Extracting accurate answers for users query to enhance user satisfaction

Kiran B. Malagi
Department of Computer Science and Engineering
K.L.E. Institute of Technology, Hubli, Karnataka, India

Veena M.

Abstract - Accurate Answer Retrieval System which provides platform for people with diverse background to share information or knowledge has become an increasingly popular research topic recently. It provides free environment where people can voluntarily ask and answer questions. Unlike the tradition automatic information retrieval this site allows peer to peer interactions and thereby answers from a large community would be more personalized and specific. Now there needs to be drift, where instead of people responding to some specific question, let the system itself answer. Unlike information retrieval systems, which return whole document or larger sections thereof, question answering systems are designed to deliver much more focused answers. In this paper we propose a novel methodology for the Automatic and accurate answer retrieval system.

Keywords – RBA- Rule based Answer Aggregation, IR- Information Retrieval

I. INTRODUCTION

Day by day internet is becoming so big that any information can be retrieved from the internet, based on the questions we ask. When we are in search of some information, we expect the data provided by the internet to be precise and accurate. But in many cases it doesn’t happen. User has to satisfy himself with the answer required by him by browsing the links provided by the search engines. The user always expects the concise and precise answer to his query. When these types of concise answers are provided by the websites, users are satisfied with the answers and confidence on the site is also increased. The time of the user is also saved by not digging into many links with irrelevant data. Question answering systems are intended to provide accurate information in reply to a query. An open domain question answering system should be able to answer a question written in natural language, for a user friendly interaction for humans.

“Do the answer given by the World Wide Web is trustable?”, Unfortunately the Answer is “No”. We have no certification to prove that the facts provided by the World Wide Web is right. Even worse, different websites often provide conflicting information, as shown in the following example.

Suppose a user is interested to know the height of Eiffel Tower and he search this in yahoo.com by giving query “What is the height of Eiffel Tower?” Among the top 20 results, one he or she will find the following facts: five websites say 324 meters, four websites say 300 meters, one says 312 meters and another one says 317.96 meters. Which answer should the user trust?. The trustworthiness problem of the Web has been realized by today’s Internet users. According to a survey on the credibility of websites conducted by Princeton Survey Research in 2005, 54 percent of Internet users trust news websites at least most of time, while this ratio is only 26 percent for websites that offer products for sale and is merely 12 percent for blogs.

There have been many studies on ranking web pages according to authority (or popularity) based on hyperlinks. The most influential studies are Authority-Hub analysis, and PageRank, which lead to Google.com. However, does authority lead to accuracy of information? The answer is unfortunately “NO”. Top-ranked websites are usually the most popular ones. However, popularity does not mean accuracy. For example, according to our experiments, the bookstores ranked on top by Google (Barnes & Noble and Powell’s books) contain many errors on
book author information. In comparison, some small bookstores (e.g., A1 Books) provide more accurate information.

In question answering it is particularly important to achieve accuracy in parsing the questions. There are often several text passages that contain an answer, so if the parser does not produce a sufficiently good parse tree for some of the answer sentences, there’s still a good chance that the question can be answered correctly based on other sentences containing the answer.

Further sections deal with Literature survey and propose the computational model and algorithm for the proposed problem. Experimental results are presented. We discuss related work and conclude this study.

II. LITERATURE SURVEY

“A Survey of Text Mining Techniques and Applications” Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, a Survey of Text Mining techniques and applications have been presented. The paper “Answer Extraction using Language Models and Data-Mining” speaks about the approach where the answers are found in the automatically built answer model from relevant documents and from language modeling. Here they are using data-mining techniques called snowball to collect the patterns for answer model. Using IR and NLP/IE techniques the answer is extracted from the pool of answer model based on priorities. The paper “Web based Pattern Mining and Matching Approach to Question Answering” by Dell Zhang and Wee sun Lee briefs about the how the textual patterns from web are cultured using TREC QA track data as training examples. These textual patterns are considered by their support and confidence which are taken from Data mining area, which helps to retrieve the appreciable answers by increasing the trustworthiness on the site. In the paper “Document Retrieval in the Context of Question Answering” the author says

That to provide a precise answer to an asked question it is important to have an efficient parser to parse the question. If it fails to find the appropriate document in first step of the process then further processing steps to retrieve the answer will certainly fail. Which leads to unsatisfied set of answers, where again the confidence on the site is loosened. Therefore they are comparing useful retrieval techniques with respect to their effective question answering.

Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, a Survey of Text Mining techniques and Applications have been presented. The paper "Structured Retrieval for Question Answering" speaks about Bag-of-words retrieval is popular among Question Answering (QA) system developers, but it does not support constraint checking and ranking on the linguistic and semantic information of interest to the QA system. We present an approach to retrieval for QA, applying structured retrieval techniques to the types of text annotations that QA systems use. We demonstrate that the structured approach can retrieve more relevant results, more highly ranked, compared with bag-of-words, on a sentence retrieval task. We also characterize the extent to which structured retrieval effectiveness depends on the quality of the annotations.

In this work, we propose a new problem called the Veracity problem, which is formulated as follows: Given a large amount of conflicting information about many answers, which is provided by multiple websites (or other types of information providers), how can we discover the true fact or answer for a question? We use the word “fact” to represent something that is claimed as a fact by some website, and such a fact can be either true or false. Also the aim is to select the best among the “true facts”. This is done by first classifying the questions into particular types and then applying set of rules set on the facts for every question. In this work, we only study the facts that are either properties of objects or the relationship between the objects. We propose a framework to defining the trustworthiness of websites for particular questions, confidence of facts, and influences between facts. We call this method as “Rule based Answer Aggregation" (RBA).
In this system, the main issue for the researchers is to provide accurate answer from huge collection of information on the Web. This system as a whole to be successful, the correct classification of the question is a must. Question Classification is technique used to extract useful information from the question by identifying its class. Then, to provide user with relevant set of answers, the appropriate answer type needs to be identified on the basis of user’s expectation. If the user asks “Who is the First Prime Minister of India?”, the user expects “Pandit Jawaharlal Nehru”. For this the answer is the name of a person that belongs to the question class “Who”. For efficient question answering, the correct classification of the question is imperative. Then, for answering user’s question, it is needed to maintain an index that considers the type of answer user expects from the system.

III. PROPOSED METHODOLOGY

The architecture of the Question Answering system is shown in Figure 1. The different phases of the system work for guiding the search to the most promising answers, removing the ones that are considered incorrect.

![Architecture of the system](http://dx.doi.org/10.21172/1.72.557)

In this architecture accurate answer are extracted from the search engine. When user asks some query, it will be analyzed by question analysis model. Here keywords are extracted from the question and that keywords are used as input to the search engine, Search engine search the extracted keywords by using RBA the algorithms. The search results will be the stored documents/web pages in which the related key words are found. Then the accurate answer will be extracted based on the similarity of the answers and their frequency of occurrence of relevant information for users query.

The class of object sought by the question

– Person (from “Who …”)
– Place (from “Where …”)
– Date (from “When …”)
– Number (from “How many …”)
– Explanation (from “Why …”)
– Method (from “How …”)

check the trustworthiness of the websites by comparing the answers provided by it with the expected confidence of facts. This can be achieved by calculating the average of facts for trustworthiness of websites $T(w)$

$$T(w) = \frac{\sum_{i=1}^{n} F_i(w)}{|F(w)|}$$

where $F(w)$ is the set of facts provided by ws. Predicting confidence of a fact is bit difficult one, the fig shows this, the confidence of a fact $f_1$ is determined by the websites providing it and other facts about the same object. Let us first analyze the simple case where there is no related fact, and $f_1$ is the only fact about object $o_1$ (i.e., $f_2$ does not exist in Fig. 2). Because $f_1$ is provided by $w_1$ and $w_2$, if $f_1$ is wrong, then both $w_1$ and $w_2$ are wrong. We first assume that $w_1$ and $w_2$ are independent. (This is not true in many cases, and we will compensate for it later.) Thus, the probability that both of them are wrong is $1 -$
t(ws1).(1 - t(ws2)), and the probability that f1 is not wrong is 1-(1 - t(ws1))(1 - t(ws2)). In general, if a fact f is the only fact about an object, then its confidence C(f) can be computed as

$$C(f) = 1 - \prod_{ws \in W(f)} (1 - T(ws)),$$

where W(f) is the set of websites providing f. In above equation, 1 - t(ws) is usually quite small, and multiplying many of them may lead to underflow. In order to facilitate computation and veracity exploration, we use a logarithm and define the trustworthiness score of a website as

$$T(ws) = -\ln(1 - T(ws)).$$

T(ws) will be between zero and infinity, and a larger T(ws) signifies higher trustworthiness. Likewise, we define the confidence score of a fact as

$$C(f) = -\ln(1 - C(f)).$$

The confidence score of a fact f is nothing but the sum of the trustworthiness scores of websites providing fact f.

IV. EXPERIMENT AND RESULT

The expected results of the proposed methodology is to check for the user satisfaction. The user is satisfied only when he gets exact answers from the search engine for the question posed by him. Hence an environment is set up to test the same. A Computer lab with 30 system with internet connection were used. The system developed in all the computers is installed and it looks like figure 3.
Totally 100 students were considered and were made into a batch of 30 each and asked to list the question of their interest and the answer they were expecting. Nearly 10 questions from each student were framed from various domains and started with the experimental session. Students entered the answers in the system and wrote down the answers obtained. Based on that the user satisfaction was calculated. The formula used was Number of right answers obtained / the number of questions posed.

The following table gives the total user satisfaction of the system based on the formula given.

<table>
<thead>
<tr>
<th>Total number of Users</th>
<th>Total number of question per User</th>
<th>Total Questions</th>
<th>Correct Results Obtained</th>
<th>Incorrect Results Obtained</th>
<th>Accuracy Obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>1000</td>
<td>683</td>
<td>317</td>
<td>68.3%</td>
</tr>
</tbody>
</table>

For the analysis of the system performance we take the categories of the question i.e., wh categorization.

<table>
<thead>
<tr>
<th>Question Category</th>
<th>Who</th>
<th>Where</th>
<th>When</th>
<th>How many</th>
<th>How</th>
<th>Why</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Questions</td>
<td>193</td>
<td>169</td>
<td>207</td>
<td>98</td>
<td>187</td>
<td>146</td>
</tr>
<tr>
<td>Correct Answers</td>
<td>160</td>
<td>133</td>
<td>193</td>
<td>76</td>
<td>45</td>
<td>76</td>
</tr>
<tr>
<td>Accuracy in %</td>
<td>82.9</td>
<td>78.7</td>
<td>93.2</td>
<td>77.6</td>
<td>24.1</td>
<td>52.1</td>
</tr>
</tbody>
</table>
The GUI is provided with a Text box for user to enter the question. When he presses the “Ask Question” button the question is tagged to Google search engine. The contents of the html pages of first five links provided by the search engine are extracted to the window. Based on the keyword identified the categorization is made the decision about what the user is searching for is taken and the answer is extracted from the HTML contents in the window.

The answer is displayed using a message box without any unwanted information.

Figure 5: Category wise questions and result obtained

Figure 6: The final result with the answer expected for the question posed.

V. CONCLUSION

We have many search engines to retrieve answers, but getting the accurate and concise answer is the big mission and the trustworthiness of the websites is also essential. Hence we have come up with a solution to provide Automatic and Accurate Answering system where Question Classification technique is used to extract required information from the question by identifying its class. The fact or trustworthiness of the answer is solved by Veracity problem by collecting large amount of conflicting information among many answers, provided by multiple websites. With “Rule based Answer Aggregation” (RBA) approach on WH questions accuracy of answer retrieved is high and user satisfaction is also good. Hence the proposed methodology works fine. In future the system has to be developed for all kinds of questions without a constraint of ‘WH’.

REFERENCES

[1] Answer discovery with multiple conflicting information providers on the web Xiaoxin Yin, Jiawei Han, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 6, JUNE 2008


