International Journal of Engineering and Techniques - Volume 4 Issue 2, April-2018

RESEARCH ARTICLE OPEN ACCESS

Malayalam Text summarization Using Vector Space Model

Kanitha D K [1], D. Muhammad Noorul Mubarak [2] & S. A. Shanavas[3]

1(Computational Linguistics, Department of Linguistics, University of Kerala)

2 (Department of Computer Science, University of Kerala, Kariavattom, Thiruvananthapuram)

3(Department of Linguistics, University of Kerala, Kariavattom, Thiruvananthapuram)

Abstract:

Automatic text summarization systems extract the significant sentences from the document and generate an accurate summary. The technique of text summarization is abstractive and extractive. Abstractive summarization understands the source text and generates new shorter text with same ideas. It requires language processing tools like Dictionaries, WordNet etc. Extractive summarization systems find the semantics of sentences and rank the semantically similar sentences and high scored sentences are selected to generate a summary. In extractive summarization statistical and linguistic methods are used to rank the sentences. The high scored sentences are selected as summary. Many techniques have been developed for summarization of text in various languages. In Malayalam, summarization systems are very few and it is in the beginning stage. This paper discusses about the semantic similarity method like vector space model and shows how ranking the sentences using this model and also gives the efficiency of proposed summarizer.

Keywords — **Natural Language Processing, Malayalam Text Summarization, Vector space model. Cosine similarity.**

I. INTRODUCTION

Now a day's numerous Malayalam documents are available from net. But finding the relevant data from various web pages is a heavy task. Reading every pages and find relevant data, it takes a lot of time and effort. At the same time user gets the summary of a document without reading the full document, it is fascinating. In this situation the methodology of text summarizer is very essential.

Text Summarization is the process of reducing the source text into shorter version preserve its information content and overall meaning [5]. Text summarization is a technique, where a text is entered into the computer and returns the summary of a text. The summary should be short and accurate. The technique has begins in 50's and wide scope in recent years. Some of the uses of summarization systems are summarize the text, summarize the legal documents, summarize the Govt. orders, summarize the foreign language text and user gets an abstract of document, summarize the online documents etc.

Text summarization methods can be classified into extractive and abstractive summarization (Hovy and Lin, 1997) [5]. Abstractive text summarization systems are same as human summarization in which system understand the original text and re-tell it in few words. Linguistic and statistical methods are used for text abstraction. Extractive text summarization extracts the significant sentences or paragraphs from the original document and concatenated into shorter form without drop the relevant information. Mainly statistical, heuristic and linguistic methods are used for extractive text summarization. The extractive summarization is simpler than abstractive summarization. Today most of the summarization systems follow extractive summarization methods rather than abstractive summarization methods. Summary generated from a single document is known as single document summarization. Summary generated from multiple documents on the same subject is known as multi-document summarization. Generic summarization systems generate summaries from the main topics of documents. Query-based summarization systems generates summary on the basis of matching of query word or key word.

Malayalam is a natural language especially used by the people of the State of Kerala in India. It is one of the scheduled languages in India and was designated a Classical Language in the year 2013. It has the official language status in Kerala and as well as in the union territories of Lakshadweep and Pondicherry. It belongs to the Dravidian family of languages. Research in Natural Language Processing for Malayalam is always challenging due to the agglutination, high ambiguity and rich morphology of words in Malayalam. The work done in the Malayalam summarization area is based on the term matching and term weight. Term matching identifies the sentence that includes the particular term and term weight the highest weighted sentences is extracted as summary.

This paper focuses to develop a tool for Malayalam text summarization based on vector space model. The road map of this paper is organized as follows. Section-2 gives a review on existing summarization methods especially concentrated on extractive summarization methods. Section-3 shows the methodology of proposed Malayalam text summarizer. Section-4 shows the analysis of result. Section-5 concludes the graft.

II. Related Work

Natural language processing begins in early when Alan Turing published paper titled as "Computing Machinery and Intelligence" and later it is called Turing Test [1]. Text summarization is an important process of NLP and it develops in early on 1950's. The first work on text summarization Luhn's method (1958) [2] considered sentence features such as word frequency and phrase frequency. Sentences are ranked on the basis of word frequency and phrase frequency. The high scored sentences are selected as summary sentences. The main drawback of the system is duplicate sentences in summary. Baxendale (1958) [3] proposed a straight forward method for sentence extraction. Sentences are selected on the basis of features of sentences such as document title, first and last sentences of a document or each paragraph. He proposed that in newspaper articles the first sentences are high chance to include in summary. But in technical papers the last sentence or concluding sections are having high chance to include in summary. On the basis of these heuristic assumptions sentences are selected as summary sentences. Lin and Hovy (1997)[5] claimed that Baxendale position method is not a suitable method for sentence extraction in different domains. Because the discourse structure of a sentence varies from different domains. The main disadvantage of this system was the summary sentences are selected on the basis of characteristics of domains. Edmundson (1969) [4] methods selects sentences on the basis of cue phrases, keywords, title words and location. Now many of the current automatic text summarization systems follow Edmunson's method. The main drawback of this system was duplication in summary. Barzilay and Elhadad (1997)[6] proposed a lexical chain method to score the sentences. The concept of lexical chain was introduced in Morris and Hirst, 1991. The lexical chain links the semantically related terms within different parts of document. Barzilay and Elhadad used Wordnet to construct the lexical chains. SweSum (Dalianis 2000) [7] was the first web based automatic text summarizer for Swedish and it summarizes Swedish news text in HTML based text. It is also available for Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi, and German Texts and it used statistical, linguistic and heuristic methods to obtain the summary sentences. The architecture of SweSum was client / server application. The web client input the original text and accepts the summarized text. The web server accepts the source text and performs tokenizing, scoring, keyword extraction and sentence ranking. The sentences are scored using statistical, linguistic and heuristic techniques such as position, numerical value, and font based feature etc. The score of each word is calculated and find the sentence score. A value is predefined and generated the desired number of summary. The query based text summarization [5] shows better result. The Summarist [4] algorithm used statistical approach for summarizing web documents. The lexical chain method [5] was used for the text connectivity or semantic relations. The lexical chains are formulated for finding the relevance of sentences used WordNet and dictionaries [12]. Text Rank [7] algorithm based on graphs theoretic approach the nodes are represents sentences and edges represents similarity between sentences. Lex Rank [9] is a graph-based algorithm same as TextRank.

Literature on text summarization clearly states that most of the current automated text summarization system used extraction method to produce summary. The extraction based systems followed some important features to be considered for including a sentence in final summary are [7]:

- Baseline: In texts the first sentence got highest score.
- First sentence: The first sentence of each paragraph of the text is ranked.
- Title: The title words held sentences got high score.
- Term frequency: The terms which are frequent in the text are more important than the less frequent terms in text.
- Sentence length: The score given to a sentence that reflects the number of words in a sentence, the length of the longest sentence is included in summary.
- Proper name: Sentences which contain proper nouns got high score.
- Average lexical connectivity: The sentences that share more terms with other sentences are scored higher.
- Numerical data: The sentences that contain any sort of numerical data are scored higher.
- Proper name: Certain types of nouns, like people's names, cities, places etc. are important.
- Pronoun: Sentences containing a pronoun (reflecting co-reference connectivity) are scored higher.
- Weekdays and months: Sentences containing names of weekdays or months are scored higher.
- Quotation: Sentences containing quotations may be important for some sort of questions, which are the input by the user.
- Query signature: When a user requires a summary on the basis of query. The query of the user affects the summary that the extracted text will be required to contain these words.

These features are the backbone of many text summarization systems. By evaluated these system summaries the semantics are very less. Some fuzzy sentences are selected as summary. At this time developers think about how to avoid these limitations and develop a good summarizer. Then authors proposed semantic similarity ranking method. One of the most commonly used semantic similarity method for information retrieval technique is the vector space model (Salton, 1975).

The vector space model is the sufficient method for extracting semantically similar sentences. Bag-of-words model is constructed and find the term and sentence frequency. Here document refers to text or text fragment, and it generally refers to an article. Term is the basic semantic unit of the document usually the words or phrases. Term weight is attached to each word denoting its importance in the document. The non-stop words that occur most frequently in the documents are treated as query. The TF value is proportional to the frequency of the word in the document. The IDF value is inversely proportional to its frequency in the documents. The term frequency and inverse document frequency (tf x idf) shows the importance of a word in a document or corpus. The tf-idf value increases proportionally to the number of times a word appears in the document. The way of ranking the documents are to measure how the vectors are close to the query vector.

Some of the limitations of vector space model are it requires lot of processing time and it cannot handle the Synonymy (Same meaning - Terms can be used to express same thing. Thus, the similarity of some relevant documents with the query can be low just because they do not share the same terms) and Polysemy (multiple related meaning- The terms can be used to express different thing in different contexts. Thus some irrelevant document has high similarities because they share some words from the query).

Bellotti T& Crook J. (2009) [14] proposed Support vector machines for extract the significant sentences.

.

III. MALAYALAM TEXT SUMMARIZATION

The proposed methodology is based on vector space model and it is used for summarizing articles in Malayalam. Some of the identified features of Malayalam are it has a rigid and vast grammar structure. It is an agglutinative in nature. It is a syllabic alphabet in which all consonants have an inherent vowel. The structure of sentences is simple, compound and complex. The morphology of language is inflectional, derivational and compounding. The main word classes are Noun, Verb, Adjectives, Adverbs, Postpositions and Conjunctions. The word order in Malayalam is Subject, Object and Verb.

The NLP in Malayalam is easy after the implementation of UNICODE. Thereafter computer understands the natural language and performs the various language processing activities. Numerous softwares are developed and implemented in Malayalam. The methodology of Text summarizer in Malayalam is explained below.

Algorithm:

- Step 1: Input the documents.
- Step2: Segment the whole text into small paragraphs.
- Step 3: Split the paragraphs into sentences and words.
- Step 4: Remove the stop words which remove the words that do not add to the individual meaning.
- Step 5: Terms are ready to processing where each unique word in a sentence is represented by the rows and sentences are represented by columns.
- Step 6: Calculate the term frequency (tf_i) of each term.
- Step 7: Calculate document frequency (df_i).
- Step 8: Calculate inverse Document frequency (idf_i = log(Total number of sentences/ df_i)
- Step 9: Calculate the term weight ($Wi = tf_i * IDF_i$) of sentences.

Step 10: Compute the similarity of sentences between the query words.

$Sim(Q,D_i) = \sum_i W_{Q,j} W_{i,j} / Sqrt(\sum_j W^2_{Q,j})* Sqrt(\sum_i W^2_{i,j})$

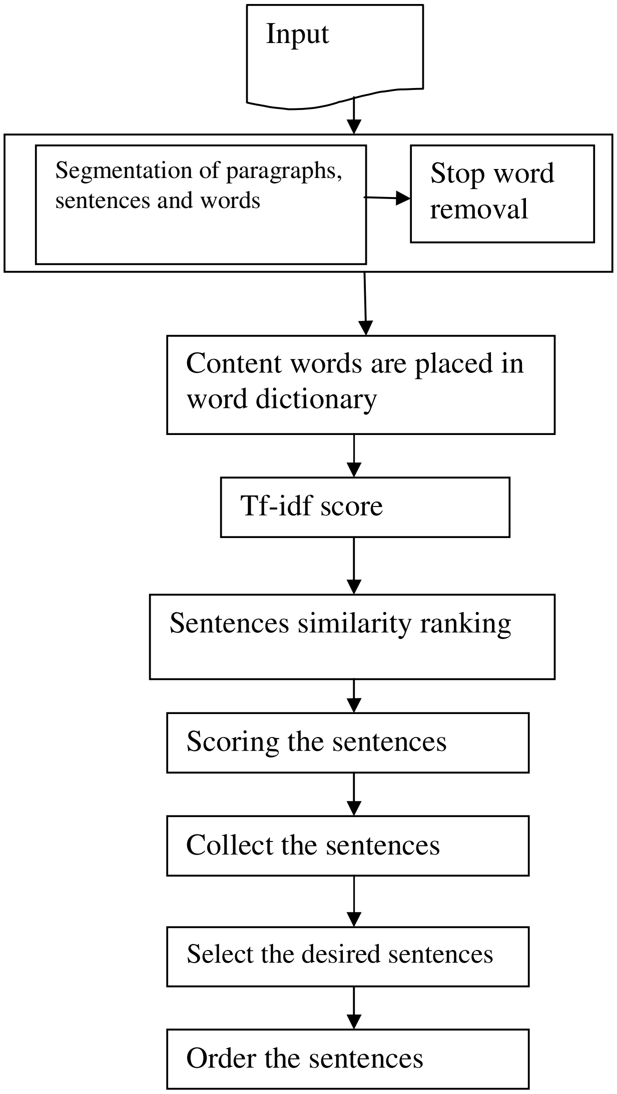
Magnitude of document=$sqrt(\sum_i W^2_{i,j})$

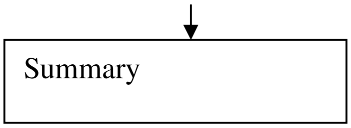
Magnitude of query= $Sqrt(\sum_j W^2_{Q,j})$

Step 11: Rank the sentences on the basis of similarity analysis.

Step 12: Collect the required number of sentences as summary.

System Architecture:


Input
Segmentation of paragraphs, sentences and words
Stop word removal
Content words are placed in word dictionary
Tf-idf score
Sentences similarity ranking
Scoring the sentences
Collect the sentences
Select the desired sentences
Order the sentences



Summary

Rank the sentences:

Query: നാസയുടെ മേവൻ പര്യവേക്ഷണ പേടകം
S1: അമേരിക്കൻ ബഹിരാകാശ ഏജൻസിയായ *നാസയുടെ മേവൻ പര്യവേക്ഷണ പേടകം* ചൊവ്വാഗ്രഹത്തിന്റെ ഭ്രമണപഥത്തിൽ വിജയകരമായി എത്തി.(rank1)
S2: ചൊവ്വായുടെ ഗ്രഹാന്തരീക്ഷത്തെക്കുറിച്ച് പഠിക്കാൻ വിക്ഷേപിച്ച *പേടകം* പത്തുമാസത്തെ യാത്രയ്ക്കൊടുവിലാണ് ചൊവ്വയിലെത്തിയത്.(rank3)
S3: ചൊവ്വായുടെ അന്തരീക്ഷത്തേക്കുറിച്ച് പഠിക്കാൻ വേണ്ടി മാത്രമായി അയച്ച ആദ്യ പേടകമാണ് *മേവൻ*.(rank6)
S4: ശാസ്ത്രലോകത്തിന് ഇതിനകം ലഭിച്ച തെളിവുകൾ പ്രകാരം ചൊവ്വായുടെ അന്തരീക്ഷത്തിൽ ഒരുകാലത്ത് ഉയർന്ന സാന്ദ്രതയില് വാതകങ്ങളുണ്ടായിരുന്നു.(rank 0)
S5: നിലവിൽ വളരെ ചെറിയ സാന്ദ്രതയില് കാർബൺഡയോക്സയിഡ് മാത്രമാണ് ചൊവ്വായുടെ അന്തരീക്ഷത്തിലുള്ളത്. (rank 0)
S6: *മേവൻ പേടകം* നൽകുന്ന വിവരങ്ങളുടെ അടിസ്ഥാനത്തില് ചൊവ്വായുടെ കാലാവസ്ഥാ ചരിത്രം മനസിലാക്കാനാകും.(rank2)
S7: ചൊവ്വായുടെ അന്തരീക്ഷത്തെക്കുറിച്ചുള്ള പഠനത്തിൽ ഇന്ത്യയുമായി സഹകരിക്കാനും വിവരങ്ങൾ ഒത്തുനോക്കാനും നാസയ്ക്ക് അതിയായ താത്പര്യമുണ്ടെന്ന് *നാസയുടെ* പ്ലാനെറ്ററി സയൻസ് ഡയറക്ടര് പറഞ്ഞു.(rank5)
S8: *നാസയുടെ മേവൻ* കുഴപ്പമൊന്നുമില്ലാതെ ചൊവ്വായുടെ ഭ്രമണപഥത്തിലെത്തിയത് ശുഭസൂചനയാണെന്ന് ഇന്ത്യൻ ബഹിരാകാശ അധികൃതർ പറഞ്ഞു. (rank4)

Cosine similarity of text

$\text{Sim}(Q,Si) = \sum iW_{Q,j} \quad W_{i,j} \;/\; \text{sqrt}(\sum jW^2{}_{Q,j}).\ \text{sqrt}(\sum iW^2{}_{i,j})$

Cosine ^{e}S1=Q.S1/|Q|.|S1|

$|S1| = \text{sqrt}(0.9031^2 + 0.6021^2 + 0.9031^2 + 0.4150^2 + 0.3010^2 + 0.9031^2 + 0.4150^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.5508$

$|S2| = \text{sqrt}(0.4150^2 + 0.0569^2 + 0.9031^2 + 0.6021^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.1485$

$|S3| = \text{sqrt}(0.3010^2 + 0.0569^2 + 0.6021^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.3130$

$|S4| = \text{sqrt}(0.0569^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.6021^2 + 0.9031^2) = 2.7759$

$|S5| = \text{sqrt}(0.0569^2 + 0.6021^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.2933$

$|S6| = \text{sqrt}(0.3010^2 + 0.4150^2 + 0.0569^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.2714$

$|S7| = \text{sqrt}(0.4150^2 + 0.0569^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.6021^2) = 3.2132$

$|S8| = \text{sqrt}(0.6021^2 + 0.4150^2 + 0.3010^2 + 0.0569^2 + 0.6021^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2 + 0.9031^2) = 2.2514$

Q.S1=0.4350 Q.S2=0.1722 Q.S3=0.0906 Q.S4=0 Q.S5=0 Q.S6=0.2628 Q.S7= 0.1722 Q.S8= 0.1722

$|Q| = \text{sqrt}(0.4150^2 + 0.3010^2 + 0.9031^2 + 0.4150^2) = 1.1183$

Cosine ^{e}S1=Q.S1/|Q|.|S1| = 0.4350/1.1183*2.5508=0.1525
Cosine ^{e}S2=Q.S2/|Q|.|S2| = 0.1722/1.1183*2.1485=0.0716
Cosine ^{e}S3=Q.S3/|Q|.|S3| = 0.0906/1.1183*2.3130=0.0350
Cosine ^{e}S4=Q.S4/|Q|.|S4| = 0/1.1183*2.7759=0
Cosine ^{e}S5=Q.S5/|Q|.|S5| = 0/1.1183*2.2933=0
Cosine ^{e}S6=Q.S6/|Q|.|S6| = 0.2628/1.1183*2.2714=0.1035
Cosine ^{e}S7=Q.S7/|Q|.|S7| = 0.1722/1.1183*2.5508=0.0604

		Sentences											Wi=tfi*idfi								
Terms	Q1	S1	S2	S3	S4	S5	S6	S7	S8	dfi	d/dfi	Idfi	Q	S1	S2	S3	S4	S5	S6	S7	S8
അമേരിക്കൻ	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
ബഹിരാകാശ	0	1	0	0	0	0	0	0	1	2	4	0.6021	0	0.6021	0	0	0	0	0	0	0.6021
ഏജൻസിയായ	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
നാസയുടെ	1	1	0	0	0	0	0	1	1	3	2.6	0.4150	0.4150	0.4150	0	0	0	0	0	0.4150	0.4150
മേവൻ	1	1	0	1	0	0	1	0	1	4	2	0.3010	0.3010	0.3010	0	0.3010	0	0	0.3010	0	0.3010
പര്യവേക്ഷണ	1	1	0	0	0	0	0	0	0	1	8	0.9031	0.9031	0.9031	0	0	0	0	0	0	0
പേടകം	1	1	1	0	0	0	1	0	0	3	2.6	0.4150	0.4150	0.4150	0.4150	0	0	0	0.4150	0	0
ചൊവ്വാഗ്രഹത്തിന്റെ	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
ഭ്രമണപഥത്തിൽ	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
വിജയകരമായി	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
എത്തി	0	1	0	0	0	0	0	0	0	1	8	0.9031	0	0.9031	0	0	0	0	0	0	0
ചൊവ്വായുടെ	0	0	1	1	1	1	1	1	1	7	1.143	0.0569	0	0	0.0569	0.0569	0.0569	0.0569	0.0569	0.0569	0.0569
ഗ്രഹാന്തരീക്ഷത്തെക്കുറിച്ച്	0	0	1	0	0	0	0	0	0	1	8	0.9031	0	0	0.9031	0	0	0	0	0	0
പഠിക്കാൻ	0	0	1	1	0	0	0	0	0	2	4	0.6021	0	0	0.6021	0.6021	0	0	0	0	0
വിക്ഷേപിച്ച	0	0	1	0	0	0	0	0	0	1	8	0.9031	0	0	0.9031	0	0	0	0	0	0
പത്തുമാസത്തെ	0	0	1	0	0	0	0	0	0	1	8	0.9031	0	0	0.9031	0	0	0	0	0	0
യാത്രയ്ക്കൊടുവിലാണ്	0	0	1	0	0	0	0	0	0	1	8	0.9031	0	0	0.9031	0	0	0	0	0	0
ചൊവ്വയിലെത്തിയത്	0	0	1	0	0	0	0	0	0	1	8	0.9031	0	0	0.9031	0	0	0	0	0	0
അന്തരീക്ഷത്തേക്കുറിച്ച്	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
വേണ്ടി	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
മാത്രമായി	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
അയച്ച	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
ആദ്യ	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
പേടകമാണ്	0	0	0	1	0	0	0	0	0	1	8	0.9031	0	0	0	0.9031	0	0	0	0	0
ശാസ്ത്രലോകത്തിന്	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
ഇതിനകം	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
ലഭിച്ച	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
തെളിവുകൾ	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
പ്രകാരം	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
അന്തരീക്ഷത്തിൽ	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
ഒരുകാലത്ത്	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
ഉയർന്ന	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
സാന്ദ്രതയിൽ	0	0	0	0	1	1	0	0	0	2	4	0.6021	0	0	0	0	0.6021	0.6021	0	0	0
വാതകങ്ങളുണ്ടായിരുന്നു	0	0	0	0	1	0	0	0	0	1	8	0.9031	0	0	0	0	0.9031	0	0	0	0
നിലവിൽ	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0
വളരെ	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0
ചെറിയ	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0
കാർബൺഡയോക്സയിഡ്	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0
മാത്രമാണ്	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0
അന്തരീക്ഷത്തിലുള്ള	0	0	0	0	0	1	0	0	0	1	8	0.9031	0	0	0	0	0	0.9031	0	0	0

നൽകുന്ന	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
വിവരങ്ങളുടെ	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
അടിസ്ഥാനത്തിൽ	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
കാലാവസ്ഥാ	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
ചരിത്രം	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
മനസിലാക്കാനാകും	0	0	0	0	0	0	1	0	0	1	8	0.9031	0	0	0	0	0	0	0.9031	0	0
അന്തരീക്ഷത്തെക്കുറിച്ചുള്ള	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
പഠനത്തിൽ	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
ഇന്ത്യയുമായി	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
സഹകരിക്കാനും	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
വിവരങ്ങൾ	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
ഒത്തുനോക്കാനും	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
നാസയ്ക്ക്	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
അതിയായ	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
താത്പര്യമുണ്ടെന്ന്	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
പ്ലാനെറ്ററി	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
സയൻസ്	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
ഡയറക്ടർ	0	0	0	0	0	0	0	1	0	1	8	0.9031	0	0	0	0	0	0	0	0.9031	0
പറഞ്ഞു	0	0	0	0	0	0	0	1	1	2	4	06021	0	0	0	0	0	0	0	0.6021	0.6021
കുഴപ്പമൊന്നുമില്ലാതെ	0	0	0	0	0	0	0	0	1	1	8	0.9031	0	0	0	0	0	0	0	0	0.9031
ഭ്രമണപഥത്തിലെത്തിയത്	0	0	0	0	0	0	0	0	1	1	8	0.9031	0	0	0	0	0	0	0	0	0.9031
ശുഭസൂചനയാണെന്ന്	0	0	0	0	0	0	0	0	1	1	8	0.9031	0	0	0	0	0	0	0	0	0.9031
ഇന്ത്യൻ	0	0	0	0	0	0	0	0	1	1	8	0.9031	0	0	0	0	0	0	0	0	0.9031
അധികൃതർ	0	0	0	0	0	0	0	0	1	1	8	0.9031	0	0	0	0	0	0	0	0	0.9031

The above examples cosine similarity is used for finding the similarity between sentences. The query held sentences got highest score than other sentences. The score of sentences are 0.1525, 0.0716, 0.0350, 0, 0, 0.1035, 0.0604 and 0.0684. The ranking of sentences are S1, S6, S2, S8, S7 and S3. The rank two approximations S1 and S6 are selected as summary. The summary gives an overall idea about the document.

IV. ANALYSIS AND EVALUATION

The most common way to evaluate the quality of summary is to compare with human summary. Numerous methods are used for predict the quality of summary. Normally the efficiency is evaluated on the basis of precision, recall and F-measure. Here the human summary is used for evaluate the quality of system summary. Other methods for summary evaluation are ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure and BLEU measure [10]. ROUGE is a recall-based measure that determines the quality of system-generated summary. BLEU is precision-based measure it shows the content present in one or more human-generated summaries.

V. CONCLUSIONS

Text summarization technique creates summary or extraction of texts. It has developed many years ago but recent years the wide use of Internet there has been great mobility in summarization techniques. The rate of information growth in Malayalam documents in WWW needs an efficient and accurate summarization system. The abstractive summarization requires heavy computational models for language generation. In such a situation the extractive text summarization produces the satisfactory result within a short span of time. The performance of statistical based extractive summarization method like vector space model shows good result in summarizing Malayalam documents. It is sufficient for finding the semantic relation between words and sentences. This method finds the summary on the basis of statistical analysis of source document and finds the representative sentence from the document.

REFERENCES

1. *Alan Turing, (1950). "Computing Machinery and Intelligence".*
2. *Luhn, (1958), "The automatic creation of literature abstracts", IBM Journal of Research Development, 2(2):159–165.*
3. *P. B. Baxendale, (1958), "Machine-made index for technical literature: an experiment", IBM Journal, 354–361.*
4. *Edmundson, H.P. (1969), New Methods in Automatic Extracting, Journal of the ACM, 16(2):264-285.*
5. *E. Hovy and C-Y Lin, (1997), "Automated Text Summarization in SUMMARIST", Proceedings of the Workshop of Intelligent Scalable Text Summarization.*
6. *Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In Proceedings of the ACL'97/EACL'97 workshop on intelligent scalable text summarization (pp. 10–17), Madrid, Spain.*
7. *Martin Hassel & Hercules Dalianis, (2000). "SweSum-Auto Text Summarizer".*
8. *Mihalcea and Tarau, (2004). "TextRank: Bringing Order into Text".*
9. *Qazvinian and Radev, (2004). "LexRank: graph-based lexical centrality as salience in text summarization". Journal of Artificial Intelligence Research 457–479.*
10. *Lin. C.Y. (2004). "Rouge: A package for automatic evaluation of summaries", Proceedings of the ACL-04 Workshop, pages 74-81.*
11. *Bellotti. T and Crook J (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302-3308.*
12. *Vishal Gupta and Gurpreet Singh Lehal, (2010)" A Survey of Text Summarization Extractive Techniques", Journal of emerging technologies in web intelligence, vol. 2, no. 3.*
13. *Sankar K, Vijay Sundar Ram R and Sobha Lalitha Devi. (2011). "Problems of Parsing in Indian Languages".*
14. *M. Pourvali and S. Abadeh Mohammad, (2012). "Automated text summarization base on lexical chain and graph using of word net and Wikipedia knowledge base," International Journal of Computer Science Issues, No. 3, vol. 9.*
15. *Nallapati. R., Zhou. B., Santos. C., Gulcehre. C and Xiang. B. (2016). "Abstractive text summarization using sequence-to-sequence and beyond". The SIGNLL Conference on Computational Natural Language Learning.*