QUESTION ANSWERING SYSTEM FOR VOICE BASED SEARCH USING NLP TECHNIQUES AND WEB SNIPPETS

Maheshwari K B
Student,IT Department,
Dr. Mahalingam college of Engineering and Technology, Pollachi.
maheshhwari97@gmail.com

Nandhinipriya J
Student,IT Department,
Dr. Mahalingam college of Engineering and Technology, Pollachi.
nandhinijoseph97@gmail.com

Shantha Lakshmi M
Student,IT Department,
Dr. Mahalingam college of Engineering and Technology, Pollachi.
. ammu731997@gmail.com

Menaha R
Assistant Professor, IT Department,
Dr. Mahalingam college of Engineering and Technology, Pollachi.
rmenahasenthil@gmail.com

***Abstract*— Question Answering System (QAS) is an information retrieval technique which provides a descriptive answer for the given question rather than web links. QAS has many applications like retrieving information from the web, online examination, education, health care, sports, and geography. For a search query the search engines like Google, Yahoo provides a list of web links as a result. Especially for the voice based search query, the Google returns the web links as well as the audio as results in English language. Though the search engine returns web links for all language search queries, the audio is given as output only for English language search queries. In order to address this problem, this paper provides a solution to provide an audio as output for all kind of search queries by using snippets of the web document. Web pages retrieval, Snippets Extraction, Answer Filtering and Text to speech conversion are the processes involved in this proposed system. For Experimental purpose, the questions related to technology, sports, biology, politics, science are queried in the Google and expected answer for the question is provided in the form of audio in our proposed system.**

Keywords— Question Answering system, Information Retrieval, Web snippets

I. INTRODUCTION

Question answering system (QAS) is dealing with the fields related to information retrieval and natural language processing (NLP) which provides an descriptive answers to the questions posed by the humans in natural language. The QAS provides answers for factoid, non factoid and boolean type of questions. But mostly current question answering system focus on factoid questions. The search engines like Google, Yahoo provides a list of web links as the results for the given query. As search engine provides answers in web links people move towards QAS as it provides descriptive answer in a minimal time.

The proposed system of our project provides answers to open domain questions in the mode of audio format. Web pages extraction, Snippets extraction, Answer identification and text to speech conversion are the processes involved in this proposed system. Then user gives the question in the user interface, the question is searched in the search engine and related web pages are retrieved from it. Web pages are stored in local files, from the web pages snippets are extracted and stored in local files. Then identifies the related answer from the web snippets and provide to the user in the form of audio as well as text.

The rest of the paper is organized as follows: Section 2 discusses some related works and Section 3 presents the design of the proposed system. The experimental results conducted are presented in Section 4 and Section 5 concludes the work and provides hints for future extension of work.

II. RELATED WORK

From the survey analysis, The QAS is classified into three domains. They are general domain, restricted domain and language dependent. The general domain or open domain [3.2] provides nearly about anything; it can provide answers for all

type of questions relies on general ontologies and world knowledge. The restricted domain or closed domain [3.1] deals with questions under a specific domain and provides a correct answer to the questions using a pre-structured database or natural language documents. Also, the closed domain might refer to a situation where only limited types of questions are accepted. Finally the language dependent QAS [3.3] which is purely dependent on the language they have used, even when we host the question in natural language.

A. Closed Domain QAS

Jaimie Kelley [10] discusses the quick response for the interactive questions but even the slow component can give the exact answers so they introduced an Ubora a design approach which can sample online queries for the second time. Lei Pang [15] presented a multimedia question answering because while a question is posted it seems comfortable for the user to understand the content if it is audio-visual than a text format. Even the question may invoke the emotional response either positively or negatively. Piero Molino [21] says that posting questions in a platform and receiving answers for it is a usual mechanism. They gather contents from five different environments and make the user comfortable with the different accurately predicting answers from larger datasets.

Aditya Kalyanpur [1] speak about decomposing factoid questions which give multiple independent facts. It is also categorized as novel decomposition framework for parallel questions which has a re-ranking component, sub queries. Michael Spranger [19] discussed based on state-of-the-art techniques that form the basis for QA system used in the field of criminal proceedings. Benefits and capabilities of special crime ontology is discussed for applying computer linguistic methods on forensic texts. Bing Zhang [8] development and design of on-line interactive QAS has provided a rather good platform for communicating sports knowledge for the users, which possesses a theoretical and practical value of sports knowledge on the network G.

Suresh Kumar [26] proposed a method to automatically construct domain ontology and to extract the concept relations from unstructured text using a syntactic and semantic probability-based Naïve Bayes classifier and it was evalved using benchmark data sets. Varsha Bhoir [28] returns short string answers or list to natural language questions related to tourism domain. Sharvari Gaikwad [24] they focus on the need for a robust domain specific question answering system targeting agriculture domain. This will help to farmers to get information about their queries related to agriculture.

Sweta P. Lende [27] describes the different method and implementation details of question answering system for general language and proposes the closed domain QA System for handling documents related to education acts sections to retrieve more precise answers using NLP techniques. Li Liu [17] introduces the mechanism of IQA, proposes a domain Ontology-based IQA framework and an incomplete question knowledge representing framework question information domain (QID).

B. Open Domain QAS

The general domain related work includes Aditya Pal [2] say about the novel problem of routing a question in a right environment (i.e.,) three main entities question, user and communities that are in right community. They introduced a knn algorithm that is a natural inclination of cut-off aggregation algorithm for ranking the answers in an order. Beomseok Hong [7] speaks about the time lag between the question and answer. So, they proposed a weighted question retrieval model that uses question titles, relationship and description for calculating question similarity in large-scale community question answering (CQA).

Melanie Herschel [18] monotonic queries (multiple perspective questions) posted by the user who can get answers to the missing questions. Also, used an algorithm called Conseil algorithm is the first hybrid algorithm. Walter S. Lasecki [30] about the blind users who can ask their question in a VizWiz platform where they can get the exact answers in a single interaction in a crowdsourcing workflow. Amit Mishra [3] early QAS is developed for restricted domains with limited capabilities. In this paper, they had done the surveyed on various QASs and based on criteria they is classified the QASs. Liang Zhenqiu [16] presented an answering system based on case-based reasoning and its main idea is to automatically retrieve existing and valid historical cases to answer the new questions.

Sa Id Alami Aroussi [22] improve a precision of Question answering systems by focusing namely on the representation of the question itself as a bag of concepts, using the Explicit Semantic Analysis (ESA). Jarang Kim [9] proposed Tree Pattern Expression (TPE) based tuple extraction method for high performance Q&A system. Wenpeng Lu [29] summarizes the classification, implementation and evaluation of question answering system (QA). Shrimai Prabhumoye [25] presents a question answer search engine prototype that uses natural language processing, natural language generation, question classification and query logs to find a precise answer to the submitted query. They describe their strategy of automatic query analysis by classifying it into nine categories and understanding the meaning of the query.

C. Language Dependent QAS

The language dependent QAS includes Archana S.M [5] they had denoted the relation to the verb or the name like these changes are made using Vibhakthi and POS tags. They focused on factoid type question answering. This paper is important in Malayalam NLP related works. Junichi Fukumoto [13] selects a clue word to decide a proper topic from retrieved documents and narrow down search space to get a proper answer using user interaction. Search space would be reduced using clue word. Jawad Sadek [11] deals with factoid questions detecting noun phrases in the text. Also, dealing with "why" and "how to" questions to present the Arabic text parser for question

answering system. Aqil M. Azmi [4] he speaks about QAS which is in Arabic language which is little difficult. They handle with the difficulty in "why" questions for Arabic using two different approaches and worked it for 100 question answer pairs.

Due to the work done related to the question answering domain, it does not provide input and gets output in the form of speech. The Google search engine provides answers for all type of search queries but in case of Google audio search occurs some sort of difficulties. Some audio search does not get right results and for some queries only audio gets reached but not with the text. For few other queries language becomes an issue and some does not give any results. So, with the available data's in Google search engine we overcome this problem.

III. PROPOSED SYSTEM

In our proposed system, the input is given in the form of text it searches in the Google search engine for the accurate results which are said as a web information retrieval. And it passes to an answer identification process where the extracted web pages are made to store on a local disk. Snippets extraction is a process where the snippets are extracted from the stored web pages using the HTML tags used. Then answer filtering is the process of filtering the answer from the web snippets. The text to speech conversion is done by passing the filtered answer text as an input and providing the answer in the form of audio.

The indexing techniques is used in snippets extraction module for extracting the required snippets from the entirely stored web pages. And, NLP techniques is used in answer filtering module which help for filtering the answers from web snippets by comparing the input with the retrieved snippets and the made to rank it according to maximum number of matches.

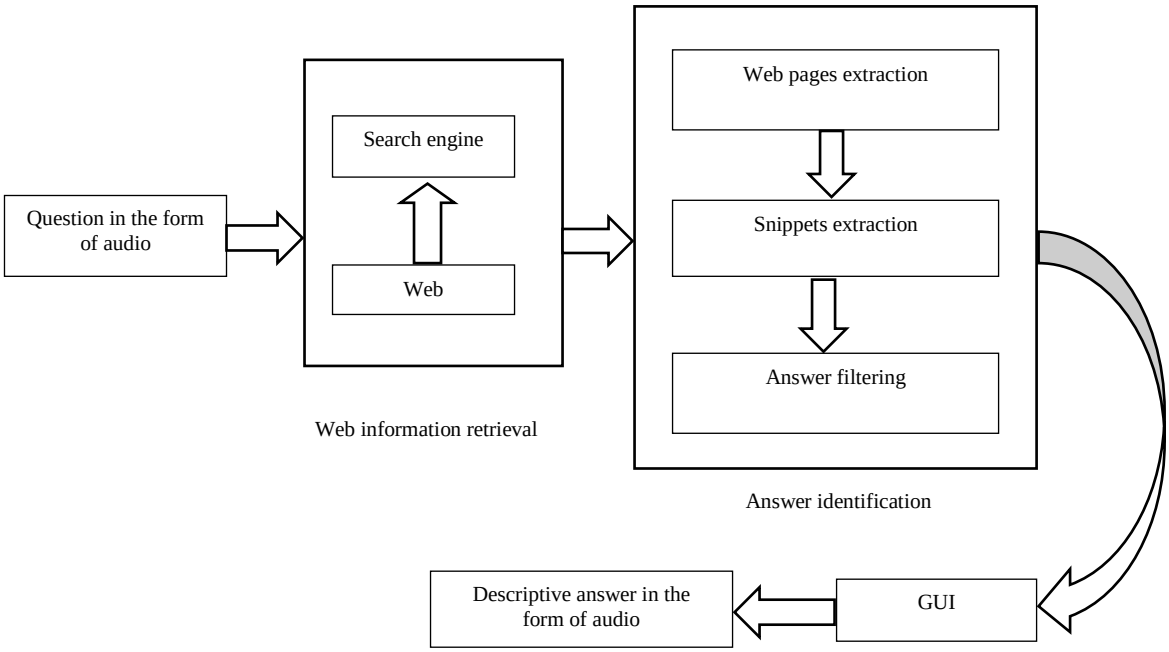

Fig. 1. Proposed system architecture

A. Speech to Text conversion

The speech to text conversion is providing the input in the form of speech and receiving the output in text format. This module involves in a conversion of speech to an text by using JAR files which helps to convert speech into the text and pass it as an input in the user interface.

B. Web pages extraction

Initially the user gives the question in the user interface. In this module, the question will be connected and then searched in the Google search engine. When searched, the web page will be extracted and a dialog box appears in the user interface. The first page of the Google web page for the given

question by the user will be retrieved. Finally the retrieved web pages are stored in the local files.

C. Snippets Extraction

From the retrieved web pages, the snippets are extracted using the indexing technique and is stored in the local files. Using line separator, snippet lines are separated as a single line for each link in web pages. In the web pages, only the desired information will be stored and retrieved. All the unwanted information like web links, pictures, and videos will be removed with the help of HTML tags like span class, div, em, strong.

Data mining is the analysis step of the knowledge discovery in databases process, or KDD. The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.

Fig. 2. Snippet for Data mining query

D. Answer Filtering

Answer filtering is a process of filtering the accurate answer for a given query from the extracted web snippets. NLP technique is used to filter the relevant answer from the extracted web snippets. This module involves in identifying the answer from the snippets which has been extracted from the Google web pages and stored in the local disk. Then all the stop words in the questions are removed and are made to compare with the snippet in a text file. By comparing all the related snippets with the question, when a large number of words gets matched then that snippet will be filtered and displayed as the relevant answer. Then that snippet will be displayed to the user within the specified answer window.

E. Text to Speech Conversion

The text to speech conversion is providing the input in the form of text and receiving the output in audio format. This module involves in a conversion of text into the form of an audio using JAVA. This is done by the filtered answer snippet in the answer filtering process. It takes the input in the form of text and it converts it into a speech using the imported JAR files and that audio will be presented to the user with the default voice in the system.

IV. EXPERIMENTS

The input is provided in the form of audio using the java synthesizer/JSyn as a JAR files to the user interface and it processes to convert the audio into the form of text where the text is made to search via search engines like Google, Yahoo etc., The web pages related to the given input query is made to search in search engines and the first page of search engine gets extracted and is stored in the form of the snippets, HTML code, web links. Then from the retrieved web pages, the web snippets are extracted separately by using the HTML tags which divides other links, pictures and videos from the snippets with NLP techniques and stored in the local disks

The line separator is used to separate each line as a link in the snippets also remove the gap at the top of the snippet file occurring for some web pages. From the extracted snippets, answer filtering process is done by fetching the words from the given input query after removing the stop words. Then compare it with all the snippets and rank each snippet according to the number of words matched with the words in the input query. And takes the maximum matched snippet as an relevant answer and display it in the user interface as an output. The speech to text conversion is then made by importing the JAR files and provides the answer in interface as an input into the text to speech conversion synthesizer process to convert the text into an audio with the default voice used in the system this system uses voice of “kevin” as an audio output.

S.No.	QUESTION	DOMAIN	ANSWER
1.	Who is the chief minister of kerala?	Politics	Pinarayi Vijayan. Pinarayi Vijayan (born 24 May 1945) is an Indian politician who is the current Chief Minister of Kerala, in office since 25 May 2016.
2.	What is siddha medicine?	Culture	Siddha Medicine (Tamil:□□□□□ □□□□□□□□□□ Citta- or Tamil-maruttuvam □□□□□ / □□□□□ □□□□□□□□□□)is a system oftraditional medicine originating in ancient Tamilakam in South India. Traditionally, it is taught that the siddhars laid the foundation for this system of medication
3	What is the national game of India?	Games	A recent RTI has revealed that **I**ndia has no national game. Earlier, field hockey (a sport India won eight Olympic gold medals) enjoyed the special status. But, how can we, a country of more than a billion people, stay without a national game.

4	Which is the highest profitable movie in India currently?	Movie	Bollywood's Most Profitable Films Of 2017 - Baahubali 2 Hindi, Badrinath Ki Dulhania, Kaabil and Jolly LLB 2 have managed to make it to the profit zone.
5	Which soil suits to sue the peanut?	Agriculture	The suit accuses a school district of endangering an elementary student who suffers from severe peanut and tree nut allergies.
6	Does WhatsApp app able to be worked on computer?	Social media	To get WhatsApp on your computer **you** first need to download and install BlueStacks App Player. This is a free program that emulates Android applications on your PC. ... If you already have WhatsApp installed on your phone it won't work, since you can only run one instance of the app per phone number.
7	Is god really exist?	Sprituality	If, on the other hand, I were neutral, and didn't already have an "a priori adherence" to a particular worldview (be it naturalistic or otherwise), the question "does God really **exist**?" wouldn't be pointless at all.
8	Illustrate perfect personality?	Personality Traits	Personality questionnaires You might have come across psychological questionnaires in newspapers and magazines. These are usually not professionally developed questionnaires, but rather a selection of questions put together to illustrate a particular issue.
9	Who invented radio?	Physics	Guglielmo Marconi: an Italian inventor, proved the feasibility of radio communication. He sent and received his first radio signal in Italy in 1895.
10	How much silver does gold contain to make jewels?	Jewellery	Information about silver jewelry; facts concerning metal alloys used to make gold accessories. Answers to ... if you have copper sensitive skin, because old European silver is .800 fine, or 80% silver/ 20% copper. Following is a listing of ... It does not matter what type of metal is "mixed" with the gold, just how much.

V. CONCLUSION

The proposed approach of an open domain question answering system using NLP techniques and web snippets is done by collecting questions from different domains and verified the accurate results. Each question is passed through the Google search engine for retrieving the web pages after pre-processing. From those web pages snippets were extracted using the indexing technique and is made to store in the local disk. Usually the snippet contains the statements which give brief introduction about that document. Sometimes those snippets may not be related to the given question. So NLP technique is used, to identify the accurate answer for the given question. The following Table evince the results of our open domain question answering system.

References

[1] A. Kalyanpur, S. Patwardhan, and B. Boguraev, "Fact-Based Question Decomposition for Candidate Answer Re-Ranking," ACM 978-1-4503-0717-8/11/10, 2011.

[2] A. Pal, "Metrics and Algorithms for Routing Questions to User Communities," ACM Transactions on Information Systems, vol. 33, no. 3, Article 14, March 2015.

[3] A. Mishra, S.K. Jain, "A survey on question answering systems with classification," Revised 9 May 2014; accepted 23 October 2014 Available online 2, November 2015.

[4] A.M. Azmi and N.A. Alshenaifi, "Answering Arabic Why-Questions: Baseline vs. RST-Based Approach," ACM-vol. 35, no. 1, Article 6, 2016.

[5] S.M. Archana, N. Vahaba, R. Thankappana , C. Raseekb, "A Rule Based Question Answering System in Malayalam corpus Using Vibhakthi and POS Tag Analysis," Procedia Technology 24 1534 – 1541, 2016.

[6] A.B. Abacha, P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," Information Processing and Management 51 (2015) 570–594. *(references)*

[7] B. Hong, Y. Kim, "A Weighted Question Retrieval Model using Descriptive Information in Community Question Answering," ACM ISBN 978-1-4503-4455-5/16/10. . . $15.00 c. 2016.

[8] B. Zhang, L. Lei, "Design and Research of On-line Interactive Q & A System on Sports Expertise," Energy Procedia 17 (2012) 373 – 378.

[9] J. Kim, K. Kim, Y.S. Choi, "TPE based Tuple Extraction Method for Question and Answering System," 978-1-4799-4441-5/14/$31.00 ©2014 IEEE.

[10] J. Kelley, C. Stewart, Y. He, S. Elnikety, "Obtaining and Managing Answer Quality for Online Data-Intensive Services,**"** ACM Trans. Model. Perform. Eval. Comput. Syst., vol. 2, no. 2, April 2017.

[11] J. Sadek, F. Meziane, "A Discourse-Based Approach for Arabic Question Answering,**"** ACM-vol. 16, no. 2, Article 11, 2016.

[12] Jovita, Linda, A. Hartawan, D. Suhartono, "Using Vector Space Model in Question Answering System," Procedia Computer Science 59 (2015) 305 – 311.

[13] J. Fukumotoa, N. Aburaib, R. Yamanishia, "Interactive Document Expansion for Answer Extraction of Question Answering System," 17th International Conference in Knowledge Based-KES2013.

[14] K. Vila, J.N. Maz´on, A. Ferr´andez, "Model-driven adaptation of question answering systems for ambient intelligence by integrating restricted-domain knowledge," Procedia Computer Science 4 (2011) 1650–1659.

[15] L pang, C.W. Ngo, "Opinion Question Answering by Sentiment Clip Localization,**"** ACM Trans. Multimedia Comput. Commun. Appl., vol. 12, no. 2, Article 31, November 2015.

[16] L. Zhenqiu, "Design of Automatic Question Answering System Base on CBR," Procedia Engineering 29 (2012) 981 – 985.

[17] L. Liu, Q. Qi, F. Li, "Ontology-based Interactive Question and Answering System," 978-1-4244-5143-2/10/$26.00 ©2010 IEEE.

[18] M. Herschel, "A Hybrid Approach to Answering Why-Not Questions on Relational Query Results," ACM, vol. 5, no. 3, Article 10, 2015.

[19] M. Spranger, D. Labudde, "Establishing a Question Answering System for Forensic Texts," Procedia - Social and Behavioral Sciences 147 (2014) 197 – 205.

[20] O. Popova, B. Popov, V. Karandey, M. Evseeva, "Intelligence Amplification via Language of Choice Description as a Mathematical Object," Procedia - Social and Behavioral Sciences 214(2015) 897 – 905.

[21] P. Molino, L.M. Aiello, P. Lops, "Social Question Answering: Textual, User, and Network Features for Best Answer Prediction," ACM Transactions on Information Systems, vol. 35, no. 1, Article 4, September 2016.

[22] S.A. Aroussi, N.El. Habib, O.El. Beqqali, "Improving Question Answering Systems by using the Explicit Semantic Analysis method," 978-1-5090-5781-8/16/$31.00 c 2016 IEEE.

[23] S. Pudaruth, K. Boodhoo, L. Goolbudun, "An Intelligent Question Answering System for ICT," 978-1-4673-9939-5/16/$31.00 ©2016 IEEE.

[24] S. Gaikwad, R. Asodekar, S. Gadia, Vahida Z. Attar, "AGRI-QAS Question-Answering . System for Agriculture Domain," 978-1-4799-8792-4/15/$31.00c 2015 IEEE.

[25] S. Prabhumoye, P. Rai, S. Sowmya Kamath, "Automated Query Analysis Techniques for Semantics based Question Answering System," 978-1-4799-4989-2/14/$31.00 © 2014 IEEE.

[26] G. Suresh kumar, G. Zayaraz, "Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems," Revised 18 November 2013; accepted 13 March 2014 available online 9,May 2014.

[27] S.P. Lende, Dr.M.M. Raghuwanshi, "Question Answering System on Education Acts Using NLP Techniques," 978-1-4673-9214-3/16/$31.00 © 2016 IEEE.

[28] V. Bhoir, M.A. Potey, "Question Answering System : A Heuristic Approach," 978-1-4799-2259-14/$31.00©2014.

[29] W. Lu, J. Cheng, Q. Yang, "Question Answering System based on Web," 978-0-7695-4637-7/12 $26.00 © 2012 IEEE.

[30] W.S. Lasecki, Y. Zhong, J.P. Bigham, "Increasing the Bandwidth of Crowdsourced Visual Question Answering to Better Support Blind Users," ACM 978-1-4503-2720-6/14/10.

[31] www.javasamples.com/showtutorial.php?tutorialid=1202

[32] https://www.youtube.com/watch?v=swuYhvwHw9w