iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4646

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Design of Multi Document Abstractive Summarization System for Online News Corpus

Dr. S. Saraswathi[1], N. Jayasudha[2], G. Kayalvizhi[3], V. Ragini[4]
[1]*Professor/IT, Pondicherry Engineering College*
[2, 3, 4]*Student Members/IT, Pondicherry Engineering College*

Abstract: Document Summarization focuses in generating integrated list of summaries for dense information that have long readability. With the awareness of lexical, priority and indexing terms, long documents are summarized with short and time saving readability through automatic learning techniques. We propose a precise and reliable summarization technique with the assimilation of Glowworm Swarm Optimization (GSO). The summarization is précised by acquiring the benefits of GSO by computing the weight of document content. The weight is computed for sentence and terms available in the document for further processing and summarization. Considering the lexical indices, the summary of the document is more close to POS tagger in the proposed approach.

I. INTRODUCTION

The World Wide Web has the capacity of supplying broad amount of online details globally. Due to this fact, the users submit their various queries on the search engine Nowadays, the trend of Information Retrieval (IR) impact on the text repossession. To understand quickly any kind of information, the users desire is to have the maximum exposure of information in a short text illustration. The search engine as a response provides information in mixture of web pages. This profuse information makes the reading density to the users. Due to the information overload problem on the web, the information searching is the critical process in IR applications. IR is the process of achieving the relevant information from the variety of information sources through a web search engine. Hence, the text summarization process is crucial in the emerging web searching field. Document summarization essentially improves the effective retrieval process from the multiple web documents. As the increasing rapid growth of today‘s information world, text summarization is the significant process to quickly interpret the text information. Summarization is the most informative way to convey all the topic relevant information within the short time perspective. The manual summarization process of the large set of documents is a very arduous task for human beings. Hence, an automatic text summarization is used to condense the abundant source text into a shorter version of a summary that ensures the original information content. The summary generation is according to the following factors: 1) the summary can be generated from a single or multiple documents, 2) the length of the summary should be less than the half of the input text, 3) the summary must incorporate the significant information. An effective summarization method targets to concatenate the significant sentences with the essential characteristics of readability, conciseness, and completeness.The summarization technologies are used in the real life systems and industry fields, for instance, Google search engine. The main goal of the text summarization is to create the concise and comprehensive summary with the most informative sentences of the original document, and to facilitate the users to understand the original context within the short-time. The summarization is investigated through Natural Language Processing (NLP) community. Therefore the summary generation is according to either focusing on users, or topic or query.

II. LITERATURE SURVEY

[1] The assignment of automatic document summarization aims at generating short summaries for formerly long documents. A good summary should cover the most significant information of the original document or a cluster of documents, while being coherent, non-redundant and grammatically readable. Numerous approaches for automatic summarization have been developed to date. In this paper they planned to give a self-contained, broad overview of current improvement made for document summarization within the last five years. Specifically, they stress on significant contributions made in recent years that stand for the state-of-the-art of document summarization, including evolution on modern sentence extraction approaches that improves concept coverage, information diversity and content coherence, as well as attempts from summarization frameworks that join together sentence compression, and more abstractive systems that are able to manufacture completely new sentences. In addition, they re-examine

improvement made for document summarization in domains, genres and applications that are deferent with traditional settings. They also point out some of the latest tendency and highlight a few possible future directions.

[2] To make known oneself with a subject area summaries play an important role. Text Summarization is a demanding problem these days. Summarization is very appealing and useful task that gives support to many other tasks as well as it takes advantage of techniques developed for linked Natural Language Processing tasks. Evaluating summaries and automatic text summarization systems are not a uncomplicated process. This review paper discusses an overview of text summarization, a variety of evaluation approach on intrinsic and extrinsic techniques. In standard, text summarization is achieved because of the logically occurring redundancy in text and because important information is spread irregularly in textual documents. Recognizing the redundancy is a challenge that hasn't been fully resolved yet.

[3] It is a familiar issue that deep learning techniques do not work well with tiny amounts of data. With some exceptions, this is, regrettably, the holder for most of the datasets accessible for the summarization task. In addition to this problem, it should be measured that phonetic, morphological, semantic and syntactic features of the language are continuously changing over the time and unfortunately most of the summarization corpus are constructed from older possessions. Another problem is the language of the corpus. Not only in the summarization pitch, but also in other pitch of natural language processing, most of the corpus are only accessible in English. In addition to the above problems, license terms, and fees of the corpus are obstacles that prevent many academic and specifically non-academics from accessing these data.

Abstractive single document summarization is measured as a demanding problem in the field of artificial intelligence and natural language processing. Temporarily and specifically in the last two years, several deep learning summarization approaches were proposed that once again attracted the attention of researchers to this field.

III. PROPOSED METHODOLOGY

The main aim of our project is to present the various online documents in short and concise manner. Based on the user needs the summary has to be produced. Most of the document summarization is extractive based summary because it is just a original text but eliminating dissimilar sentences, but we going to produce abstractive based summary which will be reshape the original document which will be in the human reasonable format. We are using techniques for producing a good summary they are GSO Algorithm. We investigate various techniques that can be used for summarization. We then focus upon a particular approach for summarization of documents belong to a scrupulous domain. The domain that we have considered is news. A new glow-worm swarm optimization (GSO) algorithm is proposed to find the optimal solution for multiple documents in the websites. In news there are a lot of events are happening, if the user wants to see the election for a exacting period of time if he searches for that, the various online news website will produce the results with dissimilar formats, to read the entire document if it is in shorter manner means it will reduce the time and the user can able to understand the things easily. Based the keywords the outcome will be formed. At first the stop words, stemming process will be done, which is nothing but a taking away of words like ‘a’, ‘to’, ‘the’, etc., Secondly the sentence will be ranked/scored, after that the words will be prepared in the same way as that of the unique text. If the ordered has been changed then the meaning of the document cant able to understand. The key words will be in the tree structure that is the words will be ready for summarization . If the event has modernized, the details in the tree structure has also been updated if the user searches for the next time. Finally, the various sets of the summary has been generated to get an optimized summary GSO has been used it has been searches for the best optimized result. It searches for next best it keep on searches.

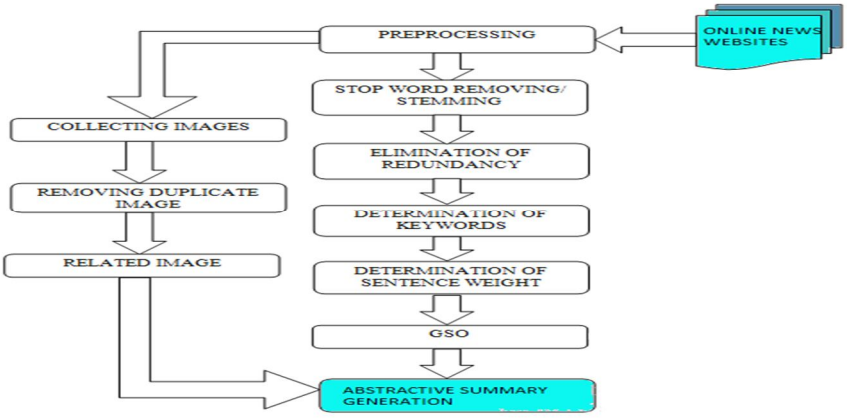


Figure 1 System Architecture

A. Preprocessing

To gather all the data from the different websites. Information retrieval is the group and repossession of information from a large number of text based documents. The pre-processing method identifies key words and associations within the text. It does this by looking for predefined sequence in the text, a process called pattern matching. The software infer the relationships between all the identified places, people, and time to give the user with carrying great weight information.

B. Stopword Removal (Stemming)

Stop words are a division of natural language. The purpose that stop-words should be removed from a text is that they make the text look heavier and less important for analyst. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pro-nouns, etc. that does not give the sense of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc. Stop words are removed from documents using POS Tagger because those words are not measured as keywords in text mining applications.

C. Sentence Ordering

Ordering information is a tricky but important task for application generating natural language texts such as multi-document summarization, question answering, and concept-to-text generation. In multi-document summarization, information is selected from a set of foundation documents. However, inappropriate ordering of information in a summary can confuse the reader and deteriorate the readability of the summary. Therefore, it is very important to properly order the information in multi-document summarization.

D. Term Weight calculation

After deleting the stop words a weight value is enabled to each individual term.

E. Sentence Weight Calculation

After enabling the weight to each term, the next process is to rank the each sentence according to their weight value. The weight of the sentence can be intended by adding the weight of all the terms in the sentence and dividing it by total number of terms in that sentence.

F. GSO

GSO, in its present form, has evolved out of several significant modifications included into the prior versions of the algorithm. GSO starts by distributing a swarm of agents aimlessly in the search space. Agents are modelled after glowworms and, hereafter, they will be called glowworms. Further, they are endowed with other behavioural mechanisms that are not found in their natural counterparts.

G. Summary Generation

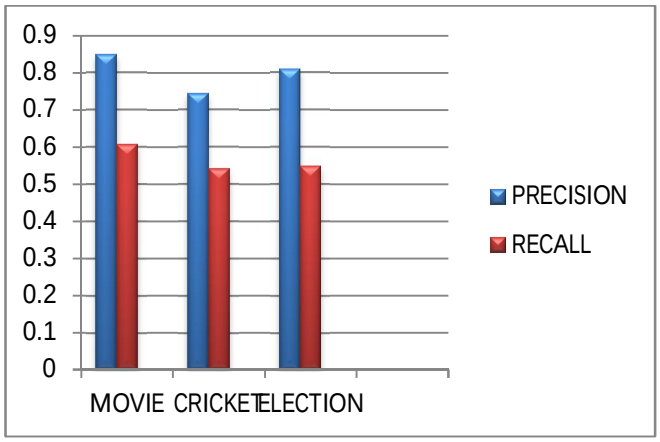
There are generally two types of extractive summarization tasks based on what the summarization program focus on. The first is generic summarization, which reveals on gaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is query relevant summarization, every so often called query-based summarization, which summarizes objects definite to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

IV. RESULT ANALYSIS

In recent years there have been enormous study that investigate further than general document summarization, addressing different use cases for document summarization with specific settings.

For example we have taken three domains namely Movies, Cricket and Election. From these domains we have obtained certain results. For those results we have calculated Precision and Recall.

Precision and Recall are the metrics used for the summary evaluation process. Precision checks how far the results are fit for the summary. Recall checks how far the extracted data is fit for the summary.



V. CONCLUSION

We have proposed a GSO integrated document summarization method for ease of retrieval and readability. GSO optimizes precision and concurrency through iterations for improving the quality of summary. The obtained summaries are close to POS tagging such that the entity-relationships are represented based on the document indices. The indices are classified through prioritized weights that projects method to be better than the existing approaches.

As a future work, we have planned to differentiate semantics of the document based on their popularity and negative feedback analyzed through machine learning.

REFERENCES

[1] Jin-ge Yao, Xiojun Wan, Jianguo Xiao "Recent Advances in Document Summarization", An International Journal of Knowledge and Information systems, Vol. 53, Issue 2 , February 11 , 2017. [Pages 297-336

[2] Deepali K.Gaikwad, C.Namrata Mahender "A Review Paper on Text Summarization", An International Journal of Advanced Research in computer and Communication Engineering Vol. 5, Issue 3, March 2016. [Pages 323-458]

[3] Pashutan Modaresi, Stefan Conrad "An Extendable Multilingual Corpus for Abstractive Single Document Summarization", International Journal of Communication Network Security, December 08, 2016. [Pages 24-27

[4] N. R. Kasture, Neha Yargal, Neha Nityanand Singh, Neha Kulkarni and Vijay Mathur "A Survey on Methods of Abstractive Text Summarization". International journal for research in emerging science and technology, Volume-01, Issue-06, November-2014. [Pages 25-31

[5] Miguel B. Almeida, Mariama S.C. Almeida, Andre F.T. Martins, Helena Figueira, Pedro Mendes, Claudia Pinto "Priberam compressive summarization corpus: A new multi-document summarization corpus for European Portuguese". Proceedings of the Ninth International Conference on Language Resource and Evaluation (LREC'14), May-2014

[6] Jishma Mohan. N, Sumitha. C, Amal Ganesh, Dr. Jaya. A, "A Study on Ontology based Abstractive Summarization". Fourth International Conference on Recent Trends on Computer Science and Engineering, May-2016. [Pages 32-37]

[7] Asha Gowda, Kargowda, Mithilesh Prasad "A survey of Application of Glowworm Swarm Optimization", International Journal of Computer Application, 2013. [Pages 226-251

[8] Siddhartha Banerjee, Prasenjit Mitra, Kazunari Sugiyama, "Multi-document abstractive summarization using ILP based multi-sentence compression". IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence, September 22, 2016. [Pages 975-978]

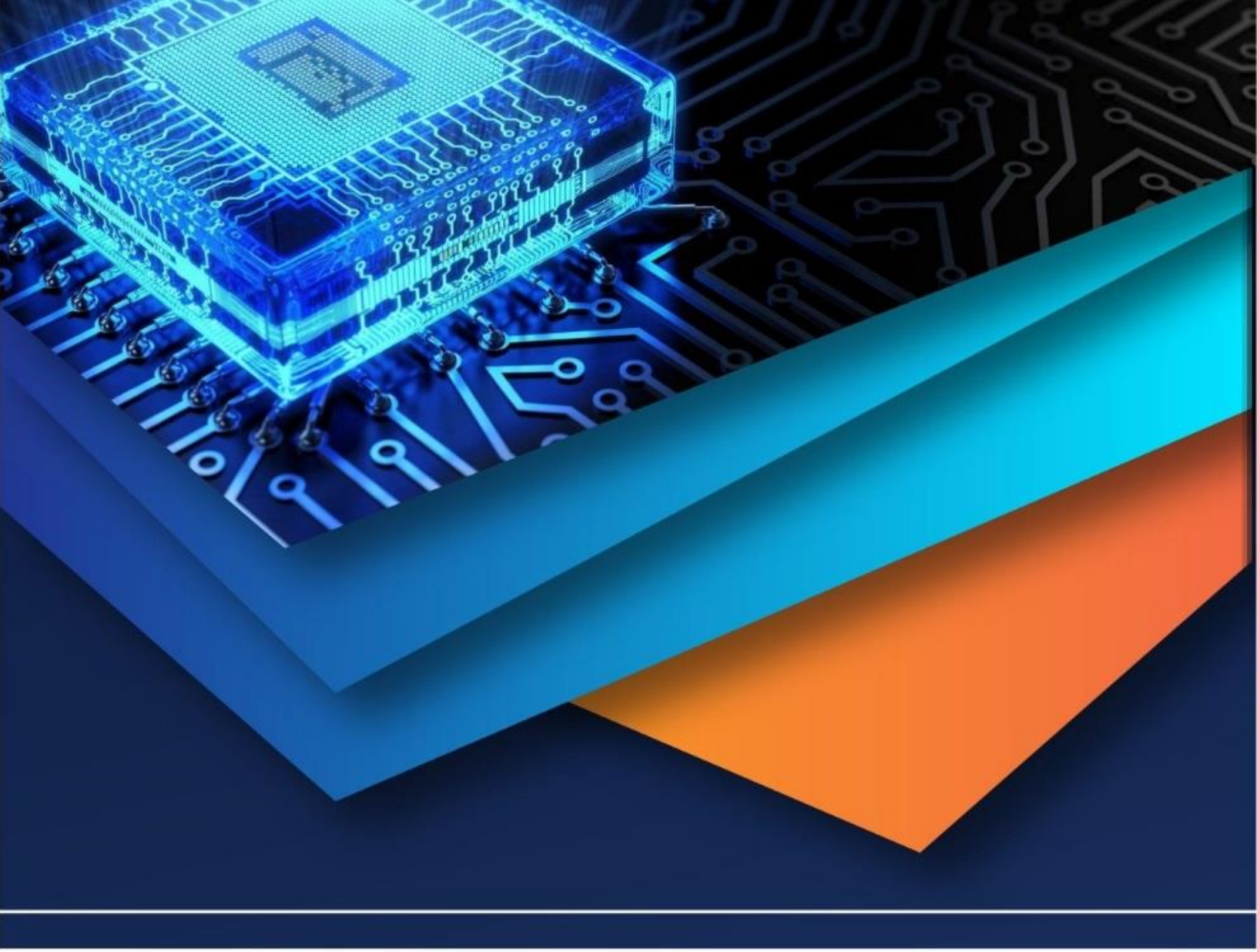
AUTHOR BIOGRAPHY

A. Dr. S. Saraswathi

She is working as Professor, in the Department of Information Technology, Pondicherry Engineering College, Pondicherry, India. She was born on 29th September, 1971 at Karaikal, Pondicherry, India. She completed her B.Tech degree in Computer Science and Engineering in the year 1993 from Pondicherry Engineering College, Pondicherry, India. She has completed her M.Tech in Computer Science and Engineering, in the year 1995 form Pondicherry University, Pondicherry, India. She has completed her Ph.D in the area of speech processing at Anna University, Chennai, India, in the year of 2008. She has been continuing in the academic streamline from 1995. Her area of interest includes Artificial Intelligence and Expert System, Speech Processing, Natural Language Processing, Agentbased Computing and Database Management System.

B. N. Jayasudha, G. Kayalvizhi, V. Ragini

They are pursing their B.Tech degree in the Department of Information Technology in Pondicherry Engineering College, Pondicherry, India.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY