QA SYSTEM PREDICTION USING DATA MINING

[1]Raghuram A S,
[1]Assistant Professor
[1]Computer Science & Engineering,
[1]ATME College of Engineering, Mysuru, India

Abstract : The Question answering system is an new standard of data fetching which gives knowledge. It allows the users to enter question and waits for answer as response, then system will predict next question based on user's past searched questions and current interaction records of the user with system. In this system, users current interactions is maintained in a question log, the question log is maintained to extract the question from the search where user enters, from log the user sessions get extracted and user sessions questions are stored in databases. Based on user sessions system gives next question on user's future interest, the performance of system will increases greatly if it is capable of predicting future question. For prediction, system makes use of a data mining technique called "Association rule''.

***IndexTerms* – Data mining, Association rule**

I. INTRODUCTION

The QA system provides answers to user's questions on behalf of taking whole file document. The extraction of information from question log, this procedure is called as "Log analysis". During interactions user search a set of queries that specify the information about what they are interested in these interactions is maintained in a specified log.

The main agenda of doing search log analysis is to selecting information for future prediction. Data mining can be applied to extract information from the log for prediction of user's interest in future.

In sequence of search log for reaping knowledge about user's interest in future, mining process is applied. The data mining technique called association rule mining is concerned with "maintaining data bases, it consists of queries, relations, transactions, associations, co- relations among item sets on basis of previous transactions". This information provides how item sets are related to one another and how they tend to group together.

This is in particular a necessary task of designing of legitimate usercentric applications in which user search behaviors are detected and taken into considerations. Describing correspondent queries for browser's engine, users can help them quickly to acquisition the desired content, recently at bottom of result page some browsers engines displays relevant keywords.

The fundamental task is that when definite query is searched to give comprehensive recommendation, recommending suitable searching special words intent for not only upgrades browser's hit rate, but also helps user to extract desired information more quickly.

Question Classification is technique used for extraction of useful information, then to provide user with relevant set of answers, the appropriate answer types to be specified on basis of user's future interest.

If user asks "Who is best cricketer?", the user expects "Sachin tendulkar" as answer which is a name of person. For this, question class "Who" is mapped to expected answer type i.e. “Person”.

2. LITERATURE SURVEY

Paper 1:

This paper presents about current browser that serve know-item search such as finding home screen. In initial search, users doesn’t know about their information required and also insufficient knowledge to search efficient queries, thus these search engine are supported by query alone .In this paper search log make an advantage for user to allow to browse beyond hyperlinks with a multi resolution topic map on the basis of search log hyperlinks with a multi resolution topic map are constructed [1].

Search log is considered to be the “footprint” that is footprint is nothing but a previous user’s interaction has been dumped. The main aim of this paper is to allow the user to browse efficiently for adhoc initial queries. So that user can easily search hyperlinks for related topics which has been stored as footprint based on the previous user interaction with the browser. For the generation of multi resolution topic map based on the search log, a data mining technique called star clustering algorithm is used [1].

By using this algorithm it can be easily find previous user interaction which is stored in server [1].

Paper -2:

This paper involves about how to increase search log with the relation semantic. To improve log search, first a method is to be described for extracting a global semantic sketch. The classification that defines question terms is a collection of semantic sketch, to perform this operation makes use of a conception of synset. A synset is group of sentences which are identical in particular association [2].

In this condition, the semantic affiliation bounded by log terms is effectively redefined as affiliation bonded on synset. The main agenda for the search log analysis is to extract the user interest, for performing the extraction of user interest a data mining technique is applied called clustering algorithm [2].

Paper -3:

This paper discuss that a system predict the queries and thus search performed with the system will be improved rapidly. According to the census, many sufficient data of users were triggered by what they have searched, that is if users is interested after reading a page then the user is interested in searching the relevant information about the search. This type search occupied can advantage both browser and users [3].

In this paper there are three kinds of technologies goes to achieve the task first after reading a page by user that is to extract all queries that user has been searched, second accord to prospect queries is effectively ranked on basis of their sequence. Third is that even if triggered by same page the search intents can be diverse and In these experimental results it as specified that proposed approach can be predicted for a given web page. To perform this prediction operation, some learning method id used [3].

Paper -4:

This paper introduces a advanced method to impose for recommendation a better queries by on clicking through data. This method is useful for the user to define a query based on previous users search experience. For executing these operations an algorithm is introduced, the recommend algorithm produces a graph bounded by queries [4].

The original query is the root of a tree with recommendation as leaves .each branches of tree represents a different from a original query .the recommendation algorithm performs by comparing with a previous user session and improves the ranking, and taking to a quality of recommendations .for the user to implement this type of recommendation algorithm is very simple and at lost [4].

3. Proposed System

Proposed system is a new QA system where user submits the question and waits for the answer as the response. Proposed QA system which predicts the user's questions in future based on current interaction records of user with system, current interaction shows the users' area of interest, for prediction system makes use mining technique called as "Association Rules".

Designing System is a means of describing parts of a system such as developing infrastructure, roles of modules, components interface and the information that send with that system. The designing phase which fulfills needed requirements for an organization, process of designing uses a bottom up level approach. This system contains of three modules which are listed below

- ➢ Administrator
- ➢ Member
- ➢ Visitor

The designing of a new QA system approach which as proposed is given.

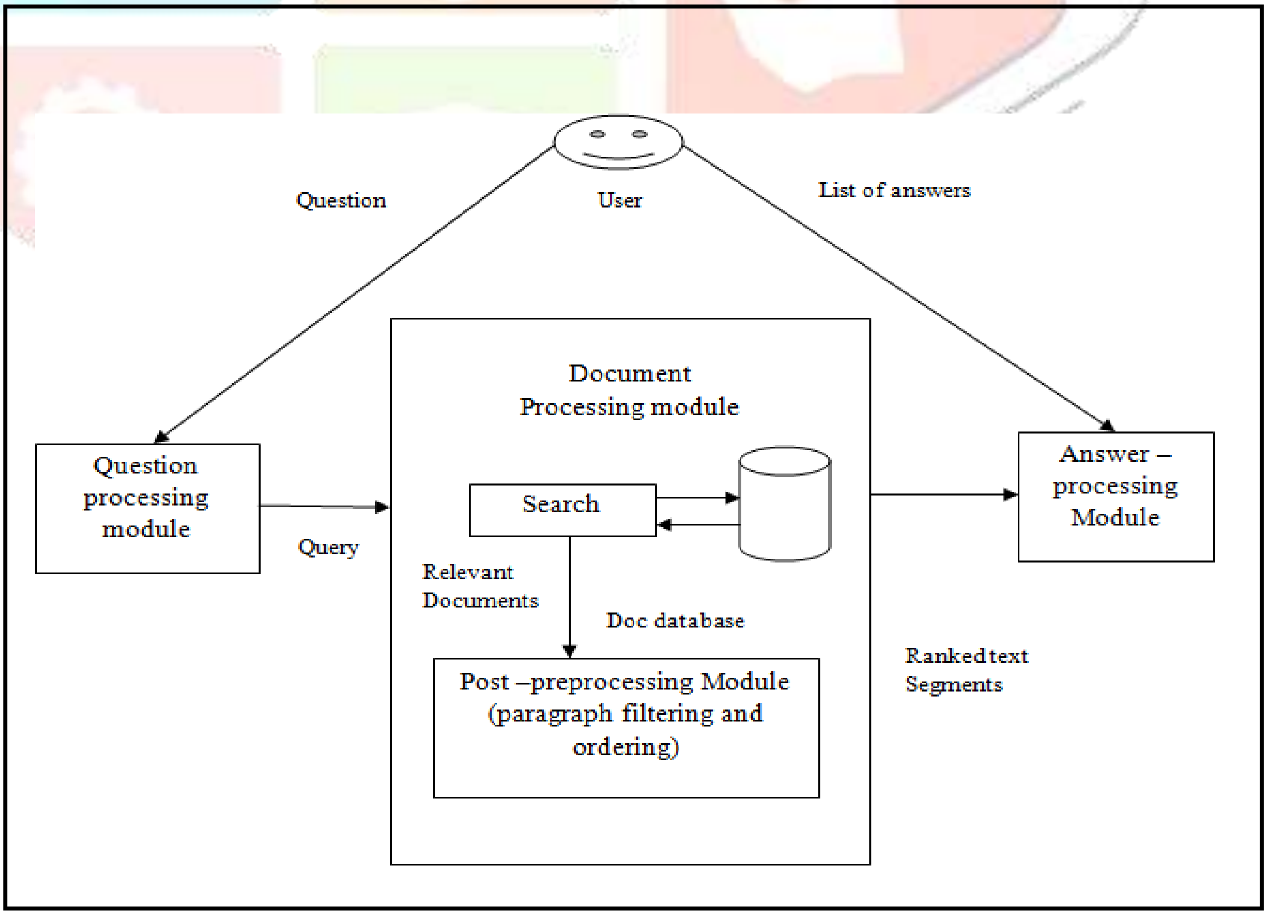
Question
User
List of answers
Document Processing module
Question processing module
Query
Search
Answer – processing Module
Relevant Documents
Doc database
Post –preprocessing Module (paragraph filtering and ordering)
Rankedtext Segments

Figure 3.1: Architecture of a QA system

In given schema, an Apriori algorithm is make used for prediction of user's further question. These questions are stored in form of question log and from this user session gets extracted .an Apriori algorithm uses a association rule technique for predicting questions.

Apriori algorithm.

Apriori algorithm manages database**,** the database contains transactions of productsets, This method achieves to detect subsets that are similar into a minimum number of item set C. The main agenda of this method is to detect rules that meet both a minimum support and minimum confidence, this method implements a "bottom up" level approach, in this way continuous subset are expanded and a single item at single point of time.

These continuous itemsets that is all groups which contain the item with minimum support is denoted by L_i for i^{th} itemsets.

4. Implementation

The Implementing of a system is main phase in developing a program, after the steps of System analysis phase, Usecase diagrams, data flow diagrams and to run this application it requires three systems and these three systems are connected through LAN. The whole system is handled by administrator. Admin manages the system by browsing the application, and performs the operations like updating, deleting, editing. Admin can register many numbers of members these information's is dumped in server, theses server is handled by admin. The admin sets the special words based on that key word members can post question into application and these posted questions and answers can be viewed, edited by admin.

Next is the member authentication, member can login into application by specifying its own id and password. Member can view posted questions along with answers. The visitor can view home page and read services provided from application. The visitor can register to application

4.1.1 Working of a New QA System

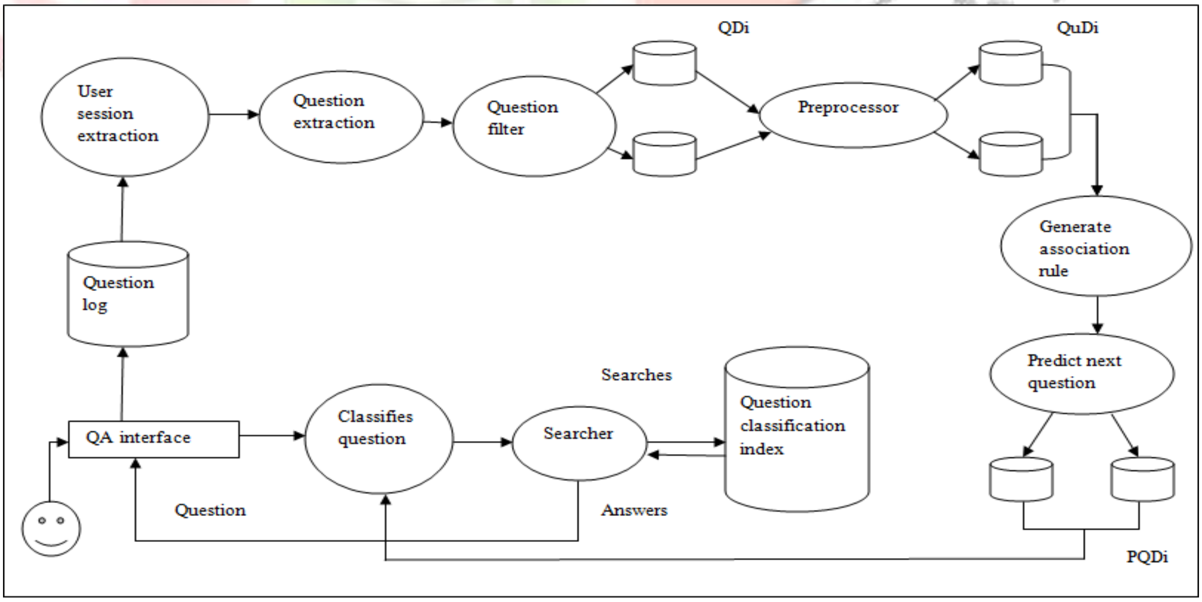

Figure 4.1: Architecture of Proposed QA system

Step by step working of a QA system**.**

- *Analysis of a search query log*

The search analysis is defined as communication between user content searches the information and browser. Main agenda of analyzing is to extract user session data.

The registered user inserts his question on interface of a QA system and search log is maintained to extract the question from the search where user enters. The question classifier classifies based on their standard query inserted and these questions get converted into query format, for each query a search is done through search module

- *Extracting sessions of login users*

After dumping period of time and date questions in search question log. Using browser's a group of user sign in sessions can taken exactly.

By considering session length as t, the extractions of user sessions from search query log for a time unit t. The main agenda of session extraction is to extract questions that have been posted within a certain point of time t units.

- *Extraction of question*

After the extraction of user session, from these sessions' questions gets extracted. For this purpose it uses question log which is maintained for interactions of users.

- *Filter question module*

For this question filter module question has been inputted to the system which is extracted from the question extraction module and these question are gets divided based on their question type.

After dividing the question type a separate data bases is created for each question type. The query filter separates questions and is stored in a database and then these questions are sent to the query preprocessor

- Query preprocessor

The main function of this step preprocessor system is to receive the query terms that form a group of elements. In query preprocessing system there are two levels: one is processing as term level and other is processing as query level. The query level processing is nothing but filtering process; first some queries are removed from format of log. Next is "bad queries" that are distilled out, means incomprehensible questions are bad queries.

In the term level preprocessing system a lexica queries can be applied, set of tokenizer is formed by dividing the each query, this module divides the question type from question and converts the remaining part of the question into query.

- *Generation Association rules*

The main agenda of implementing Association rules mining is to recognize all the rules in a market basket data analysis. This kind of classes for market basket data also referred to as a transaction data, this approach used to evaluate customers on by purchase of products in a shop or a supermarket and how these are related to one another. The group of products a customer buys is defined to as an productset and analyzing market data seeks to find relationship between purchases.

One record is builted foreach of customer transaction. The association rule method attained from group F continuous itemsets in an extraction context D. For example probability that a customer will buy “vegetables” without a “juice” is referred to as the support for the rule. The conditional probability that a customer will purchase” fruits” is taken to as confidence.

➢ *Next question predictor*

The results from association rule generation module is inputted to question prediction module. This module produces predicts question based on the association rules that have been generated. For one query database QuDi creates one separate predicted database PQDi and each predicted database contains QTi and a group of predicted questions, for example if user is searching “what is C”, then question predictor predicts that may also be interested in “what is C#”.

Similarly, if the user is searching for “how to play hockey” then result is may also be interested in “how to play throw ball and football”. These predicted questions which are stored in in predicted database is sent to Question classifier and then provides input to searcher that searches for each predicted query in Question classification based index and maintains search results as answers for later reference. In future, if user inputs a question that compares with any of predicted questions.

CONCLUSION

This is a new form of Question answering system. This approach is used to retrieve answers based on the user’s future interest as a next question. This system uses a concept called association rule mining for prediction of next question based on current interaction with system.

The next question prediction system that predicts users’ next requests based on their current interactions with the system from the search query logs, the technique of Association rule discovery is determined as one of most important techniques in field of data mining.

References

[1].X. Wang, B. Tan, A. Shakery, and C. Zhai, “Beyond hyperlinks: organizing information footprints in search logs to support effective browsing,” in Proceeding of the 18th ACM conference on Information and knowledge management, pp. 1237–1246, 2009.

[2] L. Limam, D. Coquil, H. Kosch, and L. Brunie, “Extracting user interests from search query logs: A clustering approach,”DEXA '10 Proceedings of the 2010 Workshops on Database and Expert Systems Applications, 2010.

[3] Z. Cheng, B. Gao, and T. Liu, “Actively predicting diverse search intent from user browsing behaviors,” in Proceedings of the 19th international conference on World wide web, pp.221–230, 2010.

[4] G. Dupret , and M. Mendoza, “ Recommending Better Queries Based on Click-Through Data”. LNCS, Springer, 2005.

[5] Z. Zhang, and O. Nasraoui, “Mining search engine query logs for query recommendation,” in Proceedings of the 15th international conference on World Wide Web, pp. 1039– 1040, 2006

[6] R. Mudgal, R. Madaan, A.K. Sharma, and A. Dixit, “A Novel architecture for question classification based indexing schemefor efficient question answering ,” International Journal of Computer Engineering & Applications (IJCEA), ISSN: 2321-3469, Volume-2, Issue-2, June 2013.

[7] K.H Lin, “Predicting Next Search Actions with Search Engine Query Logs,” Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference on (Volume: 1), 2011.