Information Searching Using Web Data Extraction for Acquiring Enrich MM Results

Firoz I. Kadri[1]
Student of BE Information Technology
BVCOE & RI, Nasik, India
University of Pune
firozkadri.086@gmail.com

Kaushal A. Kapse[3]
Student of BE Information Technology
BVCOE & RI, Nasik, India
University of Pune
kaushal.kapse93@gmail.com

Nitin V. Landge[2]
Student of BE Information Technology
BVCOE & RI, Nasik, India
University of Pune
nee3.1991@gmail.com

Hemant B. Shirsath[4]
Student of BE Information Technology
BVCOE & RI, Nasik, India
University of Pune
f9sunny@gmail.com

Prof. V. D. Badgujar[5]
BE Computer Engineering
BVCOE &RI, Nasik, India
University of Pune
badgujarvivek83@gmail.com

Abstract– There are many question answer sites are available now a day. In that Community Question Answering Sites are achieved lots of popularity over last years. But the drawback of available question answering system is that it can only provide multiple solutions of the textual answer and user need to select one of them. In this paper, we propose a scheme that enriches the textual answer with multimedia data. Our scheme consists of four models: QA pair extraction, answer medium selection, query generation and selection and presentation. The type media information added with the textual data is determined. The question answer pair is generated from the available Community Question Answering Sites database. Query is generated for the multimedia data. The final resulting data must be selected after re-ranking and removal operation and then present to the user.

Keyword-*Qa pair, multimedia selection, query generation, openCV, surf*

*****

I. INTRODUCTION

Now a day the social media changing as new trends rise, the popularity of various places and market are involves. As per the studies about 600 million people are use the search sites. The amount of data on the sites is irises day by day. When users search for any required data on the internet's, users gets lots of information on related data and users need to check each document in order to get match appropriate document to give proper information. This information overloading problem can be solved with the help of question answering system. In earlier QA system mainly focused in some specific domain but the question answering system provide precise

Answer to the question. Question answering system mainly divided closed domain system and open domain system. In the closed domain used structure data and convert into natural language. In open system direct access the database, as the User needs it used structure database that contain large amount of information that may be cover all information which may be help to user easily. Question answering system can efficiently handled the informative question such As what, where, when, why, like that. But it can difficult to answer the question like what and how? That may be generating proper answering still has been difficult to find the complex question and related answers. Question answering application solves this problem. It is a huge of structure information available that sharing a technical knowledge and give advice for the user for selection of proper answer of that question and other related multimedia data .In question answering, when users post a question the answer is obtained from different source with different participant, The problem in automate question answer can be replace with answer that contain human intelligence. Given question answer pair, it ensures whether the textual answer should be enriched with media information, which will categorize into four classes: text, text +image ,text +video and text +image +video. It means that the application will automatically generate images, videos or booths and enrich to the final textual Answer. Query generation for multimedia search in order to collect multimedia data from other recourses, and the most approximate query selected from this given three classes such as query extraction, medium selection, query generation and presentation. And then finally all the three models are processing on given question then removing all repeated document and ranking all display data, then finally Multimedia information must be selected proper solution and present to the user. The proper solution of related question we collect image and

 ISSN: 2321-8169
203 - 208

video on structured database and convert into natural language. In this paper approach is not to directly generate answer the question, and strategy to share the big misunderstanding between answer and question, i.e. different between text answer and multimedia answer. In this application the first preference to find appropriate textual answer we connect together by the source of intelligence of Community members, then we focus on solving to the second difference is to select appropriate multimedia database and provide proper one of them therefore, our scheme can be viewed an approach that accomplishes the community Question Answer problem by mixed together and exploring the human and computer. The existing community question answering system such as yahoo Answers ,wiki answers, Ask Metadata, stack overflow etc. problem of that Community Question Answering Sites is that, it only provide textual answer or related link that only to supplementary images or videos. Textual answers are not only sufficient to some of other questions. Example "How to cook chicken Biryani?" in the form of textual answer does not provide proper understanding. So that answer must be with the video of full procedure of coking chicken Biryani, then it will be more reliable solution to user and it may be easily sea all procedure of cooking a Chicken Biryani. So that an application introduce a Multimedia Question Answering Site . There are mainly four components.

1. Question answers pair extraction.
2. Answer medium selection.
3. Query generation.
4. Multimedia search and presentation.

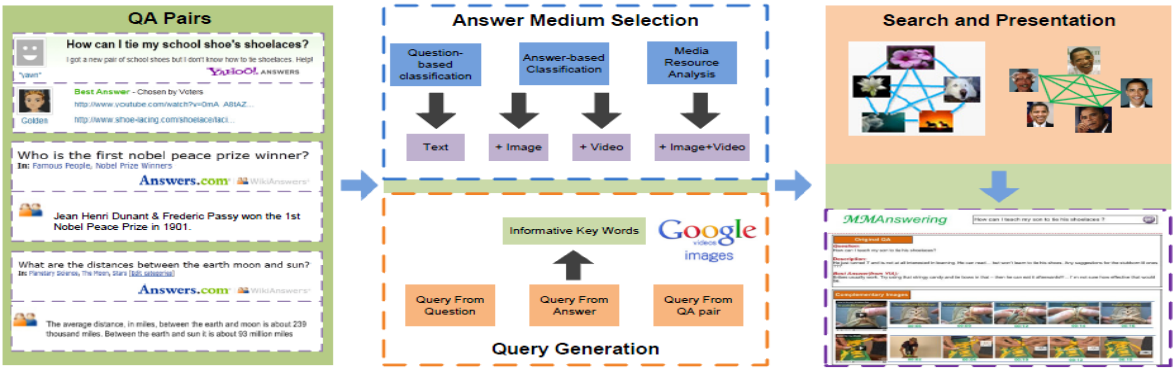

Figure 1: These schematic shows purposed of multimedia answer generation selection. And this scheme mainly three component that are query answer generation, answer medium selection, multimedia search and presentation.

II. LITRATURE SURVEY

Before developing the tool it is necessary to determine the time factor ,available recourses economy and company strength. Once these things are satisfied, then next step are to determine which operating system and language can be used for developing the application. Once the programming start building the tool of programmer need lot of external support . This support can be obtained from experience staff or programmers, and other document from Yahoo answer. Before building the system the above consideration are taken into account for developing the proposed system [1]. Along with the improvement of underlying communication technology, community QA has emerged as an extremely popular alternative to acquire information online, information gainer are able to post their specific question on any topic and obtain answer provided by other participants. For example, wiki Answer, Yahoo! answer one of the most well-known cQA systems, they are more than 15 million answered question distributed in 8,000 categories [2]. Existing cQA forums mostly support only textual answer. Unfortunately textual answers may not provide sufficient and easy-to grasp information to user. For the question what are the steps to make a weather when and what does trillion look like the answer are described by long sentence. Clearly it will be much better if there are some accompanying videos and images that visually demonstrate the process or the object [3]. Therefore, the textual answers in cQA can be significantly enhanced by adding multimedia contents, and it will provide answer gainer more comprehensive information and better experience. One definition of a question could be 'request for information'. But how do we recognize such a request in written language we often rely on question marks to denote questions [4]. However this clue is misleading as rhetorical question do not require an answer but are often terminated by a question mark while statements asking for information may not be phrased as question [5]. People can easily handle these different expressions. We mainly focus definition question which unlike factoid questions require a more complex answer, usually constructed from multiple source document ,this process is repeated until the crawlers decide to stop [6]. Collected pages are later used for other application, such as a web search engine or a web cache. As the size of the web grow, it become more difficult to retrieve the whole or a significant part of web using simple process. There are many search engines run multiple processes in parallel to

 ISSN: 2321-8169
203 - 208

perform this above task, so that download rate is maximized [9].

III. SYSTEM OVERVIEW

A. Answer medium selection

As we see in introduction which determine type of medium to be selected to enrich textual answer for example. "How to Drive a Car?" such a type of question does not need to any multimedia data or answer, only textual answer is sufficient to this type of questions. However in some cases to give this answer of that question we need to add images or video with textual answer to enrich such type of answer such as "How to overtake a Car?" it's better to provide images with textual answer and also get video to get detailed demonstration of multimedia data in the form of text, image or videos of given question such as "How connect my mobile to PC?" the selection of final answer is combination of text, image and video the choice depends on information gainer. In that all data are categorize into four classes: (a) text, only textual answer is sufficient ;(b) text +image along with text answer we also need the images ;(c) text+vedio along with text we also need the video ;(d) text+image+vedio is that all about combination of image text video. This classification is depend on user need and categorize by analyzing question and parallel all multimedia recourses analysis in question based classification contain two steps. First question can be categorized according to same or reparative word and some question directly gives the answer in text. Second classifier which perform the classification on rest of the question. The drawback of multimedia answer medium selection is does not support text+image+audio+video information.

B. Query generation for multimedia search:

Before performing multimedia search on search engine we need to collect appropriate media information such as image video from web to generate queries from proper text question answer pair. This can be done in two steps. In first step query extraction is done, with sentence of text question and answer. It extracts set of related information. Second step is query selection in that queries can be generated from question answer and combination of both. The pair generates three combination of queries first convert a question into a query which may be correct grammatical sentence in to meaningful correct sentence. In second that identify which will have most effective concept of answer and the last third option is the combination of first and second queries that will be generate from question and answer.

C. Multimedia data selection and presentation

To collect multimedia data we need to generate queries through search engine but search engine does not support only text based serial and return lots of unwanted multiple solution of that question. To avoid that problem in that scheme we must be find appropriate text answer and then after find related multimedia data using re-ranking method is used to identify the query is belong to either person or non person related query. If given query is based on person, then face detection method is used to image identification return by search engine. Non person related queries extract 428 dimensional visual features after that re-ranked performed on removal of duplicate image and provide proper solution from one of them.

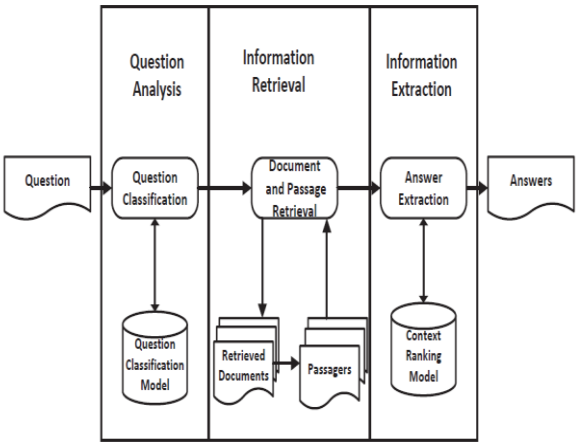

Figure 2: Flow of automatic textual question answer.

EXISTING SYSTEM:

Community question answering forums like Yahoo! Answer and wiki answer are very popular from few years that mean information fetching on the web. By posting question for other participate to answer; the information fetcher can obtain specific answer to the question. These sites are very popular and growing very rapidly. Users of popular portal like Yahoo! As well as Google!

About Yahoo! Answers:

Find out everything there is to know about Yahoo! Answers. If you are just getting any business update than it will be find immediately on Yahoo! Answers here is the place to find it.

Ask.com

Ask.com (originally known as Ask Jives) is a question answering-focused web search engine founded in 1995 by Garrett Gruener and David Warthen in Berkeley, California. The original software was implemented by Gary Chavsky from his own design Warthen, Chavsky, Justin Grant, and other built the early AskJeeves.com website around that core engine. It may provide textual answering of related post question.

PROPOSED SYSTEM:

In this paper, proposed system will gives answers for the question answering in any one of the following media formats as selected by the user based on the question enters:

 ISSN: 2321-8169
203 - 208

(a) Only text : it means that original textual answer are sufficient
(b) Text+ image: it means that textual information is not sufficient to user so image information must be added
(c) Text +video: it means that text and information and video data must be added
(d) Text+image+video: it means that we add both image and video information

As per this design we have proposed algorithm for selecting the approximate multimedia data with corresponding answer. In this paper, we propose scheme which can enrich community a part of textual answer in cQA with appropriate media data.

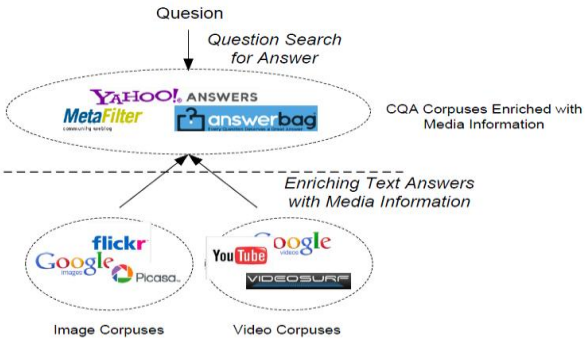

Figure 3: Accomplish MMQA by Enriching Large CQA corpuses with Media Information

IV. ALGORITHMIC STRATEGY

In machine learning, support vector machine are supervised learning models with associated learning algorithm that analyzed data recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of categories, SVM training algorithm build a model that assign new example into one category or other, making it a non-probabilistic binary linear classifier. An SVM model is representation of the example as point in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New example are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVM's can efficiently perform a non-linear classification using what is called as kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Incremental Algorithm

In the community question answering website with multimedia using following list of algorithm

- Core sentence extraction from question
  Stemming and stop-word removal on answer
- Question type based on answer medium (Naïve Bays)
- Head word extraction
- Query generation
- Query selection
  - POS feature extraction
  - Search performance prediction
  - Media resource analysis
  - Clarity score based on KL divergence
- Multimedia data selection and prediction
  - Graph based re-ranking
  - Face detection algorithm
  - Feature extraction from image
  - Key frame identification and extraction

Math Model:

We predict search performance based on the fact that, most frecuntly search result are good if the top result are quite coherent. We adopt the method proposed which define clarity score for query and collection language model i.e.

$$\text{Clarityq(Ci)} = \sum_{w \in Vci} P(w|\theta q)\log 2 \; P(w|\theta q)/P(w|\theta Ci)$$

Where vci is the entire voccabulary of the collection ci and i=1,2,3 repesent text,image,video respectivly. The term P(w/θq) and P(w/θci) are the query and collection language model, respectivly. The clarity value becomes smaller as top ranked document approch a random sample from the collection.

V. IMPLEMENTATION STRATEGY

Modules:

1. Answer Medium Selection.
2. Query Generation for Multimedia Search.
3. Multimedia Data Selection and Presentation.
4. Re-ranking.

Input and Output Design:

Input Design:

The input design is the link between the information system and the user. Without any input the system or any application does not work, so in that the pair of question is input of that system so it must be clear and understandable. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form of processing can be achieved by inspecting the computer to read data a written or printed document or it can be occur by having people keying the data directly into the system. The design of input focus on controlling the amount of input required, controlling the error avoiding delay, avoiding extra steps and keeping the process simple.

Excepted Output

A quality output is one of the most important features, which meets the requirement of the end user and presents the information clearly as per the user need. In any system result of processing are communicated to the user and to other system through outputs. In output design it is determine how the information is to be displaced for immediate need and also the hard copy of the output. It is

 ISSN: 2321-8169
203 - 208

most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making and provide approximate solution to the user.

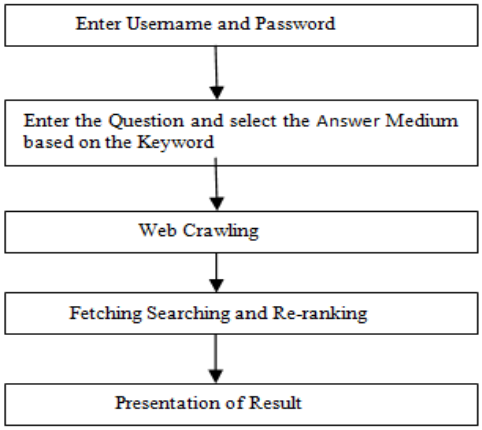

Figure 4. Block Diagram for Input and Output

VI. CONCLUSION

In this paper, we describe the inspiration and evolution of "Multimedia Question Answer System" for retrieving question answer from the available site it may be difficult to you, because of all the available Community Question Answering Sites are provide data from the unstructured database and convert into normal language so in the searching process they provide lots of solution of one question and user need to select one of them it may be difficult to the user to avoid the duplication of data. So that we introduced such application that solves that problem in our application system must be search all relevant data of given question and then removal all the duplication and re-ranking on them and provide appropriate solution from the given database. Community Question Answer for a given Question Answering pair must be selects first which type of medium is appropriate for enriching the original textual answer. Following that it is automatically generate a query based on the QA knowledge and then performs multimedia search with the query but does not provide proper solution of multimedia it just provide URL or links to enrich them. They are many different from of Multimedia data research that aims to automatically generate multimedia answer with given question, but an application system approach is built and based on the community answers, and it can deal with more common question and achieve better performance.

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] Liqiang Nie, Meng Wang, Yuegao, Zheng-Jun Zha, and Tat-Seng Chua, "*Beyond Text QA: Multimedia Answer Generation by Harvesting Web Information*,"IEEE rans. Multimedia, vol. 15, no. 2, Feb. 2013.

[3] B. Campbell, J. Rosenberg, H.Schulizrine, C. Huitema, and D. Gurle session initiation protocol (sip) extension for instant messaging, RFC 3428, 2002.

[4] P. Saint-Andre, Interdomain presence scaling analysis for the extensible message and presence protocol (xmpp), RFC internet draft, 2008.

[5] W. E. Chen Y. B Lin and R. H Lieu A weekly consistent scheme for imspresence services IEEE Transaction on Wireless Communication, 2009

[6] N. Banerjee, A. Acharya and S. K Das, Seamless sip based mobility for multiple media application IEEE network vol. 20 no. 2, pp. 618, 2006

[7] Liqiang Nie, Meng Wang , Member IEEE Yue Goo, Zheng-Jun Zha, Member IEEE Tat-Seng Chua senior member of IEEE"*Multimedia Answer Generation by Harvesting Web Information*"IEEE Transactions on multimedia vol.15 no.2 Feb 2013

[8] Richang Hong Meng Wang, China Guangda Li, Liqiang Nie, Zheng-Jun Zha, and Tat-Seng"*Multimedia Question Answering".*

[9] Apache Lucene 4 Andrej Bialecki, Robert Muir, Grant Ingersoll Lucid Imagination.

Firoz I. Kadri he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of Coding and Programming .

Kaushal A. Kapse he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of Network security .

Nitin V. Landge he is Engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of Testing & Information security .

 ISSN: 2321-8169
203 - 208

Hemant B. Shirsathhe is engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under University of Pune. His interest in the field of Database & UI Design.

V. D. Badgujar, BE Computer Engg. Was educated at Pune University. And perceiving M.tech at RGPV University, Bhopal (M.P) Presently he is working as Professor Computer Technology Department of Brahma Valley College of Engineering and Research Institute, Nasik, Maharashtra, India. He has presented papers at National and International conferences and also published papers in National and International Journals on various aspects of Computer Engineering and Networks. His areas of interest include Computer Software's Security and Advance Database.