Multimedia Answering: Enriching Text QA with Media Information

Liqiang Nie, Meng Wang, Zheng-Jun Zha, Guangda Li and Tat-Seng Chua
School of Computing
National University of Singapore
{nieliqiang, eric.mengwang, junzzustc}@gmail.com
{g0701808, chuats}@nus.edu.sg

ABSTRACT

Existing community question-answering forums usually provide only textual answers. However, for many questions, pure texts cannot provide intuitive information, while image or video contents are more appropriate. In this paper, we introduce a scheme that is able to enrich text answers with image and video information. Our scheme investigates a rich set of techniques including question/answer classification, query generation, image and video search reranking, etc. Given a question and the community-contributed answer, our approach is able to determine which type of media information should be added, and then automatically collects data from Internet to enrich the textual answer. Different from some efforts that attempt to directly answer questions with image and video data, our approach is built based on the community-contributed textual answers and thus it is more feasible and able to deal with more complex questions. We have conducted empirical study on more than 3,000 QA pairs and the results demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.5 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Question answering, CQA, medium selection, reranking

1. INTRODUCTION

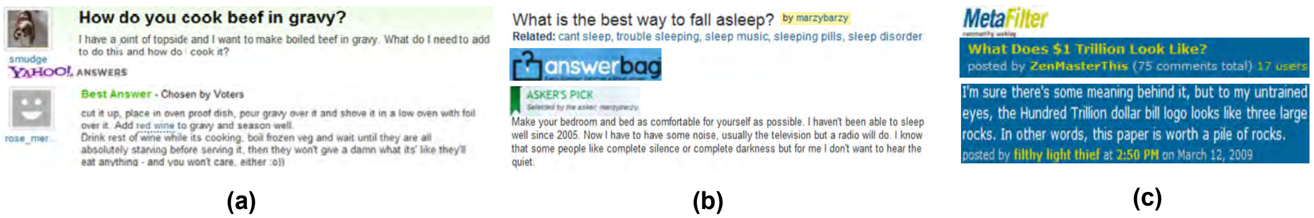
Question answering (QA) is defined as the task of automatically providing a precise answer to a natural language question posed by users [28, 26, 11, 32]. Typically, given a question, an ideal QA system is expected to find answer from certain corpuses (such as the Wide World Web) using information retrieval and natural language processing techniques. Despite great progress and encouraging results have been reported, traditional automated QA still faces challenges that are not easy to tackle, such as the deep understanding of complex questions and the sophisticated syntactic, semantic and contextual processing to generate answers. It is found that, in most cases, automated approach cannot obtain results that are as good as those generated by manual processing [2, 12].

Along with the proliferation and improvement of underlying communication technologies, community question answering (cQA) has emerged as an extremely popular alternative to finding information online, owning to the following facts. First, information seekers are able to post their specific questions on any topic and obtain answers provided by other participants. By leveraging community efforts, they are able to get better answers than simply using search engine to find them. Second, in comparison with automated QA systems, cQA usually receives answers with better quality as they are generated based on human intelligence. Third, over times, a tremendous number of QA pairs have been accumulated in their repositories, and it facilitates the preservation and retrieval of answered questions. The most well-known Internet cQA system is Yahoo! Answers (Y!A), which contains more than 1 billion QA pairs as at Oct 2009, contributed by the general public.

Despite great success has been achieved, existing cQA forums mostly provide only textual answers, as shown in Figure 1. However, a picture is worth a thousand words. In many cases, the questions cannot be well explained using only texts, and it will be much better to visualize the answers with images and videos. Figure 1 (a) illustrates such an example: for the question "*How do you cook beef in gravy*", the answer is described by several long sentences. However, users still can hardly grasp the process. Clearly, it will be much better if there is an accompanying video describing the process. Therefore, the textual answers in cQA can be significantly enhanced by adding multimedia contents, and it will provide answer seekers with better experience.

Actually in cQA corpuses, there are already many answers that directly embed hyperlinks to images or videos from which the users can get supplementary information in media form. For example, for the question "*What are the best steps to take in order to repent*", the best answer on Y!A is a URL that leads information seekers to YouTube

SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07 ...$10.00.



How do you cook beef in gravy?

I have a joint of topside and I want to make boiled beef in gravy. What do I need to add to do this and how do I cook it?

smudge

YAHOO! ANSWERS

rose_mer...

Best Answer - Chosen by Voters

cut it up, place in oven proof dish, pour gravy over it and shove it in a low oven with foil over it. Add red wine to gravy and season well.
Drink rest of wine while its cooking, boil frozen veg and wait until they are all absolutely starving before serving it, then they won't give a damn what its' like they'll eat anything - and you won't care, either :o))

(a)

What is the best way to fall asleep? by marzybarzy

Related: cant sleep, trouble sleeping, sleep music, sleeping pills, sleep disorder

answerbag

ASKER'S PICK

Selected by the asker: marzybarzy

Make your bedroom and bed as comfortable for yourself as possible. I haven't been able to sleep well since 2005. Now I have to have some noise, usually the television but a radio will do. I know that some people like complete silence or complete darkness but for me I don't want to hear the quiet.

(b)

MetaFilter

What Does $1 Trillion Look Like?

posted by ZenMasterThis (75 comments total) 17 users

I'm sure there's some meaning behind it, but to my untrained eyes, the Hundred Trillion dollar bill logo looks like three large rocks. In other words, this paper is worth a pile of rocks.

posted by filthy light thief at 2:50 PM on March 12, 2009

(c)

Figure 1: Examples of Question-Answering on popular cQA corpuses. (a) An example on Yahoo! Answer; (b) an example on Answerbag; and (c) an example on MetaFilter.

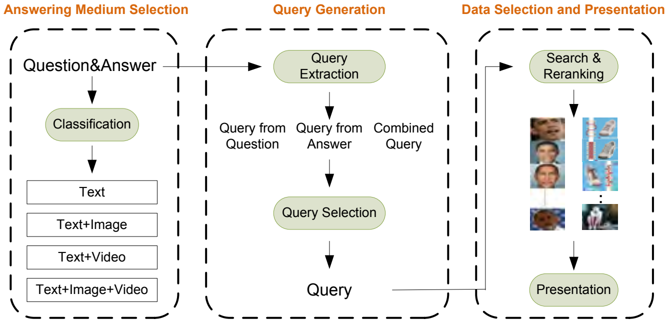


Figure 2: The schematic illustration of the proposed multimedia answering scheme.

[1]. This indicates that many answers can be enhanced by leveraging multimedia information. However, existing cQA forums do not provide adequate support on using media information.

In this work, we propose a multimedia answering scheme that is able to find appropriate image or video information to complement the community-contributed textual answers in cQA. It explores a rich set of techniques, including question/answering classification, query extraction and classification, image and video search reranking, etc. As shown in Figure 2, the scheme consists of three main components:

(1) Answer medium selection. In this work, we consider the following four cases[2] for answer media: (a) only text, i.e., the original textual answers are sufficient; (b) text + image, i.e., image information needs to be added; (c) text + video, i.e., only video information is to be added; and (d) text + image + video, i.e., we add both image and video information. We regard it as a QA pair classification problem, that is, given a question and its community-contributed answer in cQA corpus, we classify it into one of the above four classes.

(2) Multimedia query generation. In order to collect multimedia data from the web, we generate queries from each QA pair. Here we generate three types of queries from (a) question, (b) answer, and (c) both question and answer. We then choose one from the three queries by learning a classification model.

(3) Multimedia data selection and presentation. Based on the generated queries, we collect image and video data with multimedia search engines. We then perform reranking and duplicate removal to obtain a set of accurate and representative samples for presentation together with the textual answers.

It is worth mentioning that there already exist several efforts dedicated to research on automatically answering questions with multimedia data, i.e., the so-called Multimedia Question Answering (MMQA). For example, Yang et al. [36] proposed to extend text-based QA technology to support factoid QA in news video. A photo-based QA system for finding information about physical objects was presented in [37]. Li et al. [20] explored how to leverage YouTube video collections as a source to automatically find videos to describe cooking techniques. But these approaches usually work on certain narrow domains and can hardly be generalized to handle general questions in broad domains. This is due to the fact that, in order to accomplish automatic MMQA, we first need to understand questions, which is not an easy task. Our proposed approach in this work does not aim to directly answer the questions, and instead we enrich the community-contributed answers with multimedia content. Our strategy splits the large gap between question and multimedia answer into two smaller gaps, i.e., the gap between question and textual answer and the gap between textual answer and multimedia answer. In our scheme, the first gap is bridged by the crowd sourcing intelligence of community members, and thus we can focus on solving the second gap. Therefore, our scheme can also be viewed as an approach that accomplishes the MMQA problem by jointly exploring human and computer. Figure 3 demonstrates the difference between the conventional MMQA approaches and an MMQA framework based on our scheme. It is worth noting that, although the proposed approach is automated, we can also further involve human interactions. For example, our approach can provide a set of candidate images and videos based on textual answers, and answerers then can manually choose several candidates for final presentation.

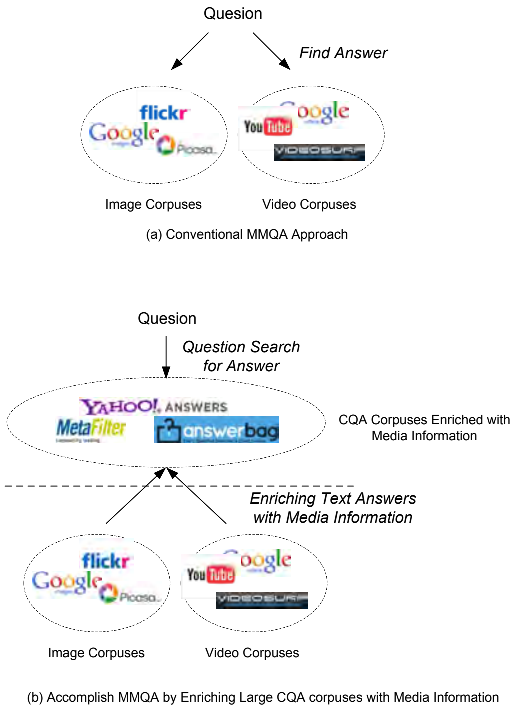
The contributions of this work can be summarized as follows:

(1) To the best of our knowledge, this is the first work on automatically enriching textual answers with image and video data for better QA experience. Different from the conventional MMQA research that aims to automatically generate multimedia answers with given questions, our approach is built based on the community-contributed answers, and it can thus deal with more general questions and achieve better performance.

(2) We investigate the prediction of appropriate answer medium. Here we want to predict whether a textual answer should or should not be enriched with multimedia information, and which kind of media data should be added.

[1] http://www.youtube.com/watch?v=zJCzrTnfPn0

[2] Here we have not considered audio because our studies show that very few users would like to answer questions using this kind medium in cQA, as most speech content can be presented in text form.


Quesion
Find Answer
Image Corpuses
Video Corpuses
Quesion
Question Search for Answer
CQA Corpuses Enriched with Media Information
Enriching Text Answers with Media Information
Image Corpuses
Video Corpuses

(a) Conventional MMQA Approach

(b) Accomplish MMQA by Enriching Large CQA corpuses with Media Information

Figure 3: The differences of the conventional MMQA approaches and MMQA based on our scheme. (a) Conventional MMQA aims to seek multimedia answers from online corpuses. (b) Our proposed scheme enriches textual answers in large cQA corpuses with image and video information (this process can be offline), and then, for a user-provided question, we can perform question search to find multimedia answers in the cQA corpuses.

(3) We propose a method to generate queries from QA pairs for multimedia search and perform query-dependent reranking for image and video data obtained from search engines by analyzing visual features.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. In Section 3 and 4, we introduce the answer medium selection and query generation components, respectively. We then introduce the media data selection and presentation approach in Section 5. Experimental results and analysis are presented in section 6, followed by the conclusion and future work in section 7.

2. RELATED WORK

The research on QA can be traced back to the 1960s, and early QA systems were mainly natural language interfaces to expert systems that were tailored to specific domains. Text Retrieval Conference (TREC) evaluations established a QA track in the late 1990s [1], and it generates extensive research interests in textual QA, such as Open-Domain QA [28], Restricted-Domain QA [26], Definitional QA [11] and List QA [32]. However, despite the great progress achieved, automatic QA still has difficulties in answering complex and realistic questions. Along with the blooming of Web 2.0, cQA becomes an alternative approach. It is a large and diverse question-answer forum, acting not only as a corpus for sharing technical knowledge but also a place where one can seek advice and opinions [2, 12]. However, nearly all existing cQA systems, such as Yahoo! Answer, Answerbag and Ask Metafilter, only support pure text-based answers, which may not provide intuitive and sufficient information.

Some research works have been conducted on multimedia QA, which aims to answer questions using multimedia data. An early system named VideoQA was presented in [36]. This system extends the text based QA technology to support factoid QA in news video by leveraging visual contents of news video as well as the text transcripts. Following this work, several video QA systems were proposed and most of them rely on the use of text transcript derived from video OCR (Optical Character Recognition) and ASR (Automatic Speech Recognition) outputs [8, 33, 18, 34]. Li et al. [21] presented a solution on "how-to" QA by leveraging community-contributed texts and videos. Kacmarcik et al. [17] explored a method of creating a non-text input mode for QA that relies on specially annotated virtual photographs. An image-based QA approach was introduced in [37], which mainly focuses on finding information about physical objects. Chua et al. [9] in 2009 proposed an approach to extend text based QA research to multimedia QA to tackle a range of factoid, definition and "how-to" QA in a common framework. Their system was designed to find multimedia answers from web-scale media resources such as Flicker and YouTube. However, literature regarding multimedia QA is still relatively sparse. As mentioned in Section 1, automatic multimedia QA only works in specific domains and can hardly handle complex questions. Different from the above mentioned works, our approach is built based on cQA. Instead of directly collecting multimedia data for answering questions, our method only finds images and videos to enrich the textual answers provided by humans. This makes our approach able to deal with more general questions and achieve better performance.

3. ANSWER MEDIUM SELECTION

As introduced in Section 1, the first component of our scheme is answer medium selection, as we first need to determine whether we need to and if so which type of medium we should add to enrich the textual answers. For some questions, such as "*what day is President Obama's birthday*", using pure textual answers is sufficient. But we need to add image or video information for some other questions. For example, for the question, "*who is Obama*", it is better to add images to complement the textual answer, whereas we should add videos for answering the question "*how to cook beef*". We regard the answer medium selection as a QA pair classification task, i.e., given a question and textual answer, we classify it into the following four classes of answer medium combinations: (1) only text; (2) text + image; (3) text + video; and (4) text + image + video. For the "only text" class, we do not need to perform more operations, and for the other cases we will need to collect appropriate data by the other two components.

There are some existing research efforts on question classification. Li and Roth [22] developed a machine learning approach that uses the SNoW learning architecture to classify questions into five coarse classes and 50 finer classes. They used lexical and syntactic features such as part-of-speech tags, chunks and head chunks together with two semantic features: named entities and semantically related words to represent the questions. Zhang and Lee [39] used

Interrogative Word	Category
be, can, will, have, when, be there, how+adj/adv	Text
what, where, which, why, how to, who, etc.	Need further classification

Table 1: A question can be judged to be answered with pure text if the interrogative word falls into a list. Otherwise, we perform classification for selecting answer medium.

linear SVMs with all possible question word grams to perform question classification. Arguello et al. [6] have investigated selecting medium type as well as search sources for a query. But there is no work on classifying QA pair according their best answer medium types. This task is more challenging as we are dealing with real data on the web, including many complex and multi-sentence questions and answers, and we need to extract rules to connect QA texts and the best answer medium types. We accomplish the task by first analyzing the question and answer separately and then combining the results.

3.1 Question-Based Classification

Since many questions contain multiple sentences (actually our statistics on Y!A show that at least 1/5 of the questions contain more than two sentences) and some of the sentences are uninformative, we first employ the method in [30] to extract the core sentence from each question.

The classification is accomplished in two steps. First, we categorize the questions based on the interrogatives (some starting words and ending words), and for some questions we can directly derive that they should be answered with text. Second, for the rest questions, we perform classification using a naive Bayes classifier.

We first introduce the categorization based on interrogative words. Questions can mainly be categorized into the following classes based on interrogative words: yes/no class (such as "*Does Java support VoIP*"), choice class (such as "*Which country is bigger, Canada or America*"), quantity class (such as "*When is the Chinese New Year*"), enumeration class (such as "*Names of the 3 wise men from the East who came to venerate Jesus Christ*"), and description class (such as "*What are the steps required to edit, publish and distribute a book*"). For example, a question will be categorized to the "quantity" class if the interrogative is "how+adj/adv" or "when". For the "yes/no", "choice" and "quantity" questions, we categorize them into the class of answering with only text; while the "enumeration" and "description" questions can need "text+image", "text+video" or "text+image+video answers". Therefore, given a question, we first judge whether it should use only textual answer based on the interrogative word. If not, we further perform classification with a Naive Bayes classifier. Table 1 shows the heuristics. For building the Naive Bayes classifier, we extract a set of text features, including bigram text features, head words and a list of class-specific related words[3].

Categories	Class-Specific Related Word List
Text	name, population, period, times, country, height, website, birthday, age, date, rate, distance, speed, religions, number, etc
Text+Image	colour, pet, clothes, look like, who, image, pictures, appearance, largest, band, photo, surface, capital, figure, what is a, symbol, whom, logo, place, etc.
Text+Video	How to, how do, how can, invented, story, film, tell, songs, music, recipe, differences, ways, steps, dance, first, said, etc.
Text+Both	president, king, prime minister, kill, issue, nuclear, earthquake, singer, battle, event, war, happened, etc.

Table 2: Representative class-specific related words. "Text+Both" stands for "Text+Image+Video".

Here head word is referred to as the word specifying the object that the question seeks.The semantics of head words play an important role in deciding answer medium. For instance, for the question "*what year did the cold war end*", the head word is "year", based on which we can judge that the sought-after answer is a simple date. Therefore, it is reasonable to use textual answer medium. We adopt the method in [16], but the key difference is that we do not use post fix as it better fit our answer medium classification task.

We also extract a list of class-specific related words in a semi-automatic way. We first estimate the appearing frequency of each phrase in the positive samples of each class, and collect all the phrases that have the frequencies above a threshold (we empirically set the threshold to 3 in this work). We then manually refine the list based on human's understanding. Examples of class-specific related words for each class is given in Table 2.

3.2 Answer-Based Classification

Besides question, the answer with rich text can also be an important information clue. For example, for the question "*how do you cook beef in gravy*", we may find a textual answer as "*cut it up, put in oven proof dish ...*". Then we can judge that the question can be better answered with a video clip as the textual answer contains many verbs and describes a dynamic process.

For answer classification, we extract bigram text features and verbs[4]. The verbs in the answer will be useful for judging whether the question can be answered with video content. Intuitively, if a textual answer contains many complex verbs, it is more likely to describe a dynamic process and thus it has high probability to be well answered by videos. Therefore, verb can be an important clue.

Based on the bigram text features and verbs, we again build a Naive Bayes classifier with a set of training data, and then perform a four-class classification with the model. The classification results are linearly combined with those of question-based classification.

[3] Actually we have also investigated other features such as unigram and trigram. Empirical study demonstrates that the combination of bigram, head words and the class-specific related words is able to achieve promising performance while maintaining good generalization ability.

[4] Actually we have investigated other features such as unigram and visually descriptive nouns. Empirical study demonstrates that the combination of bigram and verbs shows promising performance and good generalization ability.

4. MULTIMEDIA QUERY GENERATION

To collect relevant image and video data from the web, we need to generate appropriate queries from text QA pairs before performing search on multimedia search engines. We accomplish the task with two steps. The first step is query extraction. Text questions and answers are usually complex sentences. But frequently search engines do not perform well with queries that are long and verbose. Therefore, we need to extract a set of informative keywords from questions and answers for querying. The second step is query selection. This is because we can generate different queries: one from question, one from answer, and one from the combination of question and answer. Which one is the most informative depends on the question: some QA pairs embed the useful query terms in their questions, such as "*What does Disneyland tickets look like*"; some hide the helpful key words in their answers, such as the QA pair "*Q: What is the capital of Thailand; A: Bangkok*"; and some should combine question and answer to generate useful query, such as the QA pair "*Q: Who is the man in the moon; A: astronaut*", for which both of the "astronaut" and the "moon" are functional words.

For each QA pair, we will generate three queries. First, we convert question to query, i.e., we convert a grammatically correct interrogative sentence into one of the potential syntactically correct declarative sentences or meaningful phrases. We directly employ the method in [4]. Second, we identify several key concepts from verbose answer which will have the most impact on effectiveness. Here we employ the method in [7]. Finally, we combine the two queries that are generated from question and answer respectively. Therefore, we obtain three queries, and the next step is selecting one from them.

The query selection is formulated as a three-class classification task, since we need to choose one from the three queries that are generated from the question, answer and the combination of question and answer. We adopt the following features:

(1) POS Histogram. We use POS tagger to assign part-of-speech to each word of both question and answer. Here we employ the Stanford Log-linear Part-Of-Speech Tagger and 36 POS are identified[5] . For both question and answer, we then generate a 36-dimensional histogram, in which each bin counts the number of words belonging to a corresponding category of part-of-speech.

(2) Retrieval effectiveness. This is because for certain queries, existing image and video search engines cannot return satisfactory results. We adopt the method proposed in [10], which defines a clarity score for a query based on the KL divergence between the query and collection language models. We can generate 6-dimensional retrieval effectiveness features in all (three queries and search is performed on image and video search engines).

Therefore, for each QA pair, we can extract 42-dimensional features. Based on the extracted features, we train an SVM classifier with a labeled training set for classification, i.e., selecting one from the three queries.

5. MULTIMEDIA DATA SELECTION AND PRESENTATION

We perform search using the generated queries to collect image and video data with Google image and video search engines respectively. However, commercial search engines, such as Google, Yahoo and Bing, usually index web images and videos using textual information, such as titles, ALT text and surrounding texts on web pages. But frequently the text information does not fully describe content of the images and videos, and this fact can severely degrade the search relevance of web images and videos. Reranking is an approach to improving search relevance by mining the visual information of images and videos. Existing reranking algorithms can mainly be categorized into two approaches, one is pseudo relevance feedback [27, 35, 23] and the other one is graph-based reranking [25, 15, 31]. The pseudo relevance feedback approach regards top results as relevant ones and then collects some samples that are assumed to be irrelevant. A classification or ranking model is learned based on the pseudo relevant and irrelevant samples and the model is then used to rerank the images. The graph-based reranking approach assumes that the relevant images lie on a manifold in visual feature space. Generally, the approach constructs a graph where the vertices are images or videos and the edges reflect their pairwise similarities. Here we adopt the graph-based reranking method in [15]. We re-state the equation from [15] as,

$$r^j_{(k)} = \alpha \sum_{i \in B_j} r^i_{(k-1)} W_{ij} + (1-\alpha) r^j_{(0)} \tag{1}$$

where $r^j_{(k)}$ stands for the state probability of node j in the k-th round of iterations, α is a parameter that satisfies $0 \leq \alpha < 1$, and W_{ij} is the similarity between the i-th and j-th samples. Here $r^i_{(0)}$ is the initial relevance score of the sample at the i-th position, which is heuristically estimated as

$$r^i_{(0)} = \frac{N-i}{N(N-1)/2} \qquad i = 1, 2 ... N \tag{2}$$

For images, we directly estimate their similarity based on their Euclidean distance

$$W_{ij} = \exp(-\frac{||x_i - x_j||^2}{\sigma^2}) \tag{3}$$

The parameter σ is simply set to the median of the Euclidean distance of all image pairs.

For videos, we first perform shot boundary detection and then extract a key-frame from each shot using the method in [38]. Considering two videos $(v_{i,1}, ..., v_{i,m})$ and $(v_{j,1}, ..., v_{j,n})$, which contain m and n key-frames respectively, we employ average distance [24] of all cross-video key-frame pairs for similarity estimation, i.e.,

$$W_{ij} = \exp(-\frac{\sum_{i=1}^m \sum_{j=1}^n ||v_{1,i} - v_{2,j}||^2}{\sigma^2}) \tag{4}$$

Similarly, the parameter σ is simply set to the median of the Euclidean distance of all video pairs.

However, a problem with existing reranking methods is that they usually use features extracted from the whole images or video frames and they overlooked that many queries are actually person-related. Clearly, for person-related queries, it is more reasonable to use facial features instead of global visual features for reranking. For question-answering, our

[5] They are: RB, DT, RP, RBR, RBS, LS, VBN, VB, VBP, PRP, MD, SYM, VBZ, IN, VBG, POS, EX, VBD, LRB, UH, NNS, NNP, JJ, RRB, TO,JJS, JJR, FW, NN, NNPS, PDT, WP, WDT, CC, CD, and WRB.

Category	Distribution
Text	45%
Text+Image	24%
Text+Video	22%
Text+Image+Video	9%

Table 3: The distribution of the expected answer medium types labeled by humans.

Question-based classification with different features	Accuracy
Bigram	69.02%
Bigram+Head	73.60%
Bigram+Related	71.43%
Bigram+Related+Head	**74.38%**

Table 4: The performance of question-based classification with different features. Related means class-specific related words.

statistics show that more than 1/4 of the QA pairs in our data set are about person. Therefore, in this work we propose a query-depentdent reranking approach. We first decide whether a query is person-related or not, and then we use different features for reranking.

We establish the following rules to judge whether a query is person-related:

(1) If the given question starts by interrogative words who or whom, then it is categorized as person-related.

(2) We perform morphological analysis on the given question to extract information on Part-of-Speech, verb-phrase and noun-phrase. We then extract the main core terms (the strongest noun or noun phrase), followed by the other possible key terms. Following that, we employ the Name Entity extractor to identify names of persons, organizations and possible objects in the question. If a person's name appears in the query as the core term, the query is categorized as person-related.

If a query is person-related, we perform face detection for each image and video key-frame. If an image or a key-frame does not contain faces, it will be not considered in reranking (it is reasonable as we will only consider images and frames that contain faces for person-related queries). If faces are found in images or key-frames, we extract the 256 dimensional Local Binary Pattern features [5] from the largest faces of images or video frames. More specifically, the method works as follows. For person-unrelated queries, we extract 428-dimensional global visual features, including 225-D block-wise color moments generated from 5-by-5 fixed partition of the image, 128-D wavelet texture, and 75-D edge direction histogram.

After reranking, visually similar images or videos may be ranked together. Thus we perform a duplicate removal step to avoid information redundancy. We check the ranking list from top to bottom, and if an image or video is close to a sample that appears above it, we remove it. More specifically, we remove the i-th image or video if there exists $j < i$ that satisfies $W_{ij} > T$. Here we empirically set T to 0.8 throughout the work.

Answer-based classification with different features	Accuracy
Bigram	54.71%
Bigram+Verb	**57.39%**

Table 5: The performance of answer-based classification with different features.

Classification Method	Accuracy
Question-Based Classification	74.38%
Answer-Based Classification	57.39%
Linear combination	**78.29%**

Table 6: Results of linear fusion for answer medium selection.

After duplicate removal, we keep the top 10 images and top 2 videos (keeping which kind of media data depend on the classification results of answer medium selection). When presenting videos, we not only provide videos but also illustrate the key-frames to help users quickly understand the video content as well as easily browse the videos.

6. EXPERIMENTS

6.1 Experimental Settings

We randomly select 5,000 questions and their corresponding answers from the dataset used in [29], which contains 4,483,032 questions and their answers from Y!A. Here we use the best answer that is determined by the asker or the community voting[6]. Inspired by [19, 13], we first classify all the questions in cQA into two categories: conversational and informational. Since conversational questions usually only seek personal opinions or judgments, such as "*Anybody watch the Bears game last night*", in this work we only consider informational questions. In our work, there are 3333 informational questions among the 5000 questions. The questions and answers in our dataset cover a wide range of topics, including travel, life, education, etc.

Our answer medium selection and query selection need training data, and thus we split the 3333 QA pairs into two parts, a training set that contains 2666 QA pairs and a testing set of the remaining 667 QA pairs. The classifiers for answer medium selection and query selection are trained with the training set and evaluated on the testing set.

6.2 Evaluation of Answer Medium Selection

We first evaluate our answer medium selection approach. The ground truths are established by humans. Five human labelers were involved in the process. For every question, each labeler categorized it into one of the following four classes (text, text+image, text+video, text+image+video), and then voting was performed to obtain the final ground truth. For the cases that there are two classes having the same number of ballots, a discussion is carried out among

[6] There are also many research efforts on ranking community-contributed answers or selecting the best answer by machine learning and NLP technologies [29, 14, 3]. These methods can also been integrated with our work and we only need to change the best answer for each question.

Text
1. How many pairs of shoes were found when Imelda Marcos left the presidential palace in 1986?
2. When is the second generation mini cooper coming out?
3. What year was the movie Mustang Sally made?
4. what is speed limit on on california freeways?
5. Is the abbrevation for Mathematics - 'Maths' or 'Math'? Please give me a grammatically correct answer.

Text+Image
1. who won the 1992 Alice Springs cup?
2. Anybody have a picture of Anthropologie's edwarian overcoat?
3. What is the symbol of the Democratic Party?
4. What are 5 manufacturing plants around the world for reebok?
5. Largest and the highest bridge in Asia?

Text+Video
1. Does anyone have an easy recipe for butternut squash soup?
2. How do I remove wax from my refrigerator??? Please help!!!?
3. I want to go 2 for studies abroad so plz tell me the procedure how to get through it plzzzzz.?
4. What is the best way to become an Ebay Powerseller?
5. Exactly what steps do I take to get more space in my mail box?

Text+Image+Video
1. Which President had his old war horse grazing on the White House lawn?
2. What is the largest earthquake (magnitude) to strike the U.S.?
3. What was the worst event that happened in the U.S. other than wars?
4. America Drops Nuclear Bomb On Japan?
5. Who is the most popular Asian singer in Western countries like US and UK??

Table 7: The representative questions for each answer medium class. Here we do not illustrate the answers because several answers are fairly long.

Category	Ground Truth Distribution
Answer	29%
Question	49%
Combination	22%

Table 8: The Ground truth distribution for Query selection.

Classification with Different Features	Accuracy
POS Histogram	66.17%
Retrieval Effectiveness	59.41%
POS Histogram+Retrieval Effectiveness	**71.27%**

Table 9: The classification performance for query selection with different features

the labelers to decide the final ground truths. Table 3 illustrates the distribution of the four classes. We can see that, more than 50% of the questions can best be answered by adding multimedia contents instead of using purely text. This also demonstrates that our multimedia answering approach is highly desired.

For the component of medium selection, we first investigate different feature combinations for the question and answer analysis. The results are illustrated in Tables 4 and 5, respectively. It is worth noting that the stop-words are not removed for question classification since some stop-words also play an important part in question classification. But for answering classification, stop words are removed. Stemming is performed for both question and answer. From the results we can see that, for both of the two classifiers, integrating all of the introduced features is better than only part of them.

We linearly fuse the two classifiers with a grid search with optimal weighting. Table 6 illustrates the classification accuracies when combining question-based classification and answer-based classification. It can be observed that question-based classification achieves better results than answer-based classification. But fusing the classifiers is able to achieve significantly better performance with a final classification accuracy of 78.29%. Table 7 presents representative examples of classified questions for each category.

6.3 Evaluation of Query Generation

Now we evaluate the retrieval effectiveness of our query generation and selection approach. The ground truth was also obtained by the five human labelers. For every QA pair, each labeler selected one from the three queries, which are generated from the question, answer and the combination of question and answer, which will provide best information to retrieve relevant multimedia answers. These labelers are allowed to perform search on the web to compare the query effectiveness. Voting was performed to obtain the final ground truth. The distribution of the three classes is illustrated in Table 8.

We adopt SVM with RBF kernel, and the parameters, including radius parameter and the weighting parameter that modulates the regularizer term and loss term, are established with 5-fold cross-validation. Table 9 illustrates the classification results. Our approach that integrates POS histogram and retrieval effectiveness features achieves an accuracy of 71.27%.

In order to evaluate our query-dependent strategy, we first randomly selected 25 queries from the person-related ones. For each query, the top 150 images or videos are collected for reranking. We adopt NDCG@10 as our performance evaluation metric, which is estimated by

$$NDCG@n = \frac{DCG}{IDCG} = \frac{(rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i})}{IDCG} \qquad (5)$$

where rel_i is the relevance score of i-th image or video in the ranking list, $IDCG$ is the normalizing factor that equals to the DCG of an ideal ordering. Each image or video is labeled to be very relevant (score 2), relevant (score 1) or irrelevant (score 0) to a query by the voting of the five human labelers.

6.4 Evaluation of Reranking

Figures 4 and 5 illustrate the average performance comparison of our approach and the conventional method that

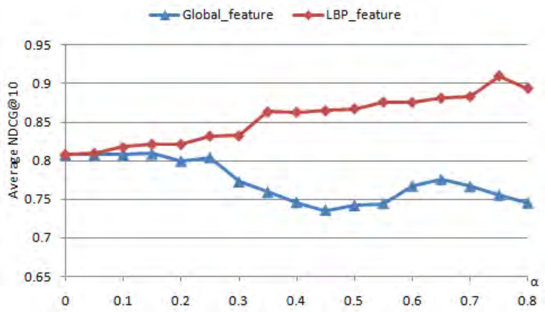


Figure 4: The image search reranking performance comparison of using global features and LBP features for person-related queries.

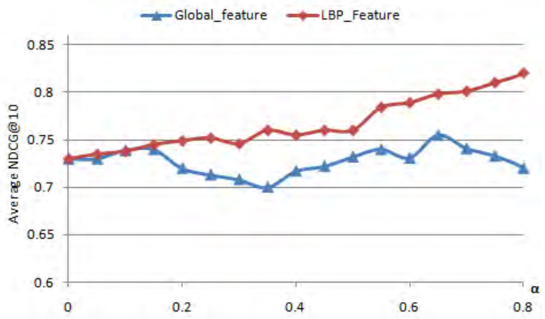


Figure 5: The video search reranking performance comparison of using global features and LBP features for person-related queries.

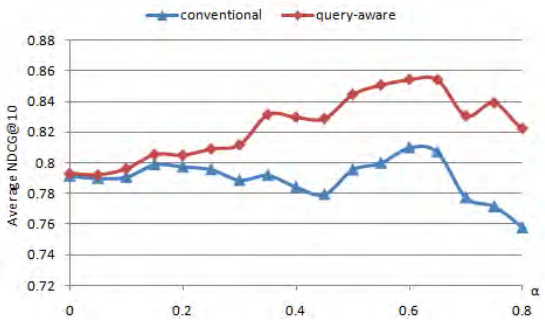


Figure 6: The image search reranking performance comparison of using our method and using the conventional method.

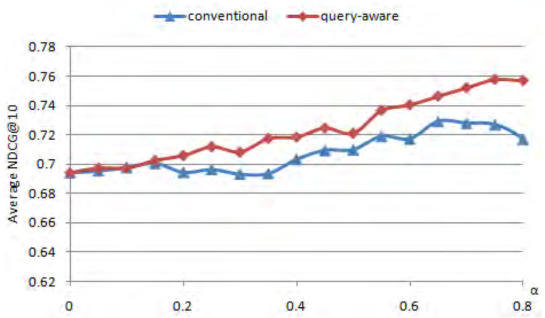


Figure 7: The video search reranking performance comparison of using our method and using the conventional method.

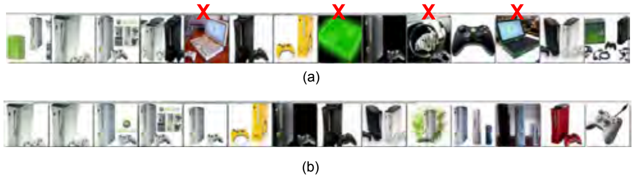


Figure 8: The non-person related image search reranking results for the query "X-Box360". (a)The top images before reranking; (b) the top images after reranking

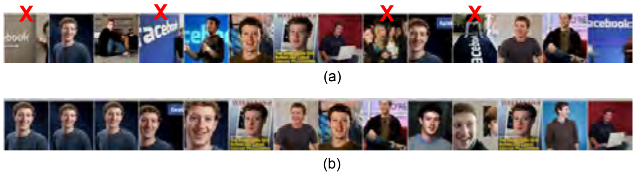


Figure 9: The person-related image search reranking results for the query "Facebook CEO". (a) The top images before reranking; (b) the top images after reranking

uses only global features for the 25 person-related queries. Here we have illustrated the performance with different values of the parameter α. Smaller α means more original ranking information is kept in reranking. When α equals 0, the reranked list will be identical to the original ranking list obtained by the text-based search. We can see that, our approach consistently outperforms the method that uses global features. This demonstrates that it is more reasonable to use facial features for person-related queries.

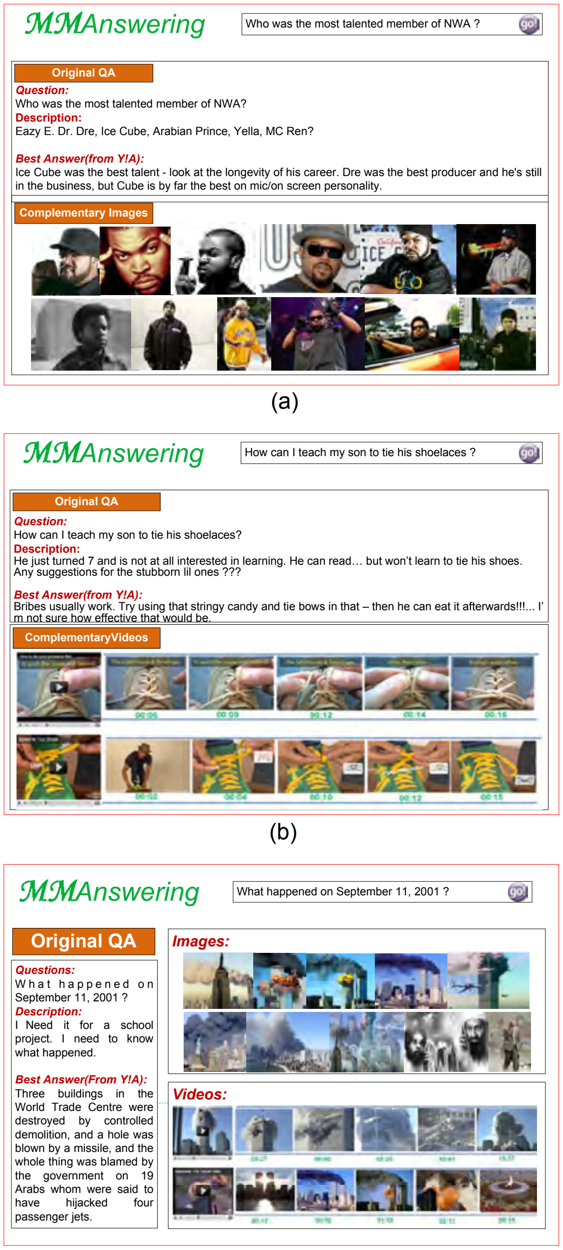
We then randomly selected 100 queries from image and video class, respectively. Figures 6 and 7 illustrate the average performance comparison. For our query-dependent method, we use global features and facial features for non-person-related and person-related queries respectively. While the conventional method only employ global features regardless the query oriented. Our approach significantly outperforms the conventional methods. We can see that, our approach can effectively improve search relevance when α varies in a wide range. Throughout our rest experiments, we set α as 0.65 and 0.8 for image and video reranking, respectively.

We illustrate the top results before and after reranking for two example queries shown in Figures 8 and 9, one about object and one about person. Figure 8(a) shows the top 14 results before reranking about "X-Box360", while (b) shows the top 14 results after reranking based on global features. Similarly, Figures 9(a) and (b) presents the top results before and after reranking with respect to person-related query "Facebook CEO". We can see that several less relevant results, marked with "X", are removed from top positions after reranking.

After reranking we perform duplicate removal and present the images or/and videos together with the textual answers, depending on the results of answer medium selection. Figure 10 shows the multimedia answer for 3 example queries.

6.5 Subjective Test of Multimedia Answering

Finally, we conduct a user study with 20 volunteers that



Who was the most talented member of NWA ?

Original QA

Question:
Who was the most talented member of NWA?

Description:
Eazy E. Dr. Dre, Ice Cube, Arabian Prince, Yella, MC Ren?

Best Answer(from Y!A):
Ice Cube was the best talent - look at the longevity of his career. Dre was the best producer and he's still in the business, but Cube is by far the best on mic/on screen personality.

Complementary Images

(a)

How can I teach my son to tie his shoelaces ?

Original QA

Question:
How can I teach my son to tie his shoelaces?

Description:
He just turned 7 and is not at all interested in learning. He can read… but won't learn to tie his shoes. Any suggestions for the stubborn lil ones ???

Best Answer(from Y!A):
Bribes usually work. Try using that stringy candy and tie bows in that – then he can eat it afterwards!!!... I'm not sure how effective that would be.

ComplementaryVideos

(b)

What happened on September 11, 2001 ?

Original QA

Questions:
What happened on September 11, 2001 ?

Description:
I Need it for a school project. I need to know what happened.

Best Answer(From Y!A):
Three buildings in the World Trade Centre were destroyed by controlled demolition, and a hole was blown by a missile, and the whole thing was blamed by the government on 19 Arabs whom were said to have hijacked four passenger jets.

Images:

Videos:

(c)

Figure 10: Results of multimedia answering for 3 example queries, "the most talented member of NWA", "tie shoelace", and "September 11". Our scheme answers the three questions with text + image, text + video, and text + image + video, respectively.

Prefer multimedia answer	Neutral	Prefer Original textual answer
45.3%	**51.5%**	3.2%

Table 10: Statistics of the comparison of our multimedia answer and the original textual answer.

Prefer multimedia answer	Neutral	Prefer Original textual answer
82.4%	11.8%	5.8%

Table 11: Statistics of the comparison of our multimedia answer and the original textual answer with exclusion of questions where only text-based answers are sufficient.

frequently use Y!A to evaluate the usability of our application. The users are asked to freely compare the conventional textual answers and our multimedia answers for different questions, and provide their ratings of the 2 systems. We adopt the following quantization approach: score 1 is assigned to the worse scheme and the other scheme is assigned with score 1, 2 and 3 if it is comparable, better and much better than this one, respectively. The average rating scores are illustrated in Figure 11. We can clearly see the preference of users towards the multimedia answering. We also perform an analysis of variance (ANOVA) test. The F-statistic of scheme factor is 18.11 and it can be derived that $p = 1.3154e - 004$. This indicates that the difference of the two schemes is significant. The F-statistic of user factor is 1.0 and it can be derived that $p > 0.5$, and this indicates that the difference among users is statistically insignificant.

We then conducted a more detailed study. For each question in the testing set, we simultaneously demonstrate the conventional best answer and the multimedia answer generated by our approach. Each user is asked to choose the preferred one. Table 10 presents the statistical results. We can see that, in about 45.3% of the cases users prefer our answer and only in 3.2% of the cases they prefer the original answers. But there are 51.5% neutral cases. This is because there are many questions that are classified to be answered by only texts, and for these questions our answer and the original textual answer are the same. If we exclude such questions, i.e., we only consider questions of which the original answer and our answer are different, then the statistics will turn to Table 11. We can see that for more than 82.4% of the questions, users will prefer the multimedia answers, i.e., the added image or video data are helpful. For cases that users prefer original textual answers, it is mainly due to the irrelevant image or video content. For several questions, the added image and video content are not only irrelevant to the question and answer but also distractive, and this thus degrades users' experience.

7. CONCLUSION AND FUTURE WORK

In this work, we have proposed a scheme to enrich text QA with media Information. For a given QA pair from cQA, our scheme first predicts which medium is appropriate to enrich the original textual answer. Next, it automatically generates a query based on the QA knowledge, and retrieves relevant image and video from search engines. Finally, query-dependent reranking and duplicate removal are performed to obtain a set of images and videos for presentation along with the original textual answer.

To our knowledge, this is the first work on enriching textual answer with multimedia information, and there is a lot of future work along this research direction. We will further improve the answer medium selection and query selection performance. We will also investigate methods to boost the relevance of the finally selected images and videos, as irrel-

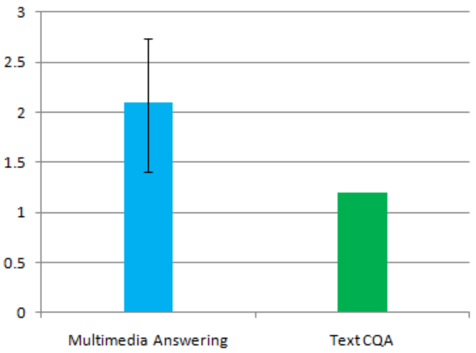


Figure 11: The mean ratings and variance of text cQA and multimedia answering.

evant multimedia content may degrade user experience. We also plan to conduct a more comprehensive user study on a larger dataset.

8. ACKNOWLEDGEMENT

We would like to acknowledge the support of "NExT Research Center" funded by MDA, Singapore, under the research grant: WBS:R-252-300-001-490.

9. REFERENCES

[1] Trec: The text retrieval conference. see http://trec.nist.gov/.

[2] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of International World Wide Web Conference*, 2008.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of International Conference on Web Search and Web Data Mining*, 2008.

[4] E. Agichtein, S. Lawrence, and L. Gravano. Learning search engine specific query transformations for question answering. In *Proceedings of International World Wide Web Conference*, 2001.

[5] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. *Proceedings of European Conference on Computer Vision*, 2004.

[6] J. Arguello, F. Diaz, J. Callan, and J. F. Crespo. Sources of evidence for vertical selection. In *Proceedings of ACM International SIGIR conference*, 2009.

[7] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of ACM International SIGIR conference*, 2008.

[8] J. Cao and J. Jay F. Nunamaker. Question answering on lecture videos: A multifaceted approach. *Proceedings of International Joint Conference on Digital Libraries*, 2004.

[9] T.-S. Chua, R. Hong, G. Li, and J. Tang. From text question-answering to multimedia qa on web-scale media resources. In *Proceedings of ACM workshop on Large-Scale Multimedia Retrieval and Mining*, 2009.

[10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of ACM International SIGIR conference*, 2002.

[11] H. Cui, M.-Y. Kan, and T.-S. Chua. Soft pattern matching models for definitional question answering. *ACM Transactions on Information Systems*, 2007.

[12] Z. Gyongyi, G. Koutrika, J. Pedersen, and H. Garcia-Molina. Questioning yahoo! answers. Technical report, Stanford InfoLab, 2007.

[13] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social qa sites. In *Proceedings of International Conference on Human Factors in Computing Systems*, 2009.

[14] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online qa sites. In *Proceedings of International Conference on Human Factors in Computing Systems*, 2008.

[15] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of ACM International Conference on Multimedia*, 2007.

[16] Z. Huang, M. Thint, and Z. Qin. Question classification using head words and their hypernyms. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2008.

[17] G. Kacmarcik. Multi-modal question-answering: Questions without keyboards. Asia Federation of Natural Language Processing, 2005.

[18] Y.-S. Lee, Y.-C. Wu, and J.-C. Yang. Bvideoqa: Online english/chinese bilingual video question answering. *American Society for Information Science and Technology*, 2009.

[19] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein. Exploring question subjectivity prediction in community qa. In *Proceedings of ACM International SIGIR conference*, 2008.

[20] G. Li, R. Hong, Y.-T. Zheng, S. Yan, and T.-S. Chua. Learning cooking techniques from youtube. In *Advances in Multimedia Modeling*. 2010.

[21] G. Li, H. Li, Z. Ming, R. Hong, S. Tang, and T.-S. Chua. Question answering over community contributed web video. *IEEE Multimedia*, 2010.

[22] X. Li and D. Roth. Learning question classifiers. In *Proceedings of International Conference on Computational Linguistics*, 2002.

[23] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *International Conference on Multimedia & Expo*, 2008.

[24] J. T. R. H. Meng Wang, Xian-Sheng Hua. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 2009.

[25] X.-S. H. H.-J. Z. Meng Wang, Kuiyuan Yang. Towards relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 2010.

[26] D. Mollá and J. L. Vicedo. Question answering in restricted domains: An overview. *Computational Linguistics*, 2007.

[27] A. P. Natsev, M. R. Naphade, and J. TešiĆ. Learning the semantics of multimedia queries and concepts from a small number of examples. In *Proceedings of ACM International Conference on Multimedia*, 2005.

[28] S. A. Quarteroni and S. Manandhar. Designing an interactive open domain question answering system. *Journal of Natural Language Engineering*, 2008.

[29] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to rank answers on large online QA collections. In *Proceedings of the Association for Computational Linguistics*, 2008.

[30] A. Tamura, H. Takamura, and M. Okumura. Classification of multiple-sentence questions. In *Natural Language Processing*. 2005.

[31] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceeding of ACM International Conference on Multimedia*, 2008.

[32] R. C. Wang, N. Schlaefer, W. W. Cohen, and E. Nyberg. Automatic set expansion for list question answering. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2008.

[33] Y.-C. Wu, C.-H. Chang, and Y.-S. Lee. Cross-language video question/answering system. *Proceedings of International Symposium on Multimedia Software Engineering*, 2004.

[34] Y.-C. Wu and J.-C. Yang. A robust passage retrieval algorithm for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008.

[35] R. Yan, A. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proceedings of International Conference on Image and Video Retrieval*, 2003.

[36] H. Yang, T.-S. Chua, S. Wang, and C.-K. Koh. Structured use of external knowledge for event-based open domain question answering. In *Proceedings of ACM International SIGIR Conference*, 2003.

[37] T. Yeh, J. J. Lee, and T. Darrell. Photo-based question answering. In *Proceeding of ACM International Conference on Multimedia*, 2008.

[38] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1993.

[39] J. Zhang, R. Lee, and Y. J. Wang. Support vector machine classifications for microarray expression data set. *Proceedings of International Conference on Computational Intelligence and Multimedia Applications*, 2003.