VOL. 13, NO. 10, MAY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

EXPERIMENTS ON DOCUMENT CLUSTERING IN TAMIL LANGUAGE

Syed Sabir Mohamed[1] and Shanmugasundaram Hariharan[2]
[1]Faculty in Computer Science and Engineering, Sathyabama University, India
[2]Department of Information Technology Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, India
E-Mail: mailtos.hariharan@gmail.com

ABSTRACT

With the rapid development of the Internet, the number of documents in electronic form is huge and grows day by day. In order to effectively address the modern information overload problem, it is extremely important to organize the documents according to the topic. Commonly, this can be achieved by using clustering techniques. Document clustering is an important tool for applications such as web search engines. This proposal deals with clustering of Tamil documents. Clustering is an un-supervised learning process that organizes documents or text files into distinct groups without having prior knowledge. This paper uses vector space model to cluster the documents. Vector space model is otherwise known as "Term-frequency approach". Stop words which are frequent, meaningless terms are removed from the input text document to decrease, the size of the document to be processed. Then the cosine similarity measure is applied to find the similarity between the input text documents. Then clustering is done using K-Medoid algorithm and optimal number of medoids and corresponding clusters are found.

Keywords: summary generation, Indian language, tamil, clustering, k-medoid.

1. INTRODUCTION

Data mining is the extraction of hidden information from large amount of data using tools such as classification, association rule mining and clustering. Clustering is the process of grouping a set of objects into classes of similar objects using keyword clusters [16, 17]. Clustering generates clusters which are used in many fields, including data mining and information retrieval which adopts genetic algorithm to find optimal k value [18]. With the rapid development of the Internet, the number of documents in electronic form is huge and grows day by day. In order to effectively address the modern information overload problem, it is extremely important to organize the documents according to the topic. Clustering is an important task and it attempts to find natural groups using existing data rather than classifying them on the basis of external criteria. By categorizing or grouping similar data items, the amount of data to be processed will become lesser.

The rest of the paper is organized as follows. Section 1 has presented some introduction on summarization tasks. The basics of clusterinf task are discussed in section. Section 3 briefs on some of the related works on clustering. Section 4 presents some experimental illustrations and finally, section 5 presents conclusions and future improvements.

2. CLUSTERING

Clustering is the process of classifying objects into different groups, or partitioning a data set into subsets or clusters. Clustering classifies data objects without consulting a known class label. Cluster analysis is a data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

Clustering is based on the principle of maximizing intra-class distance and minimizing the inter-class distance. i.e. documents with in the cluster will have high similarity value, when compared and have low similarity value for the documents in other clusters. Cluster analysis can be used to discover structures in data without providing an explanation or interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist. In clustering, to deal with large number of dimensions and large number of data items can be problematic because of time complexity and also the effectiveness of the method depends on the definition of similarity distance between Documents.

3. RELATED WORKS

Document summarization plays a vital role in the use and management of information dissemination across different languages [9]. This paper investigates a method for the production of summaries from Tamil newspaper text source. The primary goal is to create an effective and efficient tool that is able to summarize the given text documents in a form of meaningful extract of the original text document using centroid-based algorithm. The paper focuses on generating summaries using a centroid-based algorithm, which represents group of words that are statistically important for a document. Each sentence in a document is considered as a vector in a multi-dimensional space. The sentences that are nearest to the centroid value are considered as the most important sentences. The importance of a sentence is determined by three parameters the centroid value, the positional value, and the first sentence overlap. The score for each sentence is calculated and the redundancy between the sentences is eliminated using CSIS. Finally, the sentences are ranked and the sentences with highest score values are selected as summary [1].

Steganography is the ability and science of enclosed or secreted writing. The idea of steganography is masked communication to hide the existence of a message from the interfering eyes. Digital Steganography

algorithms have been developed by using text documents, image files and audio files as the cover media. This paper presents an approach for converting a secret message into a Text summary in Tamil language and that is transmitted over communication channel. The proposed method classifies the Tamil alphabets into four groups to hide secret data bits and it selects the sentence to generate a summary of the text, known as stego text. Similarly at the extraction end, the receiver extracts the characters from the stego text by following the groups and places the corresponding bits to get the secret message from the summary generated by the hiding process. This proposed method exhibits a satisfactory experimental result with the cover text chosen. [2]

Automatic summarization of text is one of the areas of interest in the field of natural language processing. The proposed method utilizes the sentence extraction in a single document and produces a generic summary for a given Malayalam document (Extractive summarization). Sentences in the document are ranked based on the word score of each word present in it. Top N ranked sentences are extracted and arrange them in their chronological order for summary generation, where N represents the size of summary with respect to the percentage of original document size (condensation rate). The standard metric ROUGE is used for performance evaluation. ROUGE calculates the n-gram overlap between a generated summary and reference summaries. Reference summaries were constructed manually. Experiments show that the results are promising. [3]

With the exponential growth of the internet, a lot of online news reports are produced on the web every day. The news stream flows so rapidly that no one has the time to look at each and every item of information. In this situation, a person would naturally prefer to read updated information at certain time intervals. Document updating technique is very helpful for individuals to acquire new information or knowledge by eliminating out-of-date or redundant information. Existing summarization systems involve identifying the most relevant sentences from the text and putting them together to create a concise initial summary. In the process of identifying the important sentences, features influencing the relevance of sentences are determined. Based on these features the salience of the sentence is calculated and an initial summary is generated from highly important sentences at different compression rates. These types of initial summaries work on a batch of documents and do not consider the documents that may arrive at later time, so that corresponding summaries need to get updated. The update summarization system addresses this issue by taking into account the documents read by the user in the past and seeks to present only fresh or different information. The first step is to create an initial summary based on basic and additional features. The next step is to create an update summary based on the basic, additional and update features. In this paper, two approaches are proposed for generating initial and update summary from multiple documents about given news. The first approach performs semantic analysis by modifying the vector space model with dependency parse relations and applying latent semantic analysis on it to create a summary. The second approach applies sentence annotation based on aspects, prepositions and named entities to generate summary. Experimental results show that the proposed approaches generate better initial and update summaries compared with the existing systems. [4]
Natural Language Processing is a vast area which has great importance when people started to interpret human language from one form to another. Summarization is one of the research works in NLP which concentrates on providing meaningful summary using various NLP tools and techniques. Since huge amount of information is used across the digital world, it is highly essential to have automatic summarization techniques. Extractive and Abstractive summarization are the two summarization techniques available. A lot of research works are being carried out in this area especially in extractive summarization. Even though more works are carried out using extractive method, meaningful summary can be attained using abstractive summary techniques which make it more complex. In Indian languages, very few works are carried out in the field of abstractive summarization and there is high need for having research works being carried out in this area. Here, we are concentrating on the various techniques available for abstractive summarization and also try to explain the limited works currently available in abstractive summary field of Indian languages. [5]

Text summarization is an emerging technique for finding out the summary of the text document. Text summarization is nothing but summarizing the content of given text document. Text summarization has got so uses such as Due to the massive amount of information getting increased on internet; it is difficult for the user to go through all the information available on web. Summarization techniques need to be used to reduce the users time in reading the whole information available on web. In this paper propose a Malayalam text summarization system which is based on MMR technique with successive threshold. Here the sentences are selected based on the concept of maximal marginal relevance. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the text document, which can be achieved by using successive threshold approach. We apply MMR approach on Malayalam text summarization task and achieve comparable results to the state of the art. [6]

Automatic text summarization is technique of compressing the original text into shorter form which will provide same meaning and information as provided by original text. The brief summary produced by summarization system allows readers to quickly and easily understand the content of original documents without having to read each individual document. The overall motive of text summarization is to convey the meaning of text by using less number of words and sentences. Summaries are of two types: Abstractive summaries and Extractive summaries. Extractive summaries involve

extracting relevant sentences from the source text in proper order. The relevant sentences are extracted by applying statistical and language dependent features to the input text. On the other hand, abstractive text summaries are made by applying natural language understanding. Human beings usually make summaries in abstractive way. Moreover abstractive summaries can also involve the words or sentences which are not present in the input text. Automatic generation of abstractive summary is more difficult as compared to producing extractive text summary. This paper concentrates on survey and performance analysis of automatic text summarizers for Indian languages. [7]

This paper proposes a summarization system for summarizing multiple tamil documents. This system utilizes a combination of statistical, semantic and heuristic methods to extract key sentences from multiple documents thereby eliminating redundancies, and maintaining the coherency of the selected sentences to generate the summary. In this paper, Latent Dirichlet Allocation (LDA) is used for topic modeling, which works on the idea of breaking down the collection of documents (i.e) clusters into topics; each cluster represented as a mixture of topics, has a probability distribution representing the importance of the topic for that cluster. The topics in turn are represented as a mixture of words, with a probability distribution representing the importance of the word for that topic. After redundancy elimination and sentence ordering, summary is generated in different perspectives based on the query. [8]

Text summarization is the process of extracting needed information from the source text and to present that information to the user in the form of summary. It is very difficult for human beings to summarize large documents of text manually. Automatic summarization provides the required solution as well as challenging task because it requires deep analysis of text. There are two types of summarization: extractive summarization and abstractive summarization. The Extractive summaries are produced by extracting the whole sentences from the source text.

Abstractive summaries are produced by reformulating sentences of the source text. This paper is about a survey of text summarization techniques for various Indian regional languages like Hindi, Punjabi, Tamil Kannada and Bengali. The proposed system is based on English language text summarization in which Naming entity reorganization and Part Of speech is used for feature extraction and graph is generated for text summarization [10].

Text Summarization is the process of generating a short summary for the document that contains the significant portion of information. In an automatic text summarization process, a text is given to the computer and the computer returns a shorter less redundant extract of the original text. The proposed method is a sentence extraction based single document text summarization which produces a generic summary for a Malayalam document. Sentences are ranked based on feature scores and Googles PageRank formula. Top-k ranked sentences will be included in summary where k depends on the compression ratio between original text and summary. Performance evaluation will be done by comparing the summarization outputs with manual summaries generated by human evaluators. [11]

This paper presents an approach to cluster multiple documents by using document clustering approach and to produce cluster wise summary based on feature profile oriented sentence extraction strategy. Related documents are grouped into same cluster using document clustering algorithm. Feature profile is generated by considering word weight, sentence position, sentence length, sentence centrality, proper nouns in the sentence and numerical data in the sentence. Based on the feature profile sentence score is calculated for each sentence. According to different compression rates sentences are extracted from each cluster and ranked in order of importance based on sentence score. Extracted sentences are arranged in chronological order as in original documents and from this, cluster wise summary can be generated. Experimental results show that the proposed clustering algorithm is efficient and feature profile is used to extract most important sentences from multiple documents. [12]

Nowadays, automatic multidocument text summarization systems can successfully retrieve the summary sentences from the input documents. But, it has many limitations such as inaccurate extraction to essential sentences, low coverage, poor coherence among the sentences, and redundancy. This paper introduces a new concept of timestamp approach with Na¨ıve Bayesian Classification approach for multidocument text summarization. The timestamp provides the summary an ordered look, which achieves the coherent looking summary. It extracts the more relevant information from the multiple documents. Here, scoring strategy is also used to calculate the score for the words to obtain the word frequency. The higher linguistic quality is estimated in terms of readability and comprehensibility. In order to show the efficiency of the proposed method, this paper presents the comparison between the proposed methods with the existing MEAD algorithm. The timestamp procedure is also applied on the MEAD algorithm and the results are examined with the proposed method. The results show that the proposed method results in lesser time than the existing MEAD algorithm to execute the summarization process. Moreover, the proposed method results in better precision, recall, and *F*-score than the existing clustering with lexical chaining approach. [13]

Automatic Text Summarization is a process of generating Summary/Head note for the text document. Text Summarization is carried out by two main methods, namely, Extraction and Abstraction. This paper utilizes the extraction process for sentence selection. Here some Feature based sentence scoring techniques also used, which played an important role in text summarization. Finally an analysis is done by comparing the Fuzzy Logic and Neural Networks techniques based upon the Precision, Recall & F-Measure. Fuzzy Logic rules were used to balance the weights between important and unimportant

features based on the feature extraction. The Experimental result shows that fuzzy Logics give an improving result than the Neural Networks. [14]

Today in the era of Big Data, textual data is rapidly growing and is available in many different languages. In the fast-moving world, it's difficult to read all the text-content. Hence, the need for text summarization is being in the spotlight. Automatic text summarization is a technique which compresses large text to a shorter text which includes the important information. There are two types of summaries: Extractive summaries and Abstractive summaries. Extractive summaries are produced by extracting the whole sentences from the source text. Abstractive summaries are produced by reformulating sentences of the source text. Several text summarization techniques have been proposed in past years for English and various European languages but there are very few techniques that can be found for native languages of India. This paper presents a survey of text summarization techniques for various Indian and foreign languages like English, European, etc. Also, an approach for summarizing Hindi text using machine learning technique has been proposed. We have also described few challenges which are still under research. [15]

4. EXPERIMENTAL SETUP, RESULTS AND ILLUSTRATIONS

This section presents some of the experimental illustrations on clustering the documents. Each of the subsections presented here give a detailed note on the steps involved in the analysis. The overall step is presented in Figure-1.

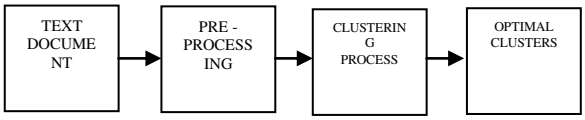

Figure-1. Overall architecture of proposed system.

4.1 Pre-processing the input text documents

The input Files are to be pre-processed before applying the clustering task, in order to reduce the size of the document and problem space. The pre-processing step gets the input document as input and all the data items are represented as vectors. Figure-2. Presents the architecture for document pre-processing.

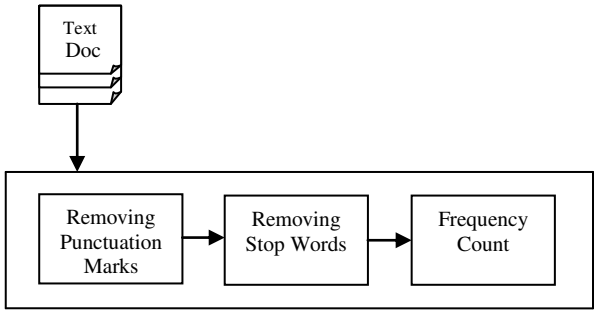

Figure-2. Document pre-processing.

4.2 Removal of stop words

Highly frequent, meaningless words are removed from the document which are called Stop Words. These words do not give any meaning and hence they can be removed from the Document. Therefore the input document size reduces and the time to compute will be reduced. Figure-3 presents the sample words with two strokes. Figures 4, 5 and 6 presents example of three strokes, four strokes and strokes higher than 4.

என	இது	அது	ஒரு

Figure-3. Stop words with two key strokes.

அவள	அவரு	அதை	அதே
அவை	அனை	அவனு	அவளு
அவர	அவன	அவங	அவற
ஆனா	ஆகும	இவை	இதை
இதே	இவர	அவர	இவனு
இவை	இதை	இவள	இவரு
இவனு	இவளு	இவர	இவன
இவங	இதன	இதற	ஒரே
ஒருவ	சிறு	தமது	வரை
வகை	அவதே	அவன	இவை
இவதே	ளுடன	எதை	எனவு
எனதே	அதற	எனது	தான
தமது	பிற	நமது	நமக
அதிக	முதல	இவற	தனது
அதிக	கூடிய	சில	அதன

Figure-4. Stop words with 3-strokes.

இதனை	அதனை	ஏனென
எப்படி	ஆகவே	உங்க

Figure-5. Stop words with 4-strokes.

ஆகியோ	இன்னொ	இப்பொ	இப்போ
அப்போ	அப்போ	இத்தை	பின்வரு
தற்போ	இத்தகை	பல்வேறு	கொண்டு
போன்ற	ஒவ்வொ	எத்தை	எப்படி

Figure-6. Stop words with 4 & higher strokes.

In English, the document is reduced to one third of the document size. In Tamil, the document size is reduced to only 12% to 18%. Removal of these stop words from the input document can reduce the noise in the file and increases the computational efficiency of the system. Here, about 175 root words are found which has the minimum number of keystrokes to identify a Stop word. It will identify Stop words of about 650 words. We derive the same root word for the example shown in Figure-7.

அவர்களுக்கு, அவர்களும், அவர்களுடன், அவர்களை, அவர்களே, அவர்களோட, அவர்களின், அவர், அவர்களிடம், அவர்கள், அவர்களால், அவரால், அவரவர்க்குரிய, அவரவர்க்குக், அவரவர்| அவரது.

Figure-7. Sample illustration for common word.

Each word is compared with the Root keys, and is removed from the input Text file. Other words which do not match with the root keys are considered as the meaningful words. After pre-processing, the size of the document is reduced.

4.3 Similarity measure

Similarity value is the distance between two objects. The similarity between two points is found based on the distance between those points. i.e. small distances correspond to large similarities between Documents and large distances correspond to small similarities. There are various measures to find the Similarity.

4.4 Clustering

Many different clustering algorithms use particular similarity measure as input and Clustering is done based on that similarity values. There are different algorithms to cluster the Documents or Dataset into set of clusters based on the principle of maximizing Intra class similarity and minimizing the Inter class similarity. The clustering algorithms are chosen based on the desired properties of the final clustering.

4.5 Optimal clustering

The quality of clustering is found by using various criteria in the relations within and between the clusters. The process of evaluating the clustering quality is called Cluster Validation. The cost function measures the average dis-similarity between the document and the Medoid of its cluster. The optimal number of clusters is found by varying the parameters like k - the number of medoids and the cost function. Table-1 presents the experimental illustrations using varying k-values and partition cost.

Table-1. Cost variation based on k-value.

K Value	Distance	Partition cost				
		M1	M2	M3	M4	M 5
1	8.71	9.71	77.02	8.71	76.02	77.02
2	6.83	8.83	48.69	7.83	47.69	50.69
3	1.52	4.52	5.31	3.52	4.31	11.31
4	0.84	4.84	4.71	3.84	3.71	16.71
5	0.49	5.49	5.24	4.49	4.24	25.24
6	0.21	6.21	6.04	5.21	5.04	36.04
7	0.05	7.05	7.00	6.05	6.00	49.00
8	0	8	8	7	7	64
9	0	9	9	8	8	81

5. CONCLUSIONS AND FUTURE IMPROVEMENTS

This paper presents some illustrative study on clustering of documents for Tamil language. The results were good and seem to generate cohesive summaries.

REFERENCES

[1] Syed Sabir Mohamed and Shanmugasundaram Hariharan. 2016. A summarizer for Tamil language using centroid approach. International Journal of information retrieval research. 6(1): 1-15.

[2] K Manimozhi, V Kalaichelvi, M Poornima, A Sumathi. 2015. An approach for text steganography: generating Tamil text summary using Tamil phonetics. International review on computers and software (irecos). 10(2): 137-143.

[3] P Krishnaprasad, A Sooryanarayanan and Ajeesh Ramanujan. 2016. Malayalam text summarization: an extractive approach Proceedings of international conference on next generation intelligent systems (icngis).

[4] S.V Kogilavani, C. S. Kanimozhiselvi and S. Malliga. 2016. Summary generation approaches based on semantic analysis for news documents. Journal of information science. 42(4): 465-476.

[5] Sunitha. C, Jaya .A and Amal Ganesh. 2016. A study on abstractive summarization techniques in Indian languages. Proceedings of fourth international conference on recent trends on computer science & engineering. pp. 25-31.

[6] Ajmal E.B and Rosna P Haroon. 2015. Summarization of Malayalam document using relevance of sentences. International journal of latest research in engineering and technology (ijlret). 1(6): 8-13.

[7] Vishal Gupta. 2013. A survey of text summarizers for Indian languages and comparison of their performance. Journal of emerging technologies in web intelligence. 5(4): 361-366.

[8] N. Shreeya Sowmya1 and T. Mala. 2011. Tamil document summarization using latendirichlet allocation. Proceedings of Tamil internet.

[9] Rahulraj M and Rosna P Harron. 2016. A study on different text summarization methods in Dravidian languages. International journal of innovations in engineering and technology (ijiet). 7(1): 323-327.

[10] Nilofar Mulla and Shital K. Dhamal. 2016. A survey of text summarization techniques for different Indian regional languages. International journal of innovative research in computer and communication engineering. 4(8): 15003-15005.

[11] Renjith S R and Sony P. 2015. An automatic text summarization for Malayalam using sentence extraction. Proceedings of 27th irf international conference. pp. 60-64.

[12] A. Kogilavani and P. Balasubramani. 2010. Clustering and feature specific sentence extraction based summarization of multiple documents. International journal of computer science & information technology (ijcsit). 2(4): 99-111.

[13] Nedunchelian Ramanujam and Manivannan Kaliappan. 2016. An automatic multidocument text summarization approach based on naive bayesian classifier using timestamp strategy. The scientific world journal. 2016(article id 1784827): 1-10.

[14] S. Santhana Megala, A. Kavitha and A. Marimuthu. 2014. Enriching text summarization using fuzzy logic. s. Santhana megala *et al*, / (ijcsit) international journal of computer science and information technologies. 5(1): 863-867.

[15] Prachi Shah and Nikita P. Desai. 2016. A survey of automatic text summarization techniques for Indian and foreign languages. Proceedings of international conference on electrical, electronics, and optimization techniques (iceeot).

[16] Hsi-Cheng Chang, Chiun-Chieh Hsu and Yi-Wen Deng Z. 2004. Unsupervised Document Clustering Based on Keyword Clusters. (ISCIT 2004).

[17] Hsi-Cheng Chang; Chiun-Chieh Hsu. 2005. Using topic keyword clusters for automatic document clustering. (ICITA.2005).

[18] A. Casillas, M.T. Gonzalae de Lena and R. Mart´ýnez. Document Clustering into an unknown number of clusters using a Genetic Algorithm.