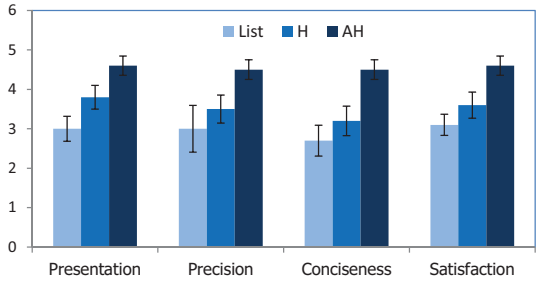

Figure 2: Subjective evaluation scores of the List, H and AH approaches, ranging from 1 (worst) to 5 (best). Our AH achieves the highest scores.

be largely enhanced. In addition, although our definition of uniqueness is straightforward, it is very effective (9% higher than the baseline). Finally, we see that for four topics, the diversity value is reduced to zero, and all the other topics have values lower than 2%, which is fairly satisfying.

Results of the user study are shown in Figure 2. The traditional list-based system has the lowest score mainly because mixing videos of all the facets creates difficulties in fully understanding the topic. Our approach AH achieves the highest scores in all the criteria. For presentation and precision, because the adaptation of the hierarchy makes it more focused for video organization, while in the meantime also can improve the precision of the assigned videos (see Table 2), it is easy to understand that the users are more satisfied in both criteria. The conciseness of AH is higher because of the near-duplicate removal function. With the good performance in presentation, precision and conciseness, the highest rating of satisfaction is rather natural.

4. CONCLUSIONS

We have introduced an approach to organize video search results to an adapted semantic hierarchy for efficient browsing. Given a topic, we started from the hierarchy extracted from its corresponding Wikipedia page, and adapted the hierarchy by analyzing the video search results. The retrieved videos were then assigned to the nodes of the adapted hierarchy according to three criteria. Results on a well-defined dataset have verified the effectiveness of our approach. One promising future work is to further enhance the adaptation method in our approach, by not only removing irrelevant nodes but also adding useful nodes. Adding nodes automatically is dangerous as many found clusters are semantically uninterpretable. Therefore, an interaction process may be designed with users in the loop to reach this goal. After all, the hierarchical organization of video search results does not need to be fully online, an efficient offline approach with minimal user interactions to pre-generate the hierarchical results for hot search topics would be a big surplus to the current video search engines.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (#61228205, #61201387), a National 863 Program (#2014AA015101), a Key Technologies Research and Development Program (#2013BAH09F01), and three grants from the Science and Technology Commission of Shanghai Municipality (#13PJ1400400, #13511504503, #12511501602).

6. REFERENCES

[1] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *ACM MM*, 2004.

[2] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225, 2008.

[3] J. Goldberger, S. Gordon, and H. Greenspan. Unsupervised image-set clustering using an information theoretic framework. *IEEE TIP*, 15(2):449–458, 2006.

[4] W. H. Hsu and S.-F. Chang. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. In *IEEE ICIP*, 2006.

[5] I. Ide, T. Kinoshita, T. Takahashi, H. Mo, N. Katayama, S. Satoh, and H. Murase. Exploiting the chronological semantic structure in a large-scale broadcast news video archive for its efficient exploration. In *APSIPA ASC*, 2010.

[6] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*. 2008.

[7] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. iGroup: web image search results clustering. In *ACM MM*, 2006.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] Z.-Y. Ming, K. Wang, and T.-S. Chua. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *ACM SIGIR*, 2010.

[10] H.-K. Tan, C.-W. Ngo, R. Hong, and T.-S. Chua. Scalable detection of partial near-duplicate videos by visual-temporal consistency. In *ACM MM*, 2009.

[11] S. Tan, Y.-G. Jiang, and C.-W. Ngo. Placing videos on a semantic hierarchy for search result navigation. *ACM TOMCCAP*, 2014.

[12] X. Wu, Y.-J. Lu, Q. Peng, and C.-W. Ngo. Mining event structures from web videos. *IEEE MM*, 2011.

[13] X. Wu, C.-W. Ngo, and Q. Li. Threading and autodocumenting news videos: a promising solution to rapidly browse news topics. *IEEE SPM*, 23(2):59–68, 2006.

[14] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *ACM SIGIR*, 2004.