International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 7 (August 2014) http://ijirae.com

Multimedia Answer Generation from Web Information

Avantika Singh
Information Science & Engg,
M S RIT-Bangalore

Abhimanyu Dua
Information Science & Engg,
M S RIT-Bangalore

Gourav Patidar
Information Science & Engg
M S RIT-Bangalore

Pushpalatha M N
Information Science and Engg,
M S RIT-Bangalore,

Abstract— Community Question Answering (cQA) services have gained a lot of popularity over the last few years. They allow people with diverse backgrounds to share their knowledge and experiences. There are a number of community based online services like Yahoo! Answers, Wiki Answers, where people answer questions posted by other people. Not only does it allow the community members to post and answer questions but on the other hand it also enables the users to seek information from a comprehensive set of well-answered questions. However, the existing cQA forums usually provide only textual answers, which are sometimes not informative enough. So here we propose a system, which can enrich community-contributed textual answers in cQA with appropriate media data. In simple words, Multimedia Answer Generation form Web information will answer questions in different media formats (text, video, and image) as selected by the user.

Keywords— cQA, Question Answering, Natural language processing, Media Selection

INTRODUCTION

Question Answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language. When compared to keyword based search systems, they are way better as they greatly facilitate the communication between humans and computer systems by stating the user's intentions through plain sentences. It will also avoid the need of painstaking browsing of a vast quantity of information returned by the search engines for the correct answers.

However fully automated QA systems face a number of challenges that are not easy to tackle, for example:

- The lack of deep understanding of complicated questions.
- The lack of sophisticated semantic, syntactic and conceptual processing to generate answers.

Basically, we found that automated QnA (Question and Answer) forums cannot obtain results that are as good as those generated by human intelligence.

This is where **Community Question Answering Services** came into the picture

Through "Community Based Question Answering" forums, people can seek answers to questions that belong to different categories and can also share their knowledge on any specific problem which is of interest to some other user. Community Based Question Answering forums give better answers to questions because unlike automated answering systems, they are based on human intelligence.

A huge amount of question and answer pairs have been accumulated in the repositories over the years. For example –Wiki Answers – one of the most well known hosts more than 13 million well answered questions in 7000 different categories (as of 2011).

Some other examples of "Community Based Question Answering" sites are Ask.com, Tutorialpoint.com, Indiabix.com, Youtube.com, Yahoo Answers.com etc.

Our system contains four main components:

- User Module – Login and Answer medium selection.
- Web Crawler
- Indexing and Searching using Apache Lucene
- Result Presentation

Problem Definition

The existing cQA forums mostly support only textual answers. Unfortunately, textual answers may not provide sufficient natural and easy-to-grasp information. The answers are described by long sentences which generally makes it very tedious to interpret. Clearly, it will be much better if there are some accompanying videos and images that visually demonstrate the process or the concept. In the existing system users usually post URLs that link to supplementary images or videos in their textual answers. Therefore we can conclude that in a way the existing cQA forums do not provide adequate support in using media information.

Existing System

Question Answer forums, today are of paramount importance to mankind. Community Question answering forums like Yahoo! Answers and Wiki Answers are very popular means of information seeking on the web. By posting questions for other participants to answer, the information seekers can obtain specific answers to their questions. These sites have been gaining a lot of popularity and are growing rapidly. Users of popular portals like Yahoo! Answers have submitted millions of questions and have also received millions of answers from other participants.
These question answer forums give us a way to:

- Request information that we do not know
- Check information that we are not sure of
- Interact with people around us and see things from their perspective and they also help us to
- Fulfill our needs at the same time.

Proposed System

Our proposed application will give answers for the questions in any one of the following media formats as selected by the user based on the question he/she enters:
(a) Only text: It means that the original textual answers are sufficient
(b) Text + image: It means that image information needs to be added
(c) Text + video: It means that only video information needs to be added
(d)Text + image + video: It means that we add both image and video information
As per the design we have proposed an algorithmic approach for selecting the accurate video, image and text for the corresponding answers -
We have named it as "Multimedia answer generation from web information"

Literature Survey

answering (QA) is a technique for automatically answering a question posed in natural language. Compared to keyword-based search systems, it greatly facilitates the communication between humans and computer 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 [1].
One definition of a question could be 'a request for information'. But how do we recognize such a request? In written language we often rely on question marks to denote questions. However, this clue is misleading as rhetorical questions do not require an answer but are often terminated by a question mark while statements asking for information may not be phrased as questions. For example the question *"What cities have underground railways?"* could also be written as a statement *"Name cities which have underground railways"*. Both ask for the same information but one is a question and one an instruction. People can easily handle these different expressions as we tend to focus on the meaning (semantics) of an expression and not the exact phrasing (syntax). We mainly focus Definition questions, which unlike factoid questions require a more complex answer, usually constructed from multiple source documents [2].
Retrieving the data for constructing an answer to the question is done by means of WebCrawlers. A Crawler is a program that downloads and stores web pages, often for a web search engine. Roughly, a crawler starts off by placing an initial set of URLs, so, in a queue, where all URLs to be retrieved are kept and prioritized.
From this queue, the crawler gets a URL (in some order), downloads the page, extracts any URLs in the downloaded page, and puts the new URLs in the queue. This process is repeated until the crawler decides to stop. Collected pages are later used for other applications, such as a Web search engine or a Web cache. As the size of the of the web grows, it becomes more difficult to retrieve the whole or a significant portion of the web using a single process. Therefore, many search engines often run multiple processes in parallel to perform the above task, so that download rate is maximized. We refer to this type of crawler as a parallel crawler [3].

Implementation

The functional aspects of the device are mainly broken into major components.
1. First, we will input the question from the user.
2. In the next step, the user will choose the correct answer medium based on the question.

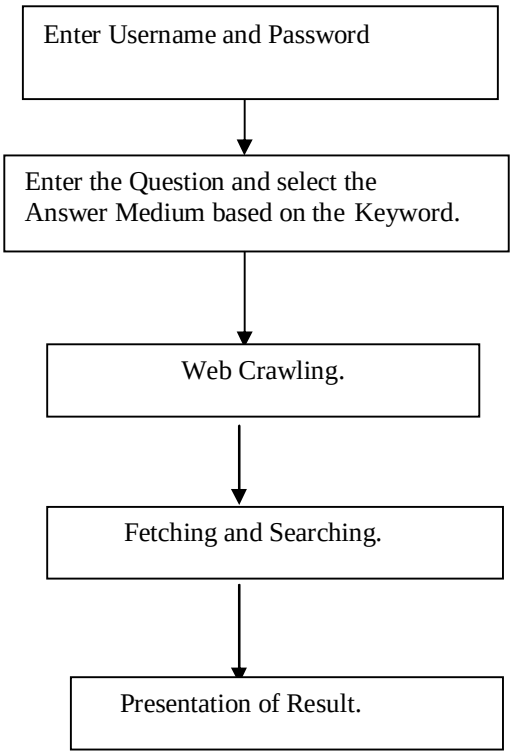
- Only text
- Text+Image
- Text+Video
- Text+Video+Image

3. Web Crawling and extracting the URL's.
4. Parsing through the data of the fetched pages.
5. Using Apache Lucene as a document Indexing and Searching Mechanism.
6. Result Presentation.

Initial Step-By-Step Description:

1. The user will first login into the system and Input the Query.
2. The system will accept the query and then the user will have to choose an appropriate answer medium based on the question.
3. The process of Web Crawling is done on the Web Pages containing the relevant data.
4. Parsing of data is performed on the pages that are visit able.
5. Apache Lucene is used as a searching mechanism.
6. Presentation of results

Block Diagram with Input and Output



Results

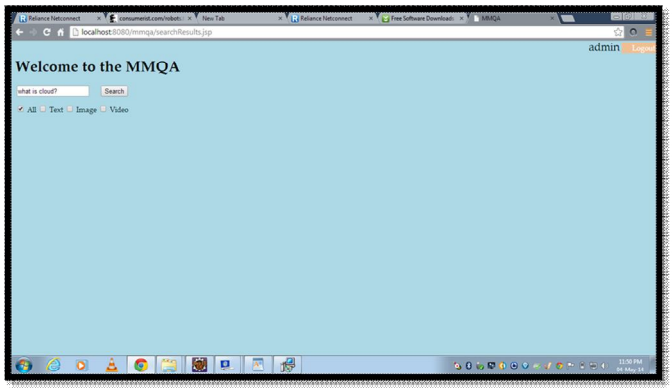


Figure 1: DoSearch-Enter Query

The doSearch.jsp is the page which is displayed as soon as the username and password have been authorized. The user is expected to enter the question in the Search bar, select any one of the following options:

- All
- Text
- Image
- Video
- Text+Image
- Image+Video
- Text+Video
- Text+Image +Video

as the desired "Answer Medium" and press the Search Button, which begins the Searching Process.

Figure 2: Presents Answer in All formats -Text, Image, Video

CONCLUSION

In this paper, we describe the inspiration behind the making and evolution of "A Multimedia Answer Generation System". As it is analyzed that the existing approaches mainly focus on narrow domains, therefore aiming at a more general approach, we developed the system to answer questions using media data thereby making the answers much more informative and easy to understand. The answers obtained are very relevant and to the point.

For a given QA pair, the user first predicts which type of medium is appropriate for enriching the original textual answer and selects the desired answer medium he/she wants. Following that, our system crawls the web and retrieves web pages of the concerned content type and stores them in the frontier for retrieval later on.

Further on, by using apache lucene we search the indexed documents to obtain the ones matching the keywords extracted from the question and display the desired results.

On the other hand we have also observed some failure cases. For example, the system may fail to generate reasonable multimedia answers if the generated queries are very complex. For several questions videos are enriched, but actually only parts of them are informative and so, presenting the whole videos can be misleading. Another problem is the lack of diversity of the generated media data, which gives a scope for future enhancement of the system.

All in all, we can safely conclude that since our system is based on community contributed answers, it can thus deal with more general questions and can achieve better performance as the answers give us the required information through the desired content type. Therefore we have successfully achieved our goal of obtaining an efficient "Multimedia Answer Generation System".

REFERENCES

[1] Liqiang Nie, Meng Wang, Member, IEEE, Yue Gao, Zheng-Jun Zha, Member, IEEE Tat-Seng Chua, Senior Member, IEEE, "Beyond Text QA: Multimedia Answer Generationby Harvesting Web Information", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 2, FEBRUARY 2013

[2] Mark Andrew Greenwood "Open-Domain Question Answering "

[3] Parallel Crawlers - Junghoo ChoUniversity of California, Los Angeles,cho@cs.ucla.edu and Hector Garcia-Molina Stanford University cho@cs.stanford.edu

[4] Richang Hong, Meng Wang, ChinaGuangda Li, Liqiang Nie, Zheng-Jun Zha, and Tat-Seng, "Multimedia Question Answering"

[5] Web-Crawled Academic Video Search Engine Dongwon Lee, Hung-sik Kim Eun Kyung Kim Su Yan Johnny chen Jeongkyu Lee+ Penn State University, USA + University of Bridgeport, USA{dongwon, hungsik,ezk112,syan,jzc160}@psu.edu+jelee@bridgeport.edu

[6] Modeling Community Question-Answering Archives - Zainab ZolaktafFaculty of Computer Science Dalhousie University zolaktaf@cs.dal.ca , Fatemeh Riahi Faculty of Computer Science Dalhousie University riahi@cs.dal.ca, Mahdi Shafiei Faculty of Computer ScienceDalhousie University shafiei@cs.dal.ca, Evangelos Milios Faculty of Computer Science Dalhousie University
[7] Apache Lucene 4 Andrzej Białecki, Robert Muir, Grant Ingersoll Lucid Imagination {andrzej.bialecki, robert.muir,grant.ingersoll} @lucidimagination.com
[8]http://www.oracle.com/us/corporate/acquisitions/sleepycat/index.html
[9] Wikipedia
[10]https://lucene.apache.org/core/4_2_1/queryparser/org/apache/lucene/queryparser/classic/package-summary.html

