Metadata Enrichment For News Video Retrieval – A Graph-based Propagation Approach

Kong-Wah WAN, Wei-Yun YAU, and Sujoy ROY
Institute for Infocomm Research
1 Fusionopolis Way, Singapore 138632
kongwah@i2r.a-star.edu.sg, wyyau@i2r.a-star.edu.sg, sujoy@i2r.a-star.edu.sg

ABSTRACT

This paper summarizes our contribution to the Technicolor Rich Multimedia Retrieval from Input Videos Grand Challenge. We hold the view that semantic analysis of a given news video is best performed in the text domain. Starting with a noisy text obtained from applying Automatic Speech Recognition (ASR), a graph-based approach is then used to enrich the text by propagating labels from visually similar videos culled from parallel (YouTube) News sources. From the enriched text, we next extract salient keywords to form a query to a news video search engine, retrieving a larger corpus of related news video. Compared to a baseline method that only uses the ASR text, significant improvement in precision has been obtained, indicating that retrieval has benefited from the ingestion of the external labels. Capitalizing on the enriched metadata, we find that videos are more amenable to the Wikipedia-based Explicit Semantic Analysis (ESA), resulting in better support for subtopic news video retrieval. We apply our methods to an in-house live news search portal, and report on several best practices.

Categories and Subject Descriptors

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

Keywords

Multimedia Search

1. INTRODUCTION

Motivated by recent advances in visual analytics, the Technicolor Challenge aims to forward the search paradigm to the retrieval of unstructured multimedia based on video queries. The challenge scenario is as follows: given a short news video, output and analyse a corpus of relevant documents such as other videos or text articles on the same or related story, summarizing and elucidating their semantic links and relationships. In this paper, we propose a metadata enrichment method to address parts of this scenario.

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508122 .

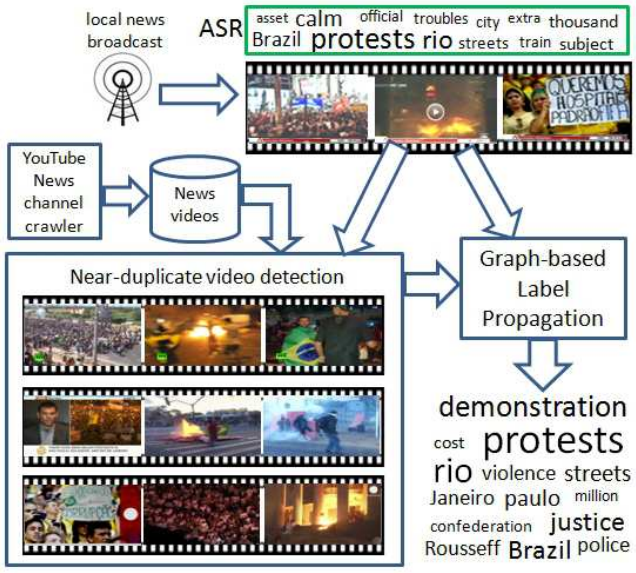

Figure 1: A short news video is used to retrieve similar videos from parallel YouTube sources, the labels of which are fed into a label propagation algorithm for textual enrichment.

The fundamental thrust in our work is that semantic analysis of content is best done in the textual domain, and that meaningful retrieval of multimedia content can only come about from the improvement in the quality of its textual description. Towards this goal, we adopt a graph-based propagation approach to enrich the textual content of a news video. Figure 1 shows the schematics of the computation in an in-house news retrieval system we have implemented. Short news video stories are first segmented from a live news broadcast. A rudimentary text representation can then be obtained by Automatic Speech Recognition (ASR). Concurrently, visually similar YouTube news videos are matched using a near-duplicate detection engine. The YouTube videos are crawled from a predefined list of official News channels, where many have proper labels, tags and short synopsis. These external labels are input to a label propagation algorithm to enrich the ASR text. The intent for this enrichment step is two-fold: missing words are recovered and important words have their prominence increased.

The enrichment procedure is our differentiation from many existing news indexing systems [6], which base their retrieval engine solely on the ASR text. While keywords are mostly picked up by ASR, personal names and important related words are often missed, and finer-grained analysis such as subtopic retrieval [9], is severely limited by the inherent noise in these ASR text. In contrast, using our enrichment process, we show how the resulting text can facilitate such semantic retrieval. In what follows, we describe details of the enrichment step in Section 2 and how it facilitates subtopic retrieval in Section 3, before concluding in Section 4.

2. METADATA ENRICHMENT

Our metadata enrichment follows the general framework of image label refinement [4], which has received considerable attention in the computer vision community. However, there have been few attempts, if any, to address video label refinement. The two main challenges are: (a) video has a temporal dimension, and labels cannot be directly associated to any segment in the video; (b) deriving a good measure of video content similarity. We circumvent the former by a bag-of-labels approach, and adopt a near-duplicate video detection approach to resolve video content similarity. We describe these in the next two sub-sections.

2.1 Graph-based Label Propagation

The basic idea in label propagation is that videos that are similar in content are more likely to also be similar in semantic meaning. A direct application of this is to first compute the pairwise similarity between videos and then to approximate the labels of an input video by its nearest neighbours. In this work, we adopt the graph-based label propagation algorithm in [11]. This is a semi-supervised learning process wherein adjacent nodes (videos) exchange an amount of information (labels) proportional to the weight (similarity) of the edges. A cost function is optimized to ensure that at the convergent state, labels are smooth amongst nearby nodes and do not deviate too much from their initial assignment.

Treat the ASR words in the i^{th} video V_i as its labels. Let the initial label assignment of n videos $\{V_i\}_{i=1}^{n}$ and m labels from a vocabulary $\{w_j\}_{j=1}^{m}$, be represented by the label matrix $Y \in \{0,1\}^{n \times m}$, such that $Y_{ij} = 1$ if the ASR word w_j is found in V_i. The goal is to produce an enriched label matrix $A \in \mathbb{R}^{n \times m}$ such that A_{ij} denotes the confidence score of the label w_j being assigned to V_i:

$$A = (1-\alpha)(\mathbb{I} - \alpha S)^{-1} Y, \tag{1}$$

where $S = D^{-1/2} W D^{-1/2}$ is the Laplacian matrix; W is the affinity matrix (see Section 2.2) where W_{ij} is a similarity score between V_i and V_j, $W_{ii} = 0$ to prevent self-reinforcement; D is the diagonal matrix $D_{ii} = \sum_{j \neq i} W_{ij}$; $\alpha \in [0,1]$ is a parameter controling the rate of propagation.

2.2 Video Similarity

Despite the abundance of literature on the subject, searching for visually similar videos remains a very challenging problem. Recent works can be broadly categorized into the use of global versus local features. In the former, global features derived mainly from color, texture, etc, are suitable for identifying near-identical videos, while local features comprising of scale-invariant keypoints are found to be more robust and effective for detecting near-duplicate videos. In this work, we adopt the local features based approach for near-duplicate video detection [10].

As the number of keypoints detected in an image can be very large, we first cluster them into a visual dictionary of size C. Each keypoint in an image is then mapped to the nearest *visual word*. We form a histogram of these words as the final bag-of-visual-words (bovw) representation of the video. The similarity between two images I_i and I_j is given by the cosine similarity of their histograms $h(.)$:

$$sim_{image}(I_i, I_j) = \frac{\sum_{k=1}^{C} h_k(I_i) h_k(I_j)}{\sqrt{\sum_{k=1}^{C} h_k(I_i)^2} \sqrt{\sum_{k=1}^{C} h_k(I_j)^2}} \tag{2}$$

We use an inverted file index to speed up the matching of the sparse histograms. Each entry in the index is a visual word and stores the links to the images containing the word. Hence, each cosine similarity is evaluated only for a small subset of the images and for the non-zero entries in the bovw.

Given two videos V_i and V_j and their keyframes, a simple way to compute the video similarity between V_i and V_j is to first exhaustively perform the image similarity scoring between all keyframe pairs, and then aggregating them. However, this does not take into account of the fact that the matching keyframe pairs must also be aligned temporally. We address this problem by a 2-D Hough Transform-like voting histogram. For each candidate video V_k matching to a query video V_q, we use k as an index along the first dimension of the 2-D histogram H_{2d}, and perform image level similarity $Sim_{image}(I_k, I_q)$ (equation 2) to retrieve all matching keyframe pairs. For each keyframe pair I_k and I_q, with corresponding timestamp t_k and t_q, we first quantize their time lag $lag = t_k - t_q$, and using it as an index along the second dimension in H_{2d}, write $Sim_{image}(I_k, I_q)$ into the bin $H_{2d}[k, lag]$. As a result, a peak $H_{2d}[k, lag]$ in the 2-D histogram indicates that keyframe segments in V_k around the timestamp lag have high matching scores in terms of local features and temporal alignment. The video similarity score is given by:

$$sim_{video}(V_k, V_q) = \max_{lag}(H_{2d}[k, lag]) \tag{3}$$

2.3 Evaluation

We perform two experiments on a dataset provided by [7]. It comprises of 127K news video + ASR-text, and 16 topical queries [1] with manually labeled ground truth. A total of 54K YouTube videos were crawled from 8 channels over the two years spanning these queries. To expedite near-duplicate detection, only the YouTube videos within a neighboring time window of each ASR video are considered for similarity matching.

2.3.1 Advantage of Label Propagation

The first experiment assesses the impact of label propagation on retrieval precision. Given an input news video, we can perform salient keyword extraction separately on the ASR text and the enriched text, to respectively generate two new text queries, denoted as Q_{ASR} and Q_{ENR}. Then, each

[1]The queries are: 1. Air France AF447, 2. Caspian Air Crash, 3. Environment Issues, 4. Italy Earthquake, 5. Pakistan Taliban, 6. US Airways Hudson River, 7. Yemenia Air Crash, 8. beijing olympics, 9. H1N1 flu, 10. mas selamat, 11. mumbai attack, 12. myanmar protest, 13. cyclone nargis, 14. obama presidential election, 15. sichuan earthquake, 16. tamil tigers.

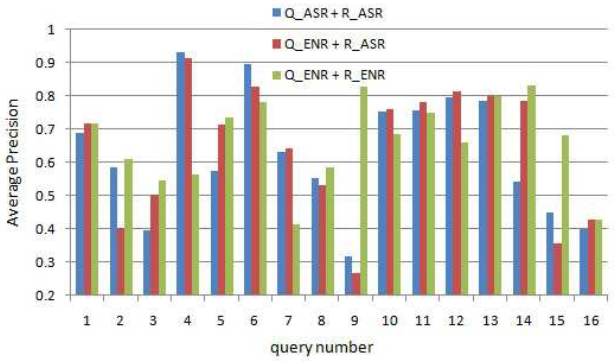

Figure 2: Comparative Average Precision results. Best viewed in color.

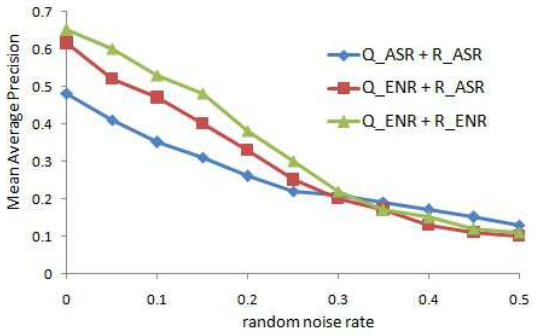

Figure 3: Comparative Average Precision results under different noise rates. Best viewed in color.

of these queries can in turn be issued to a text-based news retrieval engine, where we use the returned videos to calculate retrieval precision. The idea is that the higher the quality of a text, one would expect the better the extracted keywords and resultant query, ultimately leading to retrieved results that are more relevant. For this experiment, keywords are extracted based on a standard $tf.idf$ algorithm. Each word is weighted by $tf \times idf$, where tf is the frequency of the word in the video corpus, and idf is the inverse of the number of documents in which the word appears. The two words with the largest weight are returned as the query.

Two more additional combinations are factored in the comparison: the text-based news retrieval engine may be indexed using the original ASR text (denoted by R_{ASR}), or indexed using the enriched text (denoted by R_{ENR}). Empirically, the parameter α has little effect on the label propagation, and so is set to 0.5 for all experiments. Figure 2 shows the comparative Average Precision (AP) results. The overall Mean AP for $Q_{ASR} + R_{ASR}$, $Q_{ENR} + R_{ASR}$ and $Q_{ENR} + R_{ENR}$ are respectively 0.48, 0.61 and 0.66. As expected, the retrieval results using the ASR text issuing to an indexing system based also on ASR text performs the worst. With just label propagation, results are significantly improved, with further improvement coming from re-indexing all videos with the enriched text.

2.3.2 *Noise Resilience*

The second experiment assesses the impact of noise to retrieval performance. For each word in the video text, with a probability rate r, it is randomly deleted or a random word (chosen uniformly from the vocabulary) inserted. In Figure 3, we plot the Mean Average Precision (MAP) against noise. Even though our noise model is clearly simplistic and does not capture the underlying language model, the results are instructive. The propagation framework is resilient to small levels of noise. Beyond the noise rate of 0.25, the MAP deteriorates and the method performs worse than the ASR-based baseline. This indicates that the label propagation algorithm is very dependent on the initial label assignment. The direct implication of this result is that we should do a targeted search for videos with higher quality text that facilitate label propagation. For example, we can manually identify a few official YouTube News channels with good quality labels and synopsis from which videos are crawled.

3. SUBTOPIC RETRIEVAL

. In this section, we describe another application that has been made more amenable by the enrichment of news video text – subtopic retrieval. The problem of subtopic retrieval was introduced in [9] to study how a retrieval system can best support a user gather information about the different *aspects* or *subtopics* of a topic. It is akin to search result clustering and the removal of near-duplicates, with the goal of presenting a diversified list of results catering to the varying information needs of the users.

To elucidate subtopics in documents, many topic modeling based methods have been proposed [1]. However, a critical issue with the resulting topics is the lack of human interpretability [3]. In this work, we follow the approaches of [7] and develop an algorithm to model a news story as a distribution over Wikipedia pages. For example, a news story on "The re-capture of Mas Salemat" will be modeled by a weighted distribution over related articles like "Mas Salemat", "JI", "Internal Security Act", etc.

The motivation of our approach is three-fold. First, by interpreting each news story in terms of its affinity with a collection of Wikipedia concepts, the resulting topic model is grounded in human cognition and "explicit semantics" [5], as opposed to "latent topics" derived by Bayesian methods such as the Latent Dirichlet Allocation (LDA) [1]. Second, Wkipedia as the space of reference articles gives us the added advantage of tapping into a resource that is almost complete and up-to-date. Third, the editorial content of each Wikipedia page is organized into a taxonomic Table-of-Content (TOC) structure which not only readily elucidates the needed perspectives or aspects of the topic, but also allows a non-linear search and facilitates exploratory learning.

All that remains is to map a given video text to the related Wikipedia TOC content. To do this efficiently, each Wikipedia page is first XML-parsed to extract the TOC entries. Short TOC entries (<30 unique non-stop words) are merged with an adjacent TOC entry. An inverted index is then built that maps each word to a list of TOCs in which it appears. Given a story text, we iterate over each word, retrieves the corresponding TOCs from the inverted index, and count-aggregate them into a weighted vector of concepts, such that each feature in the vector encodes the relevance of the corresponding Wikipedia concept. To compare between two vectors, we compute the Cosine similarity $Cosine_{wiki}$. Clearly, the better the quality of the input video text, the more accurate is the mapping to the concepts.

Table 1: Comparative S-recall results (using MMR).

	S-rec@20
ASR baseline ($Q_{ASR} + R_{ASR}$)	0.277
Enrichment-1 ($Q_{ENR} + R_{ASR}$)	0.318
Enrichment-2 ($Q_{ENR} + R_{ENR}$)	0.327

3.1 Evaluation

The goal of subtopic retrieval is to include in the early ranks, documents that cover as many different subtopics as attested in the corpus. More formally, given the set of known subtopics f_i underlying the videos related to a query video V_q, and a set of videos $\{V_1, V_2, \ldots, V_k\}$ retrieved up to rank k, subtopic retrieval is evaluated by *S-recall* [9]:

$$S\text{-}rec@k = \frac{1}{n_q}\sum_{i=1}^{n_q} I(f_i \in \{V_1, V_2, \ldots, V_k\}) \qquad (4)$$

where n_q is the number of subtopics of q, and $I(\cdot)$=1 if f_i appears in any of the videos ranked 1 to k, and 0 otherwise. At any k, *S-rec*@k rewards returning a list of k videos that contain as many of the n_q subtopics as possible.

We implement a diversity retrieval model called Maximal Marginal Relevance (MMR) [2]. This is a greedy method that selects the i^{th} video V_i according to a combination of its similarity to the video query V_q and its similarity to all higher-ranked videos at position 1 to k-1:

$$\begin{aligned}\mathbf{MMR}(V_i, V_q) \quad = \quad & \beta Cosine_{wiki}(V_i, V_q) - \\ & (1-\beta)\max_{1\le j<i} Cosine_{wiki}(V_i, V_j) \quad (5)\end{aligned}$$

where $\beta \in [0, 1]$ is a trade-off parameter. We set $\beta = 0.5$.

At the time of this writing, we used as queries a small collection of 33 topic ground truths [2] provided by [8], which are based on the TRECVID-2005 dataset. For subtopic retrieval evaluation, three human assessors further annotated subtopics on these video topics [3]. As per the evaluation methodology in Section 2.3.1, we apply keyword extraction to the original ASR text and enriched text to generate two new queries Q_{ASR} and Q_{ENR}. These are issued to the two video indices R_{ASR} (based on ASR text) and R_{ENR} (based on enriched text). Table 1 shows the S-recall measures on the respective returned videos. It is clear that the text enrichment process has improved subtopic retrieval.

Figure 4 shows a snapshot of our system UI. Given a user query, returned news videos are organized according to the Wikipedia TOC (right panel). The videos in each TOC group are playback in reverse chronological order. Users can click on any TOC group for non-linear browsing.

4. CONCLUSIONS

There is a general anticipation that recent advances in visual search can bring much progress to the domain of multimedia search. However, it is our view that the semantic gap issue remains unsolved, and that text processing still holds the best prospects for progress in semantic analysis of content. In this paper, we describe a label propagation method that enriches the textual content of videos. We show that with proper targeting of external video resources with good quality labels, raw videos become more amenable to search, and finer-grained semantic analysis such as subtopic retrieval can be facilitated. Specific future works include investigating the use of bigrams in the label propagation.

[2]Sample topics: "Bush visits Canada", "Arafat-health", etc.

[3]We are in the process of creating a larger corpus of 400 topic queries, fully annotated with subtopics, and available upon request. We will also report subtopic retrieval lresults on this corpus in a later publication.

Figure 4: UI of News subtopics browser. See text.

5. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *Proc SIGIR conf on Research and development in informaion retrieval*, pages 335–336, 1998.

[3] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, 2009.

[4] L. Dong, X. Hua, M. Wang, and H. Zhang. Image retagging. In *Proc ACM Multimedia*, pages 491–500, 2010.

[5] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc Int'l Joint Conf on Artifical Intelligence*, pages 1606–1611, 2007.

[6] J. Law-To, G. Grefenstette, and J. Gauvain. Voxaleadnews: robust automatic segmentation of video into browsable content. In *Proc ACM Multimedia*, pages 1119–1120, 2009.

[7] S. Roy, M. Mak, and K. Wan. Wikipedia based news video topic modeling for information extraction. In *Proc Multimedia Modeling*, pages 411–420, 2011.

[8] X. Wu, A. Hauptmann, and C. Ngo. Novelty detection for cross-lingual news stories with visual duplicates and speech transcripts. In *Proc Multimedia*, pages 168–177, 2007.

[9] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proc SIGIR conf on Research and development in informaion retrieval*, pages 10–17, 2003.

[10] W. Zhao, X. Wu, and C. Ngo. On the Annotation of Web Videos by Efficient Near-Duplicate Search. *IEEE Transactions on Multimedia*, 12:448–461, 2010.

[11] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schlkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328, 2004.