iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** II **Month of publication:** February 2017

DOI: http://doi.org/10.22214/ijraset.2017.2096

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue II, February 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Multi-Document Summarization Using K-Medoids Clustering Approach

Dharani Dharan A C[1], Rahul Prakash S S[2], Kiruthika S[3]
[1,2]Student, [3]Assistant Professor
Computer Science Engineering, Dr. Mahalingam College of Engineering and Technology, India

Abstract: ***Multi document summarization is the process of transforming a set of documents into a single summarized document. The summarized document can give overall idea of the document collection. Summarization of text document involves intensive text processing and computations to generate the summary. Multi document summarization uses topic modeling and clustering based approach for summarizing the large text collection over MapReduce framework. Most of the systems use k-means clustering algorithm for summarization process. Summarizing large volume of text is a challenging and time consuming problem. In order to overcome the problem of time consumption and to improve the accuracy of the summarized document, K-medoids clustering algorithm is used.***

I. INTRODUCTION

Today, the data in our world is incredibly large and getting rapidly increased day by day. There is a huge amount of data available in structured and unstructured form. So it is difficult to read all the information that is available to the user. A system that automatically retrieves and summarize the documents as per the user need is required. Text summarization is one of the important and challenging problems in text mining. It gives a number of benefits to the user. A large collection of text documents are transformed to a compact and reduced document, which represents the digest of the original text collection. A summarized document helps in understanding the main idea of the large text collection quickly and also saves a lot of time by avoiding the reading of each document separately. Multi-document summarization generates a compact summary by extracting the relevant sentences from a collection of documents on the basis of topics. This system produces a summary using clustering technique to identify common themes across the set of documents. The goal is to create a summary from the document set.

II. EXISTING SYSTEM

The summarization of text document is performed in four stages. They are
Clustering of similar documents
LDA (Latent Dirichlet Allocation)
Generation of Semantic similar terms and clustering using k-means
Sentence filtering

Legal case history from the Federal court of Australia is used as the dataset. The dataset contains the file history in XML files. Clustering and topic modeling is applied on the dataset to generate summary. Initially, the data under text tag in the XML file is extracted and given as input to the first module. In the first module, using text clustering technique, similar documents will be grouped under a cluster. LDA is used in the process of topic modeling. Topic modeling is the process of assigning topic terms to document clusters. Topic terms are the words which occur frequently in the document. Before generating the topic terms, stop word elimination takes place. Stop words refers to the most common words used in a language. Mostly articles and prepositions are considered to be the stop words. A separate file is maintained for storing the list of stop words. With reference to this file, the stop words will get neglected. After generating the topic terms, semantic similar terms for those topic terms are generated using Word Net[9]. Word Net is an API which is capable of generating similar words for the given word. Word weightage for those terms gets calculated using TF-IDF technique. It provides weightage for each word and the words are clustered using K-means clustering algorithm.

Sentences containing the topic terms and their respective semantic similar terms are filtered as the summarized document.

Volume 5 Issue II, February 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

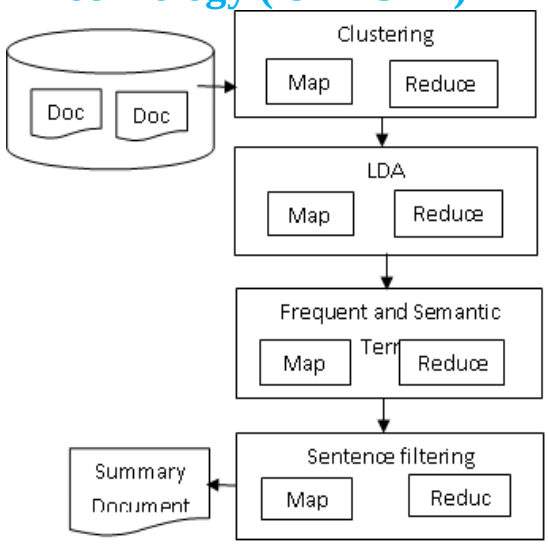


Figure 1 Methodology of Multi Document Summarization

In the block diagram, the document collection undergoes clustering process. Similar documents will be grouped under each cluster. LDA generates the topic term for each cluster using topic modeling technique. The semantic similar terms for the topic terms are generated in the next stage. The generated terms are clustered using k-means algorithm. The words with highest weightage are chosen for participating in the summary document.

III. PROPOSED SYSTEM

Summarization is performed through four steps. They are
Document Clustering
LDA (Latent Dirichlet Allocation)
Generation of Semantic similar terms and clustering using k-Medoids
Sentence filtering

A. Document Clustering

Clustering is the process of organizing one objects into groups whose members are similar in some way. Clustering of documents is done by grouping the documents based on the names of XML files. Text clustering technique is applied on the multi document collection to create the document clusters. By this way, similar text documents can be grouped under their respective clusters for making them ready for summarization as similar to the existing system.

B. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a topic modeling technique which is applied on each individual text document present in a cluster. Topic terms are the words which occur frequently in a cluster. Initially stop word elimination process will take place using the concept of Part Of Speech (POS) tagging. POS tagging is the technique which provide tags for all the words in the document based on their properties. All the stop words have their respective tags which is used to eliminate them. After eliminating the stop words, frequency of the remaining words will be calculated and then by using LDA, topic terms will be generated as similar to the existing system.

C. Frequent and Similar Terms Generation

The clustering topics which are generated from the previous module will be given as an input to this module. Semantic similar terms for those topic terms are generated using Word Net. Semantic terms are the words which have same meaning as the selected word. The frequencies of the semantic similar terms are calculated. Word weightage calculation is done by using TF-IDF concept. TF-IDF is a numerical statistic method that is intended to reflect how important a word is to a document in a collection. It is used as a weighting factor in information retrieval and text mining. After calculating the weights, clustering of words is done. In the process

of clustering the words, K-medoids clustering algorithm is used. K-medoids is a partitioning technique that clusters the data set of n objects into k clusters.

D. Sentence Clustering

Sentence Clustering is performed from each individual input text document present in the document cluster. The words which are clustered in the previous stage are given as input to this module. Clusters which are having the words with more weightage are given highest priority. The words present in the clusters with highest priority are selected for the filtering process of sentence. The sentences which are containing these words are chosen and pasted in the summary document.

IV. DATASETS

This system is also tested with Legal case history from the Federal Court of Australia. The dataset contains 100 legal case files in XML format.

Dataset link: https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports

V. CONCLUSION AND FUTURE WORKS

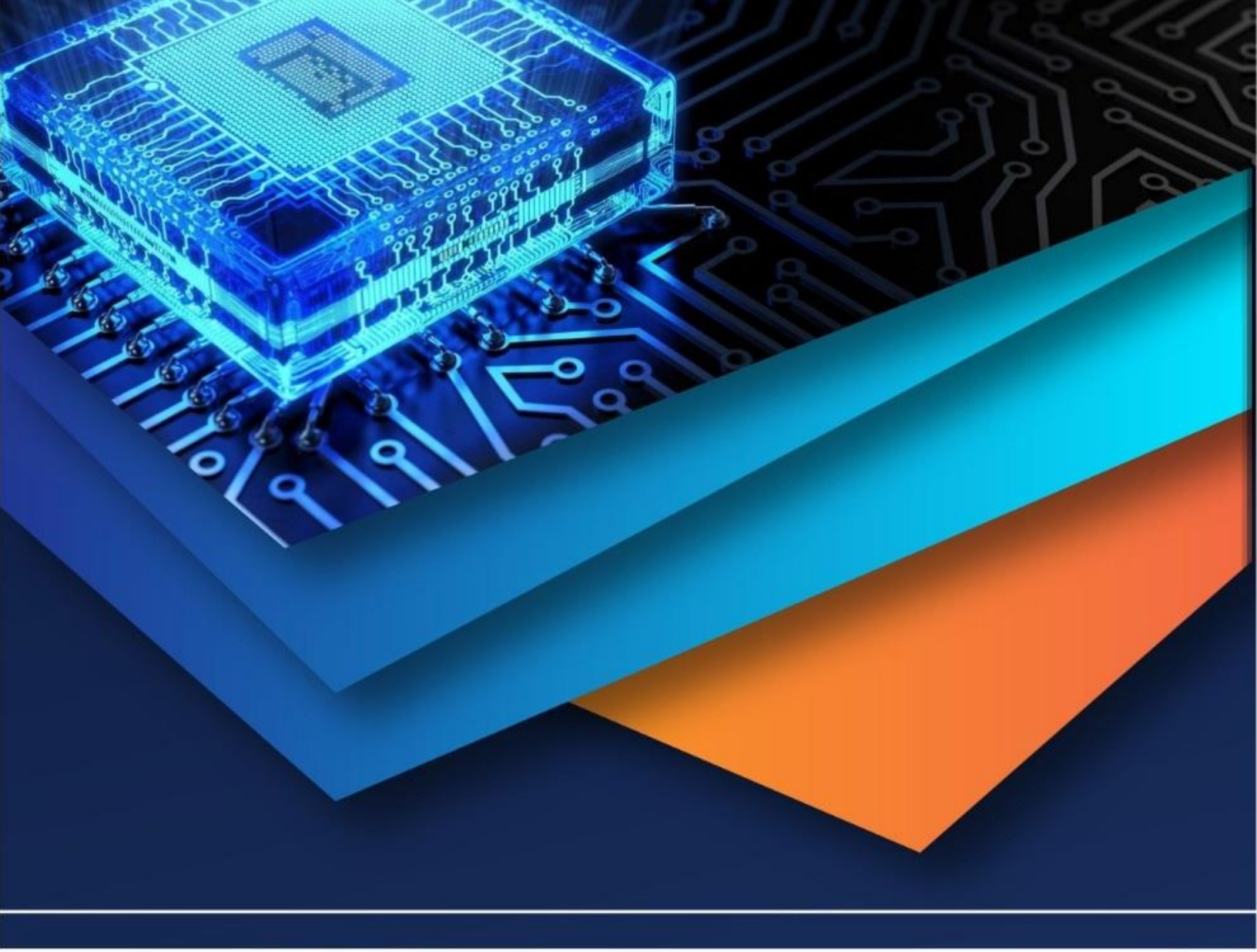
Multi document summarization system helps in reducing the time spent by them in reading the entire documents. The clustering algorithm used in the existing system (k-means) is found to be less efficient and time consuming. This problem can be rectified using k-medoids clustering algorithm. This system can be used for summarizing a large number of files than used in the dataset. Other enhancements like summarizing the documents in other languages can be done. This system can also be used in preparing fair drafts for articles.

VI. ADVANTAGES

The advantages of the proposed system are
Time consumed while clustering the keywords is less.
POS tagging makes stop word elimination easier.
Accuracy is improved.

REFERENCES

[1] Amit.S.Zore, A. D. (2014). Extractive Multi Document Summarizer Alogorithim. (IJCSIT) International Journal of Computer Science and Information Technologies .
[2] Anjali R. Deshpande, L. L. (2013). Text Summarization using Clustering Technique. International Journal of Engineering Trends and Technology (IJETT) .
[3] david M. Blei, A. Y. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research.
[4] Dingding Wang, S. Z. (2009). Multi-Document Summarization using Sentence-based Topic Models. Proceedings of the ACL-IJCNLP, (pp. 297–300). Suntec, Singapore.
[5] Giri Virat, B. M. (2010). K-Means Driven Single Document Summarization. Internal Journal of Computer Applications and Business Intelligence .
[6] Jha, M. (2015). Document Clustering using K-Medoids. International Journal on Advanced Computer Theory and Engineering (IJACTE) .
[7] Khanapure V.M, P. C. (2007). Multi-document Summarization Based on Cluster. International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering .
[8] Manjula.K.S, S. B. (2013). Extracting Summary from Documents Using K-Means Clustering Algorithm. International Journal of Advanced Research in Computer and Communication Engineering .
[9] Nagwani, N. K. (2015). Summarizing large text collection using topic modeling and clustering based on MapReduce framework. Journal of Big Data .
[10] Pankaj Bhole, D. A. (2014). Single Document Text Summarization Using Clustering Approach Implementing for News Article. International Journal of Engineering Trends and Technology (IJETT) . Rakesh Chandra Balabantaray, C. S. (2013). Document Clustering using K-Means and K-Medoids. International Journal of Knowledge Based Computer System



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY