IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834, p-ISSN: 2278-8735
PP 34-44
www.iosrjournals.org

MMQA: Multimedia Answer Generation for Community Question Answering System

Handore S. A.[1] & Ganeshwade M. M.[2]
[1,2](Comp Engg Dept., MIT Aurangabad, BAM Univ., Aurangabad(MS)India)

***Abstract* :** *Community question answering (cQA) services have accomplished quiet popularity over the past years. It helps user to get information from a comprehensive set of well-answered questions. 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 propose a scheme that is able to improve textual answers in cQA with appropriate multimedia data. MMQA consists of three components in its scheme, those are Answer medium selection, Query generation for multimedia search, and Multimedia data selection and presentation. This approach automatically figures out which type of multimedia information should be added to get an elaborated textual answer. For this, it then automatically collects data from web to enrich the answer. By processing a huge set of question- answer pairs and adding them to a dataset, our approach can set up a novel multimedia question answering (MMQA) approach as users can find multimedia answers by comparing their questions with those in the dataset. Our approach MMQA is different from rest of the MMQA research efforts as it not only provide direct question answers with image & video data but also is based on community-contributed textual answers and hence it is able to give answers of more complex questions. We have conducted huge experiments on a multi-source QA dataset. The results demonstrate the effectiveness of our approach by providing user more satisfactory answers*.
***Keywords* -** *Fitness cQA, Question Answering, Reranking*

1. INTRODUCTION

The amount of information on the Web has grown exponentially over the years, with content covering almost any topic. As a result, when looking for information, users are often bewildered by the vast quantity of results from search engines. Users usually have to painstakingly browse through a long list of results to look for a precise answer. Therefore question-answering (QA) research emerged in an attempt to tackle this information-overload problem. Instead of returning a ranked list of results as is done in the current search engines, QA aims to leverage in-depth linguistic and media content analysis as well domain knowledge to return precise answers to natural language questions. Question answering (QA) is defined as the task of automatically providing a precise answer to a natural language question posed by users. Typically, given a question, an ideal QA system is expected to find answer from certain corpuses 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. 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.

Fig. 1. Examples of QA pairs from several popular cQA forums

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 their great success, existing cQA forums mostly support only textual answers, as shown in Figure 1. Unfortunately, textual answers may not provide sufficient natural and easy-to-grasp information. Figure 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 Figure 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. 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. 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:

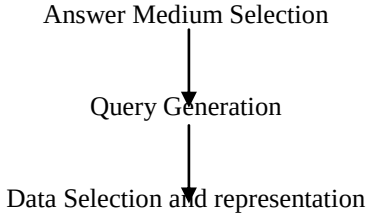

Fig 2: Proposed Multimedia answering Scheme

(1) Answer medium selection: In this work, we consider the following four cases 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.

2. RELATED WORK

An usual QA system is composed of three components: question processing, document retrieval, and answer extraction. In question processing, a given question is analyzed, and its question type is determined. This process is called "question classification". Depending on the question type, the process in the answer extraction component usually changes. Consequently, the accuracy and the efficiency of answer extraction depend on the accuracy of question classification.

Here are some existing methods for question classification. The methods are roughly divided into two groups: the ones based on hand-crafted rules and the ones based on machine learning. The system "SAIQA" used hand-crafted rules for question classification. However, methods based on pattern matching have the following two drawbacks: high cost of making rules or patterns by hand and low coverage. Machine learning can be considered to solve these problems. Li et al. used SNoW for question classification. The SNoW is a multi-class classifier that is specifically tailored for learning in the presence of a very large number of features. Zukerman et al. used decision tree. Ittycheriah et al. used maximum entropy. Suzuki used Support Vector Machines (SVMs). Suzuki compared question classification using machine learning methods (decision tree, maximum entropy, SVM) with a rule-based method. The result showed that the accuracy of question classification with SVM is the highest of all.
According to Suzuki ,a lot of information is needed to improve the accuracy of question classification and SVM is suitable for question classification, because SVM can classify questions with high accuracy even when the dimension of the feature space is large. Moreover, Zhang et al. compared question classification with five machine learning algorithms and showed that SVM outperforms the other four methods as Suzuki showed.

Following table illustrates the work done on QA systems:

Table 1: Literature survey

Sr. No.	Author	Year	Contribution
1.	Trec: Text Retrieval Conference	1990	Text based QA
2.	S.A. Quarteroni, S. Manandhar	2008	Text based QA based on type of questions Open Domain QA
3.	D. Molla & J.L Vicedo	2007	Restricted Domain QA
4.	H. Cui, M.Y. Kan	2007	Definitional QA
5.	R.C.Wang, W.W. Cohen, E. Nyberg	2008	List QA
6.	H. Yang, T S chua, S. Wang	2003	Video QA
7.	J.Cao, Y-C- Wu, Y-S Lee	2004-2009	Video QA using OCR & ASR
8.	Lot of authors	2003-2013	Content Based Retrieval

1. 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. Based on the type of questions and expected answers, we can roughly summarize the sorts of QA into Open-Domain QA, Restricted-Domain QA, Definitional QA and List QA. 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.

However, nearly all of the existing cQA systems, such as Yahoo! Answers, Wiki Answers 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 Video QA 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. Li *et al.* presented a solution on "how-to" QA by leveraging community-contributed texts and videos. Kacmarcik *et al.* explored a non-text input mode for QA that relies on specially annotated virtual photographs.

An image-based QA approach was introduced in, which mainly focuses on finding information about physical objects. Chua *et al.* 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 earlier 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.

2. 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 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.

Generally, multimedia search efforts can be categorized into two categories: text-based search and content-based search. The text-based search 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.

Further, several social media websites, such as Flicker 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 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.

3. 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. 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 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 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 pair-wise 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. 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.

3. ANSWER MEDIUM SELECTION

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, or 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 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) text + image, which means that image information needs to be added; (c) text + video, which means that only video information needs to be added; and (d) text + image + video, *i.e*., we add both image and video information.

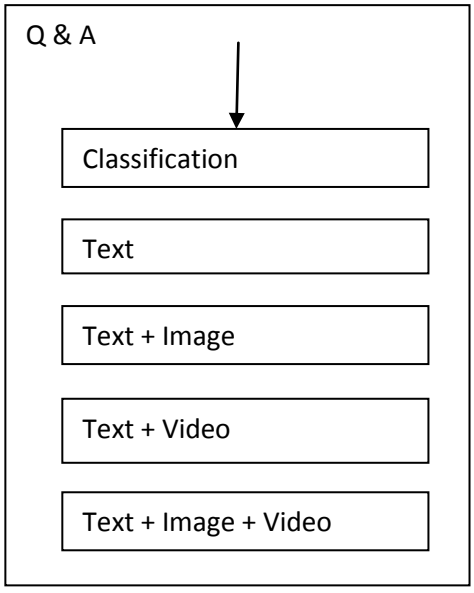

Fig. 3: Answer medium Selection

There are some existing research efforts on question classification. Li and Roth 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 used linear SVMs with all possible question word grams to perform question classification. Arguello *et al.* 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.

- 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 to extract the core sentence from each question.

The classification is accomplished with two steps. First, we categorize questions based on interrogatives 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 naïve 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 *e.g.* "Does roy jones jr have three kids",
- Choice class *e.g.* "Which country is bigger, Canada or America",
- Quantity class *e.g.* "When was the first mechanical calculator made",
- Enumeration class *e.g.* "Name of three neighboring countries of south Korea",
- Description class *e.g.* "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 "how+adj/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 "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 a classification with a Naive Bayes classifier.

Table 2 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.

Table 2: Representative interrogative words

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

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, 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 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 3.

Table3: Representative class-specific related word

Categories	**Class-Specification**

Text	population, times, country, website, number, date, age, rate, height, name, period Speed, birthday, distance, religious *etc*
Text+Image	colour, figure, band, pet, clothes, largest, photo, surface, logo, place, whom, look like, symbol, appearance ,capital *etc*
Text+Video	Said ,first, differences, recipe, dance, steps, ways, tell, film, story, invented, how to how do, how can *etc*.
Text+Image+Video	king, war, issue, kill, president, earthquake, battle, event, prime minister etc.

- 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. 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.

- 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.

We predict search performance based on the fact that, most frequently, search results are good if the top results are quite coherent. We adopt the method 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*.

$$\text{Clarity}_q(C_i)=\sum P(\omega\square\theta_q)\log_2\ p(\omega\square\theta_q) \quad (1)$$

$$\omega\ \square V_{ci} \qquad p(\omega|\theta_{Ci})$$

where Vci is the entire vocabulary of the collection Ci, and i = 1; 2; 3 represent text, image and video, respectively . The terms P(w| θq) and P(w|θCi)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 model is estimated from the top documents, R, as the following formula,

$$P(\omega\square\theta_q)=1/\square\sum_{d\square R} P(\omega\square D)\,P(q|D) \quad (2)$$

and $\square$ is defined as,

$$\square=\sum_{D\square R} P(q|D) \quad (3)$$

where P(q|D) is the query likelihood score of document D.

$$P(q|D)=\Pi_{\omega\square q}\ P(\omega|D) \quad (4)$$

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.

- 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+image", "text+video", and "text+image+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.

4. 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.

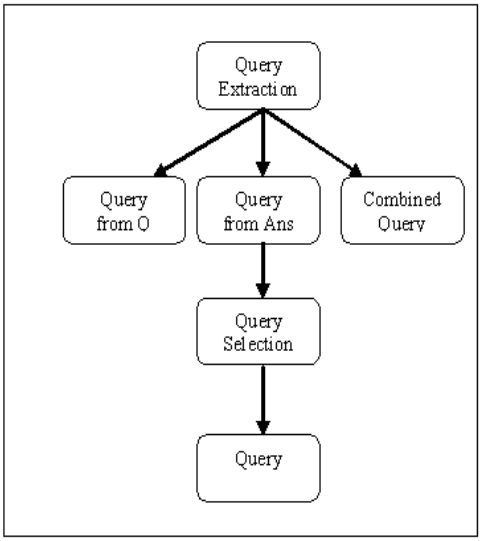

Fig.4: Query generation for multimedia search

Textual questions and answers are usually complex sentences. But frequently search engines do not work well for 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 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.

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. Second, we identify several key concepts from verbose answer

which will have the major impact on effectiveness. Here we employ the method in. 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 the 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 36 POS are identified. 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.

(2) Search performance prediction :

This is because, for certain queries, existing image and video search engines cannot return satisfactory results. We adopt the method 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.

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.

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, 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. We re-state the equation from as,

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

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 $0 <= \alpha < 1$, and Pij is the transition probability from data-point i to j. Here P is a row-normalized transition matrix obtained from similarity matrix W, and $r^{j}_{(0)}$ is the initial relevance score of the sample at the j^{th} position, which is heuristically estimated as

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

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

$$W_{ij} = \exp[(-\| x_i - x_j \|^2)/\sigma^2] \qquad (7)$$

Where, $N_K(i)$ 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*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, Considering two videos (vi;1; :::; vi;m) and (vj;1; :::; vj;n), which contain m and n key-frames respectively, we employ average distance of all cross-video keyframe pairs for similarity estimation, i.e.,

$$Wij = \exp\left(-\frac{\sum_{q=1}^{m}\sum_{p=1}^{n}\| v_{i,q} - v_{j,p} \|^2}{\sigma^2}\right) \quad (8)$$

Similarly, NK(i) denotes the index set for the K nearest neighbors of a video measured by Euclidean distance and the parameter _ 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 earlier, 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. Figure 3.2.3.1 shows our approach.

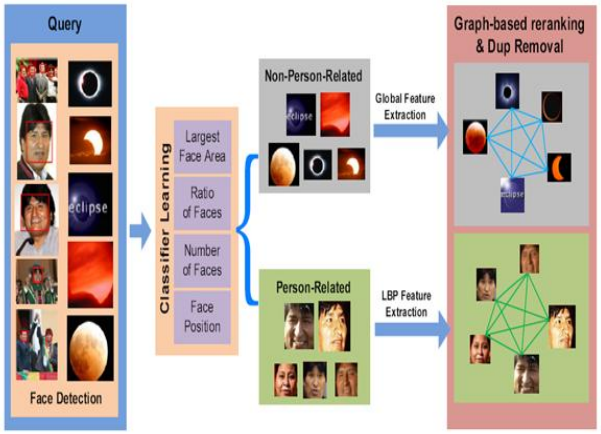

Fig.5: The schematic illustration of the query-adaptive reranking approach.

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. 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 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 i^{th} image or video if there exists $j < i$ that satisfies $Wij > T$. Here we empirically set T to 0:8 throughout the work.

After 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.

6. CONCLUSIONS AND FUTURE WORK

We have described 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, to make the enriched media data more diverse.

REFERENCES

[1] Liqiang Nie, Meng Wang, Yue Gao, Zheng-Jun Zha, Tat-Seng Chua, "Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information" IEEE Transactions On Multimedia January 2013
[2] A. Tamura, H. Takamura, and M. Okumura, "Classification of multiple- sentence questions," in Proc. Int. Joint Conf. Natural Language
[3] D. Mollá and J. L. Vicedo, "Question answering in restricted domains: An overview," Computat. Linguist, vol. 13, no. 1, pp. 41–61, 2007.
[4] Guangyou Zhou, Yang Liu, Fang Liu, Daojian Zeng, and Jun Zhao," Improving Question Retrieval in Community Question Answering Using World Knowledge", Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence
[5] R.Mervin, A.Jaya, "Web spider: A Search engine for multimedia Question answering", International Journal of Research in Information Technology, Vol. 01, Issue 01, April 2012
[6] 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.
[7] G. Zoltan, K. Georgia, P. Jan, and G.-M. Hector, Questioning Yahoo! Answers, Stanford InfoLab, 2007, Tech. Rep.
[8] L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, "Multimedia answering: Enriching text QA with media information," in Proc. ACM
[9] 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.
[10] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in Proc.
[11] J. Zhang, R. Lee, and Y. J. Wang, "Support vector machine classifications for microarray expression data set," in Proc. Int. Conf.