Prediction of the Next Question for the Question Answering System

Manvi Breja
Department of Computer Science & Engineering, Manav Rachna College of Engineering, Faridabad

ABSTRACT
User profiling, one of the main issues faced while implementing the efficient question answering system, in which the user profile is made, containing the data posed by the user, capturing their domain of interest. The paper presents the method of predicting the next related questions to the initial question provided by the user to the question answering search engine. A novel approach of the association rule mining is highlighted in which the information is extracted from the log of the previously submitted questions to the question answering search engine, using algorithms for mining association rules and predicts the set of next questions that the user will provide to the system in the next session. Using this approach, the question answering system keeps the relevant answers of the next questions in the repository for providing a speedy response to the user and thus increasing the efficiency of the system.

1. INTRODUCTION
Search engine is an information retrieval system designed to minimize the time required to find information over the vast Web. Now a days, the capability of the search engine has increased to a great extent. Rather than returning the results in the form of links of web pages, the result to the query is returned in the form of text, which describes the accurate answer to the query. Thus the search engines are advancing towards this question answering search engine.

Question answering is a computer science discipline within the fields of information retrieval and natural language processing. It is concerned with building systems that automatically answer questions which are posed in the natural language by the user. There are two types of domains of question answering—Closed domain question answering deals with questions asked from a specific domain. Open domain question answering deals with questions about nearly anything, and can only rely on general ontologies and world knowledge.

There are various components of question answering system: Question processing module analyze, classify question type, and process the question by creating appropriate representation required for other modules. Another module is IR Engine uses keyword based approach to retrieve accurate results of a submitted query and to rank those results. Document processing module includes paragraph filtering which reduce the number of documents by finding relevant documents that contain question keywords in a neighbouring paragraphs, and do the paragraph ordering to rank the paragraphs in order to obtain the correct answer. Answer processing module is responsible for identifying, extracting and ranking answers from the set of ordered paragraphs passed to it from the document processing module.
1.1. Need for Question Answering System

It is true that, search engines and Question Answering (QA) systems differ in design, objectives and processes. Search engine is designed to deliver documents from a query, a QA system is designed to deliver the exact answer to a question.

There are differences of 3 key features in both search engines and question answering systems:

a. Query mode: QA systems make use of natural language i.e. questions are written in natural language by the user. On the other hand, search engine, query is written using Boolean operators.

b. Results form: The second difference concerns what is delivered to the user. Question Answering systems deliver one or more exact answers to a question and their context whereas search engines return links of the web pages which contains the relevant information of the query.

c. Update: Usually QA systems use databases with low update rate i.e. which are not frequently updated, while search engines are tuned to the Web queries and their reference file are continuously updated.

1.2. Various Issues in Question Answering System

There are various issues which are faced, the following are some which our proposed system tried to overcome:

a. Context and QA:
Questions are usually asked within a context and answers are provided within that specific context. The context can be used to clarify a question, resolve ambiguities or keep track of an investigation performed through a series of questions.

b. Answer extraction:
Answer extraction depends on various factors:
1. The complexity of the question,
2. On the answer type provided by question processing,
3. On the actual data where the answer is searched,
4. On the search method and
5. On the question focus and context.

c. User profiling for QA:
The user profile captures data about the questioner, comprising context data, domain of interest, reasoning schemes frequently used by the questioner, common ground established within different dialogues between the system and the user, and so forth.

This paper has been organized in following sections: Section 2 highlights the proposed research method, Section 3 describes the snapshots of the results of experimental evaluation and analysis done, Section 4 concludes the discussion.

2. RESEARCH METHOD

Question Prediction is one of the essential ingredient of a user oriented question answering system. It tries to predict the next questions, given the initial question to the system. It can be predicted using many data mining techniques.

The paper proposes a novel approach keeping into account the purpose of Question Answering Search Engines and encountering the issues they have. The approach discussed uses association rule mining to predict the next questions that the user may provide to the question answering search engine. The proposed system of question prediction is divided into two modules:

1. Modules in Question Answering System
2. Modules used in data mining approach on question answering system

The proposed architecture of the Prediction of the Next Question in the Question Answering System is shown below:
The components used in the architecture are described below:

1. Modules in Question Answering System
   a. Blog Crawler
      Blog Crawler is a Web crawler that works only on blog pages or blog posts which are written by
      individuals. These pages are necessary to get the individual views about a topic. The idea is to crawl
      those pages and select the most relevant out of them.
   b. Blog Summarization module
      Blog summarization is a module used to summarize every blog pages that are given to it after
      crawling and indexing. It is a big process and is done through a series of steps. In this paper, PMI
      (Point Wise Mutual Information) is used as a foundation for computing the summary.
      The following are the steps followed:
      1) Break all the text into different sentences separated by a dot.
      2) Removing the stop words (the most frequent words) for further processing.
      3) Applying Stemming and Lemmatization for better processing of the sentence. These two steps are
         taken to avoid unnecessary process in doing repeated or unimportant words.
      4) Cutting down the whole sentence into words. This will provide the foundation to calculate PMI
         which is calculated using two words.
      5) Calculate PMI for every two adjacent words (1,2, and 2,3 and .... i,i+1) as PMI (w1,w2) =
         count(w1+w2) / (count(w1)*count(w2))
      6) Calculate TPMI (Total PMI) of a sentence as sum of PMI’s of its adjacent words.
      7) Finally the summary is obtained by calculating the total PMI’s of all sentences and then arranging
         the sentences in descending order of them.
   c. Blog Indexing Module
      Search engine indexing is the process of a search engine collecting, parses and stores data for use by
      the search engine. The actual search engine index is the place where all the data the search engine has
      collected is stored.
   d. Relevant Information Extractor
      This module is a way of getting relevant information from various data sources like Wikipedia and
      dictionary.com. These websites provides lots and lots of information about data. That is why they were
      selected as the source.
   e. Question Categorization Module


<table>
<thead>
<tr>
<th>Table 1. Question type Categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATEGORY</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Abbreviation</td>
</tr>
<tr>
<td>Reason</td>
</tr>
<tr>
<td>Definitions</td>
</tr>
<tr>
<td>Weight</td>
</tr>
<tr>
<td>Size</td>
</tr>
<tr>
<td>Terms</td>
</tr>
<tr>
<td>Period</td>
</tr>
<tr>
<td>State</td>
</tr>
<tr>
<td>Event</td>
</tr>
</tbody>
</table>

This work was started on the basis of question categorization in different categories like why, what, 
how, who, when and which. In every category, a list was made to identify which entities lie in what group.
The reason behind dividing questions in categories or defining operations that can be performed in various
categories is to ease the implementation as well as to answer every question in a correct format.

2. Modules Used in Data Mining Approach on Question Answering System
   a. Admin Module
      1) Maintaining the Question Log of the user
         A search log is an electronic record of interactions that have occurred during a searching episode
         between a Web search engine and users searching for information on that Web search engine. Since there is a
         focus on the user preference-oriented question-answering, a question log is maintained on the user basis. The
         fields which are maintained in a log are S.no.,Transaction id, Date, Time, Username, Question.

Prediction of the Next Question for the Question ... (Manvi Breja)
2) Log Extractor

Log Extractor is a module which is used to extract the important information from the question log. So, now only important fields are extracted from the question log, i.e. username, transaction id, question log, final question. Final question is the question to be worked on further.

3) Rule Miner

This module is used to extract the association rules from the question log. For this the association rule mining is applied on the log.

Figure 2: Question Log Maintained By The System For Two Users

Figure 3. Keywords Of Questions Are Extracted By The Log Extractor Module

Figure 4. Frequency of keywords found by the rule miner module

b. User Module

1) Login Interface

The user interface is plain and simple. It contains a simple 2 text box – one for the username and password and Login button so that a valid authenticated user can login to the proposed question answering system. Whenever the login button is clicked, if a valid authenticated user has login, the main question answering search engine opens.
2) Next Question Predictor

Whenever a user login, the question answering system opens up in which the user can ask the question and a button is provided which will provide the user the next probable questions and the link to their answers which the user can ask in the next session.

This module has utilized the apriori algorithm in association rule mining for predicting the next question.

Algorithm 1. The proposed algorithm for next question predictor is:

a) \( D = \{ \text{database of all questions according to user, transactions} \} \)

b) \( D_1 = \{ \text{database of all keywords of questions with transactions and count} \} \)

c) \( Q_1 = \{ \text{Initial Question asked by user} \} \)

d) \( K_i = \{ \text{Keywords of initial Question} \} \)

e) For \( i = 0; K_i \neq \emptyset; i++ \) do begin

f) Select transactions and count of \( K_i \) in \( D_1 \)

g) Find common transactions \( t_i \) of \( K_i \)

h) For all \( t_i \), find \( q_i \) from \( D \).

i) \( Q_i \) contains other predicted questions after \( Q_1 \)

j) Return \( q_i \)

3. RESULTS AND ANALYSIS

3.1. Result

The proposed system has been implemented in JSP and MS SQL Server is used as backend database. The system has worked on various classifications of the question (who, what, how, where, which). The approach taken to predict the next questions has given satisfactory results up to a great extent. The different types of questions for which the next questions are predicted are shown below:

3.1.1. "Who" Type Question

The question “who is Gandhi” is asked by both user1 (admin) and user2 (admin1) on the question answering search engine. The next question predictor module returns the following questions which are personalized for the particular user.

Thus, we could see from the above result that the user “admin1” has interest in asking questions related to political Gandhi family. Before that we have seen that the user “admin” has interest in asking questions related to “Mahatma Gandhi”. In this way the system has predicted the need of the particular user logged into and returns the correct predicted questions.

![Figure 6. Predicted next questions to initial question “who is Gandhi” for user 1 (admin)](image)

![Figure 7. Predicted next questions to initial question “who is Gandhi” for user 2 (admin1)](image)

3.1.2. “What” or “What Do You Mean by” or “What is the Meaning of” Type Question

Let say, the user asks the question “what do you mean by abstract”, the system determines the log and tries to return all the questions which are related to that term. So, here the system returns the next questions as:
3.1.3. “Where Is” Type Question
Whenever the user asks the question of the type “where”, that means the user wants to know the location of some place, country, city etc. And if we predict the related questions, that will depend on the log of the user, he may ask the other questions related to that place only or can ask the question related to the location of nearby places. The system returns the next questions to the question “where is Italy” as:

3.1.4. “How” Type Question
The user1 and user2 asked the question “How to code” and following next questions are returned by the system according to the personalized question log formed.

3.1.5. “What Events Happened on” Question Type
Whenever the user asks the question “what events happened on” type question i.e. related to all the events which happened on the particular date. Let, the user asks the question “what events happened on august 15, 1947”.

The predicted next questions are related to the questions which are programming and coding to be done in different languages according to the need of both users.
3.1.6. “Which Holidays Fall on” Question Type

By the result we could see here that the system is returning all questions which user might have asked with that question as well as all the next probable questions which are also asked after another oncoming question. Thus, the system has worked on these six types of questions and has shown accurate results.

3.2. Analysis

The proposed next question predictor module is able to provide a list of next probable questions accurately. This section calculates the relevancy of the next questions. In this, the performance is calculated for different question types. The formula for calculating performance is:

\[
\text{Performance} \% = \frac{\text{No. Of relevant next questions given in result}}{\text{Total no. of next questions returned}} \times 100.
\]

3.2.1. “Who was Gandhi” By User1 “Admin”

Because admin usually asks question related to mahatma Gandhi

The next questions returned by the system are:

a. Who was mahatma Gandhi
b. Who was Kasturba Gandhi
c. When was mahatma Gandhi born
d. When was Kasturba Gandhi born
e. When did Mahatma Gandhi died
f. When did Kasturba Gandhi died
g. Which holidays fall on October 2

All are relevant questions to the input question. Therefore, System’s accuracy = 7/7 = 100%

3.2.2. Which holidays fall on October 2

Since October 2 is Gandhi jayanti, the expected next questions must be related to mahatma Gandhi

The next questions returned by the system are:

a. When did Kasturba Gandhi died
b. When did mahatma Gandhi died
c. When was Kasturba Gandhi born
d. When was mahatma Gandhi born
e. Which holidays fall on January 13
f. which holidays fall on january 26

g. which holidays fall on october 2

h. who was mahatma gandhi

Only some are relevant questions

Therefore system accuracy = $5/12= 41.66\%$

<p>| Table 2: Accuracy of The System for Different Question Types |</p>
<table>
<thead>
<tr>
<th>Question type</th>
<th>Question</th>
<th>No. of predicted questions</th>
<th>No. of correct questions</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>Who was Gandhi</td>
<td>7</td>
<td>7</td>
<td>100%</td>
</tr>
<tr>
<td>What</td>
<td>What do you mean by abstract</td>
<td>5</td>
<td>-</td>
<td>*</td>
</tr>
<tr>
<td>Where</td>
<td>Where is Italy</td>
<td>6</td>
<td>-</td>
<td>*</td>
</tr>
<tr>
<td>How</td>
<td>How to code</td>
<td>7</td>
<td>-</td>
<td>*</td>
</tr>
<tr>
<td>Which</td>
<td>Which holidays fall on october2</td>
<td>12</td>
<td>5</td>
<td>41.66%</td>
</tr>
<tr>
<td>What events</td>
<td>What events happened on august 15,1947</td>
<td>7</td>
<td>-</td>
<td>*</td>
</tr>
</tbody>
</table>

Here, * indicates that we cannot calculate in general, the accuracy of the system for these type of questions. Because the accuracy of the system will vary according to the size of question log. Here in this paper, we have taken a small question log for which the system is showing 100\% accuracy. Here, - indicates that we cannot judge how many questions are correct out of the total predicted questions.

Today many existing systems which have the capability of showing next questions have their maximum accuracy of 30\% , which is not efficient as compared to our system.

4. CONCLUSION

The question answering system proposed has provided the accurate links of next questions as well as their relevant answers. The method adopted can increase the accuracy of many existing web search engine. Thus by the results, these important features have been seen:

1. It increases the efficiency of the system by not searching the answer for the question of the user in the index again and again rather the system predict the next questions and keep their answers in its repository to reduce the response time and complexity of the system.

2. It provides the ease to the user of not typing the next questions again and again. The system will provide the user the links of next questions to just click and see their relevant answers.

3. Information about a person, any term, location, birth date, death date, holidays, events etc. available to the user is accurate and relevant to the need of the particular user and according to the survey, many users are satisfied by it.

4. The algorithm adopted for predicting is fast in most cases. However, in many situations, researches are going on to improve its efficiency.

5. System’s main aim was to have index retrieval fast so as to the complexity of the system.

REFERENCES


[4] Zadeh A. Lotfi. From Search Engines to Question Answering Systems-the problems of world knowledge, relevance, deduction and precisiation

BIOGRAPHY OF AUTHOR

She is currently working as an Assistant Professor in Manav Rachna College of Engineering, Faridabad. She has done her M.Tech from YMCA University of Science & Technology with 9.733 cgpa. She has her research interests in Data Warehousing and Mining, Information Retrieval, Machine Learning, Business Intelligence.