ISSN NO. 2320-5407

Journal homepage:http://www.journalijar.com
Journal DOI:10.21474/IJAR01

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

QUESTION ANSWERING SYSTEM WITH REPEATED QUESTION IDENTIFICATION.

Maria Vijoy and Sangeetha Jamal.
Department of Computer Science, Rajagiri School of Engineering and Technology, Kochi, India.

Manuscript Info

Manuscript History:

Received: 12 May 2016
Final Accepted: 26 June 2016
Published Online: July 2016

Key words:
Question answering; Semantic pattern; NLP ; Repeated Answering, conceptual graph formalism

Corresponding Author

Maria Vijoy.

Abstract

This paper proposes a method to introduce the complexity of finding the solution to a given question. Question answering system is being currently used in various forms. But due to the increase in the number of users using these system, the chance of repeated searching increases.. The system implements an automatic system for answering repeated questions based on semantic question similarity. The system aims at solving one of the important issues in the information era; i.e. answering questions which are repeatedly asked in different forms. Retrieved answer along with the question is stored in a database. Repeated question is retrieved to check the current question, semantic patterns are used to match already answered questions to the new one from the database. Repeated question answering will solve the problem of wastage of resources. The system uses conceptual graph formalism will be a efficient system than normal question answering system.

Introduction:-

We live in the world of knowledge and seek for knowledge everywhere. Our thirst for knowledge enhances our integrity to ask questions on various search engines. A similar search engine is a QA system. A question answering system will automatically answer the question posted by users. With a question answering system we get the exact answer to the system. Most of the search engines provide long links and documents and its hard to dig the relevant answer the users require. An efficient question answering system will provide the appropriate answer to the question. Question answering system works closely with the field of information retrieval.

Researches on question asserting system are based on factoid questions [2]. An Example for factoid question is

"Who is Sachin Tendulkar"

the answer to this question is "cricketer".

QA systems can be open domain as well as closed domain. Closed domain QA systems provides more accuracy since they concern a specific domain. Open domain QA systems covers universal values.

Major works in QA can be seen in Text Retrieval Conference (TREC) which has showed the considerable work in this area from 1999 . The main challenge of QA systems are a combination of user demand and promising result to answer a question. Initially the system should analyze the context of the question and finally it should map to the necessary answer. This requires information retrieval and natural language techniques.

The QA systems uses yes/no questions, "Wh" questions (Who is the president of India , how high is the mount everest), indirect questions (I would like you to….) [3]and commands (Name all rivers in India). This paper deals with Wh questions. At times answers can be narrative, and in that form it comes under another related area of NLP called text summarization.

The paper we are dealing with here is the open domain QA system, here everything will be relying on general ontology as given by Mohamed Riaz[8] and word ontology. In an open domain QA system more information will be

available from which the data can be extracted. Wolfram alpha is one of the best QA systems that is presently available [10] Wolfram alpha is a computational answering engine developed by Wolfram Research. Wolfram Alpha, which was released on May 18, 2009, is based on Wolfram's earlier flagship product Mathematica, a computational platform or toolkit that encompasses computer algebra, symbolic and numerical computation, visualization, and statistics capabilities.The amount of data generated in web has been increasing every second. Millions of questions are asked in user interactive question answering [UIQA][2] systems like stack overflow, yahoo answers, Bing etc. The main problem is the users ask the same questions again and again. According to various studies in 2008 "Yahoo! answers" have acquired around 500 million questions and 40 million answers. Many of the questions are being repeatedly asked and accumulating them together would have minimised the wastage of searching again and again and also the network traffic associated with these searches.

Method:-

Repeated question answering system solve the problem of asking same question in different ways. Consider an example
"who won oscar for best actor in the year 2016"
and
"name the winner of oscar for best actor in 2016",
here the answer is same for both questions
" Leonardo DiCaprio "

.By implementing this system users will be saved from answering questions again and again.Users input a question to the system. The system checks whether the question is being repeated.In our system we do structured and semantic matching [4]. The system is mapped to the resources available in World Wide Web and relevant document are extracted. Each and every question that the user provides will be saved to database. While the question is saved it will be processed and the tokens and key nouns will be separately saved [6]. Fig 1 shows the complete working of the system. The matching of question are done using Jaccard similarity [9]. The question given undergo NLP preprocessing [6]. NLP preprocessing includes stop word removal [8], tokenising [8], and stemming [10]. In this system questions should be matched mainly by three ways 1. Direct matching 2. Context matching 3. words that match semantically with same meaning and different words. Direct matching will directly match the question previously asked, Example, "which is the largest river in the world " and " which is the largest river in the world". If the same question is asked again it then can be directly matched. In context matching the context of the question will be same but expressed in two different ways. Eg: "Who got first olympics medal from india" and "name the first athlete to secure the first olympics medal from india". Here the two questions have the same context. The next case is words with same meaning with different words. Eg: "Who assassinated Mahatma Gandhi" and "Who killed Mahatma Gandhi". Here "kill" and "assassinated" are different words with same meaning. The system is divided into 3 main A. Structure Processing B. checking for repeated questions C. Search sources for answers.

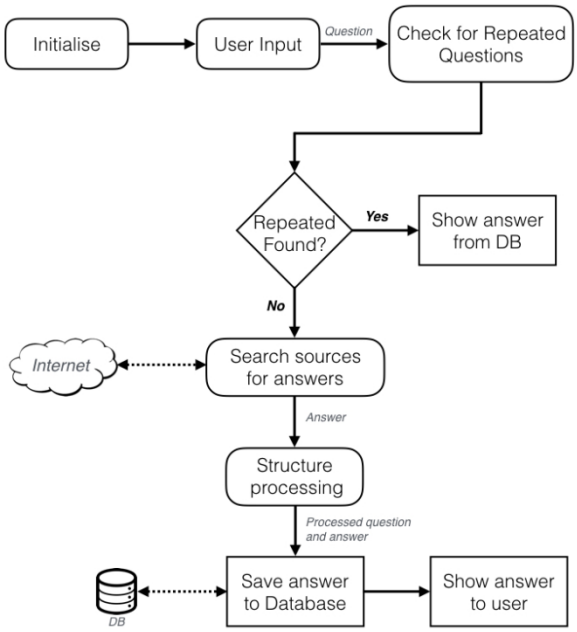
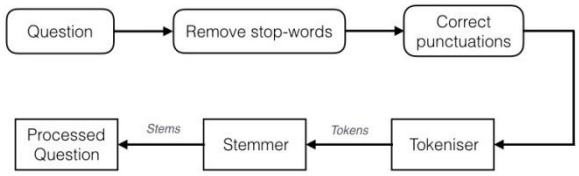


Fig 1:-

Structure Processing:-

The question the user gives will undergo preprocessing. First the stop words of the question is removed. Stop words are words which are filtered out before or after preprocessing of natural language data [11]. Some of the commonly used stop words are 'the' , 'at', 'which', 'is', 'and', 'on'.After removing stop words the words are converted into token.After tokenising stemming algorithms are applied and words are stemmed. Stemming is the process of



reducing inflected words to their word stem. The stem of word "fishing" will be "fish". Now the processed question will be saved to database. The complete working of this module is shown in Fig.2.

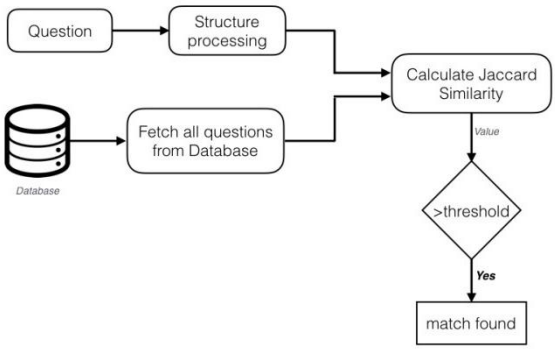
Fig 2:-

Check for Repeated Questions:-

This is the core section of the system. Each time the system checks whether user submitted question is repeated or not. For calculating the similarity between the questions Jaccard Similarity Index [12] is used, this helps in comparing the similarity and diversity of sample set. The Jaccard coefficient is defined as the size of the intersection divided by the size of Union of the sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

In Jaccard index a threshold value is set, and each and every line when matched, a value is given. This value is compared with the threshold value. The value greater than the threshold value will be 1. The new question given will be processed and all similar questions from the database and

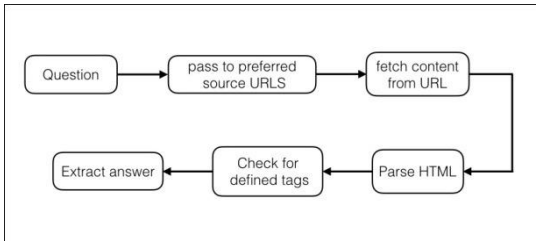


undergo Jaccard similarity check.

Fig 3:-

Search Sources for Answer:-

After checking for repeated questions and if a repeated question is not found, automatically it searches for answer sources. The question will be passed to preferred source URL. From the source URL the content is fetched and the HTML content is parsed. Finally with the parsed content it checks the defined tags and the answer is extracted. Fig.4



shows the detailed block diagram.

Fig 4:-

Experiments and Evaluation:-

Our proposed system was implemented in such a way that users can submit questions in free text. From this, system analysis the structure by relevant preprocessing. Then it search whether the question is repeated, if repeated the question is directly taken from the database and shown. If the question is not repeated, it will directly search from the answer sources. We have tried 500 random questions and from that the precision is calculated. The precision of the system was calculated by the equation as in (1). The "Correctly Retrieved answers" shows the correct answers fetched and "Total retrieved answers" showed the entire answers fetched.

Precision = (|Correctly received answers| / |Total retrieved answers|) * 100 (1)

Conclusion and Future Work:-

We have developed a system to solve repeated question answering.Users can submit their questions and system will give the exact and precise answers to the users. Since the problem of repeated answering is solved ,the problem of wastage of resources and time is solved. System have a precision of 78.2 %. Text summarisation can increase the result output in cases of "How" questions. Synonyms checking can also be implemented to retrieve relevant answers and then improve the overall result.

References:-

1. Tianyong Hao, Liu Wenyin, Automatically Answering Repeated Questions based on semantic patterns, proc. 10th IEEE, .c on cognitive informatics & cognitive computing, 2011
2. Ali Mohamed Nabil Allam, and Mohamed Hassan Haggag, "The Question Answering Systems: A Survey", International Journal of Research and Reviews in Information Sciences (IJRRIS), September 2012 Science Academy Publisher, United Kingdom
3. T.Y. Hao,, D.W. Hu,, W.Y. Liu and Q.T. Zeng. Semantic patterns for user interactive question answering, Journal of Concurrency and Computation-practice & Experience, 2007, vol. 20, pp. 1-17.
4. E.Voorhees. The TREC-8 Question Answering Track Report, NIST Special Publication of The Eighth Text REtrieval Conference TREC 8, National Institute of Standards and Technology, 1999, pp. 743-751
5. Unmesh Sasikumar ,Sindhu L, "A Survey of Natural Language Question Answer- ing System",International Journal of Computer Applications (0975 8887) Volume 108 No 15, December 2014
6. Kwok C, Etzioni O, Weld D S, (2001) "Scaling Question Answering to the Web," Transactions on Information Systems.
7. Tuffis D, (2011) "Natural Language Question Answering in Open Domains, " Computer Science Journal of Moldova.
8. Sunil A. Khillare, Bharat A. Shelke," Comparative Study on Question Answering Systems and Techniques" ,International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 11, November 2014
9. Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, Wen-Lian Hsu, "Perspectives on Chinese Question Answering Systems
10. Richard J Cooper and Stefan M R., "A Simple Question Answering System", Proceedings of the 9th Text Retrieval Conference (TREC9), NIST, 479488. News article from TIPSTER and TREC CDâĂŹs, (Pg 1-7).
11. Poonam gupta,vishal gupta, "A Survey of Text Question Answering Tech- niques ", International Journal of Computer Applications (09758887) Volume 53 No.4, September 2012
12. Moldovan, D., Harabagiu, S., Pasca, M., 2000. The structure and performance of an open-domain question answering system. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 563-570.
13. Voorhees, E.M., 2004. Overview of the TREC 2003 question answering Track. Twelfth Text REtrieval Conference, Volume 500-255 of NIST Special Publications, Gaithersburg, MD. National Institute of Standards and Technology.
14. Kan, K.L., Lam, W., 2006. Using semantic relations with world knowledge for question answering. In: Proceedings of TREC.
15. Molla, D., Vicedo, J.E.L., 2007. Question answering in restricted domains: an overview. In: Proc. of ACL.
16. Kolomiyets, O., 2011. A survey on question answering technology from an information retrieval perspective. Inf. Sci. 181 (24), 5412- 5434.