Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information

Liqiang Nie, Meng Wang, *Member, IEEE*, Yue Gao, Zheng-Jun Zha, *Member, IEEE*, and Tat-Seng Chua, *Senior Member, IEEE*

***Abstract*—Community question answering (cQA) services have gained popularity over the past years. It not only allows community members to post and answer questions but also enables general users to seek information from a comprehensive set of well-answered questions. However, existing cQA forums usually provide only textual answers, which are not informative enough for many questions. In this paper, we propose a scheme that is able to enrich textual answers in cQA with appropriate media data. Our scheme consists of three components: answer medium selection, query generation for multimedia search, and multimedia data selection and presentation. This approach automatically determines which type of media information should be added for a textual answer. It then automatically collects data from the web to enrich the answer. By processing a large set of QA pairs and adding them to a pool, our approach can enable a novel multimedia question answering (MMQA) approach as users can find multimedia answers by matching their questions with those in the pool. Different from a lot of MMQA research efforts that attempt to directly answer questions with image and video data, our approach is built based on community-contributed textual answers and thus it is able to deal with more complex questions. We have conducted extensive experiments on a multi-source QA dataset. The results demonstrate the effectiveness of our approach.**

***Index Terms*—CQA, medium selection, question answering, reranking.**

I. Introduction

QUESTION-ANSWERING (QA) is a technique for automatically answering a question posed in natural language [1]–[4]. Compared to keyword-based search systems, it greatly facilitates the communication between humans and computers by naturally stating users' intention in plain sentences. It also avoids the painstaking browsing of a vast quantity of information contents returned by search engines for the correct answers. However, fully 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 human intelligence [5], [6].

Along with the proliferation and improvement of underlying communication technologies, community QA (cQA) has emerged as an extremely popular alternative to acquire 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 engines. 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 search of answered questions. For example, WikiAnswer, one of the most well-known cQA systems, hosts more than 13 million answered questions distributed in 7,000 categories (as of August 2011).

Despite their great success, existing cQA forums mostly support only textual answers, as shown in Fig. 1. Unfortunately, textual answers may not provide sufficient natural and easy-to-grasp information. Fig. 1(a) and (b) illustrate two examples. For the questions "*What are the steps to make a weather vane*" and "*What does $1 Trillion Look Like*", the answers are described by long sentences. Clearly, it will be much better if there are some accompanying videos and images that visually demonstrate the process or the object. Therefore, the textual answers in cQA can be significantly enhanced by adding multimedia contents, and it will provide answer seekers more comprehensive information and better experience.

In fact, users usually post URLs that link to supplementary images or videos in their textual answers. For example, for the questions in Fig. 1(c) and (d), the best answers on Y!A both contain video URLs. It further confirms that multimedia contents are useful in answering several questions. But existing cQA forums do not provide adequate support in using media information.

In this paper, we propose a novel scheme which can enrich community-contributed textual answers in cQA with appropriate media data. Fig. 2 shows the schematic illustration of the approach. It contains three main components:

(1) Answer medium selection. Given a QA pair, it predicts whether the textual answer should be enriched

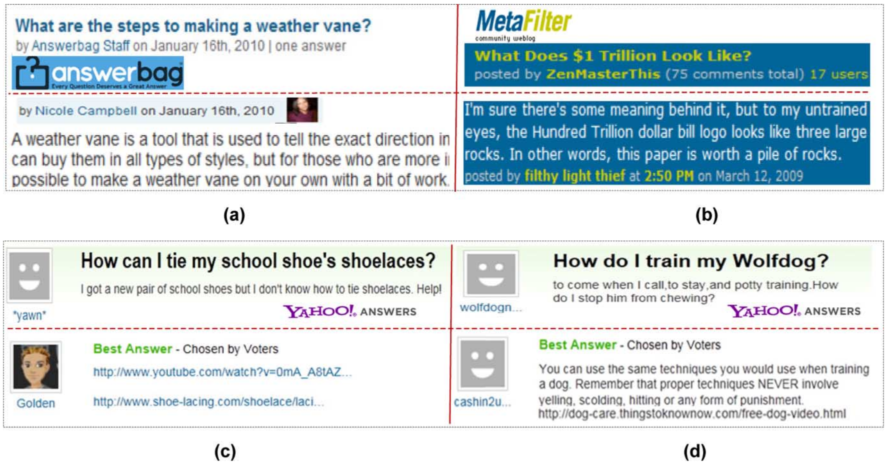
Manuscript received January 05, 2012; revised May 07, 2012; accepted July 04, 2012. Date of publication November 27, 2012; date of current version January 15, 2013. This work was supported in part by the support of "NExT Research Center" funded by MDA, Singapore, under the research grant WBS:R-252-300-001-490, and the support of the Open Project Program of the State Key Lab of CAD & CG (Grant No. A1203), Zhejiang University. The associate editor coordinating the review of this manuscript and approving it for publication was Klara Nahrstedt.

L. Nie, Y. Gao, Z.-J. Zha, and T.-S. Chua are with the School of Computing, National University of Singapore, Kent Ridge 117543, Singapore.

M. Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China (e-mail: eric.mengwang@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2012.2229971



What are the steps to making a weather vane?
by Answerbag Staff on January 16th, 2010 | one answer

by Nicole Campbell on January 16th, 2010

A weather vane is a tool that is used to tell the exact direction in
can buy them in all types of styles, but for those who are more i
possible to make a weather vane on your own with a bit of work.

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

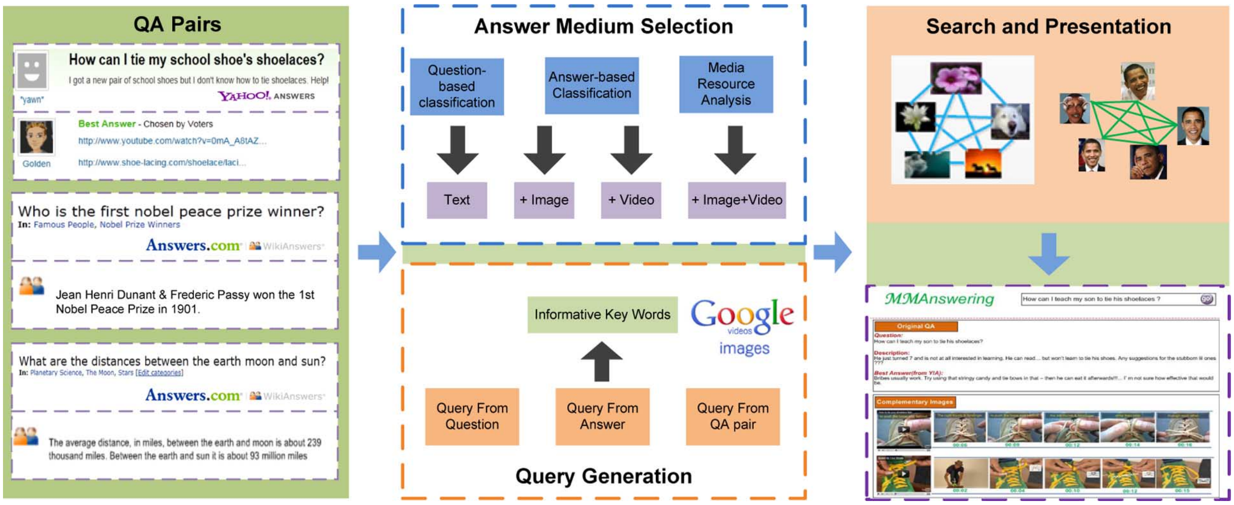
How can I tie my school shoe's shoelaces?
I got a new pair of school shoes but I don't know how to tie shoelaces. Help!
yawn

Best Answer - Chosen by Voters
http://www.youtube.com/watch?v=0mA_A8tAZ...
http://www.shoe-lacing.com/shoelace/laci...
Golden

How do I train my Wolfdog?
to come when I call,to stay,and potty training.How do I stop him from chewing?
wolfdogn...

Best Answer - Chosen by Voters
You can use the same techniques you would use when training a dog. Remember that proper techniques NEVER involve yelling, scolding, hitting or any form of punishment.
http://dog-care.thingstoknownow.com/free-dog-video.html
cashin2u...

Fig. 1. Examples of QA pairs from several popular cQA forums. (a) An example from Answerbag; (b) an example from MetaFilter; (c) an example from Yahoo! Answer that only contains links to two videos; and (d) another example from Yahoo!Answer.



QA Pairs
How can I tie my school shoe's shoelaces?
Who is the first nobel peace prize winner?
Answers.com
Jean Henri Dunant & Frederic Passy won the 1st Nobel Peace Prize in 1901.
What are the distances between the earth moon and sun?
Answers.com
The average distance, in miles, between the earth and moon is about 239 thousand miles. Between the earth and sun it is about 93 million miles
Answer Medium Selection
Question-based classification
Answer-based Classification
Media Resource Analysis
Text
+ Image
+ Video
+ Image+Video
Informative Key Words
Google videos images
Query From Question
Query From Answer
Query From QA pair
Query Generation
Search and Presentation
MMAnswering
Original QA
Complementary Images

Fig. 2. The schematic illustration of the proposed multimedia answering scheme. The scheme mainly contains three components, i.e., answer medium selection, query generation, and data selection and presentation.

with media information, and which kind of media data should be added. Specifically, we will categorize it into one of the four classes: text, $\text{text}+\text{image}$, $\text{text}+\text{video}$, and $\text{text}+\text{image}+\text{video}$.[1] It means that the scheme will automatically collect images, videos, or the combination of images and videos to enrich the original textual answers.

(2) Query generation for multimedia search. In order to collect multimedia data, we need to generate informative queries. Given a QA pair, this component extracts three queries from the question, the answer, and the QA pair, respectively. The most informative query will be selected by a three-class classification model.

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

It is worth mentioning that there already exist several research efforts dedicated to automatically answering questions with multimedia data, i.e., the so-called Multimedia Question Answering (MMQA). For example, Yang *et al.* [7] proposed a technology that supports factoid QA in news video. Yeh *et al.* [8] presented a photo-based QA system for finding information about physical objects. Li *et al.* [9] proposed an approach that leverages 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 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 contents. 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

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

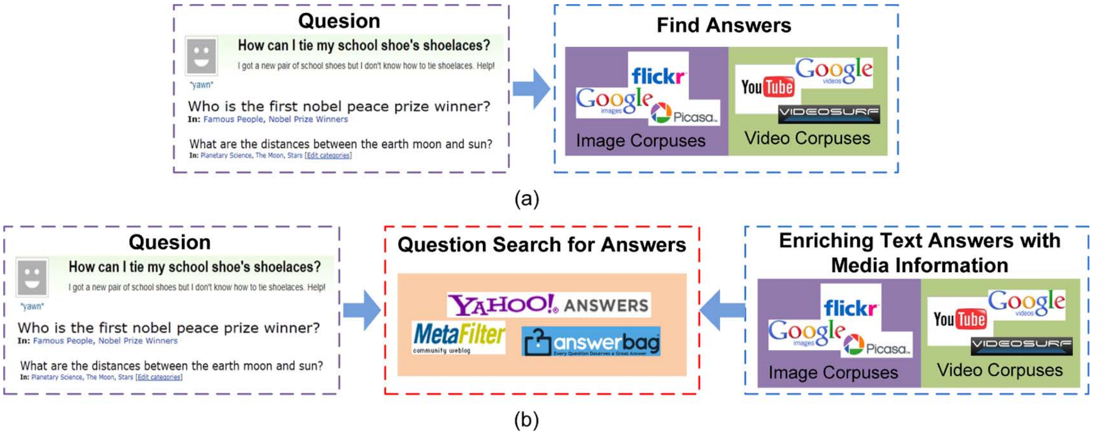

Quesion
How can I tie my school shoe's shoelaces?
I got a new pair of school shoes but I don't know how to tie shoelaces. Help!
"yawn"
Who is the first nobel peace prize winner?
In: Famous People, Nobel Prize Winners
What are the distances between the earth moon and sun?
In: Planetary Science, The Moon, Stars [Edit categories]
Find Answers
flickr
Google images
Picasa
Image Corpuses
YouTube
Google videos
VIDEOSURF
Video Corpuses
(a)
Quesion
How can I tie my school shoe's shoelaces?
I got a new pair of school shoes but I don't know how to tie shoelaces. Help!
"yawn"
Who is the first nobel peace prize winner?
In: Famous People, Nobel Prize Winners
What are the distances between the earth moon and sun?
In: Planetary Science, The Moon, Stars [Edit categories]
Question Search for Answers
YAHOO! ANSWERS
MetaFilter community weblog
answerbag
Enriching Text Answers with Media Information
flickr
Google images
Picasa
Image Corpuses
YouTube
Google videos
VIDEOSURF
Video Corpuses
(b)

Fig. 3. The differences of the conventional MMQA approaches and 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). Then, for a user-provided question, we can perform question search to find multimedia answers in the cQA corpuses.

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. Fig. 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 can manually choose several candidates for final presentation.

This framework was first explored in our previous work [10]. Compared to the preliminary version [10], we have a lot of improvements in this work. For example, for answer medium selection, we add a media resource analysis component. The results of the media resource analysis are also regarded as evidences to enable a better answer medium selection. For multimedia data selection and presentation, we propose a method that explores image search results to replace the original text analysis approach in judging whether a query is person-related or not. We introduce a new metric to measure how well the selected multimedia data can answer the questions in addition to the simple search relevance. We also investigate the cases that textual answers are absent.

The rest of the paper is organized as follows. Section II briefly reviews the related work. In Sections III and IV, we introduce the answer medium selection and query generation components, respectively. We then introduce the multimedia data selection and presentation approach in Section V. Experimental results and analysis are presented in Section VI, followed by concluding remarks in Section VII.

II. Related Work

A. From Textual QA to Multimedia QA

The early investigation of QA systems started from 1960s and mainly focused on expert systems in specific domains. Text-based QA has gained its research popularity since the establishment of a QA track in TREC in the late 1990s [11]. Based on the type of questions and expected answers, we can roughly summarize the sorts of QA into Open-Domain QA [1], Restricted-Domain QA [2], Definitional QA [3] and List QA [4]. However, in spite of the achievement as described above, automatic QA still has difficulties in answering complex questions. Along with the blooming of Web 2.0, cQA becomes an alternative approach. It is a large and diverse question-answer forum, acting as not only a corpus for sharing technical knowledge but also a place where one can seek advice and opinions [5], [6]. However, nearly all of the existing cQA systems, such as Yahoo!Answers, WikiAnswers and Ask Metafilter, only support pure text-based answers, which may not provide intuitive and sufficient information.

Some research efforts have been put on multimedia QA, which aims to answer questions using multimedia data. An early system named VideoQA was presented in [7]. This system extends the text-based QA technology to support factoid QA by leveraging the 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 [12]–[15]. Li *et al.* [16] presented a solution on "how-to" QA by leveraging community-contributed texts and videos. Kacmarcik *et al.* [17] explored a non-text input mode for QA that relies on specially annotated virtual photographs. An image-based QA approach was introduced in [8], which mainly focuses on finding information about physical objects. Chua *et al.* [18] proposed a generalized approach to extend text-based QA to multimedia QA for a range of factoid, definition and "how-to" questions. 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 I, automatic multimedia QA only works in specific domains and can hardly handle complex questions. Different from these 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 to achieve better performance.

B. Multimedia Search

Due to the increasing amount of digital information stored over the web, searching for desired information has become an essential task. The research in this area started from the early 1980s [19] by addressing the general problem of finding images from a fixed database. With the rapid development of content analysis technology in the 1990s, these efforts quickly expanded to tackle the video and audio retrieval problems [20]–[24]. Generally, multimedia search efforts can be categorized into two categories: text-based search and content-based search. The text-based search [25] approaches use textual queries, a term-based specification of the desired media entities, to search for media data by matching them with the surrounding textual descriptions. To boost the performance of text-based search, some machine learning techniques that aim to automatically annotate media entities have been proposed in the multimedia community [26]–[30]. Further, several social media websites, such as Flickr and Facebook, have emerged to accumulate manually annotated media entities by exploring the grassroot Internet users, which also facilitates the text-based search. However, user-provided text descriptions for media data are often biased towards personal perspectives and context cues, and thus there is a gap between these tags and the content of the media entities that common users are interested in. To tackle this issue, content-based media retrieval [31] performs search by analyzing the contents of media data rather than the metadata. Despite the tremendous improvement in content-based retrieval, it still has several limitations, such as high computational cost, difficulty in finding visual queries, and the large gap between low-level visual descriptions and users' semantic expectation. Therefore, keyword-based search engines are still widely used for media search. However, the intrinsic limitation of text-based approaches make that all the current commercial media search engines difficult to bridge the gap between textual queries and multimedia data, especially for verbose questions in natural languages.

C. Multimedia Search Reranking

As previously mentioned, current media search engines are usually built upon the text information associated with multimedia entities, such as their titles, ALT texts, and surrounding texts on web pages. But the text information usually does not accurately describe the content of the images and videos, and this fact can severely degrade search performance [32]. Reranking is a technique that improves 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 and the other is graph-based reranking.

The pseudo relevance feedback approach [33]–[35] regards top results as relevant samples 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 original search results. It is in contrast to relevance feedback where users explicitly provide feedback by labeling the results as relevant or irrelevant.

The graph-based reranking approach [32], [36]–[39] usually follows two assumptions. First, the disagreement between the initial ranking list and the refined ranking list should be small. Second, the ranking positions of visually similar samples should be close. Generally, this approach constructs a graph where the vertices are images or videos and the edges reflect their pairwise similarities. A graph-based learning process is then formulated based on a regularization framework.

Both of the two approaches rely on the visual similarities between media entities. Conventional methods usually measure the similarities based on a fixed set of features extracted from media entities, such as color, texture, shape and bag-of-visual-words. However, the similarity estimation actually should be query adaptive. For example, if we want to find a person, we should measure the similarities of facial features instead of the features extracted from the whole images [40]. It is reasonable as information seekers are intended to find a person rather than other objects. In this paper, we categorize queries into two classes, i.e., person-related and non-person-related, and then we use the similarities measured from different features according to the query type.

III. Answer Medium Selection

As introduced in Section I, the first component of our scheme is answer medium selection. It determines whether we need to and which type of medium we should add to enrich the textual answers. For some questions, such as "*When did America become allies with Vietnamese*", pure textual answers are sufficient. But for some other questions we need to add image or video information. For example, for the question "*Who is Pittsburghs quarterback for 2008*", it is better to add images to complement the textual answer, whereas we should add videos for answering the question "*How to install a Damper pulley on a neon*". We regard the answer medium selection as a QA classification task. That means, given a question and textual answer, we categorize it into one of the following four classes: (a) only text, which means that the original textual answers are sufficient; (b) $\mathrm{text} + \mathrm{image}$, which means that image information needs to be added; (c) $\mathrm{text} + \mathrm{video}$, which means that only video information needs to be added; and (d) $\mathrm{text} + \mathrm{image} + \mathrm{video}$, i.e., we add both image and video information.

There are some existing research efforts on question classification. Li and Roth [41] 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 to represent the questions. Zhang and Lee [42] used linear SVMs with all possible question word grams to perform question classification. Arguello *et al.* [43] investigated medium type selection as well as search sources for a query. But there is no work on classifying QA pairs according to the best type of answer medium. This task is more challenging as we are dealing with real data on the web, including 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 with two steps.

First, we analyze question, answer, and multimedia search performance. Then, we learn a linear SVM model for classification based on the results.

A. 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 at least two sentences, and the number is around $1/10$ for WikiAnswers) and some of the sentences are uninformative, we first employ the method in [44] to extract the core sentence from each question.

The classification is accomplished with two steps. First, we categorize questions based on interrogatives (some starting words and ending words), and in this way we can directly find questions that should be answered with text. Second, for the rest questions, we perform a 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 Roy Jones Jr have three kids*"), choice class (such as "*Which country is bigger, Canada or America*"), quantity class (such as "*When was the first mechanical calculator made*"), enumeration class (such as "*Name of three neighboring countries of south Korea*"), and description class (such as "*What are the ways of minimizing fan violence in sport*"). For example, a question will be categorized into the "quantity" class if the interrogative is "$\mathrm{how} + \mathrm{adj}/\mathrm{adv}$" or "when". For the "yes/no", "choice" and "quantity" questions, we categorize them into the class of answering with only text, whereas the "enumeration" and "description" questions need "$\mathrm{text} + \mathrm{image}$", "$\mathrm{text} + \mathrm{video}$" or "$\mathrm{text} + \mathrm{image} + \mathrm{video}$" 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 a classification with a Naive Bayes classifier. Table I 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.[2]

Here head word is referred to as the word specifying the object that a question seeks. The semantics of head words play an important role in determining 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 [45], but the key difference is that we do not use post fix as it better fits 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. All the phrases that have the frequencies above a threshold (we empirically set the threshold to 3 in this work) are collected. We then manually refine the list based on human's expert knowledge. Examples of class-specific related words for each class are shown in Table II.

B. Answer-Based Classification

Besides question, answer 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 answer describes a dynamic process.

For answer classification, we extract bigram text features and verbs.[3] The verbs in an answer will be useful for judging whether the answer can be enriched 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 also build a Naive Bayes classifier with a set of training data, and then perform a four-class classification with the model.

C. Media Resource Analysis

Even after determining an appropriate answer medium, the related resource may be limited on the web or can hardly be collected, and in this case we may need to turn to other medium types. For example, for the question "*How do I export Internet Explorer browser history*", it is intuitive that it should be answered using video content, but in fact video resources related to this topic on the web are hard to find on the current search engines. Therefore, it will be beneficial to take into account the search performance of different medium types. Here we only introduce our method for search performance prediction, whereas the query generation and the used search engines will be introduced in Section IV and Section VI, respectively.

We predict search performance based on the fact that, most frequently, search results are good if the top results are quite coherent [46]. We adopt the method proposed in [46], which defines a clarity score for a query[4] based on the relative entropy (or Kullback-Leibler (KL) divergence) between the query and collection language models, i.e.,

$$Clarity_q(C_i) = \sum_{w \in V_{ci}} P(w|\theta_q) log_2 \frac{P(w|\theta_q)}{P(w|\theta_{C_i})} \qquad (1)$$

where V_{ci} is the entire vocabulary of the collection C_i, and $i = 1, 2, 3$ represent text, image and video, respectively.[5] The terms $P(w|\theta_q)$ and $P(w|\theta_{C_i})$ are the query and collection language models, respectively. The Clarity value becomes smaller as the top ranked documents approach a random sample from the collection (i.e., an ineffective retrieval). The query language

[2] 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.

[3] 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] The query is generated from a given QA pair. It will be introduced in detail in Section IV.

[5] Because KL divergence only works with single medium between query and collection, we crawled the surrounding text such as titles, tags, descriptions and comments for images and videos.

TABLE I
REPRESENTATIVE INTERROGATIVE WORDS

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 II
REPRESENTATIVE CLASS-SPECIFIC RELATED WORDS

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+Image+Video	president, king, prime minister, kill, issue, nuclear, earthquake, singer, battle, event, war, happened, etc.

model is estimated from the top documents, $\mathcal{R}$, as the following formula,

$$P(w|\theta_q) = \frac{1}{\mathcal{Z}} \sum_{d \in \mathcal{R}} P(w|D) P(q|D) \tag{2}$$

and $\mathcal{Z}$ is defined as,

$$\mathcal{Z} = \sum_{D \in \mathcal{R}} P(q|D) \tag{3}$$

where $P(q|D)$ is the query likelihood score of document D. We apply the method in [47] to calculate,

$$P(q|D) = \prod_{w \in q} P(w|D). \tag{4}$$

In this work, for a query generated from a given QA pair, we use up to 20 top documents (for several complex queries, there may be less than 20 results returned) to estimate the retrieval effectiveness for each medium type, including text, image and video. Since the search performance prediction measures are used for a kind of source selection (i.e., answer medium selection), the number of used documents is not quite sensitive. It may impact prediction results, but the result of answer medium selection will not be sensitive to the number.

D. Medium Selection Based on Multiple Evidences

We perform medium selection by learning a four-class classification model based on the results of question-based classification, answer-based classification, and media resource analysis. For question-based classification, we have four scores, i.e., the confidence scores that the question should be answered by "text", "$\text{text} + \text{image}$", "$\text{text} + \text{video}$", and "$\text{text} + \text{image} + \text{video}$". Similarly, for answer-based classification we also have four scores. For media resource analysis, we have three scores, which are the search performance prediction results for text, image and video search, respectively. We regard these scores as 11-D features and thus we can learn a four-class classification model based on a training set. Here we adopt SVM with linear kernel. The training set will be introduced in Section VI.

IV. QUERY GENERATION FOR MULTIMEDIA SEARCH

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. Textual questions and answers are usually complex sentences. But frequently search engines do not work well for queries that are long and verbose [48]. 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 QA pairs. For example, some QA pairs embed the useful query terms in their questions, such as "*What did the Globe Theater look like*". Some hide the helpful keywords in their answers, such as the QA pair "*Q: What is the best computer for 3D art; A: Alienware brand computer*". Some should combine the question and the answer to generate a useful query, such as the QA pair "*Q: Who is Chen Ning Yang's wife; A: Fan Weng*", for which both "Chen Ning Yang" and "Fan Weng" are informative words (we can find some pictures of the couple, and only using "Fan Weng" to search will yield a lot of incorrect results).

For each QA pair, we generate three queries. First, we convert the question to a query, i.e., we convert a grammatically correct interrogative sentence into one of the syntactically correct declarative sentences or meaningful phrases. We employ the method in [49]. Second, we identify several key concepts from verbose answer which will have the major impact on effectiveness. Here we employ the method in [50]. Finally, we combine the two queries that are generated from the question and the answer respectively. Therefore, we obtain three queries, and the next step is to select 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. POS histogram reflects the characteristic of a query. Using POS histogram for query selection is motivated by several observations. For example, for the queries that contain a lot of complex verbs it will be difficult to retrieve meaningful multimedia results. 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.[6] We then generate a 36-dimensional histogram, in which each bin counts the number of words belonging to the corresponding category of part-of-speech.

[6] 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.

(2) Search performance prediction. This is because, for certain queries, existing image and video search engines cannot return satisfactory results. We adopt the method introduced in Section III-C, which measures a clarity score for each query based on the KL divergence between the query and collection language models. We can generate 6-dimensional search performance prediction features in all (note that there are three queries and search is performed on both image and video search engines).

Therefore, for each QA pair, we can generate 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.

V. 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, as mentioned above, most of the current commercial search engines are built upon text-based indexing and usually return a lot of irrelevant results. Therefore, reranking by exploring visual information is essential to reorder the initial text-based search results. Here we adopt the graph-based reranking method in [37]. We re-state the equation from [37] as,

$$r^{j}_{(k)} = \alpha \sum_{i \in B_j} r^{i}_{(k-1)} P_{ij} + (1-\alpha) r^{j}_{(0)} \tag{5}$$

where $r^{j}_{(k)}$ stands for the state probability of node j in the kth round of iterations, α is a parameter that satisfies $0 \leq \alpha < 1$, and P_{ij} is the transition probability from data-point i to j. Here $\mathbf{P}$ is a row-normalized transition matrix obtained from similarity matrix $\mathbf{W}$, and $r^{j}_{(0)}$ is the initial relevance score of the sample at the jth position, which is heuristically estimated as

$$r^{j}_{(0)} = \frac{N-i}{N} \qquad i = 1, 2 \ldots N. \tag{6}$$

For images, each element of the symmetric similarity matrix $\mathbf{W}$ is measured based on K-nearest-neighbor (K-NN) graph,

$$W_{ij} = \begin{cases} \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\sigma^2}) & \text{if } j \in \mathcal{N}_K(i) \text{ or } i \in \mathcal{N}_K(j) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{N}_K(i)$ denotes the index set for the K nearest neighbors of an image computed by Euclidean distance. In our work, we empirically set $K = 0.3 \times N$, where N is the number of images collected for each query. The parameter σ is simply set to the median value 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 [51]. Considering two videos $(\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,m})$ and $(\mathbf{v}_{j,1}, \ldots, \mathbf{v}_{j,n})$, which contain m and n key-frames respectively, we employ average distance [52] of all cross-video key-frame pairs for similarity estimation, i.e.,

$$W_{ij} = \begin{cases} \exp(-\frac{\sum_{q=1}^{m} \sum_{p=1}^{n} ||\mathbf{v}_{i,q} - \mathbf{v}_{j,p}||^2}{MN\sigma^2}) & \text{if } j \in N_K(i) \text{ or} \\ & i \in N_K(j) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Similarly, $N_K(i)$ denotes the index set for the K nearest neighbors of a video measured by Euclidean distance and the parameter σ is simply set to the median of the Euclidean distance of all video pairs.

However, a problem of the existing reranking methods is that they usually use query-independent global visual features for reranking. It overlooks the fact that many queries are actually person-related. As we mentioned in Section II-C, it is more reasonable to use facial features instead of global visual features for reranking the search results of person-related queries. For question-answering, our statistics show that around $1/4$ of the QA pairs in our data set are about person. Therefore, in this work we propose a query-adaptive reranking approach. We first decide whether a query is person-related or non-person-related, and then we use different features for reranking. Fig. 4 shows our approach.

Here we regard the prediction of whether a query is person-related as a classification task. A heuristic text-based rule, analyzing the textual QA information, has been proposed in [10]. But it is not easy to accomplish the task by simply analyzing the textual terms. For example, for the query "BSB", it is not easy to judge that it is the abbreviation of "Backstreet Boys" which is person-related. But from the image search results, we can find that most returned images contain several faces and thus we can determine it is a person-related query. Also, we can choose to match each query term with a person list, such as a celebrity list. But it will not be easy to find a complete list. In addition, it will be difficult to keep the list updated in time. Therefore, we adopt a method that analyzes image search results. Specifically, for each image in the ranking list, we perform face detection and then extract 7-dimensional features, including the size of the largest face area, the number of faces, the ratio of the largest face size and the second largest face size, and the position of the largest face (the position is described by the up-left and bottom-right points of the bounding box and thus there are 4-dimensional features). We average the 7-dimensional features of the top 150 search results and it forms the features for query classification. We learn a classification model based on the training queries and it is used to discriminate person-related and non-person-related queries.

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-D Local Binary Pattern features [53] from the largest faces of images or video frames. For non-person-related 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. If an image or video is close to a sample that appears above it, we remove it. More specifically, we remove the ith image or video if there exists $j < i$ that satisfies $W_{ij} > T$. Here we empirically set T to 0.8 throughout the work. After

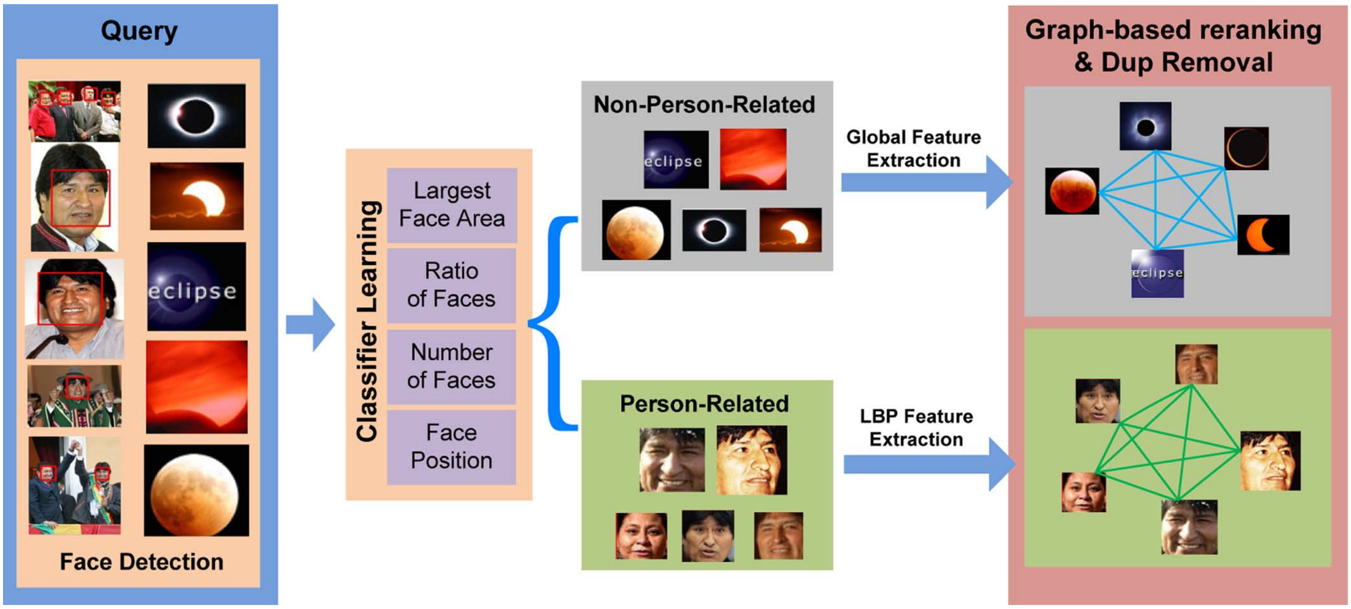


Fig. 4. The schematic illustration of the query-adaptive reranking approach. Given a query, we perform query classification based on the information clues in the image search results to decide whether the query is person-related. We then employ different image/key-frame representations according to the query classification result and perform graph-based reranking accordingly.

duplicate removal, we keep the top 10 images and top 2 videos (keeping which kind of media data depends 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 to easily browse the videos.

VI. Experiments

In this section, we introduce the empirical evaluation of the proposed scheme. We first introduce the experimental settings, such as dataset and ground truth labeling. Then, we present two kinds of evaluation. One is local evaluation which tests the effectiveness of the components in the scheme, such as answer medium selection, query selection, and multimedia search reranking. The other one is global evaluation which tests the usefulness of the enrichment of media data for question answering.

A. Experimental Settings

Our dataset for experiments contains two subsets. For the first subset, we randomly collect 5,000 questions and their corresponding answers from WikiAnswers. For the second subset, we randomly collect 5,000 questions and their best answers from the dataset used in [54], 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.[7] Inspired by [57], [58], we first classify all the questions into two categories: conversational and informational. Conversational questions usually only seek personal opinions or judgments, such as "*Anybody watch the Bears game last night*", and informational questions are asked with the intent of getting information that the askers hopes to learn or use via fact-oriented answers, such as "*What is the population of Singapore*". There are several automatic algorithms for the categorization of conversational and informational questions, such as the work in [57] and [58]. But since it is not the focus of our work, we perform the categorization with human labeling. To be specific, each question is labeled by at least two volunteers independently. In the case that the first two volunteers have different decisions about the question type, we solicit two additional volunteers to label this question again. The question will be viewed as ambiguous if the four voters cannot come to a majority classification. It is worth noting that each volunteer was trained with the question type definition as well as corresponding examples before labeling. This question type labeling process is analogous to [58]. In this way, we extract 3,333 informational questions from the Y!A subset and 4,000 from the WikiAnswers set. The QA pairs in our dataset cover a wide range of topics, including travel, life, education, etc.

The answer medium selection and query selection components need to learn classifiers based on several training data, and thus we split the 7,333 QA pairs into two parts, a training set that contains 5,866 QA pairs and a testing set of the remaining 1,467 QA pairs. The testing set consists of 800 QA pairs from WikiAnswers and 667 from Y!A. Classification models are trained with the whole training set, i.e., 5,866 QA pairs. They are tested on the 800 QA pairs from WikiAnswers, 667 QA pairs from Yahoo!Answers, or the both.

For ground truth labeling (including the ground truths for answer medium selection, query generation, and the relevance of media data), the five volunteers that have been involved in the labeling task of [10] were involved again, including two Ph.D. students and one faculty in computer science, one master student in information system, and one software engineer. The labelers

[7] There are also many research efforts on ranking community-contributed answers or selecting the best answer by machine learning and NLP technologies [54]–[56]. These methods can also been integrated with our work and we only need to change the best answer for each question.

TABLE III
THE INTER-RATER RELIABILITY ANALYSIS FOR ANSWER MEDIUM SELECTION BASED ON THE WHOLE TESTING DATASET. THE FOUR CATEGORIES ARE "TEXT", "text + image", "text + video" AND "text + image + video", RESPECTIVELY

No of Cases	No of Categories	No of Raters	Percent of Overall Agreement	Fixed-marginal Kappa
1467	4	5	0.8231	0.7445

TABLE IV
THE DISTRIBUTION OF THE EXPECTED ANSWER MEDIUM TYPES LABELED BY HUMANS

DataSet Categories	Y!A	WikiAnswers	Both
Text	45%	48%	46.6%
Text+Image	24%	21%	22.4%
Text+Video	22%	24%	23.1%
Text+Image+Video	9%	7%	7.9%

TABLE V
THE ACCURACY COMPARISON OF QUESTION-BASED CLASSIFICATION WITH DIFFERENT FEATURES. HERE "RELATED" MEANS CLASS-SPECIFIC RELATED WORDS

Testing Set Features	Y!A	WikiAnswers	Both
Bigram	71.32%	75.89%	73.81%
Bigram+Head	75.27%	78.72%	77.15%
Bigram+Related	73.59%	76.97%	75.60%
Bigram+Head+Related	**76.41**%	**80.62**%	**78.71**%

TABLE VI
THE ACCURACY COMPARISON OF ANSWER-BASED CLASSIFICATION WITH DIFFERENT FEATURES

Testing Set Features	Y!A	WikiAnswers	Both
Bigram	57.38%	61.31%	59.52%
Bigram+Verb	**59.86**%	**64.72**%	**62.51**%

TABLE VII
THE CLASSIFICATION ACCURACY COMPARISON AMONG THE PROPOSED APPROACH "INTEGRATION OF MULTIPLE EVIDENCES" WITH THE QUESTION-BASED CLASSIFICATION AND ANSWER-BASED CLASSIFICATION METHODS

Testing Set Features	Y!A	WikiAnswers	Both
Question-based classification	76.41%	80.62%	78.71%
Answer-based classification	59.86%	64.72%	62.51%
Integration of multiple evidences	**81.72**%	**84.97**%	**83.49**%

are trained with a short tutorial and a set of typical examples. We need to admit that the ground truth labeling is subjective. But a majority voting among the five labelers can partially alleviate the problem. We have also analyzed the inter-rate relability of the labeling tasks with the fixed-marginal kappa method in [59], and the results demonstrate the there are sufficient inter-rater agreements. As an example, we illustrate the labeling analysis results on the answering medium selection ground truths of the 1,467 testing points in Table III. The Kappa value is greater than 0.7, and it indicates a sufficient inter-rater agreement.

B. Evaluation of Answer Medium Selection

We first evaluate our answer medium selection approach. As previously mentioned, there are five labelers involved in the ground truth labeling process. It is worth noting that they will not only consider which type of medium information is useful but also investigate web information. For the example, for the question "*How can I extract the juice from sugar cane, at home*", video-based answer is expected. But after the labelers' practical investigation on the web, they may find that there are insufficient image and video resources related to this topic. Therefore, they would label this question as "text". Table IV illustrates the distribution of the four classes. We can see that, more than 50% of the questions can be better 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 V and VI, respectively. It is worth noting that the stop-words are not removed for question-based classification since some stop-words also play an important role in question classification. But for answer-based classification, stop words are removed. Stemming is performed for both questions and answers. From the results, it is observed that for both of the two classifiers, integrating all of the introduced features is better than using only part of them. Also the performances based on WikiAnswers outperform those on Y!A. This may be attributed to the more spelling mistakes, slang and abbreviations in WikiAnswers.

Table VII illustrates the results of question-based classification, answer-based classification, and the integration of multiple evidences. As introduced in Section III-D, the integration of multiple evidences actually learns a linear SVM based on the results of question-based classification and answer-based classification as well as the search performance prediction results. As shown in Table VII, the integration of multiple evidences achieves better results than classification based on purely questions or answers. The accuracy for answer medium selection is 83.49% on the whole testing dataset. Table VIII presents the questions classified with highest confidence scores for each category after classification.

C. Evaluation of Query Generation

Now we evaluate the query generation and selection approach. For each QA pair, three queries are generated from the question, answer and the combination of question and answer. As previously mentioned, five labelers participate in the ground truth labeling process. Each labeler selects the most informative one. They are allowed to perform search on the web to compare the informativeness of search results. The final ground truths are obtained by a majority voting. The distribution of the three classes is illustrated in Table IX.

We adopt SVM with RBF kernel, and the parameters, including the radius parameter and the weighting parameter that modulates the regularize term and the loss term, are established by 5-fold cross-validation. Table X illustrates the classification results. From the results we can see that integrating POS histogram and search performance prediction can achieve better

TABLE VIII
THE REPRESENTATIVE QUESTIONS FOR EACH ANSWER MEDIUM CLASS. HERE WE DO NOT ILLUSTRATE THE ANSWERS BECAUSE SEVERAL ANSWERS ARE FAIRLY LONG. THE CORRECTLY CATEGORIZED QUESTIONS ARE MARKED WITH "√"

Text
1. How many years was the US involved in the Vietnam War? (√)
2. When were telescopes first made? (√)
3. What year was the movie Mustang Sally made? (√)
4. what is speed limit on on california freeways? (√)
5. What is the distance between the moon and the earth? (√)
6. What is the conversion rate from British sterling pounds to the US Dollar????? (√)

Text+Image
1. Who is the final commander of the union army? (√)
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. What is mewtwos gender ?
6. 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. How to get the fire stone in pokemon emrald? (√)
6. Exactly what steps do I take to get more space in my mail box? (√)

Text+Image+Video
1. What exercises are best for tightening the muscles in the vagina? (√)
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. What is the sd card slot used for? (√)
6. What do people view on Saint Patrick's day? (√)

TABLE IX
THE GROUND TRUTH DISTRIBUTION FOR QUERY SELECTION

DataSet Categories	Y!A	WikiAnswers	Both
Answer	29%	25.5%	27%
Question	49%	53.5%	51.5%
Combination	22%	21%	21.5%

TABLE X
THE CLASSIFICATION ACCURACIES FOR QUERY SELECTION WITH DIFFERENT FEATURES. SPP STANDS FOR SEARCH PERFORMANCE PREDICTION

Testing Set Features	Y!A	WikiAnswers	Both
POS Histogram	66.89%	69.47%	68.30%
SPP	59.57%	63.34%	61.63%
POS Histogram+SPP	**72.19**%	**76.38**%	**74.47**%

performance than using merely POS histogram and retrieval performance prediction. The classification accuracies on the Y!A subset, WikiAnswers subset and the whole dataset are 72.19%, 76.38% and 74.47%, respectively. The QA pairs in WikiAnswers are well semantically and syntactically presented, which results in its higher accuracy in query generation and selection.

D. Evaluation of Reranking

To evaluate the method of judging whether a QA pair is person-related or non-person-related, we select 500 QA pairs from each dataset randomly. Table XI illustrates the statistics of the person-related and non-person-related classes based on these two subsets. Then, we learn an SVM model with RBF kernel based on 7-dimensional facial characteristics. The parameters are turned by 10-fold cross-validation. The results are illustrated in Table XII. We can see that our approach achieves fairly good performance. In the table, we also illustrate the performance of the method in [10] for comparison. The method in [10] mainly relies on the analysis of the text content of questions and answers, and thus we denote it as "text-based method".

TABLE XI
THE GROUND TRUTH DISTRIBUTION OF PERSON-RELATED AND NON-PERSON-RELATED QA PAIRS

DataSet Categories	Y!A	WikiAnswers	Both
Person-Related	26.7%	24.3%	25.4%
Non-Person-Related	73.3%	75.7%	74.6%

TABLE XII
THE CLASSIFICATION ACCURACY OF PERSON-RELATED AND NON-PERSON-RELATED QUERIES

Testing Set Method	Y!A	WikiAnswers	Both
Text-based method in [10]	82.17%	85.26%	83.86%
Our approach	**92.60**%	**95.30**%	**94.07**%

In order to evaluate our query-adaptive 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

$$NDCG@n = \frac{DCG}{IDCG} = \frac{(rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i})}{IDCG} \tag{9}$$

where rel_i is the relevance score of ith 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.

Figs. 5 and 6 illustrate the average performance comparison of our approach and the conventional method that uses only global features for the 25 person-related queries. Here we illustrate the performance with different values of the parameter α. Smaller α means more initial text-based ranking information is taken into consideration. We can see that, our approach consistently outperforms the method that uses global features. This demonstrates that, in image or video reranking, it is more reasonable to use facial features for person-related queries.

We then randomly select 100 queries from image and video class, respectively. We compare the following methods by conducting experiments on these queries:

(1) The conventional method that only uses global features. It is denoted as "conventional".
(2) Query-adaptive reranking with the text-based query classification strategy in [10]. That means, we use the method in [10] to perform query classification and

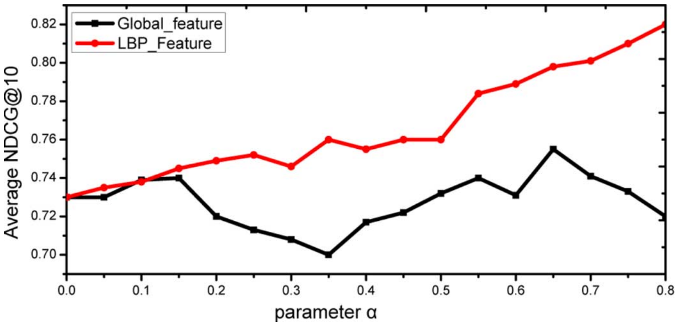


Fig. 5. The image search reranking performance comparison of using global features and LBP features for person-related queries.

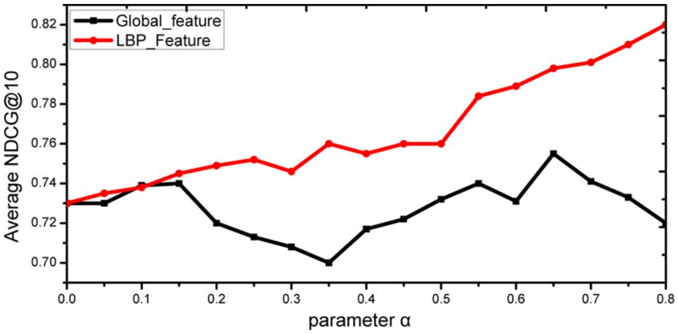


Fig. 6. The video search reranking performance comparison of using global features and LBP features for person-related queries.

then adopt query-adaptive reranking accordingly. It is denoted as “text-based query-adaptive”.

(3) Query-adaptive reranking with the proposed query classification strategy. That means, we use the proposed facial analysis method to perform query classification and then adopt query-adaptive reranking accordingly. It is our proposed approach and it is denoted as “proposed”.

Figs. 7 and 8 illustrate the average performance comparison. From the results we can see that the two query-adaptive methods consistently outperform the conventional method that uses only global features. The proposed approach performs better than the “text-based query-adaptive” method due to the more accurate classification of person-related and non-person-related queries. Throughout our rest experiments, we set α as 0.65 and 0.8 for image and video reranking, respectively.

It is worth noting that, when α equals 0, the reranked list will be the same to the initial ranking list obtained by the text-based search. We can see that the performance of text search results are not as well as the visually refined results, which implicitly substantiates our previous assertion in Section II-C.

We illustrate the top results before and after reranking for two example queries in Figs. 9 and 10, respectively, one about object and the other about person. Fig. 9(a) shows the top 7 results before reranking about “bird of prey”, while (b) shows the top 7 results after reranking based on global features. Similarly, Fig. 10(a) and (b) present the top results before and after reranking with respect to person-related query “CEO of Microsoft”. We can see that several less relevant results, marked with “X”, are removed from top positions after reranking.

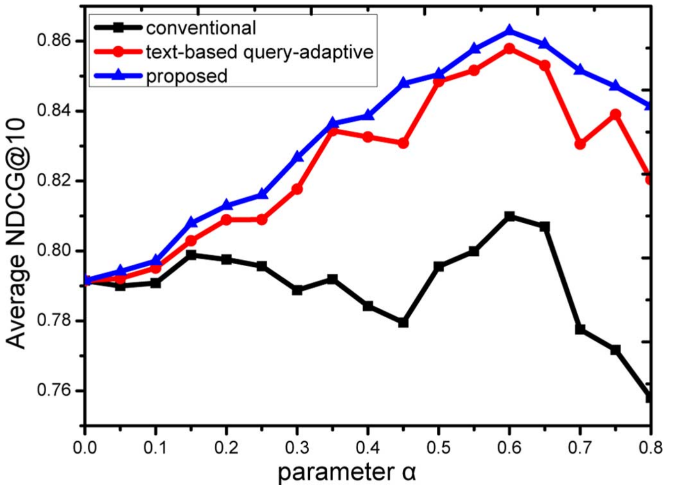


Fig. 7. The image search reranking performance comparison among different methods. It is observed that our proposed approach achieves the best performance consistently.

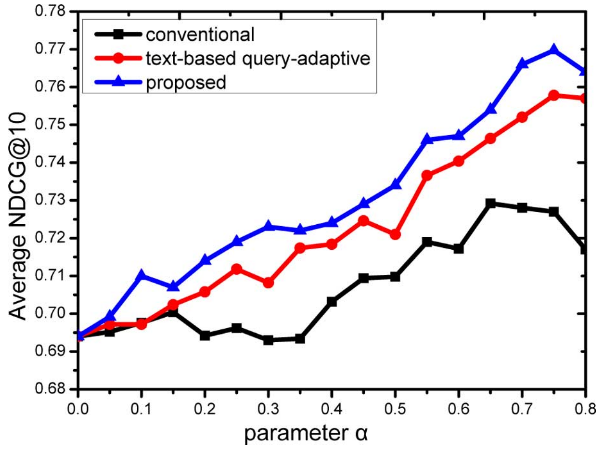


Fig. 8. The video search reranking performance comparison among different methods. It is observed that our proposed approach achieves the best performance consistently.

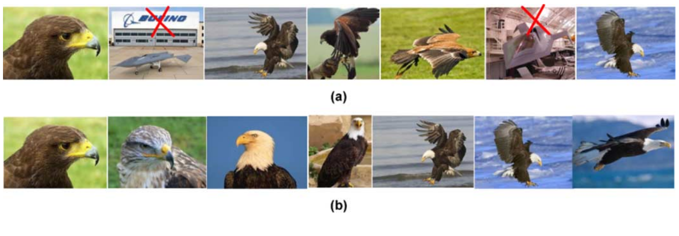


Fig. 9. The non-person-related image search reranking results for the query “bird of prey”. (a) The top images before reranking; (b) the top images after reranking.

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. Fig. 11 shows the multimedia answers for 3 example queries.

E. On the Informativeness of Enriched Media Data

In this paper, all the complementary media data are collected based on textual queries, which are extracted from QA pairs and maybe somewhat biased away from the original meanings. In

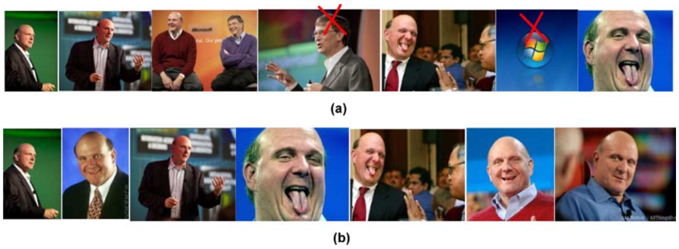
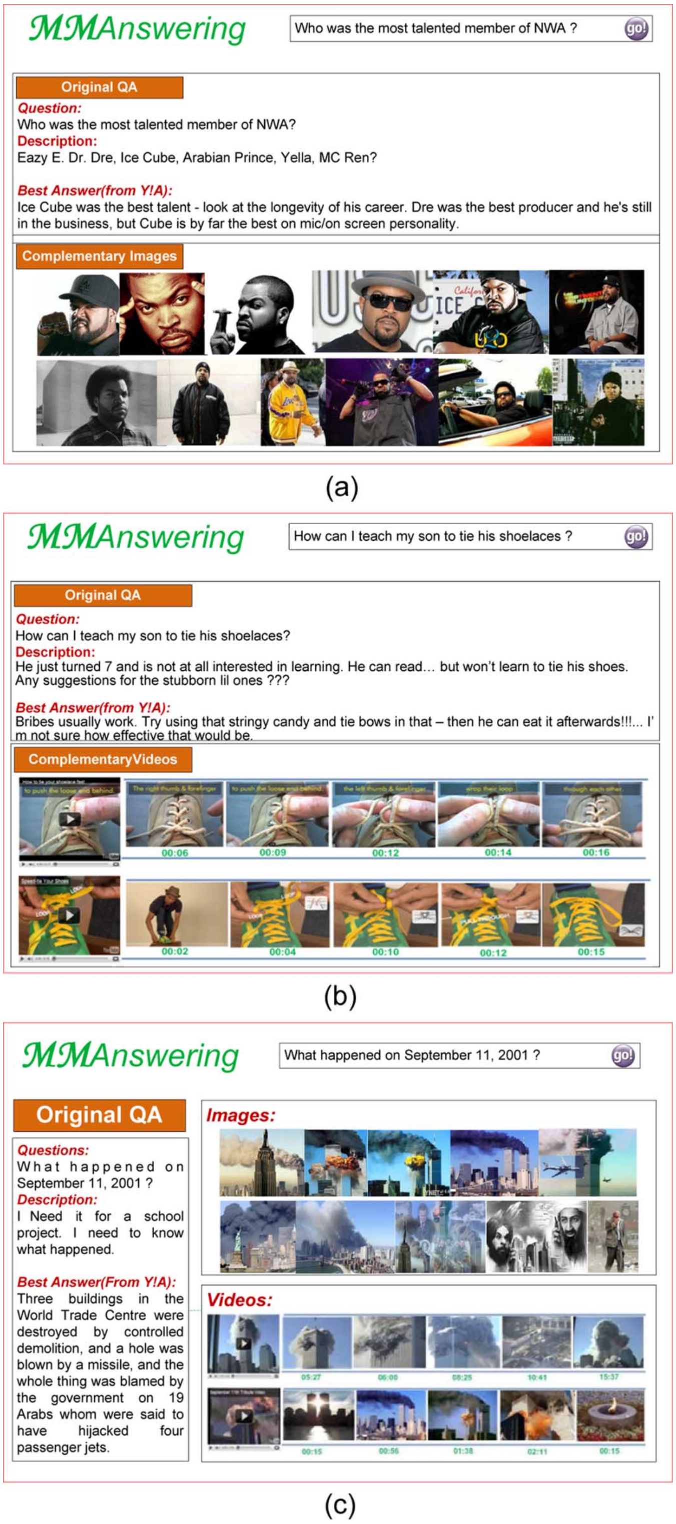


Fig. 10. The person-related image search reranking results for the query "CEO of Microsoft". (a) The top images before reranking; (b) the top images after reranking.

other words, the queries do not always reflect the original QA pairs' intention. In our above evaluation, NDCG is used to measure the relevance of the ranked images/videos to the generated query. However, it cannot reflect how well these media data answer the original questions or enrich textual answers due to the fact that there is a gap between a QA pair and the generated query. So, in addition to evaluating search relevance, we further define an informativeness measure to estimate how informative the media data can answer a question. Specifically, there are three score candidates, i.e., 2, 1 and 0. The three scores indicate that the media sample can perfectly, partially and cannot answer the question, respectively. We randomly select 300 QA pairs that have enriched media data for evaluation. For each QA pair, we manually label the informativeness score of each enriched image or video by the previously introduced five labelers. Fig. 12 illustrates the distribution of the informativeness scores. The results actually indicate that, for at least 79.57% questions, there exist enriched media data that can well answer the questions. The average rating score is 1.248.

F. Subjective Test of Multimedia Answering

We first conduct a user study from the system level. 20 volunteers that frequently use Y!A or WikiAnswers are invited. They are from multiple countries and their ages vary from 22 to 31. They do not know the researchers and also get no knowledge about which method developed by the researcher. Each user is asked to freely browse the conventional textual answers and our multimedia answers for different questions they are interested in (that means, they are information seekers in this process). Then, they can provide their ratings of the two 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. They are trained with the rules before coding: if the enriched media data are fairly irrelevant to the contextual content, users should assign 1 to our scheme, because users are distractive rather than obtaining valuable visual information; Otherwise, these volunteers should assign 1 to the original system. The average rating scores and the standard deviation values are illustrated in Table XIII. From the results we can clearly see the preference of users towards the multimedia answering. We also perform a two-way ANOVA test and the results are illustrated in the right part of Table XIII. The p-values demonstrate that the superiority of our system is statistically significant, but the difference among users is statistically insignificant.



MMAnswering

Who was the most talented member of NWA ? go!

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)

MMAnswering

How can I teach my son to tie his shoelaces ? go!

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)

MMAnswering

What happened on September 11, 2001 ? go!

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)

Fig. 11. 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 "$\text{text}+\text{image}$", "$\text{text}+\text{video}$", and "$\text{text}+\text{image}+\text{video}$", respectively.

We then conduct 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 XIV presents the statistical results. From the left part of

TABLE XIII
THE LEFT PART ILLUSTRATES THE AVERAGE RATING SCORES AND STANDARD DEVIATION VALUES COMPARISON OF TEXTUAL QA BEFORE AND AFTER MEDIA DATA ENRICHMENT. THE RIGHT PART ILLUSTRATES THE ANOVA TEST RESULTS

Average Rating Scores and Variance		The Factor of Schemes		The Factor of Uses	
MMQA	Textual cQA	F-statistic	p-value	F-statistic	p-value
$\mathbf{2.25 \pm 0.6184}$	1.15 ± 0.1342	21.09	2×10^{-4}	0.31	0.9927

TABLE XIV
STATISTICS OF THE COMPARISON OF OUR MULTIMEDIA ANSWER AND THE ORIGINAL TEXTUAL ANSWER. THE RESULTS OF LEFT PART ARE BASED ON THE WHOLE TESTING SET. WHILE THE RIGHT PART STATISTICS ARE CONDUCTED WITH EXCLUSION OF QUESTIONS WHERE ONLY TEXTUAL-BASED ANSWERS ARE SUFFICIENT

Including the questions for which textual answers are sufficient			Excluding the questions for which textual answers are sufficient		
Prefer multimedia answer	Neutral	Prefer Original textual answer	Prefer multimedia answer	Neutral	Prefer Original textual answer
47.99%	**49.01%**	3.0%	**88.66%**	6.17%	5.54%

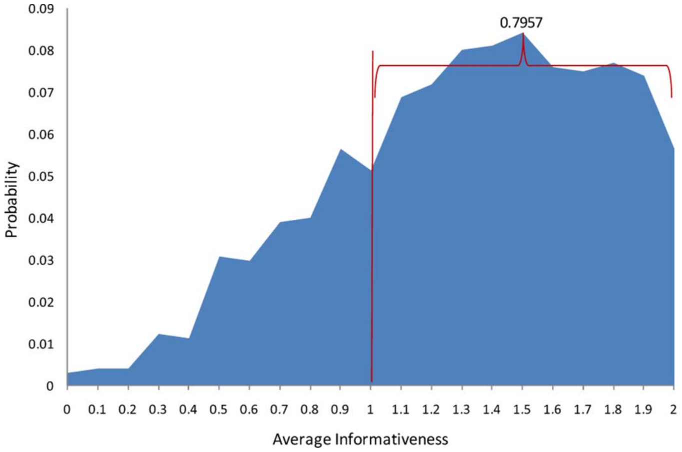


Fig. 12. The distribution of informativeness. It is observed that for more than 79% of the questions the average informativeness scores of multimedia answers are above 1.

this table, it is observed that, in about 47.99% of the cases users prefer our answer and only in 3.0% of the cases they prefer the original answers. But there are 49.01% 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 the right part of Table XIV. We can see that for more than 88.66% 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 contents.

G. On the Absence of Textual Answer

In our proposed scheme, the existing community-contributed textual answers play an important role in question understanding. So, here a question is that whether the scheme can deal with and how it will perform when there is no textual answer. For example, there may exist newly added questions that do not have textual answers yet or not well answered in cQA forums.

From the introduction of the proposed scheme in Sections III–V, we can see that it can easily deal with the cases that there is no textual answer. Actually, we only need to remove the information clues from textual answers in the answer medium selection and multimedia query generation components. Here we further investigate the performance of the scheme without textual answers.

TABLE XV
THE CLASSIFICATION ACCURACY OF ANSWER MEDIUM SELECTION COMPARISON BETWEEN WITH AND WITHOUT TEXTUAL ANSWERS

Testing Set Method	Y!A	WikiAnswers	Both
With Textual Answers	**81.72%**	**84.97%**	**83.49%**
Without Textual Answers	78.30%	82.01%	80.32%

We first observe answer medium selection. When there is no textual answer, there will only be 7-D features for classification in the integration of multiple evidences (see Section III-D). We compare the performance of answer medium selection with and without textual answers. Table XV illustrates the results, it can be observed that, without textual answers, the classification accuracy will degrade by more than 3% for answer medium selection.

When it comes to query generation, only one query will be generated from the question if there is no textual answer. So, we can directly skip the query selection step. Based on the 300 QA pairs mentioned in Section VI-E, we compare the informativeness of the obtained media data with and without using textual answers. Fig. 13 illustrates the comparison of the overall average informativeness scores. We can see that without textual answers, the score will degrade from 1.248 to 1.066. This is because, without textual answers, the generated multimedia queries cannot well reflect the question intentions in many cases. As mentioned in Section I, the textual answers can partially bridge the gap between questions and multimedia answers. Note that the approach without textual answer can be regarded as a conventional MMQA approach which tries to directly find multimedia answers based on questions. Here the results have demonstrated our approach built upon textual answers is more effective.

Finally, we conduct a user study to compare the original textual answers and the media answers generated without the assistance of textual answers. We adopt the experimental settings introduced in Section VI-F and present the user study results in Table XVI. It is interesting to see that, although the media

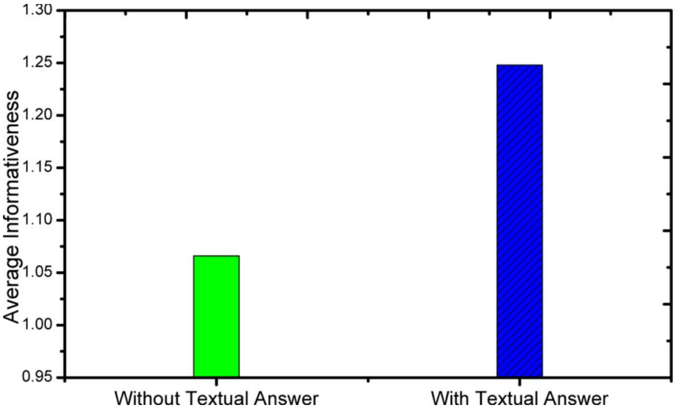


Fig. 13. Comparison of overall average informativeness scores between with and without textual answers.

TABLE XVI
STATISTICS OF THE COMPARISON OF OUR GENERATED MULTIMEDIA ANSWERS WITHOUT THE ASSISTANCE OF TEXTUAL ANSWERS AND WITH THE ORIGINAL TEXTUAL ANSWERS

Prefer pure media answer answer	Neutral	Prefer Original textual answer
50.9%	15.2%	33.9%

answers are not as informative as those generated with the assistance of textual answers, they are still very informative in comparison with pure textual answers.

Therefore, we can draw several conclusions from the investigation. First, there will be informativeness degradation for the obtained media data if there is no textual answer. Second, the performance of answer medium selection will also degrade. Third, the obtained media answers can still be useful for many questions.

VII. CONCLUSION AND FUTURE WORK

In this paper, we describe the motivation and evolution of MMQA, and it is analyzed that the existing approaches mainly focus on narrow domains. Aiming at a more general approach, we propose a novel scheme to answer questions using media data by leveraging textual answers in cQA. For a given QA pair, our scheme first predicts which type of medium is appropriate for enriching the original textual answer. Following that, it automatically generates a query based on the QA knowledge and then performs multimedia search with the query. Finally, query-adaptive reranking and duplicate removal are performed to obtain a set of images and videos for presentation along with the original textual answer. 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.

In our study, we have also observed several failure cases. For example, the system may fail to generate reasonable multimedia answers if the generated queries are verbose and complex. For several questions videos are enriched, but actually only parts of them are informative. Then, presenting the whole videos can be misleading. Another problem is the lack of diversity of the generated media data. We have adopted a method to remove duplicates, but in many cases more diverse results may be better. In our future work, we will further improve the scheme, such as developing better query generation method and investigating the relevant segments from a video. We will also investigate multimedia search diversification methods, such as the approach in [36], to make the enriched media data more diverse.

REFERENCES

[1] S. A. Quarteroni and S. Manandhar, “Designing an interactive open domain question answering system,” *J. Natural Lang. Eng.*, vol. 15, no. 1, pp. 73–95, 2008.
[2] D. Mollá and J. L. Vicedo, “Question answering in restricted domains: An overview,” *Computat. Linguist.*, vol. 13, no. 1, pp. 41–61, 2007.
[3] H. Cui, M.-Y. Kan, and T.-S. Chua, “Soft pattern matching models for definitional question answering,” *ACM Trans. Inf. Syst.*, vol. 25, no. 2, pp. 30–30, 2007.
[4] R. C. Wang, N. Schlaefer, W. W. Cohen, and E. Nyberg, “Automatic set expansion for list question answering,” in *Proc. Int. Conf. Empirical Methods in Natural Language Processing*, 2008.
[5] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, “Knowledge sharing and Yahoo answers: Everyone knows something,” in *Proc. Int. World Wide Web Conf.*, 2008.
[6] G. Zoltan, K. Georgia, P. Jan, and G.-M. Hector, Questioning Yahoo! Answers, Stanford InfoLab, 2007, Tech. Rep.
[7] H. Yang, T.-S. Chua, S. Wang, and C.-K. Koh, “Structured use of external knowledge for event-based open domain question answering,” in *Proc. ACM Int. SIGIR Conf.*, 2003.
[8] T. Yeh, J. J. Lee, and T. Darrell, “Photo-based question answering,” in *Proc. ACM Int. Conf. Multimedia*, 2008.
[9] G. Li, R. Hong, Y.-T. Zheng, S. Yan, and T.-S. Chua, “Learning cooking techniques from youtube,” in *Proc. Int. Conf. Advances in Multimedia Modeling*, 2010.
[10] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multimedia answering: Enriching text QA with media information,” in *Proc. ACM Int. SIGIR Conf.*, 2011.
[11] Trec: The Text Retrieval Conf. [Online]. Available: http://trec.nist.gov/.
[12] J. Cao, F. Jay, and J. Nunamaker, “Question answering on lecture videos: A multifaceted approach,” in *Proc. Int. Joint Conf. Digital Libraries*, 2004.
[13] Y.-C. Wu, C.-H. Chang, and Y.-S. Lee, “Cross-Language Video Question/Answering System,” in *Proc. IEEE Int. Symp. Multimedia Software Engineering*, 2004, pp. 294–301.
[14] Y.-S. Lee, Y.-C. Wu, and J.-C. Yang, “Bvideoqa: Online English/Chinese bilingual video question answering,” *Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 509–525, 2009.
[15] Y.-C. Wu and J.-C. Yang, “A robust passage retrieval algorithm for video question answering,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 10, pp. 1411–1421, 2008.
[16] G. Li, H. Li, Z. Ming, R. Hong, S. Tang, and T.-S. Chua, “Question answering over community contributed web video,” *IEEE Multimedia*, vol. 17, no. 4, pp. 46–57, 2010.
[17] G. Kacmarcik, Multi-Modal Question-Answering: Questions Without Keyboards, Asia Federation of Natural Language Processing, 2005.
[18] T.-S. Chua, R. Hong, G. Li, and J. Tang, “From text question-answering to multimedia QA on web-scale media resources,” in *Proc. ACM Workshop Large-Scale Multimedia Retrieval and Mining*, 2009.
[19] H. Tamura and N. Yokoya, “Image database systems: A survey,” *Pattern Recognit.*, pp. 29–43, 1984.
[20] V. Kobla, D. Doermann, and K.-I. D. Lin, “Archiving, indexing, and retrieval of video in the compressed domain,” in *Proc. SPIE Conf.*, 1996.
[21] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, “Query by humming: Musical information retrieval in an audio database,” in *Proc. ACM Int. Conf. Multimedia*, 1995.
[22] M. Wang and X. S. Hua, “Active learning in multimedia annotation and retrieval: A survey,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 10–31, 2011.
[23] Y. Gao, M. Wang, Z. J. Zha, Q. Tian, Q. Dai, and N. Zhang, “Less is more: Efficient 3d object retrieval with query view selection,” *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1007–1018, 2011.

[24] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, 2012.

[25] I. Ahmad and T.-S. Jang, "Old fashion text-based image retrieval using FCA," in *Proc. ICIP*, 2003.

[26] M. Wang, X. S. Hua, R. Hong, J. Tang, G. J. Qi, Y. Song, and L. R. Dai, "Unified video annotation via multi-graph learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 5, pp. 733–749, 2009.

[27] M. Wang, X. S. Hua, T. Mei, R. Hong, G. J. Qi, Y. Song, and L. R. Dai, "Semi-supervised kernel density estimation for video annotation," *Comput. Vision Image Understand.*, vol. 113, no. 3, pp. 384–396, 2009.

[28] J. Tang, R. Hong, S. Yan, T. S. Chua, G. J. Qi, and R. Jain, "Image annotation by KNN-sparse graph-based label propagation over noisy-tagged web images," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, pp. 1–15, 2011.

[29] J. Tang, X. S. Hua, M. Wang, Z. Gu, G. J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 2, pp. 409–416, 2009.

[30] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[31] S. K. Shandilya and N. Singhai, "Article: A survey on: Content based image retrieval systems," *Int. J. Comput. Appl.*, vol. 4, no. 2, pp. 22–26, 2010.

[32] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *Proc. ACM Int. Conf. Multimedia*, 2008.

[33] A. P. Natsev, M. R. Naphade, and J. TešiĆ, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *Proc. ACM Int. Conf. Multimedia*, 2005.

[34] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. Int. Conf. Image and Video Retrieval*, 2003.

[35] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li, "Learning to video search rerank via pseudo preference feedback," in *Proc. Int. Conf. Multimedia & Expo*, 2008.

[36] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Trans. Multimedia*, vol. 12, no. 8, pp. 829–842, 2010.

[37] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. ACM Int. Conf. Multimedia*, 2007.

[38] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang, "Visual query suggestion," in *Proc. ACM Int. Conf. Multimedia*, 2009.

[39] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua, "Visual query suggestion: Towards capturing user intent in Internet image search," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 6, no. 3, pp. 1–19, 2010.

[40] L. Nie, M. Wang, Z. Zha, and T.-S. Chua, "Oracle in image search: A content-based approach to performance prediction," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, pp. 13–13, 2012.

[41] X. Li and D. Roth, "Learning question classifiers," in *Proc. Int. Conf. Computational Linguistics*, 2002.

[42] J. Zhang, R. Lee, and Y. J. Wang, "Support vector machine classifications for microarray expression data set," in *Proc. Int. Conf. Computational Intelligence and Multimedia Applications*, 2003.

[43] J. Arguello, F. Diaz, J. Callan, and J. F. Crespo, "Sources of evidence for vertical selection," in *Proc. ACM Int. SIGIR Conf.*, 2009.

[44] A. Tamura, H. Takamura, and M. Okumura, "Classification of multiple-sentence questions," in *Proc. Int. Joint Conf. Natural Language Processing*, 2005.

[45] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in *Proc. Int. Conf. Empirical Methods in Natural Language Processing*, 2008.

[46] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. ACM Int. SIGIR Conf.*, 2002.

[47] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proc. ACM Int. CIKM Conf.*, 1999.

[48] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua, "Harvesting visual concepts for image search with complex queries," in *Proc. ACM Int. Conf. Multimedia*, 2012.

[49] E. Agichtein, S. Lawrence, and L. Gravano, "Learning search engine specific query transformations for question answering," in *Proc. Int. World Wide Web Conf.*, 2001.

[50] M. Bendersky and W. B. Croft, "Discovering key concepts in verbose queries," in *Proc. ACM Int. SIGIR Conf.*, 2008.

[51] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Syst.*, vol. 1, no. 1, pp. 10–28, 1993.

[52] M. Wang, X.-S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, 2009.

[53] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *Proc. Eur. Conf. Computer Vision*, 2004.

[54] M. Surdeanu, M. Ciaramita, and H. Zaragoza, "Learning to rank answers on large online QA collections," in *Proc. Association for Computational Linguistics*, 2008.

[55] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online QA sites," in *Proc. Int. Conf. Human Factors in Computing Systems*, 2008.

[56] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. Int. Conf. Web Search and Web Data Mining*, 2008.

[57] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein, "Exploring question subjectivity prediction in community QA," in *Proc. ACM Int. SIGIR Conf.*, 2008.

[58] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: Distinguishing informational and conversational questions in social QA sites," in *Proc. Int. Conf. Human Factors in Computing Systems*, 2009.

[59] S. SiegelCastellan, *Nonparametric Statistics for the Social Sciences*, 2nd ed. New York: McGraw-Hill, 1988.

Liqiang Nie Liqiang Nie received the B.Sc. degree from Xi'an Jiaotong University of China, Xi'an, in 2009. He is pursuing the Ph.D. degree at the School of Computing, National University of Singapore.

His current research interests include multimedia content analysis, search, large-scale computing as well as multimedia applications such as multimedia question answering, image reranking and expert mining. Various parts of his work have been published in top forums including ACM SIGIR, ACM MM, TOIS and TMM etc. Mr. Nie has been a Reviewer for various journals and conferences.

Meng Wang (M'09) is a professor in the Hefei University of Technology, China. He received the B.E. degree and Ph.D. degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, respectively. He previously worked as an associate researcher at Microsoft Research Asia, and then a core member in a startup in Silicon Valley. After that, he worked in the National University of Singapore as a senior research fellow. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He has authored more than 100 book chapters, journal and conference papers in these areas. He received the best paper awards successively in the 17th and 18th ACM International Conference on Multimedia and the best paper award in the 16th International Multimedia Modeling Conference.

Yue Gao received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2005, and the M.E. degree and Ph.D. degree from Tsinghua University, Beijing, China, in 2008 and 2012 respectively. He had been a visiting scholar at Carnegie Mellon University and worked with Dr. Alexander Hauptmann from Oct. 2010 to March 2011, a research intern at National University of Singapore and Intel China Research Center, respectively. He is currently a Research Fellow with the School of Computing, National University of Singapore.

His research interests include large scale image/video retrieval, 3D object retrieval and recognition, and social media analysis. He is the author of 30 journals and conference papers in these areas. He is a Special Session Chair of the 2012 Pacific-Rim Conference on Multimedia and the 18th International Conference on Multimedia Modeling 2013.

Zheng-Jun Zha (M'08) received the B.E. degree and the Ph.D. degree in the Department of Automation from the University of Science and Technology of China (USTC), Hefei, China.

Dr. Zheng-Jun Zha is currently a senior research fellow in the School of Computing, National University of Singapore. His current research interests include multimedia content analysis, computer vision, as well as multimedia applications such as search, recommendation, and social networking. He received Microsoft Research Fellowship in 2007, President Scholarship of Chinese Academy of Science in 2009, and the Best Paper Award in the 17th ACM International Conference on Multimedia (ACM MM). He is a member of ACM.

Tat-Seng Chua (SM'06) is the KITHCT Chair Professor at the School of Computing, National University of Singapore (NUS). He was the Acting and Founding Dean of the School of Computing during 1998–2000. He joined NUS in 1983, and spent three years as a research staff member at the Institute of Systems Science (now I2R) in the late 1980s. Dr Chua's main research interests are in multimedia information retrieval, multimedia question-answering, and the analysis and structuring of user-generated contents. He works on several multi-million-dollar projects: interactive media search, local contextual search, and real-time live media search.

Dr. Chua has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the conference co-chair of ACM Multimedia 2005, CIVR (Conference on Image and Video Retrieval) 2005, and ACM SIGIR 2008, and the Technical PC Co-Chair of SIGIR 2010. He serves in the editorial boards of: ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He sits in the steering committee of ICMR (International Conference on Multimedia Retrieval), Computer Graphics International, and Multimedia Modeling conference series; and serves as member of International Review Panels of two large-scale research projects in Europe. He is the independent director of 2 listed companies in Singapore.