An Approach to Summarize the Text Document to Retrieve Answers for Non-Factoid Queries

[1]Manjula A K, [2]Ramya R S, [3]Dr Venugopal K R

[1]Student, [2]Research Scholar and PG Guide, [3]Vice chancellor of Bangalore University
Department of Computer Science & Engineering
University Visvesvaraya College of engineering, Bangalore University, Bangalore.

***Abstract*: This paper formulates a document summarization method which extracts the answers of passage size for non-factoid queries, which are referred as answer-influenced summaries. In this paper three methods are demonstrated for optimization they are question-biased, answer-biased CQA and extended question biased here extension terms are formulated from associated CQA data. The methods of ranking are also demonstrated which consists of features taken from CQA data. The quality of CQA data will affect the precision of the summary of optimization based. On the other hand the method of ranking is not much influenced by the quality of CQA. Suggestion is provided for the best use of the three optimization methods with respect to the different CQA quality of answers. In further research the accuracy is judged by other dataset which are open source. Most of the community information sharing website will contain answers which are shared according to user experience, these answers are collectively taken to find a better answer and one big answer will contain various information. The answers are used to expand the questions and answers among the search websites; this unique feature makes this approach for perfect answer generation.**

***Keywords*: Summarize text document, Answer influenced summaries, document summarization, CQA data.**

Chapter 1

INTRODUCTION

Present answering engines will provide efficient single line answers which give direct information for the question, this answers are mainly direct for the popular questions which are related to particular fact. The user on the website will be more satisfied with the appropriate answers for the search results without reading the entire passage of answer.

1.1 Overview:

Present search engines provide direct answers to user's query. Direct answers
will improve efficiency of the search results. The users who have short screen size and low bandwidth is most beneficial for direct answers. Non-Tidbit queries are the usual questions asked on the search engine. In this work it examines the idea of retrieving a summary from each document which contains the answers to a non-factoid query this method is been motivated by the previous work on non-factoid queries. CQA websites such as Yahoo! Accumulates diverse number of questions and answers in the user's community. The information present in CQA websites are needed for the non-factoid queries.

In this paper the following research questions are analyzed
Research Q1: To extract good summaries from the documents can we use CQA contents which are related?
Research Q2: Does the accuracy of the summaries produced are affected in terms of quality of CQA data?
This section provides the overview about how the project is formulated when a user submits a direct query he is supposed to get a direct single line answer but when a user submits a big query according to his needs he is supposed to get answer from various community question answering websites called as CQA. The answers generated must be very relevant and short, to achieve this aim of providing relevant and short answers this project is been formulated.

Chapter 2

PROBLEM DEFINITION AND ARCHITECTURE

2.1 Proposed System Architecture:

The system architecture shows how the application is built up and how it functions in a step by step way, following is the step by step process how the software works as a whole when merged.

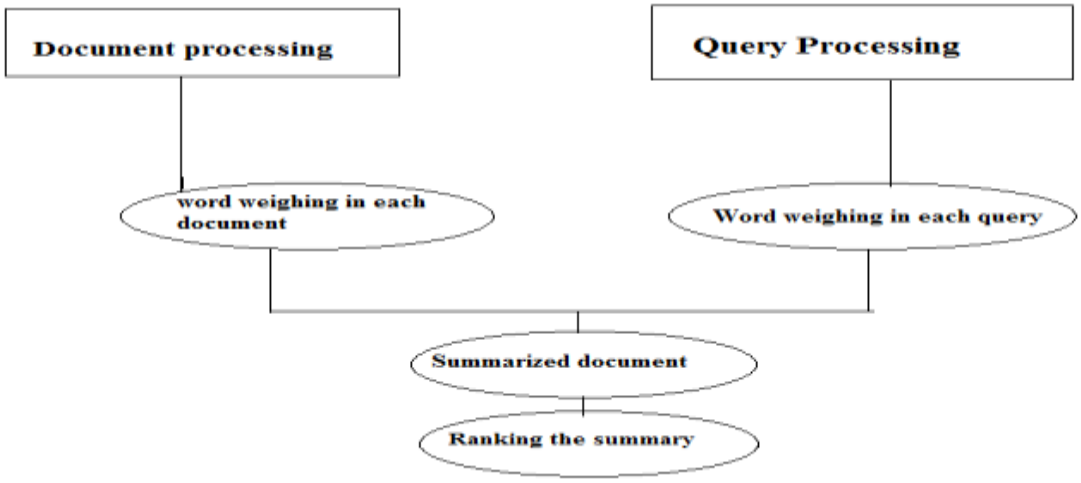


Figure 1: Architecture of the summary accumulation from the document

Figure 1 shows the system architecture of the proposed system. The Archive taken from the WebAp dataset will be prepared by expelling the stop words like uncommon images and the words which are minimum vital after expulsion of stop words every word are isolated from the record and recurrence of each word is known, the word with most noteworthy recurrence is the most vital word in the report through which recovery should be possible effortlessly. The query taken from the WebAp dataset will be processed in a same way as document is processed. Every word after stop words removal from each document will be weighed according to its occurrences in the document. It will be done same as word weighing for each document. The document will be summarized making the big document to a passage size summary

2.2 Architecture of Question Answering System:

A factoid QA is providing succinct facts for example "Who is the father of our nation?" A non factoid QA is providing detailed question according to the users need and the query length is not restricted. Example includes "What is the need to learn python?" The question is submitted and it is analyzed in the retrieval system of the document, the document is collected from various CQA and web answers and it is stored in the database through which answers are generated. The architecture below mentions the answering system for the submitted query.

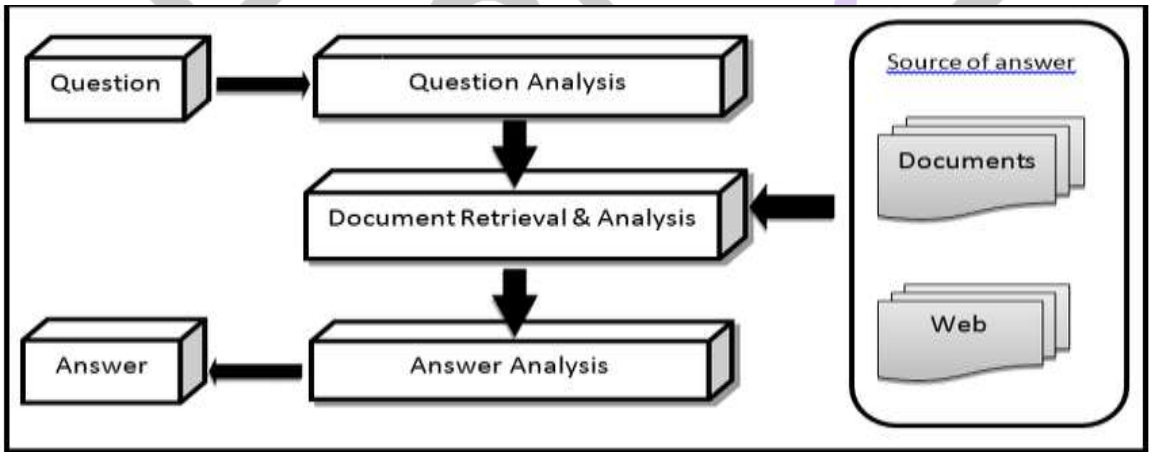


Figure 2: Architecture of the QA module.

2.3 Contribution:

It Propose CQA content in summarization algorithm for positioning answer holding sentence in the document. It recommends learning to grade based methods and three optimization based method for non-factoid queries providing relevant answers. Proper analyzing is done on the changes in quality of the associated CQA data on the propose approach, then best used methods are recommended with respect to the various quality levels of CQA answers. There are two data sets on which this experiment is conducted they are WebAp and MSMARCO.

2.4 Advantages:

CQA answers accuracy is very less compared to the accuracy of the answer biased summaries. Answer biased summaries are generated from the related CQA answers. The ranking method will improve the accurateness of the summaries for the associated CQA answers that are not accessible.

2.5 Disadvantages:

This process is time consuming as it involves the word weighing of each document and also it calculates document frequency and inverse document frequency which can take more time than expected.

Chapter 3

LITERATURE SURVEY

3.1 Extracting People Information in Web Search Result Page:

L. B. Chilton and J. Teevan [1] formulated this paper. Web search engine connects the search query to the user along with the search results for similar answers. In this paper query search log analysis approach is used where users query is analyzed among many web users, the data importance can be predicted in this approach when search result of same query is repeated. Interaction among users can also generate answers from the user.

3.2 Direct Answer Generation:

M. S Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz[3] formulated this paper which extends the answers for every information asked on web. A large collection of answers which expresses a direct meaning are not so much in demand individually but when all the direct answers are put together substitute a great need in information centers. The problem analyzed in this approach is tail answers meaning short answers may not be clear to user, short answer will sometimes prove best answer but sometimes it will be in need of some more data in it for concise results, to avoid such answers question-word queries are used which express answers in informal language.

3.3 Finding Answers from Passages:

Q. Liu et al.,[19] formulated this paper . Web browsers have increased rapidly throughout the world as search results are very precise and accurate, to know the user level satisfaction of the retrieved results this approach has been introduced. To gain the user satisfaction results on web .This approach break down an expansive number of web looks through that outcome in a visit to a prominent CQA site, and distinguishes one of a kind attributes of searcher fulfillment in this setting, specifically, the impacts of inquiry lucidity, inquiry to-address match, and reply quality. We at that point propose and assess a few ways to deal with anticipating searcher fulfillment that endeavor these attributes.

Chapter 4

IMPLEMENTATION

4.1 Optimized Proposed Method:

This method is the proposed method to summarize the document and in the later section it improved its methodology with the three existing methods, it covers the maximum important words from the document to improve efficiency and removes redundancy from the document making the answer in a passage form. The relevance of the document is taken as λ. In this approach the less redundancy documents are mentioned with lower λ. This method is based on optimization of the sentence and is not related to ranking of the answer.

$$(1-\lambda)\sum_j w_j z_j + \lambda \sum_i \left(\sum_j w_j a_{ij}\right) x_i \qquad (1)$$

$$s.t. \sum_i c_i x_i \le K \; ; \quad \forall j, \sum_i a_{ij} x_i \ge z_j \; ;$$

$$\forall i, x_i \in \{0,1\} \; ; \; \forall j, z_j \in \{0,1\}$$

Here wj is the word weight in the document and ej is the weight of each word in the document; the binary value here is zj that covers the word in a summary. aij is also the binary value but it covers the word in sentence which is denoted by si. The Binary value xi denotes the choice of sentence. The first condition here is $\Sigma ci\ xi$ <= K indicates the summary size will not exceed K, K indicates words which is set to a limit of 50 words, this condition guarantees that each summary will be lesser than or equal to 54 words. The second condition is $\Sigma aij\ xi >= zj$ this condition relates to each word coverage in the passage level answer. In the first model, words are weighed in view of a record content. The TF-IDF recipe is utilized in Eq 2 for word weighing, this unique strategy is called DocOpt which uses this improvement of Eq 1 and decides the significance of words in view of report content.

$$w_j = tf_{j,doc} \times idf_j \qquad (2)$$

Where t ,doc is the recurrence of word ej in the archive; and $idfj$ is the opposite record recurrence of word ej in the GOV2 web gathering, figured as pursues:

$$idf_j = ln(1 + \frac{n}{df_j}) \quad (3)$$

Here n is the aggregate amount of records in the network collection and dfj is the quantity of archives in the network gathering that contain word . In this investigation, a step was prepared and finished by evacuating stop words and stemming utilizing the Krovetz stemmer on record and also associated CQA answers. In this approach, three variations are proposed to this underlying DocOpt synopsis display:

4.2 QueryOpt (question one-sided):

This strategy adjusts the above model to create query-one-sided synopses. Naturally, it produces summaries that cover the same number of vital inquiry words as possible and at the same time limits excess. Words in the report are weighted in view of their events in the inquiry, as pursues:

$$w_j = tf_{j,query} \times idf_j \quad (4)$$

Where t , query is the recurrence of word ej in the question. Accordingly, sentences in the archive include question terms. So this technique is more successful when the inquiries (or inquiries) are long. The computed weights of words are then fused into the model in Eq 1 that is advanced to create outlines. This method is exceptionally effective since the advancement organize considers only question words which includes words and sentences. Then it sets λ in Eq 1 to 0.1, following the best setting is made.

4.3 AnswerOpt (CQA-reply one-sided):

This technique recognizes words that are probably going to be utilized in these answers, and after that produces rundowns that cover whatever number of these words as could be expected under the circumstances. Instinctively, words that show up in many associated CQA answers are allocated a high weight (see Eq 5). As mentioned in area 3, the associated CQA answers for a standard particular inquiry comprise of the best responses for the main 10 coordinating inquiries. The response for a coordinating question recovered in the lower position is probably going to be less applicable. Subsequently, a punishment of the record of an answer's situation in the CQA query item list is connected:

$$w_j = (\sum_{p-1}^{|CQA|} \frac{tf_{j,answer_p}}{\ln(1+p)}) \times idf_j \quad (5)$$

Where t ,$swerp$ is the recurrence of word ej in the response at pth position; |CQA| is the aggregate number of associated CQA answer for the inquiry (i.e. greatest of 10). We played out a 9-overlay cross approval (CV) to enhance λ in Eq 1 in range [0.0,… , 1.0] with venture of 0.1, that augments the ROUGE-2 score. The decision of 9-crease makes an adjusted subsection of the 45 queries. The normal ideal λ esteem was 0.2.

Chapter 5

DATA FLOW DIAGRAM

5.1 Text Document Summarization Workflow:

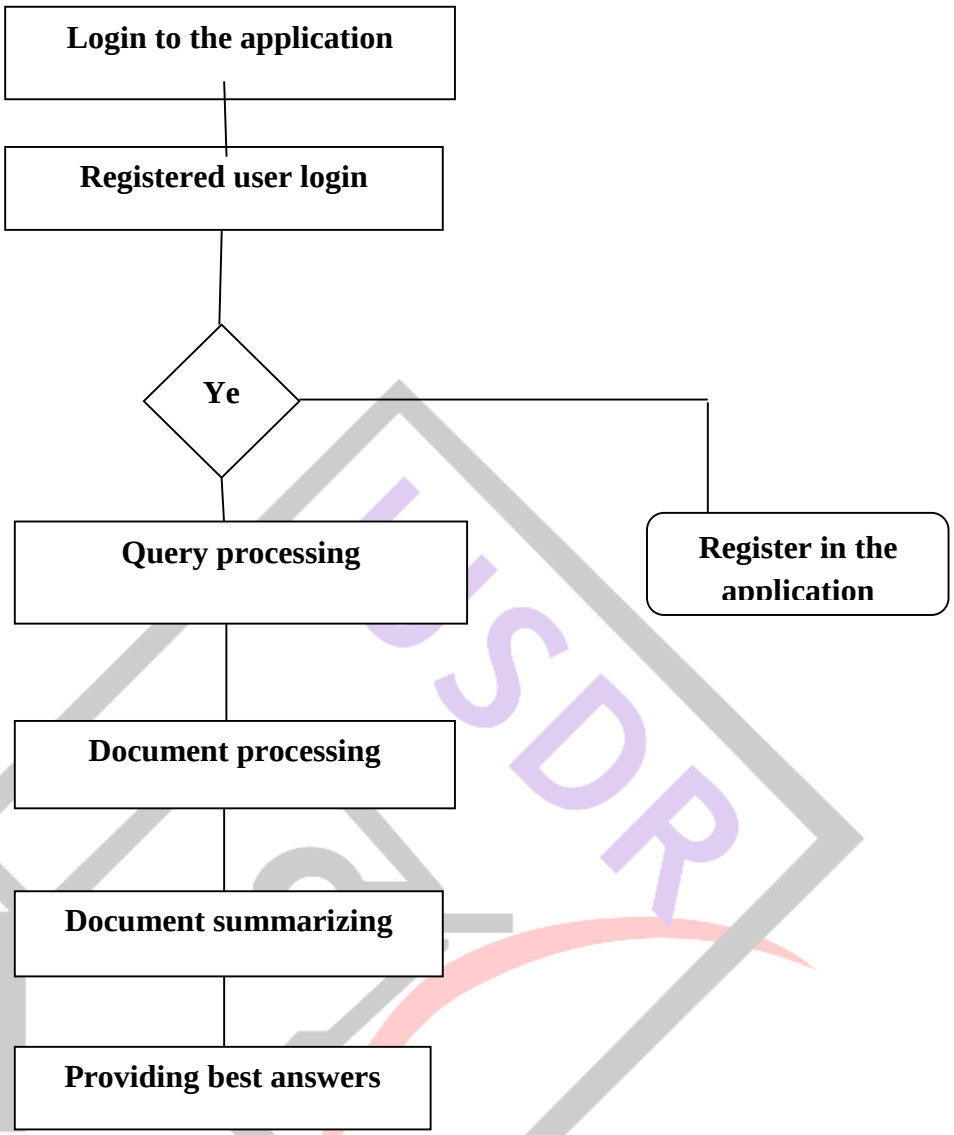


Figure 3: DFD of Text Summarization.

Figure 3 shows the DFD of text summarization where query is processed by removing the stopwords, once the query is processed the document is also processed and both the query and the document is weighed and frequency count is calculated.

5.2 Document and summary:

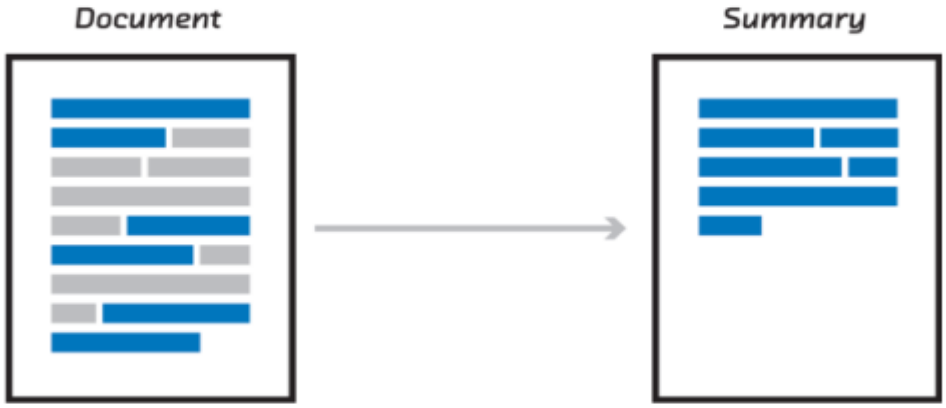
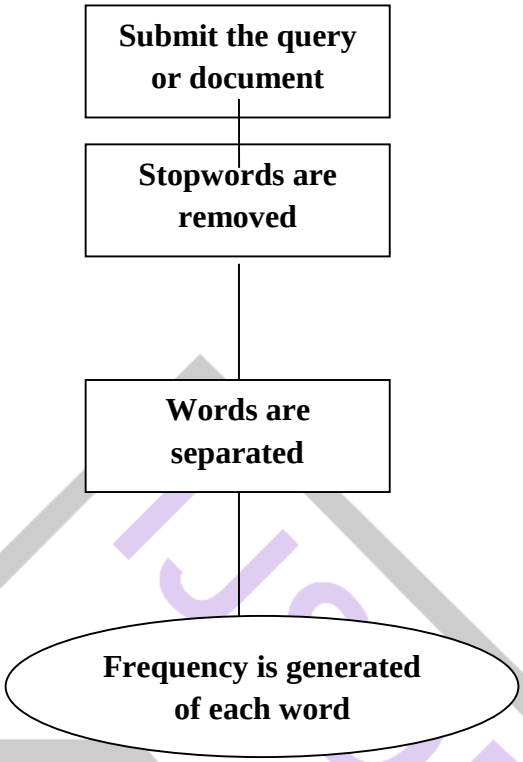


Figure 4: Document and summary.

5.3 Query and document processing:

Once the query or document is submitted the stopwords are removed and the important words are separated and frequency count is calculated of each word.

Figure 5: processing of document and query.



Chapter 6

RESULTS AND DISCUSSION

This approach used query one sided method where the summaries are generated with the best answers which cover maximum number of query words. The answers generated will be annotated with excellent, good and average answers. The answer annotated with excellent will be on the top searched result and it will cover the maximum query words as possible. The answer annotated with average and good will cover the query words less than the excellent answer.

Figure 6: Searched Answers in the Application

6.1 Word weighing of each word in the Document:

Words are weighed and document frequency count is calculated with the required formula and inverse document frequency is calculated with the log formula used.

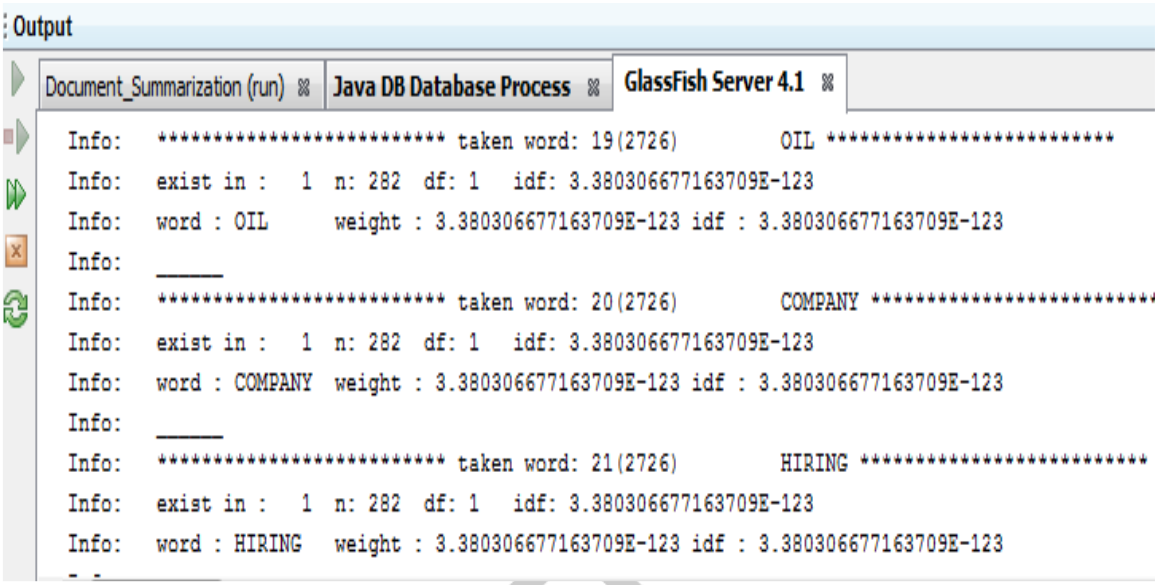


Figure 7: Word Weighing of Each Word.

6.2 Processing of the Query:

In this process query processing will take place which removes the unwanted words and special symbols which are not used to retrieve answers for the query.

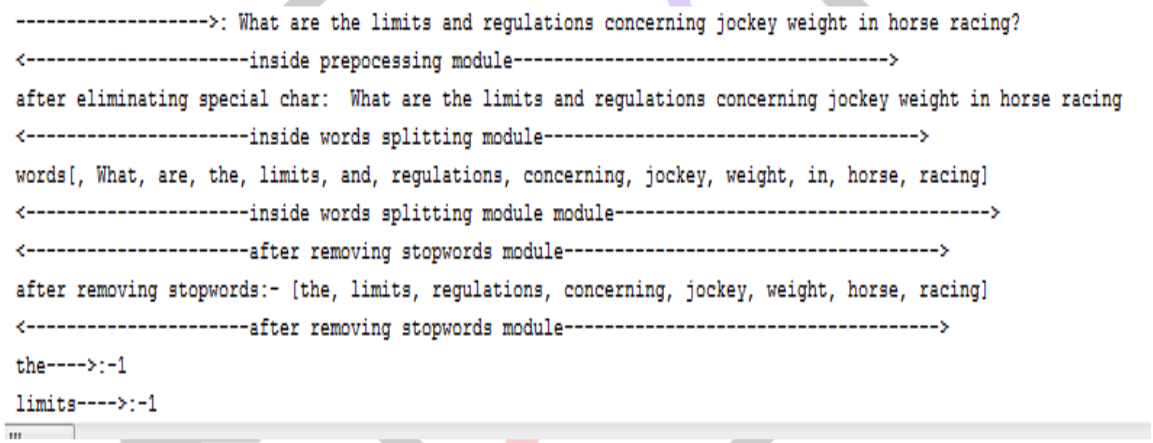


Figure 8: Query Processing.

Chapter 7

CONCLUSIONS AND FUTURE WORK

It propose to utilize outer data from interrelated CQA substance to control the mining of an answer-one-sided synopsis from each recovered record. Three development-based techniques to figure out how to-grade-based strategies were proposed. These outcomes demonstrate that the associated CQA information that doesn't really hold culminates reply to the inquiry, are valuable to extricate enhanced answer-one-sided synopses from records. A large portion of the web and flow examine depends on extractive multirecord rundown, current synopsis frameworks are generally used to condense news and other online articles. The vast majority of the early strategies were manage based while the present one applies factual methodologies. Inquiry based procedure offer thought to client inclinations. It is a approach which facilitates precise answers without encouraging the bad answers for the user search results.

Further it enhances with a feature where the user can submit if they are satisfied with the retrieved answers. The optimization trick is well suited to acquire relevant information with respect to the user search questions; this proves the answer to be gathered in a direct way among the web users.

REFERENCES

[1] Referred the paper for literature survey, authored by L. B. Chilton and J. Teevan, "Addressing People's Information Needs Directly in a Web Search Result Page," in Proc. 20th Int. Conf, pp. 27–36, World Wide Web, 2011.

[3] Referred the paper for literature survey, authored by M. S. Bernstein, J. Teevan, S. Dumais, D. Liebling, and E. Horvitz, "Direct Answers for Search Queries in the Long Tail," in Proc. SIGCHI Conf. Human Factors Compute. Syst., pp. 237–246. 2012.

[19] Referred the paper for literature survey, authored by Q. Liu et al., "Predicting Web Searcher Satisfaction with Existing Community based Answers," in Proc. 34th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 415–424, 2011.