IJCSE International Journal of Computer Sciences and Engineering Open Access
Research/ Review/Survey Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Malayalam Question Answer System Based On Inverted Index And N-Gram Model

Teena Shereef [1*] and Rosna P. Haroon [2]

[1*] Dept. of Computer Science, ICET, KTU, Muvattupuzha, Kerala, India
[2] Dept. of Computer Science, ICET, KTU, Muvattupuzha, Kerala, India

e-mail: teenashereef@gmail.com, rosna.haroon@gmail.com

*Corresponding Author: teenashereef@gmail.com, Tel.: +91 9895309915

Available online at: www.ijcseonline.org

Received: 01/04/2018, Revised: 00/…./2017, Accepted: 18/May/2017, Published: 30/Aug/2017

***Abstract*—** In data recovery, Question Answering (QA) is a procedure of offering answer to an inquiry which is postured by human in Natural Language (NL) using an arrangement of put away archive. This paper proposes the closed domain Question Answering System for taking care of the Malayalam News Articles and information about current undertakings to recover more exact answers. This framework tries to extricate the exact answers from ontology perception base for the inquiry identified with current issues asked by client. Keywords from queries and answer database plays an essential role for extracting answer. Ontology predicated testing is a standout amongst the most unfurling subfields of Information Extraction. To indicate a relation between substance to verb or to different elements some morphophonemic changes are incorporated in Malayalam sentences. These transmutations are refined using vibhakthi. Here we proposed a QA framework which gives word level answer using a few tenets and vibhakthi investigation of recovered report words, and the words in the given inquiry set.

***Keywords*—** NLP, Ontology, Information Retrieval, Inverted Index, N-GramModel

I. INTRODUCTION

All web indexes accessible these days have extraordinary thriving and abilities, however the difficulty with these web search tools is that in lieu of giving an exact response to the client's inquiry they give a list of documents. The essential goal of Question Answering system is to process requests in natural language form and to give the exact short reactions to them.

Building a fully capable QA system depends on two steps: Information Retrieval (IR) and Natural Language Processing (NLP). IR is enhanced by integrating NLP functionalities at a enormous scale. The possible NL document collections used for QA systems include: a local collection of reference texts, a set of Wikipedia pages and a subset of World Wide Web pages.

Closed-domain question answering:

Closed-domain question answering refers to specific domain related questions and can be seen as an easier task because NLP systems can provide domain-specific knowledge. It has very high accuracy but limited to single domain.

O*pen-domain question answering:*

Open-domain question-answering deals with the questions which are related to every domain. In this systems, mainly have more data available from which the system extract the answer. It can answer any question related to any domain but with very low accuracy as the domain is not specific.QA System has the basic modules as:

- Question Processing
- Document Retrieval
- Answer Extraction

The question processing module identifies the semantics of question which is asked by the user. The main tasks of this modules are:

a) Determining the question type
b) Determining the answer type
c) Extracting query words

After getting the keywords ,all the documents containing these keywords are going to get retrieve in the document retrieval module. Document retrieval is defined as the

matching of some verbalized user questions with a set of text records.

Answer extraction (AE) module is a key component in QA system which will extract the more precise answer from the retrieved documents. Document retrieval module calculate the homogeneous attribute rate between the user's question words and each document. Answer Extraction techniques are applied to the top ranked document and extract the final answer. The design and implementation of question answering system consist of the question analysis, query preprocessing, document retrieval, answer extraction and answer ranking.

Rest of the paper is organized as follows, Section II contain the related work of QA systems, Section III contain the proposed method , Section IV explain the solution methodology, section V describes the experimental results with graph and then Section VI concludes research work with future directions.

II. RELATED WORK

A. Efficient natural language pre-processing for analyzing large data sets: Belainine Billal, Alexsandro Fonseca and Fatiha Sadat: This paper describes an efficient pipeline for pre-processing the big data for NLP by using traditional machine learning and NLP techniques. . It also based on graph theory and semantic relationship between different words. Here this paper take twitter content as an example for big data. It integrates various traditional NLP methods.

B. A Rule Based Question Answering System in Malayalam corpus Using Vibhakthi and POS Tag Analysis: Archana S.M, Naima Vahab and C. Raseek: proposed a question answer system which gives an answer by differentiating vibhakthi of question words. This paper first check the question word and check the vibhakthi of the question word by identifying the suffix using sandhi splitting and if there is a word showing the same vibhakthi in answer set, retrieve that as answer.

*C. Question Answering System on Education Acts Using NLP Techniques : Sweta P. Lende, and Dr.M.M. Raghuwanshi :*This paper which describes the different methodology and implementation details of question answering system for general language and also proposes the closed domain QA System for handling documents related to education acts sections to retrieve more precise answers using NLP techniques. Here they made an inverted indices and used jaccard similarity for finding score.

*D. Intelligent Information Retrieval within Digital Library using Domain Ontology : Thinn Mya Mya Swe :*Ontology is also denote as a concept based search. This paper proposed an effective way to retrieve a set of result from digital libraries by proposing an ontology framework for semantic retrieval. Domain based ontology gives a better performance than traditional information retrieval. Here they used Query expansion method.

*E. Information Retrieval based on a Query Document using Maximal Frequent Sequences : Ricardo Merlo-Galeazzi, J. Ariel Carrasco-Ochoa, and J.Fco. Martínez-Trinidad :*This paper proposes an information retrieval process to retrieve a set of similar documents related to a query documents by the depiction based on Maximal Frequent Sequences. After testing the method this system gives a good performance in obtaining results.

*F. Efficient Information Retrieval Using Domain Ontology : Pratibha S. Sonakneware :*Search engine understand the sense of an applicant's query and the connections among the ideas in a document connect to a specified domain. This paper proposed an information retrieval system which receives a homogenous collection of documents that related with the user's query meaning by forming a SPARQL query.

*G. Sandhi Splitter for Malayalam Using MBLP Approach : Nisha M, Reghu Raj P C :*This paper presents a method to find the suffixes or modified words of Malayalam. Here they implemented a Memory Based Language Processing (MBLP) method for Malayalam word identification. MBLP is an approach to language processing according to the model storage during learning and similarity reasoning during processing.

H. A Classification of Sandhi Rules for Suffix Separation in Malayalam: Mary Priya Sebastian, Sheena Kurian K and G. Santhosh Kumar: In this paper, various rules designed for the suffix separation process in the English Malayalam SMT are presented. A classification of these rules is done based on the Malayalam syllable preceding the suffix in the modulated form of the word (check letter). By examining the check letter in a word, the different suffix separation rules that related with the check letter can be directly applied to extract the root words.

*I. A Rule Based Approach For Root Word Identification In Malayalam Language: Meera Subhash, Wilscy. M and S.A Shanavas :*The Root Word Identifier proposed in this paper a rule based approach which will eliminate the suffix part and derive the root word using morphophonemic rules. The system will undergo a number of suffix stripping steps.

*J. Literature Review: Stemming Algorithms for Indian and Non-Indian Languages : R. Vijaya Lakshmi, IIDr. S. Britto Ramesh Kumar :*Stemming is an important preprocessing step in NLP methods. This paper gives a brief idea about

stemming methods of various languages. Stemming is the process of deriving the root word from the affected forms after applying some root word extraction methods.

*K. Context- Aware Restricted Geographical Domain Question Answering System : Amit Mishra, Nidhi Mishra and Anupam Agrawal :*This paper deals the development of a geographic domain QA system deals with the mapping abilities help us to add location related with the query keywords. Hence this paper will be useful in retrieving the geographical features and gives a satisfactory performance.

*L. A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository : Pragisha K. and Dr. P. C. Reghuraj :*This paper describes the design and execution of a QA System in Malayalam. The system extract answers of Malayalam truthful questions by checking a corpus of Malayalam documents. It is maintained as a closed domain system that contain four classes of Malayalam questions. It uses the main traditional techniques in NLP.

*M. Question Analysis for a Closed Domain Question Answering System : Caner Derici, Kerem C¸ elik, Ekrem Kutbay and Yi˘git Aydın :*This is another paper evaluates the means by which developed a closed domain QA system that is used for high-school students to support their education. This paper deals two difficulties like focus extraction and question classification.

*N. An Effective Reasoning Algorithm for Question Answering System : Poonam Tanwar, Dr. T. V. Prasad and Dr. Kamlesh Datta:*Knowledge representation (KR) is one of the advisable area for research to make one system perceptive. The aim of this paper work is to make the constructive knowledge representation method for constitute the general knowledge and an algorithm for QA system (QAS) work as information, so that suitable knowledge can be theorize from the system.

*O. A Reasoning Methodology for CW-Based Question Answering Systems : Elham S. Khorasani, Shahram Rahimi, and Bidyut Gupta :*The paper of Computing with Words put forward a mathematical tool to conventionally indicate and cause with responsive information. This paper gives a QA system for a restricted domain Computing with Words system. This method takes the query using the constraints and represent it as a tree structure. This propagation tree generate a proper method to find the answer.

*P. Effective Question Answering Techniques and their Evaluation Metrics : Jaspreet Kaur, Vishal Gupta :*Here paper gives an overall idea about question answering system. The main modules in a question answering system are question processing, document retrieval and answer extraction. In query preprocessing module first perform some preprocessing techniques like tokenization and get the keywords. Then next module will retrieve the documents that contain these key words. Answer extraction module will extract the precise answer from the document.

*Q. Inverted index and interval lists for keyword search : J Giridharan, and Dr. S. V. Vairavan :*Documents are always store as list, but in inverted indexes store words in a tabulated form and that inverted index will show the keyword and the documents in which the keyword appears and the number of occurrence of each keyword. This paper presents the searching of words on inverted indexes, their types and techniques.

*R. Ginix: Generalized Inverted Index for Keyword Search : Hao Wu , Guoliang Li, and Lizhu Zhou :*Inverted lists used to retrieve the stored documents based on a number of keywords efficiently. This paper describe an accurate index structure, the Generalized INverted IndeX (Ginix), in which the system combines consecutive IDs in inverted lists to save storage space.

*S. A Document Retrieval System with Combination Terms Using Genetic Algorithm : S.Siva Sathya, and Philomina Simon :*This paper proposed a system which contain three methods to deal with the keyword and information retrieval from database. First draw out the keyword then make a selection of genetic algorithms. Third step is make use of the genetic algorithms in information retrieval process to get the correct result.

T. A Sandhi Splitter for Malayalam : Devadath V V ,Litton J Kurisinkel ,Dipti Misra Sharma and Vasudeva Varma :This paper shows various rules for Sandhi splitting. Sandhi splitting helps us to find the suffix present with the root word. Hence by identifying root word and suffix present in a sentence helps to identify different vibhakthis. This paper using the phonological changes that take place in the words while joining. This resulted in a combined method which identifies the split points and splits using predefined character level linguistic rules.

III. PROPOSED METHOD

After analyzing the literature survey of many papers, we can understand that the closed domain QA System gives a precise answer than the open domain QA System as it only defines a categorical domain to give efficient answer for user query. If we are going through the scenario of queries cognate to Malayalam news articles, there is no such a QA system, which ascertains the correct answers. The user generally asks questions in any form regarded to current affairs, and all these questions should get an answer related to Malayalam news articles.

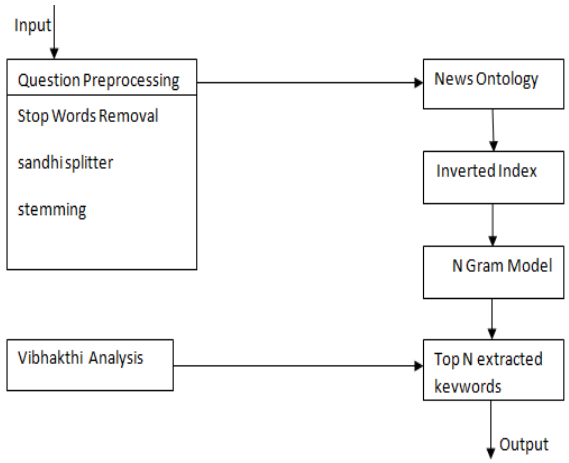

Fig. 1 Proposed Architecture of QA System

Information retrieval (IR) is the method of pursuing an information from an amassment of information resources by utilizing some keywords search. In other words, it is the detection and revelation of concrete information from stored data. Mostly the traditional way of information retrieval was done by the keyword search. Here this keyword search will perform in an inverted index to retrieve the documents where this Inverted Index engenders from an Ontology of News.

IV. SOLUTION METHODOLOGY

Collection and Study of relevant data set:

The initial design phase of the proposed work contain the Collection and Study of pertinent data. For the verbally expressed work the applicable data set is the records of Malayalam news articles. The needed data set is accumulated and studied from different news paper sites that handle worldwide news.

A) Preprocessing

Once we have accumulated the information it is indispensable to perform some preprocessing operations in each single file of documents. Then only the creation of Ontology will get complete. The preprocessing steps here applied are Stop Words removal, Sandhi Splitter and Stemming.

a. Stop words Removal:

Stop words are words which strain out before or after processing of natural language data. The often used stop words in documents are prepositions and pro- entities like “is, for, the, in, etc” . Stop words are abstracted from dataset because those are not taken as keywords in IR applications.

b. Stemming:

Stemming is the process of abstracting all the suffixes appended to it and converting that words in to its root word. A stemming algorithm converts the words "fishing", "fished", and "fisher" to the root word, "fish".

c. Sandhi splitter

Sandhi splitting is one of the main preprocessing method in Malayalam like language. Sandhi splitter splits the joined words into individual words. In this paper sandhi splitting avails to identify the suffix joined with the root word. This suffix will decide the vibhakthi assigned in this question.

B) Creation of Corpus (Ontology based)

After the amassment of all data next step is to engender a corpus. Corpus is a database which contain all the document that we have accumulated from different websites that cognate to news. Here we are utilizing a News based ontology (corpus), as domain. An Ontology is a formal denominating and definition of the types, properties, and interrelationships of the entities that authentically subsist in a particular domain.

Design the domain concrete ontology following steps are consider.

Step 1: Accumulate all details regarding the domain.
Step 2: List paramount terms in ontology.
Step 3: Identify the classes and subclasses can be created .
Step 4: Third step is to identify the properties and characteristics subsists in classes and subclasses.
Step 5: Bind the appropriate properties to congruous class.
Step 6: Developed ontology.

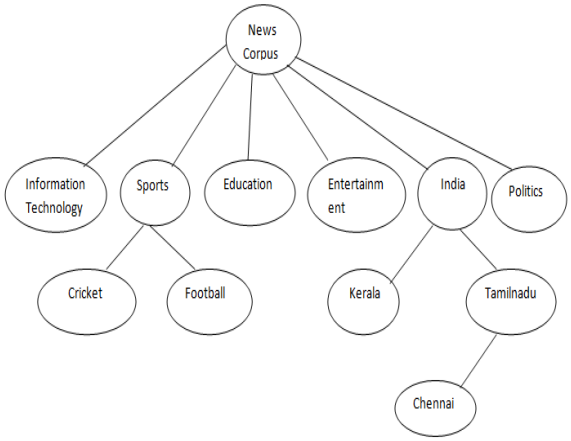

Fig. 2 News Ontology For QA System

C) Inverted Index

An inverted index is a construction which cache the frequency and the occurrences of the keywords in a selection of documents. It contain the frequency of words in a document, so that it will reduce the computation process at the time of retrieval of data.

The following algorithm is used to generate an Inverted Index of the document extracted.

Algorithm CreateIndex (*heap, stemmer, onto*) {

- For Each Record *rec* in *heap*{

- *rec_entry* = *index*.addRecEntry(*rec*.id);
- For Each Word *wrd* in *rec*.fullText {
- If *wrd* is compound word *wrd_ont* in *onto* {
- *wrd* = *stemmer*.stem(*wrd_ont*);
- } Else {
- *wrd* = *stemmer*.stem(*wrd*);
- }
- *rec_entry*.countOccurence(*rec ,wrd*)
- If *rec_entry* is not in *index*.termEntries {
- *index*.addTermEntry(*wrd*);
- }
- *term_entry* = *index*.getTermEntry(*wrd*);
- *term_entry*.countOccurence(*rec,wrd*);
- }
- }
- *index*.computeWeightTerms();
- Returns index;
- }

D) IR From Inverted Index Using N-Gram search

In the fields of computational semantics , an ***n*-gram** is a neighboring series of *n* items from a text or speech. There are *n*-gram with size of 1,2,3 known as unigram, bigram and trigram respectively. Larger size are mentioned by the value of *n*.

Table 1: Sample N-Gram Model

N=1: T his is a sentence	**unigram**	**this,** **is,** **a,** **sentence**
N=2: T his is a sentence	**bigram**	**this is,** **is a,** **a sentence**
N=3: T his is a sentence	**trigram**	**this is a,** **is a sentence**

At the point when a user question(uq) is given check the coordinated terms in Inverted Index and afterward recover the best positioned reports which contain the terms in uq. The picked terms will sort identified with the recurrence and best n-keywords will recover. The conditional probability for uq and all the retrieved keywords will calculate.

Here, conditional probability for bigram is calculated for the keywords/terms {*t(1,1), …., t(1,n)*}, {*t(2,1), …., t(2,n)*}, …. {*t(r,1), …., t(r,n)*} when *uq* is given. For predicting the 3rd keyword, the first and the predicted second keyword will select and probabilities are added which will give a keyword higher frequency in one or more documents. Then this documents are stored as separate list and check for the intersected document. This intersection will help to know the probabilities of keywords in document. The top frequency keyword will select from intersected document and calculate the conditional probability.

ALGORITHM

•Step1: For each user query.
•Step2: Query term is compared for matching terms from the engendered corpus of inverted index
•Step3: Documents containing matched terms and having maximum frequency are culled
•Step4: Threshold tr is set for culling top-r documents from the culled documents list
•Step5: Extract all the terms or keywords {t1,t2,t3,…..,tn} from the culled top-r documents and sorted depending on frequency
•Step6: Threshold tn is set to consider the extracted top-n keywords
•Step7: Calculate conditional probability for the top-n extracted keywords t1,t2,t3,…..,tn} to presage next word
• Step8: Rank and suggest top n presaged words to the user.

E) Vibhakthi Analysis

Ashtadhyayi is a celebrated grammer message in Sanskrit presented by Panini , an antiquated Sanskrit philologist. Vibhakthi is usually utilized for communicating the connection between thing to verb in a sentence or between two things. As indicated by Ashtadhyayi 7 vibhakthi frames are there;

"നിർദേശിക"	(Nirdheshika)
"പ്രതിഗ്രാഹിക"	(Prathigraahika)
"സംയോജിക"	(Samyojika)
"ഉധേഷിക"	(Udheshika)
"പ്രയോജിക"	(Prayojika)
"സംബന്ധിക"	(Sambandhika)
"ആധാരിക"	(Aadhaarika)

This proposed framework initially inspects the query by user, and choose which question word is utilized. For recognizing the vibhakthi , the postfix added ought to be stripped first. After N-gram seek there will be an arrangement of top of the line keywords and prepared to think about the vibhakthi of

these words. The framework will recover the keyword that has vibhakthi same as the vibhakthi of question word and retrieve as answer.

Table 2. Vibhakthi analysis of question words

Case	Suffix	Question Word
നിർദേശിക	No Suffix	ആര് , ഏത് , എന്ത്
പ്രതിഗ്രാഹിക	എ	ഏതിനെ, ആരെ, എന്തിനെ
സംയോജിക	ഓട്	ആരോട്, എന്തിനോട്
ഉധേഷിക	ക്ക്, ന്	ആർക്ക്
പ്രയോജിക	ആൽ	ആരാൽ
സംബന്ധിക	ഉടെ , ന്റെ	ആരുടെ, ഏതിന്റെ
ആധാരിക	ഇൽ, കൽ	ആരിൽ, എന്തിൽ

V. EXPERIMENTAL RESULTS

In this experiment we have generated an Ontology which contains 50 documents and 3000 terms that related with worldwide news from different websites. Consider the terms t1 be കേരളം(Kerala) t2 be മന്ത്രി(minister) etc. and the corresponding postings are given in an order. If a user enters a query related to കേരളം(Kerala) then keyword will get after tokenization and it will check the matched document from Inverted Index and then the corresponding posting list to be retrieved. *"<2,2> <8, 1> <14, 1> <24 , 6> <32 , 8><48,28>..."* These are the postings from retrieved from Inverted index and it also has the position of word കേരളം(Kerala) and then select top 'r' documents. At that point select top 'n' keywords from each documents by considering the frequency of occurrence. Hence the number of keywords 'n' from 'r' documents will be n*r.

Table 3:A Sample posting list

Term	1-Gram sample	Postings list
t1	കേരളം(Kerala)	*<2,2>,<16,1>,<20, 1>,<24,6>,<25, 8>*
t2	മന്ത്രി(minister)	<10,19> ,<7,8>,<13,12>,<14,7>,<19,9>

The conditional probability is computed for the best n*r watchwords with the word കേരളം(Kerala) . The conditional probability is characterized as the likelihood of an occasion (A) guaranteed that another occasion (B), which has just happened and is figured as,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (1)$$

The calculated probability of the keywords are sorted and highest weighted keyword will select as the second predicate term which joined with the first word കേരളം(Kerala).

Table 4:The calculated probability of sample of 2-grams

Keyword	Suggested 2-grams	Weighted Probability
കേരളം(Kerala)	വിദ്യാഭ്യാസം(Education)	0.878
	മുഖ്യമന്ത്രി(Chief Minister)	0.444
	ചരിത്രം (History)	0.328
	ഭൂമിശാസ്ത്രം (Geography)	0.285
	ആഘോഷങ്ങൾ(Festivals)	0.257
	ടൂറിസം (Tourism)	0.232
മന്ത്രി(minister)	ആഭ്യന്തരം(Home Affairs)	0.536
	വിദ്യാഭ്യാസം(Education)	0.514
	ധനകാര്യം (Finance)	0.354
	വൈദ്യുതി (Electricity)	0.232

Consider if user has given മുഖ്യമന്ത്രി(Chief Minister) as the next term in query, then the posting related with മുഖ്യമന്ത്രി(Chief Minister)will retrieve. These two lists compared and intersected documents will retrieve and this is carry out to get the keywords that correlate with കേരളം(Kerala). Again start with the same process for 3rd term and select another set of top ' r ' documents. Calculate the 3rd gram conditional probability of words A,B,[C using equation,

$$P(C|AandB) = \frac{P(A \cap B \cap C)}{P(B)P(A|B)} \qquad (2)$$

For assessing the performance, we have computed the documents from the Inverted Index list rundown and the vast majority of them are physically checked. We have utilized the standard execution measures exactness and review and F-

measure for assessing the proposed approach by utilizing following conditions.

$$Precision = \frac{Number\ of\ relevant\ document\ retrieved}{Total\ documents\ retrieved} \quad (3)$$

$$Re\ call = \frac{Number\ of\ relevant\ document\ retrieved}{Total\ number\ of\ relevant\ documents} \quad (4)$$

$$F\ Measure = \frac{2\ Precision * Recall}{Precision + Recall} \quad (5)$$

The test comes about shows exactness and review against number of intersected keywords recovered after n-gram search in Fig.3 and Fig.4. It is seen that the Precision rate will has a tendency to run high with increment in the quantity of intersected documents in n-grams. Thus, it is watched that when the quantity of n-grams builds, the recall rate diminishes.

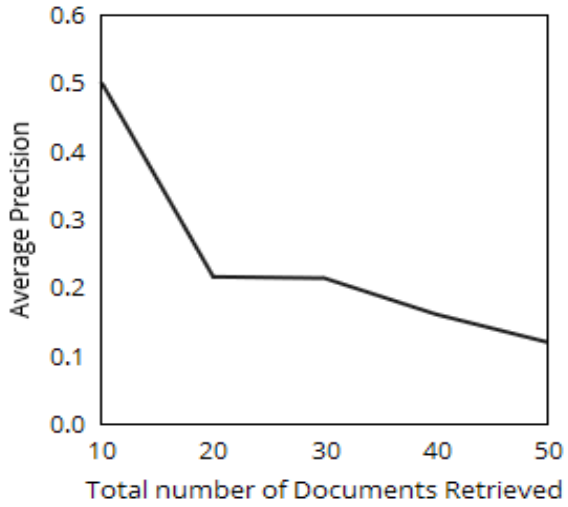

Fig. 3 Average precision For Retrieved Keywords

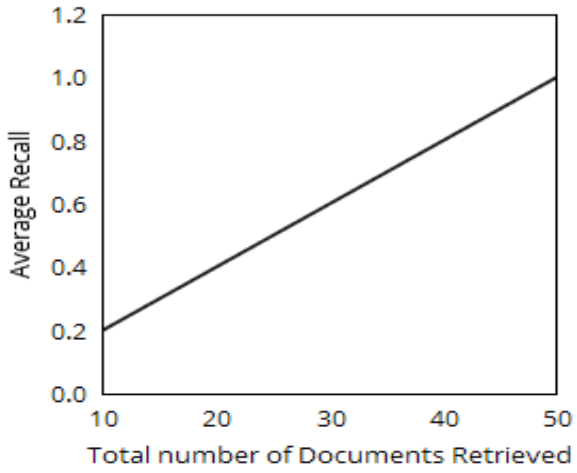

Fig. 4 Average recall For Retrieved Keywords

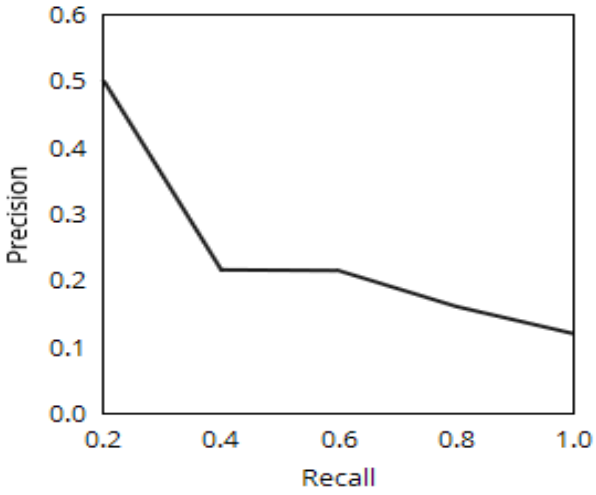

Fig. 5 Average Precision vs. Recall for Retrieved keywords

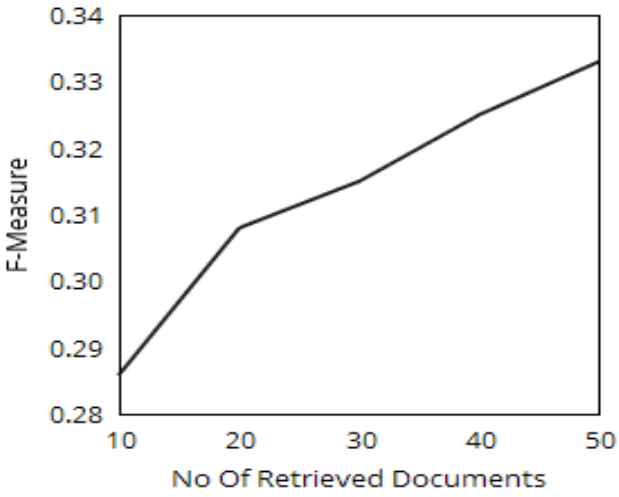

Fig. 6 F-measure for Retrieved Keywords

At that point plot F-Measure by thinking about both review and exactness into account and is given in Eq. 4. From Fig 6, it is watched that the quantity of document recovered increases, the F-measure additionally increments.

VI. CONCLUSION

In any web crawler framework, the question handling part is a critical part. The question refinement system is considered as an appropriate instrument for enhancing the user inquiry to acquire better accuracy of recovery. In this paper, we have proposed a way to deal with make an Ontology interfacing news and made a modified Index posting for refining the user's question. This method creates the corpus in an effective way such that only the keywords in the documents along with their frequency of occurrence are used. During the creation of n-gram corpus, the conditional probability is

calculated for matching the pattern. After the determination of top of the line keywords Vibhakthi investigation have done and recovered the exact answer. The proposed approach is tried utilizing the reports. From the test result, we found that the proposed approach gives better exactness of recovery.

, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

REFERENCES

[1] Belainine Billal, Alexsandro Fonseca and Fatiha Sadat "*Efficient natural language pre-processing for analyzing large data sets*",in IEEE International Conference on Big Data (Big Data),December 2016.

[2] Archana S.M, Naima Vahab and C. Raseek , "*A Rule Based Question Answering System in Malayalam corpus Using Vibhakthi and POS Tag Analysis*", in Elsevier Ltd 2212-0173 © ScienceDirect,Procedia Technology 24 (2016) 1534 – 1541.2016

[3] Sweta P. Lende, and Dr.M.M. Raghuwanshi, "*Question Answering System on Education Acts Using NLP Techniques*," in IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare, 2016

[4] Thinn Mya Mya Swe, "*Intelligent Information Retrieval within Digital Library using Domain Ontology*",in Proceedings of the International Conference on Applied Computer Science, 2008

[5] Ricardo Merlo-Galeazzi, J. Ariel Carrasco-Ochoa, and J.Fco. Martínez-Trinidad, "*Information Retrieval based on a Query Document using Maximal Frequent Sequences*" in 32nd International Conference of the Chilean Computer Science Society, 2013.

[6] Pratibha S. Sonakneware, "*Efficient Information Retrieval Using Domain Ontology,"* in International Conference for Convergence of Technology - 2014.

[7] Nisha M, Reghu Raj P C, "*Sandhi Splitter for Malayalam Using MBLP Approach"*, in Elsevier Ltd ScienceDirect, Procedia Technology 24 (2016) 1522 – 1527.2016

[8] Mary Priya Sebastian, Sheena Kurian K and G. Santhosh Kumar, *"A Classification of Sandhi Rules for Suffix Separation in Malayalam"*, in Proceedings of Fourth International Conference on Information Processing,Bangalore, India (2011)

[9] Meera Subhash, Wilscy. M and S.A Shanavas, "*A Rule Based Approach For Root Word Identification In Malayalam Language*", in International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 3, June 2012

[10] R. Vijaya Lakshmi, IIDr. S. Britto Ramesh Kumar, "*Literature Review: Stemming Algorithms for Indian and Non-Indian Languages*", in International Journal of Advanced Research in Computer Science & Technology Vol 41, No 8, IJARCST 2014

[11] Amit Mishra, Nidhi Mishra and Anupam Agrawal, *"Context-Aware Restricted Geographical Domain Question Answering System",* In International Conference on Computational Intelligence and Communication Networks March 2014 ISSN 2104 – 0635 2010

[12] Pragisha K. and Dr. P. C. Reghuraj, *"A Natural Language Question Answering System in Malayalam Using Domain Dependent Document Collection as Repository."* International Journal of Computational Linguistics and Natural Language Processing Vol 3 Issue 3 March 2014

[13] Caner Derici, Kerem C¸ elik, Ekrem Kutbay and Yi˘git Aydın, *"Question Analysis for a Closed Domain Question Answering System",* in Springer International Publishing Switzerland Part II, LNCS 9042, pp. 468–482, 2015..

[14] Poonam Tanwar, Dr. T. V. Prasad and Dr. Kamlesh Datta, *"An Effective Reasoning Algorithm for Question Answering System",* in Science and Applications,Special Issue on Natural Language Processing, Volume 29, 2 014.

[15] Elham S. Khorasani, Shahram Rahimi, and Bidyut Gupta *"A Reasoning Methodology for CW-Based Question Answering Systems",* in Springer-Verlag Berlin Heidelberg, WILF 2009, LNAI 5571, Volume 25– No.14 pp. 328– 335, 2009.

[16] Jaspreet Kaur, Vishal Gupta, "*Effective Question Answering Techniques and their Evaluation Metrics"*, in International Journal of Computer Applications (0975 – 8887) Volume 65– No.12, March 2013.

[17] J Giridharan, and Dr. S. V. Vairavan, *"Inverted index and interval lists for keyword search",* in Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on 2014.

[18] Hao Wu , Guoliang Li, and Lizhu Zhou *"Ginix: Generalized Inverted Index for Keyword Search ",* in tsinghua science and technology i s s nll1 0 07 - 0 2 14ll1 0/ 1 2llp p 7 7-8 7 Volume 18, Number 1, February 2013.

[19] S.Siva Sathya, and Philomina Simon, *"A Document retrieval System with Combination Terms Using Genetic Algorithm"*, in International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010 1793-8163

[20] Devadath V V ,Litton J Kurisinkel ,Dipti Misra Sharma and Vasudeva Varma, "*A Sandhi Splitter for Malayalam"*, in Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003.

