ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Advanced Multimedia Answer Generation by Scraping Information through Web

Ramakrishnan.R[1], Jayalakshmi.A[2], Priyadharshani.S[3]

Associate Professor, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry, India

PG Student, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry, India

PG Student, Department of MCA, Sri Manakula Vinayagar Engineering College, Puducherry, India

ABSTRACT: Advanced Multimedia Answer generation by scraping information through web emerged as an extremely popular alternative to existing community question answering. Unfortunately, textual answers may not provide sufficient natural and easy to grasp information. It will be much better if there are some additional videos and images that visually demonstrate the process. Therefore, the textual answers in cQA can be significantly enhanced by adding multimedia contents, and it will provide answer seekers more widespread information and experience. In our system the user question is analyzed, from that the requirement of the user has been identified, then the appropriate medium of answer should be searched from the online and then it will be presented to the user. Model

KEYWORDS: Scraping, CQA, MMQA, E-VSM, POS Tagger.

I. INTRODUCTION

Question Answering (QA) is a technique for automatically answer a question posed in natural language. It defined as the task of searching for and extracting the text that contains the answer for a specific question, stated in natural language, from a collection of text documents or a corpus (Molla and Vicedo, 2007). QA systems can be classified in two types, depending on the application domain: Open-Domain Question Answering (ODQA) or Restricted-Domain Question Answering (RDQA) systems. While the first type is concerned with a wide variety of questions (e.g. *who is the prime minister of India?*), the second type is properly adapted to a particular area (e.g. *what antibiotic was needed for fever person?* in a medical domain), thus obtaining more precise results regarding a specific topic.

Although aQA had achieved significant progress with the year-by-year evaluation exercises, they largely focused on short and simple questions; more complex questions were generally less studied. Besides factoid, list and definitional questions, there are other types of questions commonly occur in the real world, including the ones concerning procedures ("how"), reasons ("why") or opinions, etc. Different from the simple questions, the answers to these complex questions may not locate in a single part of a document, and it is not uncommon that the different pieces of answers are scattered in the collection of documents.

The next technology emerged was community question answering (cQA). It mostly supports only textual answers. But, textual answers may not provide sufficient information. For the questions "*What are the steps to be follow to install SQL server 2005*" and "*How to prepare coffee*", the answers are described by long sentences. Clearly, it will be much better if there are some videos and images that visually demonstrate the process. Therefore, there is a need to enhance the textual answer with multimedia content such as image and video.

Then the next scheme emerged was novel multimedia question answering (MMQA)which was proposed in our paper. This technique can enrich community-contributed textual answers in cQA with appropriate media data. It consists of three components:

1. Answer medium selection
2. Query generation for multimedia search.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

3. Multimedia data selection and presentation.

The rest of the paper describes the working flow of the project from textual QA to multimedia QA.

II. LITERATURE REVIEW

The early investigation of QA started from 1960s. In open domain interactive QA[1], question processing, document retrieval and answer extraction were used and an additional component is the User Modelling (UM) component, which is introduced to overcome the traditional inability of standard QA systems to accommodate the users' individual needs. Automated QA does not provide worse answers for complex questions.In knowledge sharing and yahoo answer [3], a user posts a question, and other usersreply directly to that question with their answers. Besides, knowledge sharing, it also allows users to share advice, opinions, etc.

In community Question Answering system, it provides only pure text based answers, which may not provide sufficient information for the users.so there is a need for multimedia answers.

Some research efforts have been put on multimedia QA,which aims to answer questions using multimedia QA. The approach photo based QA was presented in 2008 which was mainly focused on finding images of the physical objects and presenting answers to the user with text and images. Still there is lacking in video answers which had required for some questions.Then the research had performed on video retrieval and it was implemented in 2011.It tried to find the appropriate video from the youtube, google videos, etc. for the user questions.

Each research provided answers with multimedia content but all are implemented separately which had not fully satisfied the user. So there is a need to provide the combination of multimedia answers.

Then the technology emerged was called Multimedia Question Answering (MMQA) in 2013. The main goal of this approach is to enrich the textual answer with multimedia content such as images or videos or images and videos.

As previously introduced, both traditional automatic textual QA, community based QA have achieved great success. The former approaches mainly addresses the simple and factoid questions, while the latter approaches makes it possible to answer verbose and complex questions via utilizing the intelligence from Internet users. However, to date, QA research has largely focused on text and the existing MMQA research work either can lightly handle certain questions in narrow domains or only support question-independent monolithic media type, such as pure video and pure image. Further, the existing should not deeply understand the needs of the users, which is the key to handle complex and general questions in broad domains. In addition to, answer medium determination, answer availability prediction, and media answer selection are all not exactly used before.

III.RESEARCH APPROACH

Existing system uses a novel scheme to answer questions using media data by using 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, based on the users question the answers should be presented with text or text and image or text and image and video.

In this project, we propose a novel scheme which can enrich community-contributed textual answers in cQA with appropriate media data.

It contains three main components:

1. Answer medium selection
2. Query generation for multimedia search.
3. Multimedia data selection and presentation.

A. Answer medium selection:

Given a QA pair, it predicts whether the textual answer should be enriched 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+videos,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

text+images, and text+images+videos. It means that the scheme will automatically collect images, videos, or the combination of images and videos to enrich the original textual answers.

B. 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.

C. Multimedia data selection and presentation:

Based on the generated queries, we collect image and video data through search engines. We then perform interactive re-ranking and duplicate removal to obtain a set of accurate and representative images or videos to enrich the textual answers.

III. DESIGN OF THE METHODOLOGY AND TECHNIQUE

The various techniques and algorithm used in the project are

1. Stemming Algorithm & stop word removal.
2. Naive Bayes
3. Bigram text classification
4. POS Histogram

A. Stemming Algorithm:

A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form,

For example,
Connection
Connections
Connective ---> connect
Connected
Connecting

It is important to appreciate that we use stemming with the intention of improving the performance of IR systems. It is not an exercise in etymology or grammar. In fact from an etymological or grammatical viewpoint, a stemming algorithm is liable to make many mistakes. In addition, stemming algorithms - at least the ones presented here - are applicable to the written, not the spoken, form of the language.

B. Stopwords removal:

It has been traditional in setting up IR systems to discard the very commonest words of a language - the stopwords - during indexing. A more modern approach is to index everything, which greatly assists searching for phrases for example. Stopwords can then still be eliminated from the query as an optional style of retrieval. In either case, a list of stopwords for a language is useful.

Getting a list of stopwords can be done by sorting a vocabulary of a text corpus for a language by frequency, and going down the list picking off words to be discarded.

The stopword list connects in various ways with the stemming algorithm:

The stemming algorithm can itself be used to detect and remove stopwords. One would add into the irregular_forms table something like this,

"", /* null string */
"Am/is/are/be/being/been/" /* BE */
"Have/has/having/had/" /* HAD */
"Do/does/doing/did/" /* DID */

so that the words `am', `is' etc. map to the null string (or some other easily recognized value).

C. Naive Bayes:

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculation the posterior probability, P(c/x), from P(c), p(x) and P(x/c). Naïve Bayes classifier assumes that the effect of the value of a predictor(X) on a given class© is independent of the values of other predictors. This assumption is called class conditional independence.

$$P(c|x)=P(x|c)P(c) / P(x)$$

$P(c|x) = P(x_1|c) * P(x_2/c) * P(x_3/c) * P(c)$

P (c|x) is the posterior probability of class given predictor

P(c) is the prior probability of class

P (x|c) is the likelihood which is the probability of predictor given class

P(x) is the prior probability of predictor

D. Bigram text classification using E-VSM bigram frequency count algorithm:

A bigram is every sequence of two adjacent elements in a string of tokens, which are typically letters, syllables, or works; they are n-grams for n=2. The frequency distributions of bigrams in a string are commonly used ofr simple statistical analysis of text in many applications, including in computational linguistics, cryptography, speech recognition, and so on. Skipping bigrams are word pairs which allow gaps(perhaps avoiding connecting words, or allowing some simulation of dependencies.

E-VSM stores all the positions of every word where it is repeated in the document. Length of position vector is nothing but the appearance frequency of that particular word. These position vectors can be exploited to find the frequency of ngram in document. Bigram is most commonly used n-gram in text processing. Following is algorithm to calculate the bigram frequency referred form [7].

E. POS Histogram:

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.

IV. EXPERIMENTAL DESIGN

The following experiments had conducted to evaluate various approaches.

A. Evaluation of answer medium selection approach

We first evaluate our answer medium selection approach. It consists of two techniques: Question based classification, Answer based classification.

1) Question based classification:

The set of 25 different questions should be prepared on various type of question such as yes/no type, choice type, quantity type, enumeration type, description type. Then the experiment should be conducted with 50 members. Based on the results provided by the members the analysis and calculation had been made on the results. Finally the keyword used to find the answer medium for user question had found.

The following table describes the category and class specific related keywords.

TABLE I

Table 1: Keyword used for question based classification

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,looklike,who,image,pictures,appearance,largest,band,photo,surface,capital,figure,what is a, symbol, whom, logo, place, etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

TEXT+VIDEO	Howto,howdo,howcan,invented,story,film,tell,songs,music,receipe,differences,ways,steps,dance,first,said,etc.
TEXT+IMAGE+VIDEO	President,king,primeminister,kill,issue,nuclear,earthquake,singer,battle,event,war,happened ,etc.

2) Answer based classification:

The answer based classification performed using bigram feature and verb form present in the answer. The combined usage of both bigram and verb achieves better result.

B. Evaluation of query generation approach:

Next we evaluate the query generation and selection approach. The three queries are generated from question, answer and combination of question and answer. After generating three queries one most informative query should be selected from them. This can be performed using POS (Part Of Speech) technique.

V. RESULTS AND DISCUSSION

Some questions and the corresponding results obtained for that question had been discussed below

1) The question which require text as output

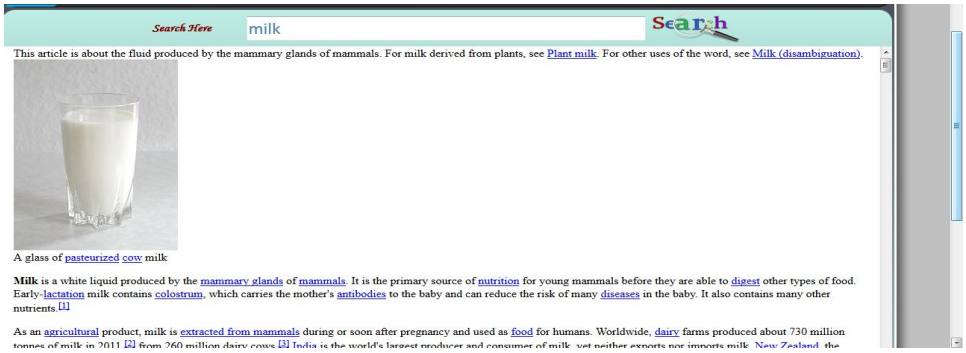

Fig 1: Question which require text as output

2) The question which require imageas output

Fig 2: Question which require image as output

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

For the question what is banana our approach shows the images as result.

2) The question which require Text+video has output

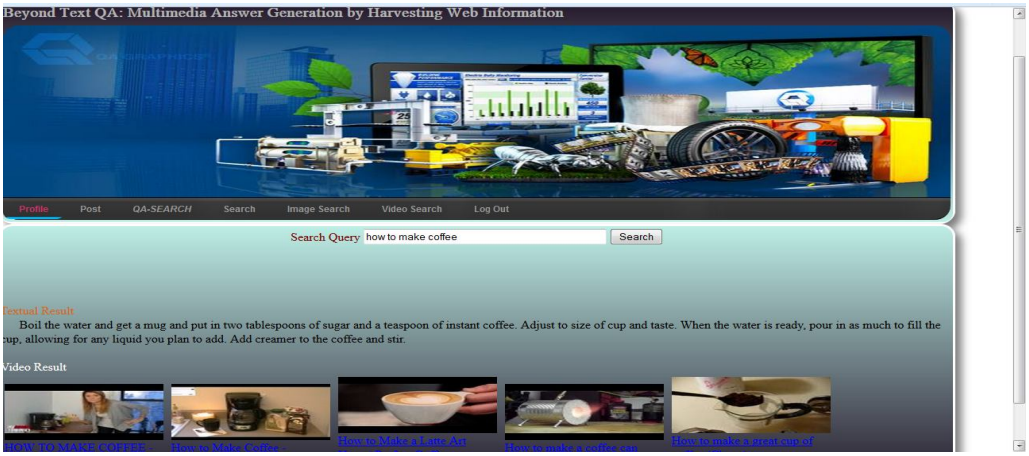

Fig 3: Question which require text and video as output

The above question is how to make coffee. From the question our system predicts the user expectation and provides the text and video as result to the user.

VI. CONCLUSION

In this paper we describe a schematic approach that enriches the textual answer in CQA with corresponding media data. For that our scheme Advanced MMQA decides what category of results should be presented to the user in order to make user more understandable. Our scheme should search the appropriate multimedia answer through the web. The algorithm used to calculate the bigram frequency count using E-VSM (Enhanced-Vector Space method makes answer based classification more effective when compared to general approach used in the existing system.Our approach produces the good results for complex query too.

REFERENCES

[1] D. Mollá and J. L. Vicedo, “Question answering in restricted domains: An overview,” Computat. Linguist., vol. 13, no. 1, pp. 41–61, 2007.
[2] 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.
[3] L. A. Adamic, J. Zhang, E. Bakshy, andM. S. Ackerman, “Knowledge sharing and Yahoo answers: Everyone knows something,” in Proc. Int. World Wide Web Conf., 2008.
[4] T. Yeh, J. J. Lee, and T. Darrell, “Photo-based question answering,” in Proc. ACM Int. Conf. Multimedia, 2008.
[5] G. Li, H. Li, Z. Ming, R. Hong, S. Tang, and T.-S. Chua, “Question answering over community contributed web video,” IEEEMultimedia, vol. 17, no. 4, pp. 46–57, 2010.
[6] 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.
[7] Ankit Bhakkad, S. C. Dharamadhikari, Parag Kulkarni,.” Efficient Approach to find Bigram Frequency in Text Document using E-VSM”. International Journal of Computer Applications Volume 68– No.19, April 2013
[8] Zheng-Jun Zha,Meng Wang,, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua,”Interactive Video Indexing With Statistical Active Learning”,IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 14, NO. 1, FEBRUARY 2012
[9]Eugene Agichtein Steve Lawrence Luis Gravano, “Learning Search Engine Specific Query Transformationsfor Question Answering”, ACM
[10] R.Manju1, “Beyond Text QA Multimedia diverse relevance ranking based Answer Generation by Extracting Web”, International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014
[11] Nandhini.N, Ramya.K , Sandeepa.P” Multimedia QA generation by using search diversification”,International Journal of Computer Science and Mobile Computing, Vol.3 Issue.2, February- 2014