Efficient and Intelligent QUEST

Shikha Dongre¹, Swati Singh Lodhi²
¹,²CSE, Sanghvi Innovative Academy, Indore, India

Abstract— Searching on Internet growing very rapidly. Extracting useful data from a large web database is challenging task. From many years researchers and developers are trying to implement a system that should be able to retrieve only useful data and ignoring insignificant data, they have used many techniques and algorithms to implement above described system, some of them technique are Semantic web and Ontology. Semantic Web is a marvelous concept of Internet. Semantic web represent only serviceable data and conceal remaining data so that user can easily get required result. Ontology is too auxiliary in searching, because it classify the data into different categories so that retrieval becomes easy. Using Semantic web and Ontology together improve searched result.

Improving the searched result is now becomes primary need of different systems because so many times user get large amount of data but does not found relevant data. There are so many question answering systems that tries to provide appropriate result to user but they are still not able to answer proper. We design a system, that is semantic web and ontology base efficient and intelligent question answering system, that provides semantic results to the users. We have used PHP Framework, Web Scrapping, Ajax and MySQL queries to implement the system, that is able to generate proper and meaningful answer in the form of paragraph.

Keywords—Ontology, OWL, Question Answering, RDF, Semantic Web, SPARQL, Web Data.

I. INTRODUCTION

Now a days, huge amount of collected information is preserved in data storage devices. After centuries people have realized importance of accessing information from stored devices. For accessing these information people used World-Wide Web (WWW), which is also known as Internet. Internet becomes very popular in world and everyone preferred to get data and information from the web, in place of books or any other source. Even, people saved their important data and information on computer and web. Much software has been developed to store data and information on computer and WWW. Very large size of data can be stored on the web. People use cloud storage to preserve their important information on the web. This becomes tough to access large amount of data, that has been stored on web. More the data is stored means complexity of retrieving information increase.

Semantic Web is first invented by Tim Berners Lee for web data. He has defined Semantic Web as “The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation”. Semantic web means, meaningful search on web. With the help of semantic web, we can protect search of senseless data from “Web of Data”. The user can get only useful and matched result on the place of un-useful and irrelevant data. In other words we can define Semantic Web as “Semantic Web is a design of getting collection of related application for data on the web in such a manner, that form set of ideas from web data”. Different technique of semantic web are described below:

A. RDF, which is stands for Resource Description Framework is a syntax, developed by World Wide Web Consortium(W3C) in 1999. RDF is a framework for creation and modification dates of metadata. RDF is described in terms of subject, predicate and objects, for using these three terms we say this as RDF triples. In these triples 'subject' belongs to 'source', 'object' belongs to 'target' and 'predicate' belongs to 'verb', which shows the relation between subject and object. In a RDF statement, Subject and Predicates are uniform resource identifiers (URIs) and Object is either a URI or a literal value. RDFs can be represented by different syntaxes like, RDF/XML, N3, Turtle, RDFa etc. It is standard for semantic web. RDF is represented in the form of graph, that described properties and class of RDF based resources for semantic web.

B. Web Ontology Language is a language for Ontology of semantic web, which was defined by W3C. The Web Ontology Language (OWL) inherited the properties of RDF. Understanding OWL is difficult, but it has more desirable interpreting ability than RDF. The reason behind the better OWL system is, it added few more knowledge into the RDF related to class, relation between classes, their principles, characteristics, synonyms and joins. Therefore using OWL and RDF together provide more semantic results than using them alone.
C. SPARQL stands for SPARQL Protocol And RDF Query Language. SPARQL is like SQL query, it is also related to database. SPARQL is set of rules(protocol) of semantic web and a query language that is used to search RDF documents. The output of SPARQL may be set of RDF documents. SPARQL have very important place in semantic web technologies. Similar to RDF, SPARQL uses query triples conjunction, disjunction and optional pattern.

Ontology is a concept based visualization of objects, their relationships and vocabulary. Ontology becomes a very important part of Semantic web. With the help of ontology, we can retrieve information and share knowledge on the web. Ontology for any Country is shown in Figure 1.

![Figure 1: Ontology for Countries](image)

II. LITERATURE REVIEW

A question answering system, answer user’s question according to their requirements. These systems are not like search engines like Google, Ask, Yahoo etc., while these are like exam paper solution. The above sentence meant to say that, question answering systems are similar to exam paper and its solution in exam; For example, students are given question and they have to write correct or related answer for that question. Similarly in question answering system, the system has to retrieve correct or related answer to the user.

The main goal of these types of systems is to avoid ambiguous, irrelevant data, and provide relevant and accurate data according to the question’s requirements. Many researchers have been focused on implementing this type of systems, and they have implemented it. They implemented these systems by using only semantic web, only ontology and semantic web and ontology together. They have used many more different techniques to implement question answering systems.

A. AQUA-Ontology based question answering system

AQUA[18] was an experiment based question answering system, that combines ontology, natural language processing and information retrieval technology in a similar framework. For this similarity, this system uses similarity algorithm that is very important feature of this system. This similarity algorithm finds the relationship within question and parsed query designed by ontology. AQUA is able to answer the questions related to organizations and academics.

AQUA used RDF and RDFs to provide basic environment for representing data about the data (called “metadata”) on the web. In AQUA, ontology is used to rectification of entered natural language query, consideration procedure, and in similarity algorithm. In this system, ontology endues an intelligent redevelopment of asked natural language query with its meaning to minify the probability of failure to answer.

B. QUERIX: A natural language interface to query ontologies based on clarification dialogs

The major problem of natural language is ambiguity and some systems are not able to adopt new domain. These two issues are main problems in old system, which is founded by Esther, Abraham, Renato[17] and gives idea them to design a system like Querix[17] which was domain independent NLP interface for semantic web. Querix is not any intelligent system that interprets and perceives the input queries; but it deputed a reduced collection of NLP tools. It is very simple system so that it provide a portable interface.

C. PowerAQUA: A multi ontology based question answering system

PowerAqua[15] was a multi ontology based question answering system, that takes input in natural language and is not limited to single ontology. Usually, ontologies are overlapped, disjoint domains that can be similar or different. Ontology selection was a challenge, because overlapping and disjoints. Understanding the terminology and provide the answer, for example: Aishwarya Rai and Aishwarya Rai Bachchan are same, so they need to make a system that will be able to understand the terminology and their meaning. Sometimes single answer may require information from multiple source so the system should be able to find the answer from the different sources that refers to same entity or object.
The working of PowerAqua is shown in Figure 2, the system has 3 main parts, Linguistic Component, Power map and Triple similarity services. Linguistic component generate Linguistic Triple, that is subject, predicate and objects. Power Map, ontology discovery takes place to find the required set of ontologies. By entity mapping table, every table links a query term with a collection of logics mapped in different domain ontologies.[15] Triple Similarity Services takes input from Entity mapping table and find the similarity between the Linguistic and generated ontologies. And the final triple is send to the Triple Mapping table, where relationship is generated and the final answer is send to the user.

D. Information Retrieval: improving question answering systems by query reformulation and answer validation

Retrieving an exact answer is very big problem of QA systems, therefore Information retrieval [14] used syntax and semantic relation for Penetrative asked questions and mobile pattern for finding actual answer. It worked on ontology domain and uses the concept of “weather forecasting” information based pattern to retrieve domain specific question answer.[14] This improved its functionality as compare to other system by avoiding factoid form of question answer. Main aim of Information retrieval system was to develop a different model for classification, regeneration and answer indorsation in a system by using weather forecasting, NLP, syntax and semantic relation between words, mobile patterns and formal information about particular domain.

Information retrieval system was the set of documents and able to retrieve domain specific answer for the question asked in natural language. There were 2 types of domain had been used, one is open domain that was related to all types of domains and another was closed domain that was related to particular field like(Computer, Medicine etc.).

Whenever user asks any question, first it parses into syntax components and than keywords. The question patterns are designed according to English grammar and find the pattern verb. Pattern verb may be main verb, auxiliary verb or model verb and it also finds the question word used in the question to process the document.[14]

E. SMART


SMART was an easily available to provide simple knowledgeable tool for scientists to show, unite, handle the query. The main goal of SMART was to develop an intelligent system that can be able to know the meaning of query means semantic. This system integrates with bio-informatics and generates a strong knowledge.

The SMART system consists of SMART Mediator, Ontology Repository, Ontology Indexer, Hybrid Reasoner, SPARQL query engine and Graphical User Interface. Java is used by SMART to implement the system, because it is an open source technology. With Java it have used another helpful technologies which are Wonder Web, OWL, API, Pellete, Jena, Protege, AJAX and SPARQL. GUI of SMART is like Google search engine.

F. QASYO: A question answering system for YAGO Ontology

QASYO [10] is a question answering system, which uses combination of natural language, YAGO ontology and information retrieval in a uniform manner. In this system, query asked in natural language by user and YAGO ontology is used as input to the system and it access answer from available source.

YAGO was the reference for the QASYO, it was a small ontology that joins high coverage with high quality.[10] It joins tree structure of WordNet vocabulary and Wikipedia. YAGO was free to store data in particular format to generate simple text files and minimal overhead that used as internal format.
Each relation and files of YAGO are managed by storing them into a folder. Architecture for QASYO is shown in Figure 3. This architecture was designed with reference to waterfall model of software engineering, in which a natural language question is transformed into collection of intermediate triple representation, and query-triple converted into ontology compatible triples. QASYO takes input as natural language and yago ontology and generate the output as required by user.

![Figure 3: QASYO Architecture](image)

<table>
<thead>
<tr>
<th>Different QA Systems</th>
<th>Input Applied</th>
<th>Result Percentage</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANTO</td>
<td>Natural Language</td>
<td>88.05%</td>
<td>Used small ontologies and No database is used</td>
</tr>
<tr>
<td>QUERIX</td>
<td>Natural Language</td>
<td>86.08%</td>
<td>Not able to accept new ontologies and can not solve the problems of accuracy</td>
</tr>
<tr>
<td>ORAKEL</td>
<td>SPARQL+F-Logic</td>
<td>93%</td>
<td>Can not use multi-ontology</td>
</tr>
<tr>
<td>QACID</td>
<td>Natural Language</td>
<td>80%</td>
<td>Domain independency make it costly</td>
</tr>
<tr>
<td>SMART</td>
<td>SPARQL</td>
<td>80.3%</td>
<td>Not interacted with user</td>
</tr>
</tbody>
</table>

S. Kalaivani and K. Duraiswamy improved their QAAL system by using Semantic search and Graph Matching Ontology. They define a simple and clear architecture for QAAL system that is shown in Figure 4. user interaction User gives input using search engine and get answer. In server part, the question is processes using parsing, analysis, question reformulation and classification. And then from database, using semantic search, retrieve the required query result, and if not get the answer, the query is sent to search engine to get relevant document to show answer to the user.

**H. Intelligent Semantic Question Answering System**

Many of the search engines, such as Google; implements QA system, used parsing and ranking of searched records while other search engine like Ask have storage of question and answer given by users and experts.
Google gives the answer of multilevel questions, is calculated as expensive, therefore the system require crawling the page and provides ranking dynamically and ask.com provides the answer is become limited, because large amount of data cannot be stored at runtime.

Erfan Najmi, Khayyam Hashmi, Fayez Khazalah and Zaki Malik found two major problems in old question answering systems that was the calculation cost of answer of system should be cheap, and the system should be able to update itself dynamically according to the requirements. There are many types of questions starts with what, where, how, whose and many more are classified by different techniques that are factoids classification, classify the question in two categories one is true/false type questions and second is “wh” words type questions and one special case of ‘how’. Another type of classification is Keyword classification, they used 2 tools for keyword classification, one is NLP but it have issues of time complexity and lake of knowledge while another is N-Gram approach, that is faster and have knowledge. This system uses N-Gram approach.

Approach used after question classification is Question Translation that converts the classified question into RDF triple. In their algorithm they distribute the algorithm into 2 parts, one to convert NLP into RDF triple and another to expand algorithm to handle complex questions. Let the question is “Where does Taj Mahal Located?” The final algorithm the created is shown below:

1. From the question classifier we know that the question is related to location/place.
2. Looking for the verb in question is “located”.
3. Remove Wh word and extra word, result will become “Taj Mahal Located”.
4. Group the words with subject and verb, so the subject becomes “Taj Mahal” and verb “Located”.
5. Now the Triple is [Taj Mahal, Located ? (Location/Place)].
6. To find result run SPARQL query SELECT? (Location/Place) FROM <RDF repository.RDF> WHERE {<http://name#_Taj_Mahal> act:located?(Location/Place)}.

I. Intelligent Question Answering System

Intelligent question answering system designed by V.S. Babanne, Dr. S. T. Patil and D. J. Joshi. This system solves the problems of search engines that throws a broad amount of data to the user and user select the solution of his/her question/query. They designed a rule based automatic question answering system, that was based on if-than rule and uses keyword based search and uses knowledge base to simplify the result retrieval. For knowledge base they have used Artificial Intelligence.

Architecture for intelligent question answering system is shown in Figure 5.

III. PROBLEMS IDENTIFICATION

AQUA [18], Querix [17], PowerAQUA [15], SMART [13], QASYO [10] and QAAL [9][8] all the Question Answering systems used Semantic Web, Ontology or both to make their system relevant and superior than other systems. As we have studied above named different question answering systems in literature survey, they are having some problems while retrieving answer for user's question. Those systems were not able to provide exact and meaningful answer according to question. Table 2 describes a comparison and limitation of different question answering system that we have studied.
<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name</th>
<th>Technique Used</th>
<th>Algorithm</th>
<th>Result</th>
<th>Limitations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AQUA</td>
<td>I. Ontology</td>
<td>I. Query</td>
<td>Retrieved answer validate by AKT reference Ontology</td>
<td>I. Less Library Knowledge II. Multi Ontology</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. NL queries</td>
<td>II. Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>III. QLL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Querix</td>
<td>I. NL queries</td>
<td>I. Jena</td>
<td>Including un useful and stupid question answer 78.6% and after removing them answer 87.11%</td>
<td>I. Highly dependent on Vocabulary II. Should use Semantic Technique</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. SPARQL</td>
<td>II. SPARQL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PowerAqua</td>
<td>I. Multi Ontology</td>
<td>I. Power</td>
<td>Contains 7000 ontologies 800K Entities 16L. Relationships</td>
<td>I. Query range is limited II. decomposition of query III. User Interface</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. NLP</td>
<td>II. Similarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>III. Semantic Web</td>
<td>III. RSS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SMART</td>
<td>I. Semantic Web</td>
<td>I. Semantic Query Answering</td>
<td>Good GUI that provide Yeast biological knowledge</td>
<td>I. Truth Maintenance II. SPARQL Expansion III. Knowledge Composition</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>II. RDF</td>
<td>II. Parsing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>III. OWL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IV. SPARQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The problems we found in old systems is shown below:

I. Some of systems used rule based system, the problem occurs in that type of system is that it cannot be dynamically changes according to the situation.

II. If a question asks by user is not a valid question or the system is not able to search the answer of the system than what should be done.

III. In some systems when user asks the question and the system start the process to search the result the question not been saved as a template for future use.

IV. There is no suggestions are available when user type any question.

V. Some systems don’t answer the list questions.

VI. There are no metadata available in some previous system.
IV. PROPOSED ARCHITECTURE

Figure 6 shows the proposed architecture for question answering system. Our Efficient and Intelligent QUEST contains different parts and step to generate final answer. The flow of our system is to Type Question=>Search for appropriate Answer=>Generate Final Answer. The steps and parts involved in our system to complete the system are as follows:

A. Search For Question

In this part user interact with system using browser and type question in search box, then click on the search button to search answer of the question user have entered.

B. Question Classification

Typed question of user than go the classification part of question, here question is classified according to the question word(QW), Subject, Object and Verb. Question classifier contains the knowledge of English grammar and minimizes the question into simple query so that answer can be easily searched by the system.

C. Search for Question

Simplified question is transferred to the search for Questions, it is an intermediate step between the simplified question and Database, where simplified question is arranged in a proper way.

D. Database

Simplified arranged question is than send to the local database to search answer into database. Database apply content based search and search the answer for question using RDF and ontology. If it found the answer, it retrieve the answer send to next step and if answer is not found in database, then the simplified question send to network/Web to get answer.

E. Internet

Simplified question is send to Internet from database to search the correct answer from web. That simplified question is sent to Google, and by using Web Scrapping it retrieves answer from web data.

F. Web Data

Web data is web database that contains the data of every world. It contains every data that is stored on web.

G. Retrieved Answer and Retrieved Answer From Web

The set of answers that have been searched from database of from web data is stored in the form of paragraph into retrieved answer or retrieved answer from web.

H. Filter

Question and retrieved set of answer from database or retrieved answer from web is sent to the filter to find the appropriate answer for question asked by user.

I. Final Answer

Filtered answer is our final answer for the question that user want to search and the final answer in the form of paragraph has been sent to the user.

Algorithm for our question answering system can be given as:
1. User type a Question.
2. Classify question according to question words (Who/Whom/Whose->Person, Where->Location, When->date and time).
3. Minimize the Question by removing Question words.
4. Search for answer of question into local database.
5. If(answer found)
   { Show result; }
else
   { goto Function SearchOnWebdata(question); }
6. Function SearchOnWebdata(question){
   search answer on Internet until the answer is found
Google Page Rank Algorithm.}
7. Save Internet searched data into our local database
8. Take user’s feedback.
9. end

V. IMPLEMENTATION

Figure 7. Search on Browser

Figure 8: Retrieved Answer

If we apply the Algorithm Implemented by us will work as follows:
1. Question: Who is Barack Obama?
2. Found question word: Who.
3. who related to 'person'.
4. simplify the question by removing question word
   and is, now the question will be 'Barack Obama'.
5. search for Barack Obama into 'person' table, If
   found in database, then it will show the result in the
   form of paragraph.
6. If answer not found into database, then it will
   search answer from Internet Web data and show the
   answer.
7. For future search and less time it stores the data
   into local database so that user can found result
   fast.

VI. RESULT ANALYSIS

We have searched too many answer for particular
question and calculated accuracy percentage that is
shown below in the table:

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Question</th>
<th>Answer Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>When was Mahatma Gandhi Born?</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>Which laptop is better?</td>
<td>60%</td>
</tr>
<tr>
<td>3</td>
<td>To whom Shreya Ghoshal Married?</td>
<td>50%</td>
</tr>
<tr>
<td>4</td>
<td>What is the Semantic Web?</td>
<td>97%</td>
</tr>
<tr>
<td>5</td>
<td>Who is Arvind Kejriwal?</td>
<td>99%</td>
</tr>
<tr>
<td>6</td>
<td>Where is Taj Mahal Located?</td>
<td>100%</td>
</tr>
<tr>
<td>7</td>
<td>How to implement stack?</td>
<td>87%</td>
</tr>
<tr>
<td>8</td>
<td>Why C++ is better than C?</td>
<td>98%</td>
</tr>
<tr>
<td>9</td>
<td>Whose birth date is 11 October 1942?</td>
<td>0% (Not Answer)</td>
</tr>
<tr>
<td>10</td>
<td>When the first world war happens?</td>
<td>85%</td>
</tr>
<tr>
<td>11</td>
<td>Prime Ministers of India</td>
<td>100%</td>
</tr>
<tr>
<td>12</td>
<td>Who is Chief Minister of Madhya Pradesh?</td>
<td>100%</td>
</tr>
<tr>
<td>13</td>
<td>Where the Lions Live?</td>
<td>100%</td>
</tr>
<tr>
<td>14</td>
<td>To whom India Government awarded with Padm Vibhusham on 26 january 2015?</td>
<td>50%</td>
</tr>
<tr>
<td>15</td>
<td>How to run Java Program?</td>
<td>87%</td>
</tr>
</tbody>
</table>
We have calculated total efficiency of our system is 80.87% that is better and accurate than any of other question answering system. We have taken some of the parameters for percentages and question-words. Maximum accuracy for system can be 100%, we have inputted some of basic questions by using different question words like when, which, why, what, who, whom, how, where, whose etc. and get different answers. We notice our system's accuracy for answer is variable according to question words. If we draw a graph for accuracy of system answering between the accuracy and question words we will get the output as shown in Figure 9.

![Figure 9: Graph for result accuracy](image)

By the execution time of different questions, we have analysis that If the answer searched from web database, it take time while searching from Local database it takes very less time. If answer for any question not found in local database, than it searched to the web database, and the retrieved result from web database have been saved to our local database therefore in future answer for the same question will be searched from our local database.

VII. CONCLUSION AND FUTURE WORK

After result analysis of Efficient and Intelligent QUEST, we can conclude that the system does not perform exactly as proposed system, but 90% similar to the proposed system. In result analysis, we input different types of questions and get answers as output. We notice that our system answers many of the questions, approx 70%-80% of answers are accurate, rest of 10% are nearly accurate to answer. User rating has become an additional feature for our system, with which we can interact with user and can provide answers according to their requirements. As we have found the answer accuracy of old semantic and ontology based question answering systems were not able to answer this much accurate answer to user.

None of the things in this world are 100% accurate, therefore our system also have some limitations that can be solved if we try our best. We have discussed in conclusion, or system provides 70%-80% accurate answers and 10% nearly accurate answer because of some inaccuracy in programming logic that can be improved. Another problem is execution time is higher when data is retrieved from web because of too many condition and loops have been used. If the loops and condition will be reduce, the system will perform faster.

Acknowledgment

This is to acknowledge that I am Thankful to my guide Ms. Swati Singh Lodhi, HOD, Department of Computer Science & Engineering, Sanghvi Innovative Academy, Indore, for guiding me and showing me the path whenever I was in difficulty.

REFERENCES


