The Use of Topic Evolution to Help Users Browse and Find Answers in News Video Corpus

Shi-Yong Neo, Yuanyuan Ran[1], Hai-Kiat Goh, Yantao Zheng, Tat-Seng Chua, Jintao Li[1]

School of Computing, NUS, [1]Institute of Computing Technology, CAS

{neoshiyo, gohhaiki, yantaozheng, chuats}@comp.nus.edu.sg, [1]{ranyuanyuan, jtli}@ict.ac.cn

ABSTRACT

Earlier research in news video has been focusing mainly on improving retrieval accuracies given the limited amount of extractable video semantics. In this paper, we propose an enhancement to news video searching by leveraging extractable video semantics coupled with relevant external information resources to support event-based analysis; leading to discovery of topic hierarchy for browsing key events and supporting question answering (QA). We introduce topic browsing based on news structures obtained through hierarchical clustering and threading, with emphasis on interesting events determined by measuring the amount of "*web activities*" on these events on Blog sites. For QA, we employ extensive query analysis to obtain various query features in addition to the topic hierarchical structures to answer both context-oriented and visual-oriented questions. Our main contributions includes: (a) combining multimodal event information extracted from news video, web news articles and news blogs to support event analysis, (b) introducing topic evolution browsing based on users' interest and (c) extending QA on top of topic hierarchy to handle various types of specialized video queries. Experiments performed on 70 hours of multilingual news from TRECVID 2005 dataset shows that the proposed approach is effective and appealing to users.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Search Process

General Terms: Design, Experimentation

Keywords

Video Question Answering, Video Analysis, Event Evolution

1. INTRODUCTION

Effective multimedia information retrieval is becoming increasingly important especially with the ever-increasing amount of multimedia data. Such a need calls for systems that are capable of retrieving and integrating relevant information from multiple sources and in different media in order to present users with a richer set of answers with different perspectives to enhance understanding. This is especially true for the more structured type of information such as the news where abundance information in both text and video is available. It has been shown in [6] that most users tend to pay attention only to the top few search results. Further, researchers in streamsage.com [11] observed that while "the cost of returning an extra paragraph of text is trivial, retrieving extra media contents not only waste a significant amount of user's time but may cause the user to skip the rest of the media altogether". Thus most existing video retrieval research has been focusing on improving retrieval precision. However, accuracy is not the only requirement end-users are looking for [6], it is also important to "embed intelligence" into the system such that users can have some controls over individualizing their information needs especially when faced with myriad of information from different sources and medium types. It is therefore necessary to provide a substantial amount of analysis between the users and system to bridge the information gap. A typical user query for news video usually comprises of either a **query topic** (*Arafat, US election, etc*) and/or a **specific aspect of that topic** ("*Shots related to the death of Arafat*", "*victims in earthquake*" etc). It can be interpreted that the first type of question is looking for a specific set of news materials relevant to the topic while the second type is looking for exact answers.

Topic Retrieval. Most users tend to search for online news periodically and they often center among topics of their interest (e.g. sports, entertainment, financial, etc) or issues which may have an direct effect on them (e.g. election, annual budget, etc). Special news relating to disaster, crimes, international trials are also prime news that most users would like to see. At present, news video search engines display lists of candidate results arranged in order of relevance to the users (see for example, commercial site such as www.streamsage.com). Such an arrangement might be good for collecting data related to the topic, but may be too data-overwhelming for most users. Looking ahead, retrieval can be more appealing if users are first presented with an overview containing various interesting stages of a news topic at that time instance, before they decide on which one to explore further. For example: given a query on "*Arafat*" on November 2004 where Arafat has just passed away, it would be good to present a hierarchical overview of reports arranged in a chronological order with sub-topics like "Arafat is hospitalized", "Arafat has fallen into coma", and "Arafat is pronounced dead by the Palestinian officials". This overview is time-dependent, and the very same query when posted in December 2004 might show another group of results like "report of his funeral", "who has taken over his position", "effects of his death" etc. This kind of grouping of search results is similar to text clustering done in a commercial search system named Vivisimo [26]. The added advantage of such a presentation is that it shows the evolution of news topic when arranged chronologically.

Question Answering (QA). In contrast to topic retrieval, QA aims to find multimedia answers targeting at specific aspect of a topic. A user initially looking for news on "*Arafat*" may have following-on questions like "*when was Arafat hospitalized?*",

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

"*which hospital did he go?*" An effective combination of query analysis as well as concise understanding of a given news video must be available to handle such queries accurately. Beside the usual context-oriented questions, users may also be interested in looking for visual details like "*shots containing Arafat*" or "*shots on Arafat's funeral*". This type of questions will require certain visual intrinsic semantics for effective retrieval.

For both types of retrievals, we aim to provide users with a multimedia answer that is more engaging, fun and informative to watch. At present, the limited amount of video semantics obtainable from within the news video contents is not sufficient to perform such detailed analysis. This is because news video is often presented in a summarized form and various important contexts may not be available, or is often erroneous such as the ASR (automatic speech recognition) text. In this paper, we propose the fusion of news video with various external news resources in an event-based fashion to enhance news video search. In particular, we utilize relevant online news articles automatically harvested from the Web to supplement the limited context and content of news video. This is done by performing hierarchical clustering using event-entities extracted from both news video and news articles to thread events and derive the *topic hierarchy*. In addition we utilize the statistics obtained from news blogs to gauge the "*level of interestingness*" of events by measuring the "*related web activities*". The user can then browse through the generated *query topic-graph* (a sub-graph of the topic hierarchy). For video QA, we extend the techniques developed in open-domain text QA [24, 31]. In particular, the system uses the resultant query topic-graph, with the help of extensive query analysis, to effectively fuse available multimodal features.

Our contribution is three folded: (a) We combine multimodal event information extracted from news video, Web news articles and news blogs to support event analysis. (b) We introduce topic evolution browsing based on users' interest. (c) We perform QA on top of topic hierarchy to support various types of specialized video queries. The resulting system is tested on 70 hours of TRECVID [25] 2005 news video and is found to be effective and user appealing.

2. RELATED WORK

The main emphasis of this work is in leveraging event information from external information resources to support topic evolution browsing and QA on news video. Our work is related to other research on video retrieval and text-based QA. One such related work is the well-known *Informedia* project, which covers most aspects of feature extraction, segmentation and retrieval of news video [8]. Similar to our approach, they also utilized news transcripts and external news articles. In particular, they used the extracted name lists harvested from news transcripts and external news articles to improve the accuracy of video OCR [27]. Grouping of events by their relative similarities and differences also helps to track events across time. This has been introduced in text-based topic detection and tracking (TDT) [2], which uses lexical similarity of document texts to generate coherent clusters, in which elements in the same cluster belongs to the same topic. The leverage of such topic/event structures from news video provides excellent partial semantics for retrieval as well as news video story boundary detection and threading [15]. [18] showed that the recognition errors in ASR can significantly degrade the performance of news video clustering. It is therefore necessary to supplement the incomplete and erroneous ASR with external information sources in order to obtain a clustering result that is more representative and descriptive of the underlying distribution of news events/stories. [19] further integrated HLFs (High-Level Features), such as fire, car, face etc, to provide additional context and knowledge about the events in news stories and shots. For example: if there are shots containing the HLF "*fire*", it can strongly indicate stories on topics like "*forest fire*", "*fire breakout*", "*explosion*", etc. As the same events tend to yield similar visual and context, [19] showed that it is appropriate to base retrieval on clusters at the pseudo story level and subsequently re-rank them at shot level.

QA has been extensively researched in the domain of text processing where systems were developed to answer open domain questions [24]. State-of-the-art QA systems are able to answer more than 65% of queries correctly over a corpus consisting of over 1-million documents [4]. With this hindsight, [30] extended the techniques from text processing to video QA by modeling the Web and linguistic knowledge for effective QA. As video itself is multimodal, it is necessary to consider the various multimodal features during fusion. Here [29] proposed 4 different query-classes within the domain of general news videos to perform class-dependent multimodal fusion. The four classes are Named-person, Named-object, General-object and Scene. Given a query, they performed query analysis to categorize the query and employed appropriate query-dependent model to fuse the multimodal features. To handle visual oriented queries, [20] proposed incorporating HLFs using WordNet [12] in pinpointing exact shots. [1] also automatically mapped query text to HLF models trained using SVM. The weights are derived by co-occurrence statistics between ASR tokens and detected concepts as well as their correlations. They found that the use of HLF is very effective for answering visually oriented queries. From these related works, we apprehend the importance of several key semantic features in retrieval of news video.

3. EVENT AND NEWS VIDEO

In this research, we define an event as happenings that occur at a given time and place; and a topic as a grouping of related events occur across temporal domain. Our definition of an *Event* follows that used in the context of text QA [31] as depicted in Figure (1). News reports are typical depictions of important and interesting events that the users might want to see. They usually contain various aspects and entities like: *Location*, *Time*, *Subject*, *Object*, *Quantity*, *Description* and *Action*, etc. In addition, news video also contains multimedia depictions that reveal the visual aspects of an event. By considering visual concepts that occurs within these news video, we can further obtain visual event information which is previously not available from text. These visual concepts or HLFs (High-level features) can be added to the basic event definition shown in Figure (1).

event : == { event_element }
event_element : == time |location |subject |object |quantity |description| action |others
Event_Template : == {event_element} + {HLF}

Fig 1: Event definitions

To extract the HLF semantics from news video, recent work has taken a machine learning approach, where a detector for each

HLF is trained against an annotated corpus of video clips as discussed in TRECVID HLF task [25]. Another well-known HLF set is the LSCOM [17], which contains approximately 1,000 concepts catering to support news video retrieval. In our system, we leverage 50 HLFs from our previous work [20].

3.1 News Video Event Space

Here, an event is considered to be a distinct point in the multi-dimensional event space. Figure (2) presents a 4-D view of an event, which contains the information about the time, location, action and its HLF. One example of such a 4-D event can simply be a video of President George Bush giving a speech at the White House on 3rd of Nov regarding presidential election. Given such complete information, it will be possible to handle other enquiries related to this event such as "*Where did Bush present his speech on presidential election on the 3rd of November?"* However, due to the erroneous nature of ASR, it is normally difficult to obtain all the necessary entities relating to an event. This limitation prompts the use of relevant external resources to supplement news video. The innate associations among elements in an event make it possible to leverage mutual information [16] to group known event elements together, and even subsequently predict missing entities.

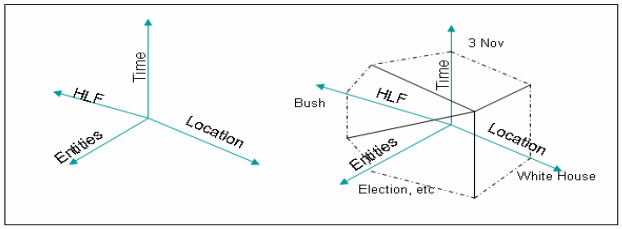

Fig 2: Illustration of elements in news event

Prior to constructing the feature space, the first task is to define a suitable basic unit for event representation as news video is continuous. Even though shot boundary detection [21] rates are excellent, we choose to base on pseudo story segments as they provide more coherent textual and visual semantics. [2] has also shown that story level segments are better in terms of semantic coherence than shot level segments. We use the story boundaries provided by [15], and enrich the boundaries by utilizing information on anchor-person shots to perform second level segmentation [7]. The main reason for doing this is that we prefer over-segmentation rather than under-segmentation as the latter tends to cause the merger of different stories to overlap more frequently.

The extraction of event features from news video is critical to the overall process of discovering event evolution as well as question answering. The speech stream of video, make available through ASR, is primarily responsible for the event entities in news video events. However, as ASR is not perfect, as well as the unavoidable translation errors from Machine Translation (MT) for non-English news video, there is a strong need for supplementary sources of information which are reliable and relevant to news video. For this reason, we employ external news articles as in [9], which are automatically collected from within the same period as the news video, to provide the necessary additional context. As these news articles are grammatically correct, it is possible to apply morphological analysis to obtain the Part-of-Speech (POS), which allows for more complete analysis of the document. These POS-tagged documents are then passed to the extractor module used in [31] to extract various Name Entities (NE) such as the person's name, location, etc. Each news article will be represented by its set of NEs with respect to time. As ASR often contain incomplete sentences, it is not possible to obtain the POS accurately. We therefore only apply the NE rule-lists in [31] to match known NEs directly within the designated pseudo story boundary. The HLFs which are deemed to have a high confidence of appearing in the news video stories is also added. As a result, each news video story is represented by a list of NEs and HLFs. The news video stories and external news articles are then combined together for clustering. The overall framework of the proposed system is shown in Figure (3).

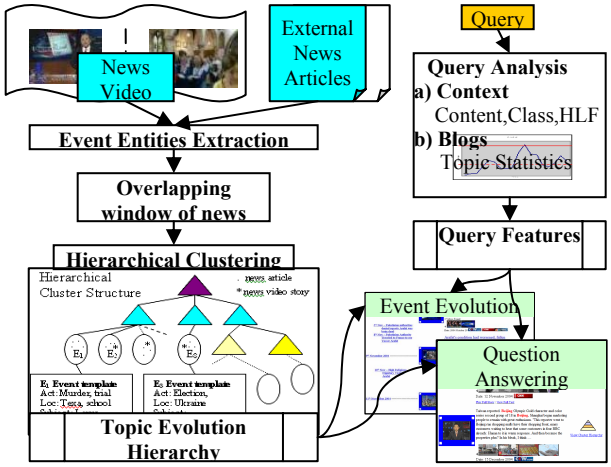

Fig 3: Overall framework for news videos searching

3.2 Hierarchical Event Clustering

[27, 32] have shown that the NE lists harvested from parallel news articles are useful in recovering missing person names in the OCR and ASR transcripts. However, in our system, other aspects of events are also important. For example in an event like "*plane crash*", the number of casualties might be the key important numbers that people are interested to know. As only limited amount of information can be derived from ASR, it is unlikely that the full aspects of event can be obtained. Furthermore, broadcast news is often presented in summarized form, and certain aspects of events might also be unavailable at all. We therefore propose a strategy by combining the news articles and news video stories in a same clustering space. The intuition here is to leverage innate relationships between event entities in both sources of information to complement each other. For example, two similar news videos of a same event having missing key entities due to ASR errors may be clustered wrongly into different clusters. However, the parallel news articles may provide the semantic bridge by providing the mutually overlapping entities between the two news videos.

Performing a single clustering on the entire corpus [19] is straightforward and simple. However, such clustering process has two major drawbacks. First, it might not deliver satisfactory clustering results, as there exist a considerable number of outlier news stories that are reported only once and these outliers may compromise the clustering process leading to incoherent clusters. Second, the clustering of entire corpus is computationally expensive. For example, the upper bound of k-means complexity

[14] is $O(n^{dk})$, where n is the number of samples, d is the number of dimensions and k is the number of clusters. The tremendous amount of news video and stories in the corpus (around 10,000) and its high dimensional feature space (around 1,500) makes k-means infeasible even on a high-performance machine. In order to tackle the above two issues, we choose to perform clustering on an overlapping sliding temporal window that takes into consideration interesting events tend to have multiple news reports within a short period of time. In addition, rare events or single-reporting news events can be modeled more appropriately in the sliding window. We experimentally set the sliding window size to five days with an overlap of two days. To further ensure clustering efficiency and accuracy, we propose an *asymmetric hierarchical k-means clustering*. One problem with traditional hierarchical k-means clustering is that it overlooks the cluster size and cluster density because it simply performs the hierarchical clustering until the required depth is reached. The asymmetric hierarchical k-means clustering is more suitable for our implementation as it constraints the size of a cluster to be S. Furthermore, it checks the cluster density D at every iteration and stops the recursive clustering on partitions where the threshold is reached. This clustering approach can reasonably ensure the quality of cluster by using cluster density maximization so that the sporadic outliers will have less probability of being clustered together with major clusters. We achieve this by setting the number of clusters k = 2 for each k-means. With k = 2, k-means behaves like a dichotomizer [5] and clusters the data samples based on its distance from two cluster centroids. The distance measurement use is the cross feature-vector cosine similarity. As news events of same topic tend to have similar feature values, they are highly probable to be assigned to the same cluster at each level of the k-means clustering. The total time taken for clustering the whole corpus is about an hour. In the experiments, we define the cluster density D for cluster C_j with centroid c_j in Eqn (1):

$$D = \frac{1}{n_j} |\mathbf{S_w}| \qquad (1)$$

where $\mathbf{S_w} = \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{c_j})(\mathbf{x} - \mathbf{c_j})^T$, $\mathbf{c_j} = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$ and n_j is the size of cluster j. Consequently, the number of clusters will be less than 2^d as d increases. More importantly, the resulting clusters may not be in the same level of the hierarchy and the large clusters with many outliers can be further split without dividing the desired coherent small-sized clusters of the same or similar news events.

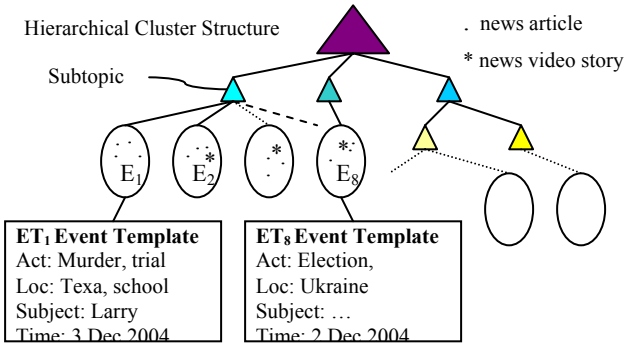

Fig 4: Hierarchical cluster structure with event-templates

Besides efficiency and accuracy, the asymmetric hierarchical k-means clustering is able to generate a more natural hierarchical organization of news events which can be leveraged to provide additional structure information as shown in Figure (4). Each node in the hierarchy consists of news articles and/or news video.

3.3 Event Template and Topic Hierarchy

After obtaining the hierarchical clusters for each sliding windows, we need to find a mechanism to link and thread these clusters together so that they can be used coherently to support retrieval. To achieve this, we employ a template filling and matching approach by making use of the events entities contained within the elements of each cluster. A new video Event-template (ET) containing the list of entities is shown in Figure (1). There are 3 main reasons for extracting an ET at the cluster level. First, news about the same event tends to have several news videos or reports, either from the same news source or from multiple sources. Second, different news reports contain entities that complement one another. Third, by merging event elements at the leaf nodes and propagating upwards, we can obtain event elements which describe the entire cluster belonging to the same parent node.

To fill an ET effectively for each cluster, we need to find the minimal cover which can best describe the cluster. We first consider all the event entities within the cluster and group them separately into respective entity type such as time, location, subject, etc. Thereafter, we assign a confidence score for each of these entities based on their frequency and rareness within each cluster. This is similar to the TF.IDF ranking often used in text retrieval [10]. The formula for the confidence score is given by:

$$Confidence\ (e_j \in C_k) = Freq(e_j \in C_k) * Log_2(\frac{M}{m}) \qquad (2)$$

where e_j is an event element in cluster C_k, M is the total number of instances in C_k and m is the number of instances containing e_j. Only event elements within each entity type with a confidence score of above a pre-defined threshold δ will be added to the template.

Next, we will make use of the template from each cluster to thread highly similar clusters across various temporal hierarchical clusters as shown in Figure (5). The similarity between two templates ET_a and ET_b is computed as in Eqn (3).

$$Sim(ET_a, ET_b) = \sum_i \alpha_i \cdot CommonEnti\ ties_i^j \qquad (3)$$

where $\sum \alpha_i = 1$. $CommonEntities^j$ checks for overlapping terms of the same type (for example: to check if elements in the location or subject type are the same). The time similarity is computed based on the difference between two timed events normalized by the duration of the event. Generally, α-values for location, subject and time have higher weights as compared to the rest of the entities. A high *Sim()* score will intuitively mean that the two events are closely related.

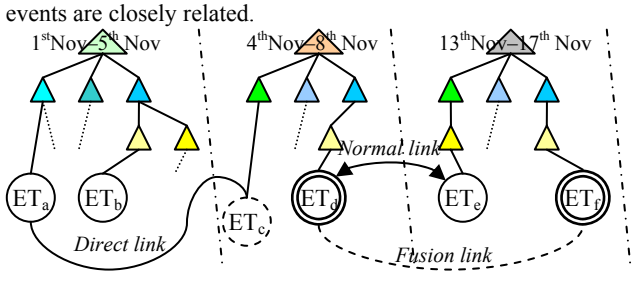

Fig 5: Topic hierarchy based on event threading

In our implementation, 3 types of relationship linkages are considered: "identical", "near-duplicate" and "high-similarity".

The first type, "identical", occurs when all the news video elements within two different cluster templates are the same. This happens when an event extends over two overlapping sliding windows. In this case, a *direct link* will be created between the 2 hierarchies (which is equivalent to linking the hierarchies at the non-leaf nodes or events). The second type, "near-duplicate", occurs when *Sim()* score is above δ_n (where δ_n=0.95). Here the two clusters have many overlapping news video instances, and a *fusion link* is created for this type of relation. The third type, "high-similarity" occurs when clusters that have a similarity value of above preset threshold δ_s but below δ_n. In this case, a *normal link* is created.

4. TOPIC EVOLUTION

The topic hierarchy shown in Figure (5) simply details the range of possible events within topic. While some events are important and interesting to the users, many are trivial and uninteresting. To understand what key events within the topic hierarchy are likely to of interest to the users, we leverage other sources of news related information, such as the number of news posting, total news video broadcast duration, or blog activities related to that event.

4.1 Blog Analysis of Key Events in Topic

In this research, we propose to leverage on unrefined "collective intelligence" available on the Web. The numerous online news website commentaries and news web blogs provide valuable resources to obtaining such intrinsic information about the interestingness of news events. Given the availability of such information, we attempt to identify important time periods in the topics using news blogs. There is a growing mass of people expressing their views and ideas on events happening around them in the form of web blogs. The events they commented on range from their everyday life, current news, animal rights issues, to rumors on celebrities. A typical web blog consist of text, images, videos and links etc related to the topic. When a particular event that has great interest to many people happens, we can observe a sharp rise in "*web activity*" on that event and its related topics. For example: "*the capture of Saddam Hussein*", which triggered a huge number of blog postings and news articles relating to him in December 2003. It was only in 2003 where worldwide blogging has just picked up. In addition, we can also see an overwhelming amount of "*web activities*" when Saddam was put to death in early 2007. Thus a sharp increase in postings on a topic usually suggests that an important event has occurred in that topic. According to this phenomenon, there is an implicit but direct correlation between "*web activities*" and the stages of evolution of events/topics.

It will be useful for event analysis if we know the dates of important events for a particular topic. To achieve this, we first retrieve the news blog postings relating to the topic from Technorati.com [23], which is the largest online web blog index search engine. Next, we employ a crawler to retrieve related blog postings that are within a single-link away from the initial retrieval results. The reason for performing a single-link crawl instead of multiple-link crawl is to prevent linking to unrelated topics as these blog sites usually contain many external links (e.g. online commercial, link to other blogs etc). To further ensure that these postings are relevant to the topic, only those that have overlapping topic key phrases in the title will be considered. Third, we extract the date of each posting. A statistical plot of "news of Arafat's" for a period in Nov 2004 is shown in Figure (6a).

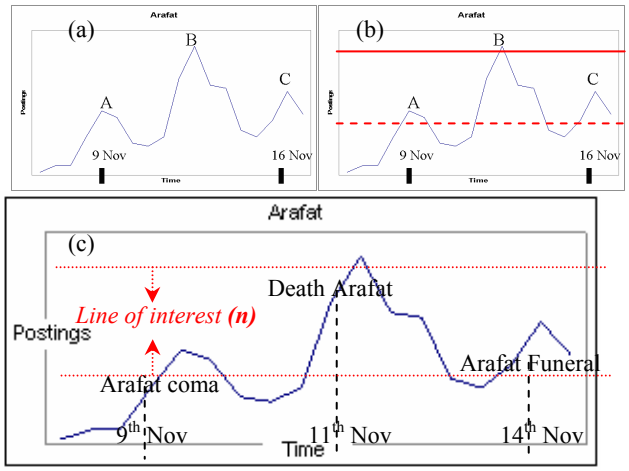

Fig 6: Topic Statistics by blogs for "Arafat" in Nov 2004.

From Figure (6a), we can see a number of fluctuations in the number of postings. The total number of relevant postings collected for the entire month is about 5,000, and the difference between the days with the highest (304) and lowest (33) number of postings is close to 300. The first peak (A) on the left happens when Arafat was reported to be critically ill in the hospital. The highest peak (B) occurs on the third day after Arafat is pronounced dead by the officials. News often has a life-cycle after it has been published; it requires a time period for the public to get interested and start the discussion. After this period, it will begin to quiet down until the next wave of discussion triggered by some new related events. These statistics accurately reveal the dates of events which are of great interest to the public. In addition, it also facilitates the extraction of views at different granularities across time. For example in Figure (6b), by projecting the line of interest at different blog postings level (represented by the red dotted line), we can find other dates that might have other interesting events going on. As the number of postings varies greatly across time, we model this posting number using a logarithmic function to smooth the curve. An interestingness factor ε (simulates the level of interest) is added to calculate the periods of high online activities as shown in Eqn (4).

$$Interestingness(t) = \log(P_t) - \frac{\max(\log(P_t))}{\varepsilon} \qquad (4)$$

where P_t is the number of postings at the given time t and factor ε can be adjusted to return positive value on dates which are deemed to have high interest. As the date of high Blog activities usually lags behind the date of actual event, it is important to backdate appropriately to obtain the actual event date. From observations, we find that this time-lap is usually one to two days after the occurrence of the event. In our implementation, we empirically set the date of actual event to be the nearest date of news report relevant to that event before the Blog date. In addition, the density of news reports is also considered and priority is given to dates with a sharp rise of news relating to the topic.

4.2 Extraction of Query Topic-graph

Given a query from the user, the system first performs text retrieval to obtain the relevant video stories and/or articles that contains the topic terms from the corpus. These retrieved documents are then mapped to the respective nodes containing them in the topic hierarchy as shown in Figure (7). The resultant sub-graph is the *query topic-graph*.

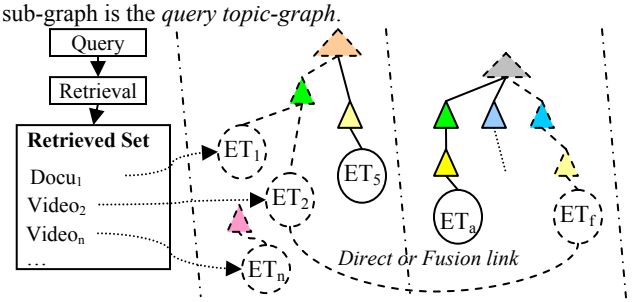

Fig 7: Query topic-graph (denote by dashed lines)

To further leverage on threading information, nodes that can be reached by a single direct or fusion link are also added as part of the query topic-graph. A set of representative words is then generated to represent each node. These keywords (denote by *headers*) will provide a short context summary of the news materials and they are generated by retaining prominent key terms and noun phrases using both news article titles and ASR of news video in each node. In particular, we limit the header for each node to 12 words. The hierarchical structure allows the parent nodes to contain all the traits of the child nodes.

4.3 Displaying News Videos in Topic-graph Space to support Browsing

To display the most important stages or interesting events of the topic, we rank videos V_i within the query topic-graph using highly interesting dates t_{dates} (from Eqn 4) and user query Q as:

$$Score(V_i, Q, t_{dates}) = \beta \cdot Text(V_i, Q) + \gamma \cdot Sim_q(ET(V_i), Q') + \delta \cdot Time(V_i, t_{dates}) \quad (5)$$

where the $\beta+\gamma+\delta=1$ are trained weights. *Text()* is the text retrieval score computed using Cosine similarity formula. $Sim_q()$ is modified from Eqn (3), which compute the similarity based on overlap between the expanded query Q' to the event-template ET of V_i without considering the various entities type. Here Q' is a list of terms with high mutual information [16] occurring together with the original query terms Q derived from parallel news articles (more details in Section 5.1). *Time()* measures the time similarity and is computed based on the difference between two timed events normalized by the duration of the event.

Fig 8: Topic evolution for "Arafat" in Oct/Nov 2004

The user will be presented with a ranked list of videos with their descriptive headers in as shown in Figure (8). In addition, the user is also able to view other relevant materials to his topic by traversing the topic-graph in a hierarchical (▲) fashion.

5. VIDEO EVENT QA

While the topic evolution browsing provides the overview of a topic at the global level, the event-based QA aims to return precise video answers to questions posed over the news video topic. It can be used in a personalized video setting in which a user may request for **details** of different aspects of news such as "*How old was Arafat when he passed away?*" (looking for a number relating to age) or *"When did Arafat pass away?"* (looking for a date). For this type context-oriented question, users expect the system to return short video segments with speech containing the answers. The existing text-based QA system in TREC [24] rely on the inferred *answer-type* (whether the query is looking for a name, time, location, etc) to return the possible answer candidates. In addition to context-oriented questions, users may also be interested in visual-oriented questions like "*Find shots containing Yasser Arafat*" or "*Find shots of people holding signs or banner*". These questions would require segments of videos containing the required visual elements. Each type of question requires a different retrieval strategy for maximum efficiency. Applying this concept to video event QA would mean returning different aspects of an event.

5.1 Query Analysis

In order to return the most relevant segments for the user, it is necessary to interpret the user's question correctly and retrieve the most relevant segment given the various multimodal features. We first perform detail analysis on the given question to extract query content, query class, query answer target and query-HLF. This analysis is crucial as it helps us to understand the users' intention and differentiate between the topic and the constraints.

5.1.1 Query Content

Content keywords are taken as the topic of the query, which are normally the strongest nouns or noun phrases which itself can also be name entities (e.g Bill Clinton). As the original question is usually short and contains little contextual information, it is hard to only rely on these few query-terms to retrieve relevant video stories precisely. In our implementation, we induce additional query terms Q' by generating a list of keywords which has high mutual information with the original query terms Q. This is performed by expanding the query using the set of news text articles collected in the same period as the video data. Query expansion has shown to be useful in [20, 30].

5.1.2 Query Class

Query class [28] is the next important feature that has been shown in many prior works to be effective in fusing multimodal features. In this work, we employ nine query-classes [29] as follows: {Person, Sports, Finance, Weather, Disaster, Health, Politics, Military, General}. The General-class is created to accommodate the queries that do not belong to any of the first eight classes. The main reason for this classification scheme is to create an explicit mapping of query-class to video program-genre. For example: the answer shots for sports questions are normally found in sport news; and similarly for financial news and weather news. These

nine classes are also chosen because they can be easily classified by using simple heuristic rules based on textual and color information. This is important as it is not possible to perform complex query classification for short text queries. In addition, we allow a user question to be classified into a maximum of 2 classes. This is reasonable as there are questions which can belong to more than one class. One such typical question is: "*Financial crisis in Thailand?*" can be classified into Finance or Politics.

5.1.3 Query Answer Target

Answer-type refers to the entity types of required answers which could be name of a person, location, date, etc. With the knowledge of answer-type, the system can effectively narrow down the search range by looking for the presence of such entities and rank answers in a systematic way. We make use of a rule-based classifier developed in [31] to perform answer typing. Table 1 illustrates some of the questions and their respective answer-type. This classifier is not comprehensive to cover *all* questions but is sufficient to handle most event entity targeted questions.

Table 1. Sample questions and their event answer-target

Question	Rule	Answer-type
What is the name of the serial killer?	name+person	Per-name
Who is the President of U.S?	Who	Per-name
Which team won the Stanely Cup?	Which+team	Org-name
Where is Osama?	Where	Location
Which are the states which suffered tornadoes?	Which+state	Loc_state
Which teams played in the Stanely?	Which+team	Org-name

5.1.4 Query-HLF

The query-HLF plays an important role in answering visual-oriented queries as it is evident that text alone is not sufficient to pin-point exact shots or frames which contains the required visual evidence. The query-HLF measures the importance of a HLF with respect to a query by performing descriptive lexical matching. This is done by performing morphological analysis using the WordNet [12] lexical database on both the HLF feature descriptions and the query-terms. The details of this matching can be found in [20]. However, having a close lexical relation may not imply that the particular HLF or query-term may actually be appearing together. After all, lexical similarity may not necessarily translate to visual co-occurrence. For example: cars, planes and ships are all modes of transportation but it is quite unlikely to find two such modes appearing together in a single image or video shot. In order to know how to effectively tell whether certain visual objects or concepts are likely to appear together, we utilize another valuable online resource Flickr [13], which houses millions of user tagged photos, to predict the visual co-appearance between concepts. From Flickr statistics, it is possible to know how frequently certain visual objects can coexist within a single image. For example, Flickr statistics shows the four most frequently occurring tags with "*sky*" are "*blue, cloud, sea, ocean*". It is therefore reasonable to assume that these four concepts are more likely to coexist with "sky" than other concepts. Eqn (6) shows the similarity score of query Q_j to HLF_k.

$$Sim_{HLF}(Q_j, HLF_k) = \eta Lex(Q_j, HLF_k) + \chi Flickr(Q_j, HLF_k) \quad (6)$$

where $\eta + \chi = 1$. *Lex()* make use of the information-content metric of Resnik [22] to compute the similarity between pair of words based on WordNet concept hierarchy. *Flickr()* determines the likelihood of co-occurrence between visual concepts using Flickr statistics.

5.2 Retrieval and Answer Extraction

Given a new question from the user, the system first performs an initial round of retrieval (similar to Section 4.2) to obtain the query topic-graph. Alternatively, the same query topic-graph can be employed if the user specifies the question as a follow-on question to a topic query. Using the query topic-graph, different ranking strategies are used to answer the context and visual oriented questions.

5.2.1 Answering Context-oriented Questions

The algorithm for answering context-oriented questions employs a density-based ranking (using minimal distance between the matched words) to measure the answer candidates in terms of: locality context, answer-type and the proximity between the answer and the query-terms using Eqn (7). A question "*when did Arafat pass away*" can be answered more confidently if the correct sub-topic trees are employed during matching. For example, relevant sub-topics are those concerned with his "*death*" rather than his "*funeral*" in the Arafat topic-graph. Additional weights can be given to answer candidates found in the correct subtopic trees. A score will be given to every possible answer candidate within an event-template ET^R in the topic-graph R.

$$Score(entity_{k,l} \in ET_k^R) = \sum_i \alpha_i \cdot F_{k,l}^j \quad (7)$$

where $\sum \alpha_i = 1$. The various features F used are: F^1 when the predicted answer genre matches the named entities genre (1 if there is a match and 0 otherwise); F^2 for density-based ranking (% of term overlap and distance between event terms Q' and event terms in the template); F^3 when the query-class matches story genre class (1 if there is a match and 0 otherwise); and F^4 represents the proximity of date (number between 0 to 1 with 1 signifies the required date of query). F^5 is the subtopic similarity (% overlap between query terms and the subtopic headers).

Fig 9: Result of QA (answers in red), (a) Where was Arafat taken for treatment? (b) Which are the candidate cities competing for Olympic 2012?

The news video with the highest scoring answer will be returned (number of results can be varied by user as question might have multiple answers) together with the promising answer candidates highlighted in red. Figure (9) shows the sample outputs of 2 queries.

5.2.2 Answering Visual-oriented Questions

While most users are satisfied with one correct contextual answer, they would usually be interested to see more than one relevant video-footage. This is because footages of the same event across different news stations may be different due to broadcasting rights. The query topic-graph can be leveraged to provide semantic event cluster information, as similar context events tend to have similar footages. To ensure higher recall, we further expand the original query topic-graph by considering immediate clusters or nodes in query topic-graph *R* that are one-"normal" link away. The news videos in the expanded query topic-graph *R'* are then used as the initial retrieval set for finding the answer candidates. We choose shots to be the unit for visual-oriented QA as it is the smallest addressable video unit meaningful to users, as in TRECVID [25]. The shots *S* within the topic-graph *R'* are re-ranked using Eqn (8) modified from [20]. As HLF detection may be incorrect and erroneous, it is necessary to adopt appropriate strategies to fuse them. We make use of the query-HLF as well as the detection confidence of HLF in the shot for fusion.

$$Score(Q,S) = \psi_c \cdot Text(Q, ET(S)) + \varphi_c \cdot \sum_{HLF_k \in S} [Conf(HLF_k) \times Sim_{HLF}(Q', HLF_k)] \quad (8)$$

where $\psi_c + \varphi_c = 1$ are the weights trained for each query-class *c*, *Text()* is the Cosine similarity score. $Conf(HLF_k)$ is the confidence of HLF_k appearing in the shot *S*. Figure (10) shows the results of a visual oriented query, "*Find shots containing fire or explosion*".

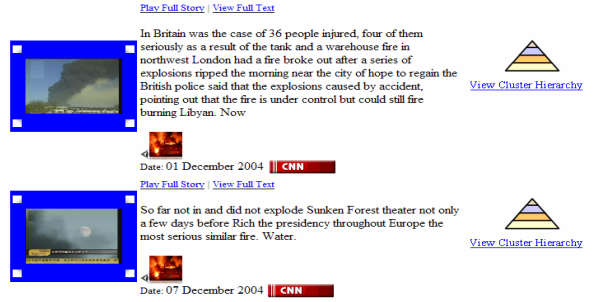

Fig 10: Sample result for visual-oriented queries

6. EXPERIMENTS

To evaluate the effectiveness of the proposed system, we employ the TRECVID 2005 testing dataset which consist of 70 hours of broadcast news in English, Chinese and Arabic recorded in late 2004. This TRECVID dataset is one of the most widely used dataset for testing news video retrieval performance. We divide the experiment into three parts to test different retrieval aspects mentioned in our system. The first part test on the accuracy of the clustering as the quality of cluster is essential. The second part is a user-based performance rating assessment to judge the quality of event evolution topic retrieval. The third part uses 150 questions to determine the QA accuracy.

6.1 Clustering Accuracy

We manually annotate 15 clusters of distinct topics within a time frame of 2 weeks in the video corpus. This annotation is roughly equivalent to 10 hours of video news. The 15 topics are listed in Fig (11).

US Presidential Election, flood, disaster, Iraq war, explosion, Ukrainian Presidential Election, nuclear weapons, North Korea News, Olympics, Yasser Arafat, Taiwan Politics, Same Sex marriage, European Union, South America News, Russian News

Fig 11: List of Manually tagged topics

We carry out a series of test runs to investigate our clustering quality as well as the effects of various features. The performance measure is the standard recall and precision. *Run1* considers only the use of significant text unigram terms determined by TF.IDF term weighting scheme. *Run2* uses event-entities, and *Run3* adds the use of HLFs. Table 2 shows the results of the experiment.

Table 2: Performance of Clustering using various features

Run	Mean Recall	Mean Precision	F_1 Measure
1.Significant unigrams	0.435	0.399	0.416
2.Event entities	0.567	0.599	0.583
3.Run2 + HLFs	**0.588**	**0.605**	**0.597**

From Table 2, we can draw the following observations. First, the selection of features for clustering is important as we observe statistically significant improvement in F_1 measure from Run1 to Run2. This is mainly attributed to the nature of news video which is event-oriented in nature. A significant difference in event entities (different persons or organizations) usually means that it is a different event. Thus event entities are more discriminating than words. Second, we see that the addition of HLFs further improves the overall performance.

6.2 Performance on Topic-evolution Browsing

We design 50 topic questions relating to news event in the news corpus for this series of experiments. These 50 topics are generated based on key events which happened during the months of October, November and December in 2004 obtained from Wikipedia News [28]. Five such sample topics are: *Conflicts in Iraq, Presidential Elections, Death of Yasser Arafat, Israeli-Palestinian conflict, Ukraine presidential election.*

A user pool of fifteen students who have experiences with online news/ news video retrieval has been asked to try the system using the list of 50 topics. The students will select 15 out of the 50 topics for the first round of retrieval and then continue the second round using 5 self-generated topics not within the list of 50. In addition, they are required to assess the following seven questions in the scale from 1-5 (1–Strongly Disagree, 2–Disagree, 3–neutral, 4–Agree, 5–Strongly Agree).

Table 3: Summary of Assessment

Assessment Question	1	2	3	4	5
1) Rated quality of the retrieved results	0	1	3	**4**	**7**
2) The event evolution display helps in locating what I want	0	1	2	**8**	**4**
3) The clusters contain sensible results	0	2	5	**6**	**2**
4) The topic graph makes retrieval easier	0	0	1	**9**	**5**
5) The system can return better results in terms of interestingness	0	0	0	**6**	**9**
6) The event evolution display is better than traditional listing display	0	0	1	**9**	**5**
7) There is significant performance difference between 1st round of retrieval and 2nd round of retrieval	**1**	**9**	2	2	1

Table 3 summarizes the user's assessment scores. We see that more than half of the user pool gave positive responses to questions 1-6. This indicates that the users feel strongly that the system is effective and is capable of returning results that they are interested. In particular, Question 5 which is concerned with the "interestingness" obtained the best response. This shows that the system is able to present events of interests to the users. Almost all users (except 1) think that the event evolution approach to topic retrieval is better than traditional listing display. For the purpose of unbiased testing, Question 7 allows the users to feedback if there are significant performance differences between using the 50 pre-defined topics and their own topics. In particular, we verified at ($p<0.05$) level that there is no significant difference.

6.3 Performance of Question Answering

The third series of test is targeted at evaluating the effectiveness of QA. In this experiment, we employ a total of 150 questions related to the chosen news video corpus. The questions consist of 126 context-oriented queries and 24 visual-oriented queries. The 126 context queries are question modified from past TREC [24] QA tasks, while the 24 visual oriented queries are queries used in the search task of TRECVID 2005. A partial list of questions is given in Figure (12).

1) What is the name of the serial killer
2) Which countries are competing for Olympic 2012
3) Which team won the Stanely Cup
4) Which team won the NBA title
5) What is Hong Kong unemployment rate
6) What is the US consumer price index
7) What is the name of the new drug that fight AIDS
8) Result of the basketball game last night
9) *Find shots of Iyad Allawi, the former prime minister of Iraq*
10) *Find shots of tennis players on the court 2 players visible at same time*
11) *Find shots of a helicopter in flight*
12) *Find shots of people with banners or signs*

Figure 12: Partial list of questions (*1-8 – context, 9-12 – visual*)

6.3.1 Context-oriented Questions

The system is used to return exact video segments or speech containing the answers. In addition, promising answers detected within the ASR will be highlighted in red and displayed as shown in Figure (9b). The assessors are asked to evaluate the relevance of returned video segment based on first 15 seconds or 30 seconds of segments. For assessment, as long as the correct answer is contained within the video segments, we consider the answer to be correct. Two series of runs are conducted: (a) QA ranking based on the list of retrieved news video, and (b) QA based on news video from query topic-graph. Table 4 tabulates the results.

Table 4: Context-oriented QA performance (126 queries)

Video	*Run Type*	*Correct*	*Accuracy*
15 Seconds	w/o query topic-graph	75	59.5%
	w/ query topic-graph	81	64.3%
30 Seconds	w/o query topic-graph	77	61.1%
	w/ query topic-graph	85	67.5%

The results from Table 4 indicate that the system is able to obtain 85 correct questions out of the 126 questions. The performance at 30 seconds interval is better that at the 15 seconds interval because for some answer segments, the correct answers appear towards the end, after the wrong ones. This is especially true for long video segments which contain a large number of sub-segments of same entity types, like locations or names. We will be exploring techniques to overcome this problem to improve the QA accuracy. In addition, runs based on topic-graph yield better results because: (a) clustering can improve the recall by retrieving relevant news videos with missing key entities due to ASR errors; and (b) leveraging subtopics structures can increase the precision.

6.3.2 Visual-oriented Questions

To assess the performance on a comparative scale, we follow the evaluation standard as in TRECVID automated search task [25]. The participants in this task are required to return a ranked list of shots (maximum of 1000) arranged according to their degree of relevance. The performance measure used is the mean average precision (MAP) which is widely used for system evaluation in information retrieval over large corpuses where the recall rate is hard to determine. We designed 3 runs: *Run1* which only uses the text event entities at the shot level; *Run2* adds the use query-HLF to Run1; and *Run3* includes both query-HLF and a combination of multimodal features like Video-OCR as in our previous work [20]. We also compare our performance with the best automated run submitted by NUS [9] in TRECVID 2005 which obtained a MAP of 0.126. The results are displayed in Table 5.

Table 5: Visual-Oriented QA Performance (24 queries)

	Run1	Run2	Run3[20]	NUS [9]
w/o topic-graph	0.086	0.118	0.130	0.126
w/ topic-graph	0.095	0.123	0.133	0.126
w/ expanded topic-graph	**0.096**	**0.124**	**0.138**	**0.126**

Table 5 clearly illustrates that the appropriate use of query-HLF can significantly improve the performance of retrieval. The most significant jump in improvement comes from Run2 where we added the use of query-HLF over Run1. The system returns an excellent result of 0.124 which is comparable to the best submitted result in TRECVID 2005 automated search task. Runs that make use of topic-graph clearly show better result than those without using the topic-graph. In particular, the expanded topic-graph Run3 yields an MAP performance of 0.138 which is much better than the best reported run in TRECVID 2005. This observation confirms our earlier hypothesis and also validates the positive effects brought by event clustering and threading

6.4 Discussion of Results

In developing such a system that integrates technologies and research from many fields, many sources of errors may incur and needs to be tackled. Although cares have been taken to minimize errors, many errors do occur and affect the quality of results. We still experience under-segmentation in the news video even with the 2nd level segmentation. For example: in sports, we under-segment baseball news from golf or football news. The cumulative error is about 10%-15% for this type of error. Further improvements should be done to ensure the accuracy of story boundary detection.

The event evolution requires online processing of news blogs which takes time. However, this can be tackled by daily crawling of interesting topics collected from the user pool. The results of question answering task on persons' names tend to yield worse results than locations or organizations as they are more vulnerable to speech recognition and machine translation errors.

7. CONCLUSION

We introduce a framework to news video searching by leveraging on usable semantics and relevant external information resources to support a series of event-based analysis in order to discover event evolution and perform question answering (QA). The newly developed system is capable of displaying interesting key stages in the evolution of news topic from multiple news sources, as well as performing precise question answering based on event entities. The main contributions of this research are: (a) We combine multimodal event information extracted from news video, web news articles and news blogs to support event analysis. (b) We introduce topic evolution browsing based on users' interest. (c) We extend question answering on top of topic evolution to handle various types of video queries. We tested the systems on 70 hours of TRECVID 2005 news video, with over 200 questions and found that the system is effective. For future work, we will refine our event topic model and explore the use of other online resources and event features to support more implicit event modeling and analysis.

8. ACKNOWLEDGMENTS

We would like to express thanks to the reviewers and TPC for their kind and valuable suggestions in improving the paper.

9. REFERENCES

[1] J. Adcock, M. Cooper, A. Girgensohn, and L. Wilox. "Interactive Video Search Using Multilevel Indexing" *CIVR 2005*, Singapore, 205-214, July 2005.

[2] J. Allan, R. Papka and V. Lavrenko, "On-Line New Event Detection and Tracking" *SIGIR 1998*, Melbourne, Australia, 37-45, 1998.

[3] A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M.R. Naphade, A.P. Natsev, J.R Smith, J. Tesic, T. Volkmer, "IBM research TRECVID- 2005 video retrieval system" *TRECVID 2005 Workshop*, NIST, USA Nov 2005.

[4] AQUAINT, http://www.ldc.upenn.edu/Catalog/docs/

[5] A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup, G. Schneider, " A Hierarchical clustering Approach for Large Compound Libraries," *Journal of CIM*, 807-815, 2005.

[6] H. Bruce, "Perceptions of the Internet: What People Think When They Search the Internet for Information" *Internet Research: Electronic Networking Applications & Policy,* Vol 9, 187-199. 1999.

[7] L. Chaisorn, T.S. Chua and C.H. Lee, "The segmentation of news video into story units" *ICME 2002*, Ischia, Italy, Jul 2002.

[8] M.G. Christel, A.G. Hauptmann, H.D. Wactlar and T.D. Ng, "Collages as Dynamic Summaries for News Video" *ACM Multimedia 2002*, 561-569, Juan-les-Pins, France, Dec 2002.

[9] T.S. Chua, S.Y. Neo, H.K. Goh, M. Zhao, Y. Xiao, G. Wang, "TRECVID 2005 by NUS PRIS" *TRECVID 2005 Workshop*, NIST, USA Nov 2005.

[10] W.B Croft, H.R. Turtle, D.D. Lewis, **"**The use of phrases and structured queries in information retrieval" *ACM SIGIR 1991*, Chicago, USA, 32-45, 1991.

[11] A. Davis, P. Pennert, R. Rubinoff, T. Sibley, E. Tzoukemann, "Retrieving what is relevant in audio and video: statistics and linguistics in combination". *RIAO'2004*.

[12] C. Fellbaum. WordNet: An Electronic Lexical Database.MIT Press 98.

[13] Flickr, http://www.flickr.com

[14] S. Har-Peled, B. Sadri, "How Fast Is the k-Means Method?" *Algorithmica* 41, Vol3, 185-202, Jan. 2005.

[15] W.H. Hsu, L. Kennedy, S.F. Chang, M. Franz, and J. Smith, "Columbia-IBM News Video Story Segmentation in TRECVID 2004", *Columbia ADVENT Technical Report*, New York 2005

[16] C. Kenneth and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *ACL*, 1989

[17] LSCOM Lexicon http://www.ee.columbia.edu/dvmm/lscom

[18] J. McCarley, M. Franz. "Influence of speech recognition errors on topic detection". *SIGIR 2000*, 342-344, New York, NY, USA, 2000.

[19] S.Y. Neo, Y. Zheng, T.S. Chua, Q. Tian, "News Video Search with Fuzzy Event Clustering using High-level Features" *ACM Multimedia 2006*, Santa Barbara, USA, 23-27 Oct 2006.

[20] S.Y. Neo, J. Zhao, M.Y. Kan, T.S. Chua, "Video Retrieval Using High-level features, "Exploiting Query-matching and Confidence-based Weighting", *CIVR 2006*, Arizona, USA, 143-152, July 2006.

[21] C. Petersohn. "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System" *TRECVID 2004 Workshop*, NIST, US, Nov 2005

[22] P. Resnik, "Semantic similarity in a taxonomy: An information- based measure and its applications to problems of ambiguity in natural language" *Journal of Artificial Intelligence Research*, 95–130, Nov 1999.

[23] Technorati, http://www.technorati.com

[24] TREC, Text Retrieval Conference, http://trec.nist.gov

[25] TRECVID, TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid

[26] Vivisimo, http://www.vivisimo.com

[27] H.D. Wactlar, A.G. Hauptman, M.G. Christel, R.A. Houghton, and A.M. Olligschlaeger, "Complementary video and audio analysis from Broadcast News Archives" *Comm. of ACM*, Vol 43. No. 2, 42-47, Feb 2000.

[28] Wikipedia News, http://en.wikipedia.org/wiki/Nov_04

[29] R. Yan, J. Yang, and A. G. Hauptmann, "Learning Query-Class Dependent Weights for Automatic Video Retrieval" *ACM Multimedia 2004*, New York, Oct 2004.

[30] H. Yang, L.Chaisorn, Y. Zhao, S-Y. Neo, T.S. Chua, "VideoQA: question answering on news video" *ACM Multimedia 2003*, Berkeley, USA, 632-641, Nov 2003.

[31] H. Yang, T. Chua, S. Wang and C. Koh. "Structured use of external knowledge for event-based open-domain question-answering" *SIGIR 2003*, Canada, Jul 2003.

[32] M. Zhao, S.Y. Neo, H.K. Goh, T.S. Chua, "Multi-Faceted Contextual Model for Person Identification in News Video" *MMM 2006*, Beijing, China, 2006.