From Text Question-Answering to Multimedia QA on Web-Scale Media Resources

Tat-Seng Chua, Richang Hong, Guangda Li, and Jinhui Tang
School of Computing
National University of Singapore
{chuats, hongrc, tangjh}@comp.nus.edu.sg, g0701808@nus.edu.sg

ABSTRACT

With the proliferation of text and multimedia information, users are now able to find answers to almost any questions on the Web. Meanwhile, they are also bewildered by the huge amount of information routinely presented to them. Question-answering (QA) is a natural direction to address this information over-loading problem. The aim of QA is to return precise answers to users' questions. Text-based QA research has been carried out for the past 15 years with good success especially for answering fact-based questions. The aim of this paper is to extend the text-based QA research to multimedia QA to tackle a range of factoid, definition and "how-to" QA in a common framework. The system will be designed to find multimedia answers from Web-scale media resources such as Flicker and YouTube. This paper describes the architecture and our recent research on various types of multimedia QA for a range of applications. The paper also discusses directions for future research.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**] : Web-based Services

General Terms

Algorithm. Documentation. Experimentation

Keywords

Multimedia Question-Answering.

LS-MMRM'09, October 23, 2009, Beijing, China.

1. INTRODUCTION

The amount of information on the Web has been growing at an exponential rate. An overwhelming amount of increasingly multimedia contents are now available on almost any topics. When looking for information on the Web, users are often bewildered by the vast quantity of information returned by the search engine, such as the Google or Yahoo. Users often have to painstakingly browse through large ranked lists of results in order to look for the correct answers. Hence question-answering (QA) research has been evolved in an attempt to tackle this information-overload problem. Instead of returning a ranked list of documents as is done in current search engines, QA aims to leverage on deep linguistic analysis and domain knowledge to return precise answers to users' natural language questions [9].

Research on text-based QA has gained popularity following the introduction of QA in TREC (The Text Retrieval Conference) evaluations in the late 1990s [20]. There are many types of QA, depending on the type of questions and the expected answers. They include: factoid, list and definitional QA, and more recently, the "how-to", "why", "opinion" and "analysis" QA. Typical QA architecture include stages of: question analysis, document retrieval, answer extraction, and answer selection and composition [9]. In factoid and list QA, such as *"What is the most populous country in Africa?"* and *"List the rice-producing countries"*, the system is expected to return one or more precise country names as the answers [7]. On the other hand, for definitional QA, such as *"What is X?"* or *"Who is X?",* the system is required to return a set of answer sentences that best describe the question topic [8]. In a way, definition QA is equivalent to query-oriented summarization, in which the aim is to provide a good summary to describe a topic. These three types of QA have attracted a lot of research in the last 10 years [9]. They provide fact-based answers, often with the help of resources such as the Wikipedia[1] and WordNet [3]. In fact, Factoid QA has achieved good performance and

[1] Wikipedia: the free encyclopedia that anyone can edit (http://www.wikipedia.org/)

commercial search engines have been developed, such as the Powerset [14] that aims to return mainly factoid answers from Wikipedia.

More recently, attention have been shifted to other types of QA such as the "how-to", "why" and "opinion" type questions. These are harder questions as the results require the analysis, synthesis and aggregation of answer candidates from multiple sources. To facilitate the answering of "how-to" questions, some recent research efforts focus on leveraging the large question-answer banks available in community QA sites such as the Yahoo!Answers to provide the desired answers. Essentially, the system tries to find equivalent questions with readily available answers in Yahoo!Answers site, turning the difficult "how-to" QA into a simpler similar question matching problem [11].

Given that the vast amount of information on the Web is now in non-textual media, it is natural to extend the text-based QA research to multimedia QA. There are several reasons why multimedia QA is important. First, although most media contents are indexed with text metadata, most such metadata, such as those available in YouTube, is noisy and incomplete. As a result, many multimedia information contents are not retrievable, unless advanced media content analysis techniques are developed to uncover the contents. Second, many questions are better explained in or with the help of non-textual medium. For example, in providing textual answers to a definition question such as *"What is a thumb drive?",* it is better to also show image or video of how thumb drives look like. Third, media contents, especially videos, are now used to convey many types of information as evident in sites such as YouTube and other specialized video/image sharing sites and blogs. Thus many types of questions now have readily available answers in the form of video. This is especially so for the difficult "analysis" and "how-to" type questions. Answering such questions is hard even for text because further analysis and composition of answers are often needed. Given the vast array of readily available answers, it is possible now to find video for "how-to" questions such as *"How to transfer photos from my camera to the computer?"*. From user's point of view, it would be much clearer and instructive to show them video detailing the entire transfer process, rather than a text descriptions of the steps involved.

Multimedia QA can thus be considered as a complement to text QA in the whole question-answering paradigm, in which the best answers may be a combination of text and other medium answers. Essentially, multimedia QA includes image, video and audio QA. They all aim to return precise images, video clips, or audio fragments as answers to users' questions. In fact, the factoid QA problem of finding precise video contents at the shot level has partially been addressed by TRECVID [21], a large scale public video evaluation exercise organized yearly in conjunction with TREC. This is done in the form of automated and interactive (shot) video retrieval, where the aim is to find a ranked list of shots that visually contain the desired query target, such as finding shots of *"George Bush*". An early system specifically designed to address the multimedia factoid QA is presented in [7] for news video. It follows a similar architecture as text-based QA, with video content analysis being performed at various stages of QA pipeline to obtain precise video answers. It also includes a simple video summarization process to provide the contextual aspects of the answers. Other than factoid QA, as far as we know, no research in the equivalent of multimedia definition and "how-to" QA has been attempted.

The aim of this paper is to extend the text-based QA research to multimedia QA to tackle a range of factoid, definition and "how-to" QA in a common framework. The system will be designed to find multimedia answers from Web-based media resources such as Flicker and YouTube. This paper describes the architecture and our recent research on various types of multimedia QA for a range of applications. It focuses only on visual media such as image and video. The paper also presents initial results of our research based on YouTube videos.

Briefly, the outline of the paper is as follows. Section 2 presents related work; and Section 3 describes the architecture and overview of multimedia QA. Sections 4 and 5 describe recent works on definition and how-to multimedia QA respectively. The summary with discussion of trends in multimedia QA is presented in Section 6.

2. RELATED WORK

With the proliferation of multimedia contents on the Web, research on multimedia information retrieval and question-answering are beginning to emerge. The early work that addresses the issues of QA in video in a system named VideoQA is reported in [7]. This system extends the text-based QA technology to support factoid QA in news video by leveraging on visual contents of news video as well as the text transcripts generated from ASR (Automated Speech Recognition). Users interact with the system using short natural language questions with implicit constraints on contents, duration, and genre of expected videos. The system comprises two stages. In the preprocessing stage, it performs video story segmentation and classification, as well as video transcript generation and correction. During the question answering stage, it employs modules for: question processing, query reinforcement, transcript retrieval, answer extraction and video summarization.

Following this work, several video QA systems were proposed with most of them relying on the use of text transcripts derived from video OCR (optical character

recognition) and ASR outputs. [10] developed a lexical pattern matching-based ranking method for domain-dependent video QA. [26] designed a cross-language (English-to-Chinese) video QA system based on retrieving and extracting pre-defined named entity entries in text captions. The system enables users to query with English questions to retrieve the Chinese captioned videos. The system was subsequently extended [28] to support bilingual video QA that permits users to retrieve Chinese videos through English or Chinese natural language questions. [27] presented a robust passage retrieval algorithm to extend the conventional text-based QA to video QA.

As discussed earlier, shot retrieval as proposed in TRECVID can also be regarded as a kind or base technology for factoid video QA. For example, if the user issues a query *"Who is Barack Obama?",* the shot retrieval system would aim to return a video that visually depict the query subject. In this sense, the body of work done on shot retrieval [2] as part of TRECVID efforts can be considered as research towards factoid multimedia QA. The first step in shot retrieval is to extract relevant semantic information for the shot. This includes ASR text, as well as possible presence of high level concepts, such as the face, car, building etc [19]. Given a query, most shot retrieval systems follow similar retrieval pipeline of: query analysis, shot retrieval, shot ranking and answer selection [18]. Query analysis performs query expansion and inference of relevant high-level concepts by considering the correlation between query text and concepts. In order to cover concept relations that cannot be inferred from statistics, knowledge-driven approaches to relating high-level concepts to queries have been incorporated. Given the expanded query, a combination of text and high-level concept matching is performed to retrieve relevant list of shots. A multi-modal approach is then employed to re-rank the shots for presentation to the users [2].

Few works have been done on image-based QA except the one presented in [22] that describes a photo-based QA system to find information about physical objects. Their approach comprises three layers. The first layer performs template matching of query photo to online images to extract structured data from multimedia databases in order to help answer questions about the photo; it uses question text to filter images based on categories and keywords. The second layer performs searches on internal repository of resolved photo-based questions to retrieve relevant answers. In the third human-computation QA layer, it leverages community experts to handle the most difficult cases.

Overall, it can be seen that work on factoid multimedia QA has just been started, whereas little work has been done on the more challenging and practical tasks of definition and "how-to" QA.

3. OVERVIEEW OF MULTIMEDIA QA

The aim of question-answering is to present precise information to users instead of a ranked list of results as is done in the current search engines. Most text-based QA system follow the pipeline as shown in Figure 1 comprising of: question analysis, document retrieval, answer extraction, and answer selection [9]. Question analysis is a process which analyzes a question to extract the question context in the form of list of keywords, and identify the answer type and target in order to formulate question strategy. Document retrieval is a step that searches for relevant documents or passages from a given corpora. Answer extraction then extracts a list of answer candidates from the retrieved documents, by selecting sentences that cover the expanded queries terms and contain the expected answer target. Finally, answer selection aims to pin-point the correct answer(s) from the extracted candidate answers. For definition QA, the last step also involves composing the selected answer sentences into a coherent summary. Step 2 on document retrieval aims to achieve high recall, while the last 2 steps aim to identify precise and correct answers. The last 2 steps often involve the use of deep linguistic analysis together with the use of domain knowledge.

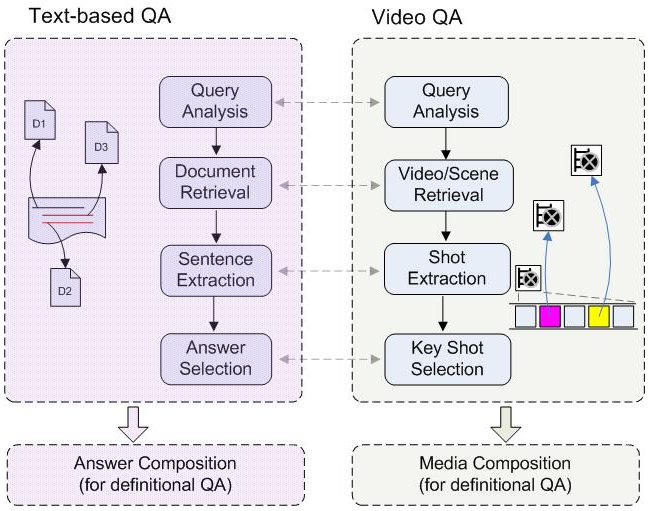

Figure 1: The retrieval pipelines of text and multimedia QA. We can see that the procedures of the two QAs are analogous to each other.

Multimedia QA uses a similar retrieval pipeline as that in text-based QA as depicted in Figure 1. For the case of video, we can draw the analogy between text and video by equating: word to video frame, sentence to shot, paragraph to scene and document to video sequence. The first step on query analysis aims to expand query by identifying context of text search terms and inferring key high-level visual

Figure 2: Examples of video summary generated by our video definition QA system. They are "Columbia Space Shuttle Disaster", "G20 Summit", "Handover HongKong 1997" and "September 11 attacks". We can see that the resulting video skims incorporate both the importance of the sub-events and timeline constraint.

concepts in query. In step 2, the expanded query is matched against the meta-data of stored video sequences to retrieve the relevant ones. For the case of YouTube video, the metadata includes user-assigned tags and categories; in addition, we will include automatically generated high level visual concepts [19] of a pre-defined concept set. In step 3, a combination of multi-modal analysis involving visual content, high-level visual features and metadata text is normally performed to identify good shot candidates. This step is similar to the processing done in most shot retrieval algorithms used in TRECVID video retrieval evaluations [2]. Re-ranking algorithm involving addition content features or domain knowledge is applied to produce the final ranking of shots in Step 4. For the case of definition QA, further processes are performed to identify the set of key shots and the sequencing of these shots into a coherent summary that meets both the semantic and temporal constraints.

In the next two Sections, we will discuss the issues and challenges faced in multimedia definition and "how-to" QA research.

4. DEFINITION QUESTION-ANSWERING

Definitional QA was first introduced to the Text REtrieval Conference (TREC) QA Track main task in 2003 [23]. Questions like "*what is X" and "where did X happened"*, which correspond to the event/entity definition, account for a large number of queries submitted to the Web search engines [1]. To answer such questions, many search engines such as Google and Yahoo tend to rely on existing online definition resources, such as Google Definitions, Wikipedia, Biography.com, s9.com etc, depending on the types of queries. However, the definition of entities often changes over time, with many new ones being introduced daily. Therefore, automatic definitional QA systems are needed that can extract definition sentences that contain descriptive information about the target entity from multiple documents and summarizes these sentences into definition.

A good overview of text-based definition QA can be found in [8]. It differs from factoid QA in Step 4 in the way that it selects that relevant sentences that depict the diverse key aspects of the target entity. Most approaches select sentences to meet criteria of good information coverage, diversity, as well as exhibiting good definition patterns. [5] added human interest model to the selection criterion by preferring those sentences that are of greater interests to the users, and showed good performance on TREC QA dataset. For semi-structured text such as Wikipedia, [17] incorporated knowledge of structure by leveraging on links between concepts and infobox to extract good definition summary for Wikipedia pages. Most techniques employed in text definition QA are applicable to community-contributed videos such as those available in YouTube.

In recent years, we have witnessed the exponential growth of community contributed social media on Flickr and YouTube, etc, where users collaboratively create, evaluate and distribute vast quantity of media contents. Take YouTube as an example, which is one of the primary video sharing sites. Studies have shown that it serves 100 million distinct videos and 65,000 uploads daily; and traffic of this site accounts for over 20% of the web in total and 10% of the whole internet, covering 60% of the videos watched on-line [13]. The prevalence of Web 2.0 activities and contents has inspired intensive research to exploit the freely available metadata in multimedia content analysis. In the scenario of definitional video QA, it is important to exploit both visual and textual metadata information for selecting good video shots and generating high quality video summaries.

However, given an event type query, the retrieved videos from current popular video sharing sites tend to be diverse and somewhat noisy. For example, from the retrieved list of " September 11 attacks" from YouTube, we will see not only relevant video excerpts from news TV, but also re-

assembled excerpts of news video clips produced by general users, as well as many irrelevant ones that are extensions to the event such as interviews to politicians etc. To navigate the mass of information, we need to be able to identify shots representing key sub-events while removing those auxiliary shots, similar to the text-based approach of identifying key relevant sentences while removing those irrelevant. Fortunately for the case of YouTube, we found that most video retrieved tend to share many video shots depicting key sub-events. In fact, recent studies [25] on video sharing sites have shown that there exists a significant amount of over 25% of duplicate videos in the search results. We categorize the content redundancy on web videos into two classes: near duplicate and overlap. The former indicates that most of the frames from the two videos are duplicates and the latter indicates that the video pair shares some near duplicate frames. We focus on the case of overlap and look at it from a different perspective. We plan to exploit such content overlap in Web video sharing system to automatically answer definitional questions of events. For a given event or entity, the few scenes that convey the main messages, such as the principal sub-events or key shots, tend to be re-used in multiple news reports, and copied in many other self-assembled videos. Thus we can identify such shots by performing near duplicate detection on the set of retrieved videos.

Key shots identified through near duplicate detection meet the criteria of being salient and popular as they are re-use in multiple video sources. Further textual analysis of meta text may also yield information on key sub-events. Together, they form the basis for definition QA for video for event and entity queries.

Based on the above analysis, we design and implement a video definition QA system. The goal of the proposed system is to give users a quick overview for a definition query by leveraging on the content overlap of the Web video search results. The proposed framework consists of four main stages, namely, web videos and meta data acquisition; visual processing of key shots (key shot linking, ranking and threading); semantic analysis of key shots (tag filtering and key shot tagging); and summary generation.

In Stage 1, given an event query, a ranked list of videos and the corresponding metadata for each video (tags, descriptions and titles) are retrieved from YouTube.

In Stage 2 on key shot processing, we first perform shot segmentation and extract the keyframes from the shots. For each keyframe we then extract its local point feature for matching; where we extract the scale-invariant feature transform [4] features. To reduce the computation cost, local point features are mapped to a fixed dimension and some keyframes are filtered by offline quantizing the keypoint descriptors. The keyframe pairs with similarity value above a given threshold are retained as near duplicate keyframes and their corresponding shots are defined as key shots. We then rank the key shots according to their informative score, which is defined as the linear combination of relevance and normalized significance. After that, the shots are chronologically threaded. Here, we utilize the chronological order lies in the original videos and formulate the threading as the minimization of the time lag between the key shot pairs in different original videos.

In Stage 3, we perform tag analysis using visual similarity based graph. Based on the tags associated with the videos, we define the metrics of representativeness and descriptiveness to measure the ability of the tags to represent and describe the event or sub-events. Based on that, the noisy and less informative tags with respect to the query are removed. We then perform the random walk [24] on visual similarity-based graph to spread the tags to other key shots connected by near duplicate keyframe links.

Stage 4 performs the summarization in the form of video skim by sequencing the selected key shots using a greedy algorithm. We embed the tag descriptions into the key shots to help users better comprehend the events. Both the duration of the video skims and the number of frames are flexible according to users' requirements.

Further details of this implementation can be found in [16].

To test our system, we selected 20 news events as definition QA queries. Examples of queries include: *September 11 Attacks; China Sichuan Earthquake 2008; Space Shuttle Columbia Disaster; G20 Summit;* and *Opening Ceremony Beijing 2008* etc. These queries, after query analysis, are used to retrieve videos from YouTube. The total number of videos down-loaded is 1,780 with an average of 89 videos per query. The total duration of the videos is 155 hours. We conduct user based evaluations to validate the effectiveness of our definition QA by generating video summaries of different durations. Some examples of the generated video summaries or skims are shown in Figure 2. Our studies show that our approach is feasible and effective.

5. HOW-TO QUESTION-ANSWERING

Beyond definition QA, the next set of challenges in question-answering is to handle the "how-to" and "why-type" questions. Example of an "how-to" questions is *"How to transfer pictures in my digital camera to computer?"*. The ability to answer such questions requires understanding of the relevant contents, and often involves the generation of specific answers. This is beyond the capability of current technologies unless it is for a very narrow domain. Because of the strong demands for such services, community-based QA services, such as Yahoo!Answers (YA) [29], has become very popular. Through YA services, people ask a variety of "how-to"

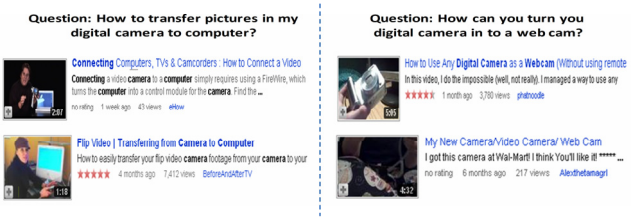

Figure 3: Output of our "how-to" video QA system. The top two ranked video answers for each question are presented to the user.

questions and obtain answers either by searching for similar questions on their own or waiting for other users to provide the answers. As large archives of question-answer pairs are built up through user collaboration, the knowledge is accumulated and is ready for sharing. To facilitate great sharing of such knowledge, one emerging research trend in text-based QA is to develop techniques to automatically find similar questions in YA that have ready answers for the users [11].

However, even when the best text-based answer is presented to the users, say, for the "*picture transfer"* question, the user may still have difficulty grasping the answers. This is because from the textual answers, the users may still have no idea on how to deal with USB cable, from such answer as *"... connect your digital camera though USB cable ..."*. However, if we can present visual answers such as videos, it will be more direct and intuitive for the users to follow. Overall, in addition to normal textual references or instructions, visual references or instructions such as videos should be an ideal complementary source of information for users to follow.

Some commercial websites like ehow.com [6] do provide "how-to" videos. They do so by recruiting general users to produce problem-solving videos, so that other users can easily search or browse them. However the coverage of topics in this Web site is limited, as only carefully selected videos by certain photographers will be published on their website [12]. On the other hand, community video sharing sites, such as YouTube and Yahoo Video, contain huge collections of videos contributed by the users. Many videos in such sites do provide "*how to"* instructions on a wide variety of popular topics in the domains of electronics, traveling, cooking etc. This makes such video sites ideal sources for offering answers to many popular "how-to" questions.

In general, metadata tags on community-shared videos tend to be sparse and incomplete. Hence attempt to use original user text queries to retrieve such videos from sites such as YouTube tend not to be effective. Also, we should exploit the richness of visual contents within the video in conjunction with textual information mentioned above to identify the best video answers. On the other hand, successful techniques have been developed to find readily text-based answers for the same questions from Yahoo! Answers [11]. Hence a natural approach is to combine both text-based and multimedia approaches to leverage on the strength of both approaches to find video answers. In other words, we utilize the text answers from Yahoo! Answers to support the search of "how-to" videos. The overall stages of our prototype "how-to" QA system for video consist of 2 main stages as follows.

Stage 1 focuses on recall-driven related video search. It aims to find similar questions posed using different language styles and vocabulary from Yahoo! Answers site; where the corresponding answers tend to contain the desired instructions. We then use the similar questions found to improve the coverage of the topic terms covered in the original query. In a way, this is a text query expansion step where terms related to the topic commonly used by users are extracted. However, community video site like YouTube can only take in precise queries; we therefore extract only key phrases from these questions and use that to formulate multiple search queries to ensure high recall of search results.

Stage 2 is the precision-driven video re-ranking step, where related videos based on their relevance to the original questions are re-ranked. We utilize 3 sources of information to perform the re-ranking: (a) the redundancy of video through near duplicate detection as is done in Section 4; (b) the presence of key visual concepts in the video; and (c) community viewers comments. In our preliminary work, we focus on electronic domain and we pre-define a set of visual high-level concepts such as the *camera, computer* etc., that are important in this domain. We manually select training images for these concepts using Google Image Search; and perform salient object recognition based on image matching techniques to detect the presence of these concepts in the video. We also analyze community viewers' comments to assess the community's opinion on video's popularity. In a way, this is similar to opinion voting. Finally, a rank fusion scheme is adopted to generate a new ranking list based on the evidences from visual cues, opinion voting and video redundancy.

Further details of our approach can be found in [15]. We evaluate the system on 24 "how-to" queries mined form Yahoo!Answers that have corresponding video answers in YouTube. Our initial test shows that our approach is effective. Figure 3 show the screen shots of our "how-to" QA.

6. SUMMARY

With the proliferation of text and multimedia information, users are now able to find answers to almost any topics on the Web. On the other hand, they are also overwhelmed by the huge amount of information routinely presented to them. Question-answering (QA) is a natural direction to overcome this information over-loading problem. The aim of QA is to return precise answers to users' questions. Text-based QA research has been carried out in the past 15 years with great success especially for answering fact-based questions with commercial offerings. This paper extended the text-based QA research to multimedia QA to tackle a range of factoid, definition and "how-to" QA in a common framework. The system was designed to find multimedia answers from Web-based media resources such as Flicker and YouTube. This paper described our preliminary research in performing definition QA on events, and "how-to" QA on electronic domain. Our research results showed that it is feasible to perform factoid, definition and "how-to" QA by leveraging on large community-based image and video resources on the Web.

The research is preliminary. Several follow-on research directions can be identified. First, there is an urgent need to setup large test corpora to promote multimedia QA research, especially on definition and "how-to" QA. Second, there is a need to develop better techniques for visual matching and visual concept detection. Visual concept detection techniques are important to un-covering additional visual contents in image/video clips. To ensure scalability of such techniques to Web-scale problems, we need to exploit the various online visual databases with comprehensive visual concept coverage and visual examples, such as the Wikimedia and online visual dictionaries. Finally, we need to demonstrate effectiveness of our approaches to general domains.

7. REFERENCES

[1] A. P. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, Semantic concept-based query expansion and re-ranking for multimedia retrieval, ACM Multimedia, pp. 991–1000, Augsburg, Germany, 2007.

[2] Cees G. M. Snoek and Marcel Worring, Concept-Based Video Retrieval, Foundations and Trends in Information Retrieval, vol. 4, iss. 2, 215-322, 2009.

[3] C. Fellbaum, ed., WordNet: An Electronic Lexical Database. Cambridge, USA: The MIT Press, 1998.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer, 60:91–110, 2004.

[5] Dave Kor and Tat-Seng Chua. Interesting Nuggets and Their Impact on Definitional Question Answering. ACM SIGIR 2007. Amsterdam, Netherlands. July 2007. 335-342.

[6] eHow: http://www.ehow.com/videos.html

[7] Hui Yang and Tat-Seng Chua, Shuguang Wang and Chun-Keat Koh. Structured use of external knowledge for event-based open-domain question-answering. 26th Int'l ACM SIGIR Conference' 03. Canada, Jul/Aug 2003. 33-40.

[8] Hang Cui, Min-Yen Kan and Tat-Seng Chua. Soft Pattern Matching Models for Definitional Question Answering. ACM Transactions on Information Systems (ACM TOIS). Vol 25(2), April 2007. 30 pages.

[9] John M. Prager: Open-Domain Question-Answering. Foundations and Trends in Information Retrieval 1(2): 91-231 (2006)

[10] Jinwei Cao, Jay F. Nunamaker. Question Answering On Lecture Videos: A Multifaceted Approach, ACM/IEEE Joint Conference on Digital Libraries, 2004.

[11] Kai Wang, Zhaoyan Ming, Tat-Seng Chua. A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services. To appear in ACM SIGIR 2009, Boston, Massachusetts, USA.

[12] K. –Y. Chen, L. Luesukprasert, and S. T. Chou. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE transactions on knowledge and data engineering. 19(8):1016-1025. 2007

[13] M. Cha, H. Kwak, P. Rodriguez, YY. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. San Diego, California, USA. 2007.

[14] Powerset: a commercial factoid-based search engine that was acquired by Microsoft. See http://www.powerset.com/

[15] Guangda Li, Zhaoyan Ming, Haojie Li, Yantao Zheng, Tat-Seng Chua. Video reference: question answering on YouTube. To appear in ACM Multimedia 2009.

[16] R. Hong, J. Tang, H. Tan, S. Yan, C. –W. Ngo, T. –C. Chua. Event driven summarization for web videos. Submitted to ACM Multimedia 1st Workshop on Social Media (ACM-MM-WSM 2009).

[17] Shiren Ye and Tat-Seng Chua and Jie Lu. Summarizing Definition fromWikipedia. To appear in ACL'09, Singapore.

[18] S. -Y. Neo, J. Zhao, M. -Y. Kan, and T. -S. Chua, Video retrieval using high level features: Exploiting query matching and confidence-based weighting, in CIVR, (H. Sundaram et al., eds.), pp. 143–152, Heidelberg, Germany: Springer-Verlag, 2006.

[19] S. -F. Chang, W. Hsu, W. Jiang, L.S. Kennedy, D. Xu, A. Yanagawa and E. Zavesky. Columbia University TRECVID-2006 video search and high-level feature extraction. Proceedings of TRECVID Workshop, Gaithersburg, USA, 2006.

[20] TREC: The Text Retrieval Conference. See http://trec.nist.gov/.

[21] TRECVID: a video evaluation forum organized in conjunction with TREC. See http://trecvid.nist.org/.

[22] T. Yeh, J. J. Lee, T. Darrell. "Photo-based Question Answering", ACM Multimedia, 2008.

[23] E. M. Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. In Proceedings of TREC.

[24] W. H. Hsu, L. S. Kennedy, and S. F. Chang. Video search reranking through random walk over document-level context graph. In Proceeding of ACM 14th international conference on Multimedia, Augsburg, Germany, October 2007.

[25] X. Wu, A. G. Hauptmann, and C. –W. Ngo. Practical elimination of near-duplicates from web video search. Proceedings of the 15th international ACM conference on Multimedia, Augsburg, Germany. 2007

[26] Y. C. Wu, Y. S. Lee, C.H. Chang. "CLVQ: cross-language video question/answering system", The 6th IEEE international symposium on multimedia software engineering, 2004.

[27] Y. C. Wu, J. C. Yang. "A Robust Passage Retrieval Algorithm for Video Question Answering", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 18, No. 10, Oct. 2008.

[28] Y. S. Lee, Y. C. Wu, J.C. Yang. "BVideoQA: Online English/Chinese Bilingual Video Question Answering", Journal of the American Society for Information Science and Technology, 60(3):509–525, 2009.

[29] Yahoo alpha search: http://au.alpha.yahoo.com/.