International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Multi-Documents Extractive Text Summarization using Node Centrality

Anish Mathew Kuriakose, V. Umadevi

Abstract*: The advancement of technologies produce vast amount of data over the internet. The massive amount of information flooded in the webpages become more difficult to extract the meaningful insights. Social media websites are playing major role in publishing news events on the similar topic with different contents. Extracting the hidden information from the multiple webpages are tedious job for researchers and industrialists. This paper mainly focuses on gathering information from multiple webpages and to produce summary from those contents under similar topic. Multi-document extractive summarization has been developed using the graph based text summarization method. Proposed method builds a graph between the multi-documents using the Katz centrality of nodes. The performance of proposed GeSUM (Graph based Extractive Summarization) is evaluated with the ROUGE metrics.

Keywords: Extractive text summarization, node centrality, Katz centrality.

I. INTRODUCTION

The volume of electronic reports on the Internet have speed up due to the quick advancement of Internet evolution. Individual internet user can obtain and share the information from different sources due to the rapid growth of social media websites. The Internet as of now gives access to billions of archives. As this builds each second, an unpredictable development of information has been seen over a brief timeframe. Based on the user's inputs, the web crawlers or search engines such as Google, Bing, etc. are able to retrieve most related web pages. Even though the search engines which are having high performance computing are not able to integrate the retrieved document meaningfully. The search engines are immature and lack of ability to provide the information they crawled. This issue lead the path to develop the necessary tools to process the data [1]. For collecting and aggregating the huge number of web documents or any type of documents, the most popular method "text summarization" is being used. Notwithstanding the most recent advancement found in text summarization, the issue has not yet been completely settled. The most significant data in the huge collection of text document is discovered by automated text summarization techniques and it able to consolidate and summarize the contents to be read by the users [2]. Also, it is important to acquire certain highlights by preparing information to abbreviate the time spent getting to data. These issues have expanded enthusiasm for the field of text summarization methods.

Undoubtedly, human life are improved by the abundant data that is being generated and available in the internet, but it makes fast access of data and yet it is difficult to summarize it [3]. The advancement of automated text summarization are progressively intended by industrialists and researchers to accomplish more prominent efficiencies through upgraded techniques. In the text mining research field, most of the work has being carried out in text summarization area and lot of new techniques are proposed for the improvement of text summarization, still the hype of the research is not decreased [1]. The above statement clearly indicates that, text summarization is a hot topic which comes under Natural Language Processing (NLP). It has the goal to provide huge content in a packed and conceivable structure. The most known categories for the text summarization are as follows [2]:

- Extractive summarization which comprises of three phrases: portrayal of writings, sentence scoring, and sentence choice
- Abstractive summarization decipher the principle substance of reports by utilizing NLP methods, and afterward rewrite the content with its own understandings

Based on the number of documents, the summarization can also be categorized as single document summarization and multi-documents summarization. Data deluge in the internet leads to have abundant amount of information. These information are available in multiple webpages which are referring to a similar topic. This leads, the researchers and industrialist to consider the multi-document summarization to produce summary from multiple documents [4]. Summaries can likewise be sorted as either nonexclusive or question highlighted [5]. Most by far of studies directed by scientists have been based on nonexclusive summarization. In this sort of outline, a couple of constrained suppositions about the reason for shaping the summarization are made, and the general substance is kept up, while attempting to cover however much data as could reasonably be expected.

Revised Manuscript Received on February 15, 2020.
* Correspondence Author
Anish Mathew Kuriakose*, Research Scholar, Department of Computer Science, Jairams Arts and Science College Karur affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India.
Dr. V. Umadevi, Director, Department of Computer Science, Jairams Arts and Science College Karur affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India.

Question highlighted summarization is focused on inquiries controlled by the client and data identified with the subject in the content is returned [6].

This research paper is organized as follows: Background study is discussed in the section 2. In section 3, the proposed algorithm is given with a neat architecture. In the proposed structure for Multi-Document Summarization which are utilizing graph theory methods. The experimental results of the proposed graph based extractive text summarization method is presented in section 4. To end with the paper is concluded with future work and limitations in section 5.

II. RELATED WORK

In this literature study, a graph-based non-specific, extractive and multi-document summarization techniques are discussed to extract correct summaries from the given large text of web documents. Extractive summarization method is an easiest way for the learners who are having difficulties to read large content [7]. The ranking of word frequency is an important task in the text summarization with the part of speech tags such as nouns, proper nouns, places and objects [8]. In the text summarization most of the works have been carried out using the graph-based approach.

In another research work, the author [9] have considered the other features of the content such as boldened, italics, underline and quotes as the important properties to find the significance of the sentence. The author [10] have considered the words that were presented in the title to extract the summary. The word in the title are most relevant to the summary. The authors in the paper [11] have discussed about the sentence scoring methods. The position of the sentence with highest occurrence of words in the document are determined by the sentence ranking. Ranking the phrases is also an important way of study to extract the summary from the large text. The author [12] have discussed the different types of scoring methods such as word, sentence and graph based scoring. Graph is the easiest way to find the link between one sentences to another sentences. It defined the structure by showing the relationships between the sentences. Graph represents the semantic flow of the contents. The link between one nodes to another depicts the continuity of the meaning of the documents. The strongest and weakest relationship denotes the relevancy of the sentence flow [16]. Most of the graph-based text summarization have been carried out using the LexRank and TexRank method. TexRank which contains the edges between the vertices are used to find the relationships among the different sentences [13]. Eigenvector based node centrality is used along with LexRank to produce the summary [14]. The authors [15] have discussed the advantage of PageRank algorithm in the text summarization. Maximum of the researchers have concentrated on extractive text summarization. This research work also focusing extractive text summarization using graph-based approach to increase the accuracy of ROUGE.

III. PROPOSED SUMMARIZATION TECHNIQUE

The proposed multi-documents extractive text summarization aims to provide a summary that are collected from the multiple webpages under a similar topic. It aims, to improve the accuracy of graph-based summarization approach. A research flow diagram of the proposed summarization technique is shown in figure 1.

A. Big Data Pre-processing

Huge amount of texts are collected through the web crawler which is developed by using python 3.7. The web crawler collect the texts based on the event passed by the user. Depends on the topic / event passed as input, the web crawler extract the similar contents from the multiple webpages. These documents contain high amount of noisy data. The raw texts are then clean by the big data pre-processing methods. Audios, videos, images, URL, etc in the raw texts are removed and the text contents are stored in a database. The cleaned texts are used for the further process. Algorithm 1 shows the pseudo code for the big data preprocessing.

Algorithm 1: *Big Data Preprocessing*
Input: *Event name, e*
Output: *Cleaned text documents, D*

1 $u_i \in U$
2 $d_i \in D$
3 e = {}
4 initialize i =0
5 for u_i in U:
  a if e in u_i:
  b remove images, audios, videos in d_i
  c $d_i = u_i$.text()
  d increase i by 1
6 end for
7 return D

The first task of proposed summarization technique is to collect the huge texts which are relevant to the given event name. Search engines retrieve all the data which are relevant to the given event name. The most of content are not necessary, since it returns the content which are discussed by the blog writers and fake news producers. To avoid these issues, the proposed technique considered only the selected news article websites. In the algorithm 1, *'u'* denotes the URL link of the selected news article websites. *'e'* denotes event name. The given event name is searched in the contents that are available in all the given URL links. Noisy data such as images, audios, videos, etc which are present in the webpages are removed by the preprocessing function that is written using *'BeautifulSoup'* of the python library. The cleaned text are stored as single document with a document identity for the further references. *'d'* denotes text document.

B. GeSUM Technique

The proposed GeSUM (Graph based Extractive Summarization) text summarization technique consists of two main phases that are as follows:

- Keyword ranker
- Topic modeler

Keyword Ranker

This is the first phase of the proposed technique. The crawled texts from the multiple webpages are stored in separate document.

Algorithm 2, represents the complete flow of the proposed GeSUm technique.

The important keywords that are available in the documents are identified by the ranker formula. The most popular keyword ranker is tf-idf method.

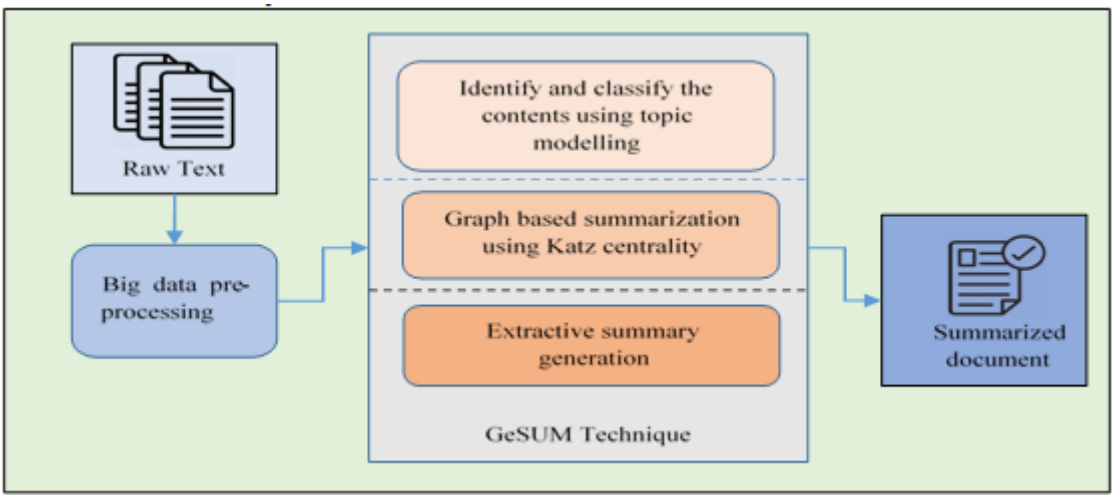

Figure 1 Research flow of Multi-Document Extractive Text Summarization

This method outperformed well for referring and comparing between multiple documents. In the proposed work, it has to rank the keyword within the document at first stage. Hence it is recommended to use the Zipf's law. This law produced better ranking for identifying the most frequent word within the document. Zipf's law states that the frequency of any word is inversely proportional to its rank in the frequency table [17]. The equation 1 denotes the definition of the Zipfs' law.

$$Prob(r) = \frac{Freq(r)}{N} \quad \text{..... Eq. (1)}$$

where

Freq (r) = frequency of word occurs at rank r in the text document and

N = total number of words in the text document.

The important keywords are ranked and stored in a separate list. Word with the following characters: capitalized, italicized, bold, underlined which are not categorized under the top ranked words are also considered. These characters are represent the importance of the word [18]. The ranked keyword are send to the second phase.

Topic Modeler:

In the final phase, word cohesions between the sentences or phrases are represented graphically by using the topic modelling and node centrality measures. Topic modelling plays an important role for summarizing the large documents. The top most keywords that are identified in a document is compared across all other documents using the LDA (Latent Dirichlet Allocation) method. This identifies the filtered keywords across the multiple documents under the same topics and dumped the data into a single document. LDA accomplishes the following tasks for topic modelling on each every document d:

1. Consider there are 't' topics in the collected documents similar to the same event 'e'.
2. Given these 't' topics from a document d by assigning topic to each word.
3. Consider the topic 't' is incorrect to the other word in the document which is allocated as correct topic for each word in the web documents.
4. Find the probability of correctness across the topics
5. The above steps has to be repeated until the document reaches it end.
6. Based on the probability value extract the topics

The LDA model is used to tokenize the sentences based on the topics across the multiple documents. The tokenized sentences are used for generating summary by applying graph-based technique. In algorithm 2, line 9 generates graph G = [V, E, W], where V = set of vertices equivalent to sentences indicated by word count vector, E = set of edges and W = set of weights that are associated with each edge 'e'. The tokenized sentences in the document are ranked by using 'SenRank' which is calculated by the graph-centrality of degree method. The node centrality is used for finding the relationships among the connected sentences. Node centrality is found by using the Katz centrality measures [18].

Katz Centrality:

Let A be the adjacency matrix of a graph which is to be generated. Features (aij) of A are variables that consider a value 1 if a node 'i' is connected to node 'j' and 0 otherwise. In simple term, if a sentence is related to any other sentences in the document 'D' then the value for A is 1. The powers of 'A' specify the occurrence (presence or absence) of relationship between two nodes through mediators. For example, in matrix 'A5' , if feature (a6, a25) = 1, it specifies that node 6 and node 25 are linked over some route of length 5. The length of the node is used to ranking the sentences. The Katz centrality of a node 'i' is denoted mathematically by the equation 2.

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \propto^{k} (A^{k})ji \quad \text{..... Eq. (2)}$$

IV. RESULTS AND DISCUSSIONS

In this research work, multiple document which were extracted from top news article such as NDTV News, IbnLive, BBC News India, Indian Express, India Today, Reuters India and ZeeNews websites. Experimental methods were accomplished using a computer with an Intel Core i7-5th generation CPU with 3.60 GHz and 8 GB memory using Python 3.7. Beautiful soup library was used to crawl the web pages from the internet. Spacy 2.0 library was used for the textual processing.

ROUGE (Recall - Oriented Understudy for Gisting Evaluation) metrics are the most prevalent valuation metrics used in text summarization classifications. These metrics are assessed based on word sequences and n-grams. Rouge is measured based on the evaluations between computer generated summary method and human generated summary. The scores produced by the ROUGE are range from 0 to 1. The rank of the sentences was calculated by the Katz centrality measures. The more connected sentences are considered for generating the summary. The sentences that are linked to each other represents the importance of the content that is to be summarized and it would be more informative. The effect of node centrality played a major role to identify the flow of the contents. The sentence in one document was related to a sentence in another document. Relationships between the sentences are calculated based on the similarity of the word vectors. The proposed summarization technique is compared with the baseline TexRank algorithm. Table 1 represents the comparative analysis of the proposed algorithm with the TexRank algorithm.

Figure 2, 3 and 4 shows the comparative chart of the algorithm TexRank and GeSUM for the metrics Rouge 1, Rouge 2 and Rouge L respectively. The results shows the efficiency of the proposed algorithm. It is clearly visible that there is increase in the accuracy of the proposed work when compared to the existing work. The graph generated using the Katz centrality eliminates the sentences which are having low degree of relationships with the higher sentences. This generated the summary of the multi-document.

Table 1. Comparative results of TexRank and GeSUM

	TexRank			**GeSUM**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
ROUGE-1	0.6123	0.629	0.6076	0.6471	0.6389	0.6429
ROUGE-2	0.5948	0.5891	0.5919	0.6258	0.6198	0.6228
ROUGE-L	0.6748	0.6698	0.6722	0.7326	0.7299	0.7312

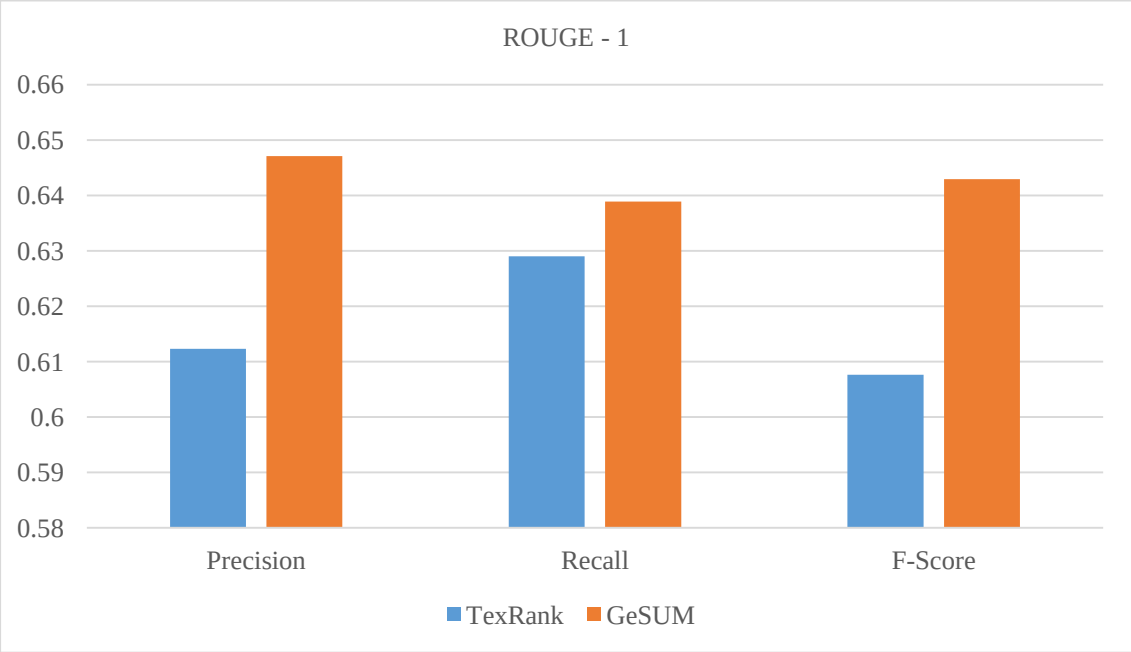

Figure 2 Comparative results of TexRank and GeSUM for ROUGE - 1

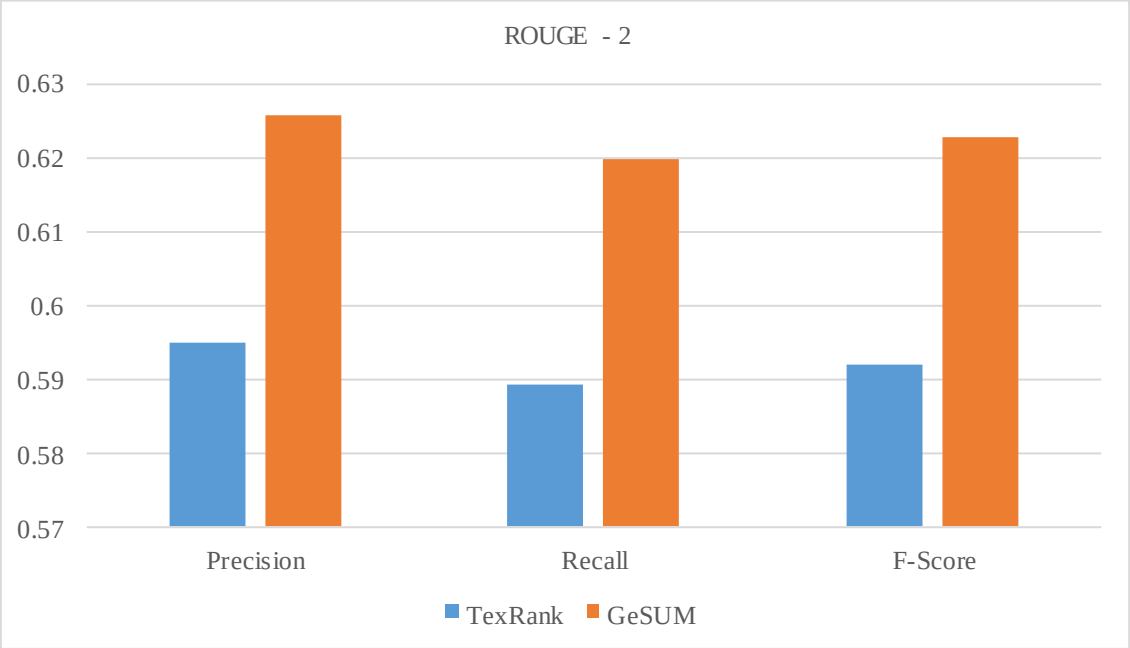

Figure 3 Comparative results of TexRank and GeSUM for ROUGE - 2

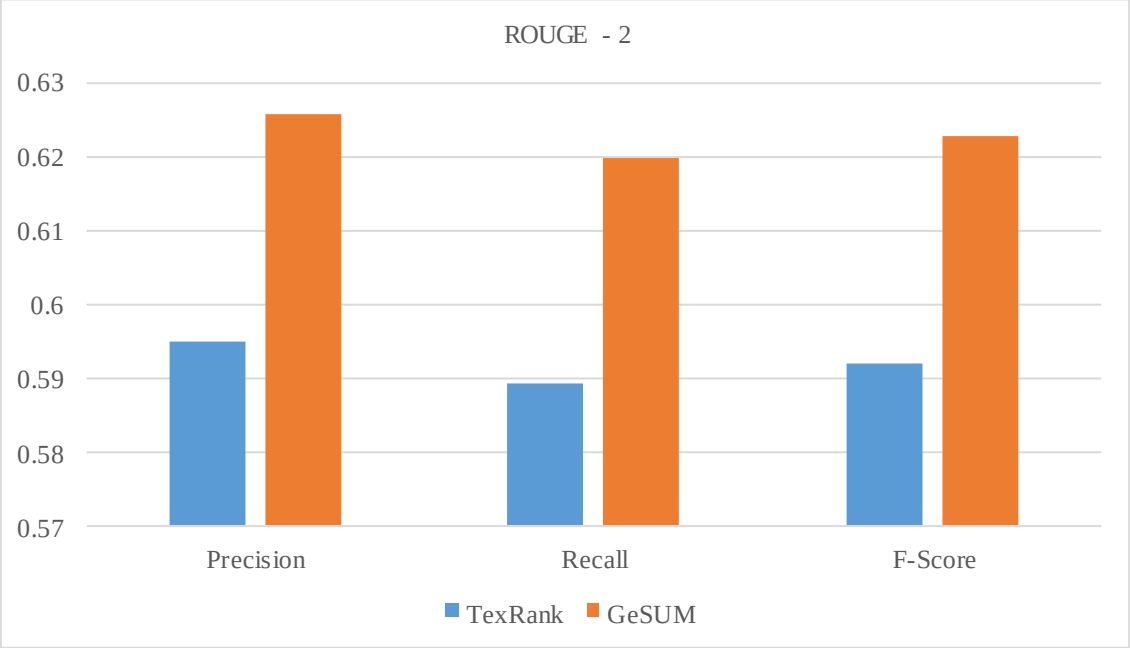

Figure 4 Comparative results of TexRank and GeSUM for ROUGE - L

V. CONCLUSION

The proposed GeSUM technique outperformed well when compared to the existing TexRank method. The pre-processing algorithm cleans the raw text and created an easy-to-use text for the extractive summarization. This research work produced an effective summary from the multiple documents. The relationship between the sentences and documents are measured by using the Katz node centrality. The Zipf's law which was used to rank the keyword played a key role for the topic modelling. LDA based topic model also supported to identify the similar topic which was discussed by the news channel. This helped to extract the good summary from the multi-documents. The limitations of this work are, the proposed algorithms are generic and it is to be worked for the preloaded web pages. In future, the work is like to be expanded for the abstractive text summarization.

REFERENCES

1. Ge Yao J, Wan X, Xiao J. "Recent Advances in Document Summarization". Knowledge Information Systems 2017; 53(2): 297–336.
2. Ermakova L, Cossu JV, Mothe J. "A Survey on Evaluation of Summarization Methods". Information Process Management 2019; 56(5):1794–814.
3. Mao X, Yang H, Huang S, Liu Y, Li R. "Extractive Summarization using Supervised and Unsupervised Learning". Expert System Applications 2019; 133:173–81.

4. Joshi A, Fidalgo E, Alegre E, Fernández-Robles L. “SummCoder: an Unsupervised Framework for Extractive Text Summarization Based on Deep Auto-Encoders”. Expert System Applications 2019; 129:200–15.
5. Gambhir M, Gupta V. “Recent Automatic Text Summarization Techniques: A Survey”. Artificial Intelligence Revolution 2017; 47 (1):1–66.
6. Sarkar K, Saraf K, Ghosh A. “Improving Graph Based Multidocument Text Summarization using an Enhanced Sentence Similarity Measure”. In: 2015 IEEE 2nd International Conference Recent Trends Information System ReTIS 2015 – Proceedings p. 359–65.
7. Van Lierde H, Chow TWS. “Query-Oriented Text Summarization Based on Hypergraph transversals”. Information Process Management 2019; 56(4):1317–38.
8. Xiong S, Ji D. “Query-Focused Multi-Document Summarization using Hypergraph-Based Ranking”. Information Process Management 2016; 52(4):670–81.
9. Erkan G, Radev DR. “Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization”. J Artificial Intelligence Res 2004; 22:457–79.
10. Coe DM. “A Comparative Study of Hindi Text Summarization Techniques”. Genetic Algorithm and Neural Network, 2015.
11. Gupta V. “Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents”. In: Mining Intelligence and Knowledge Exploration. Springer; 2013. p. 717–27.
12. Gupta V, Lehal GS. “A Survey of Text Summarization Extractive Techniques”. J Emerging Technology Web Intelligence 2010; 2(3):258–68.
13. Parveen D, Ramsl H-M, Strube M. “Topical Coherence For Graph-Based Extractive Summarization”. In: Proceedings of the 2015 conference on empirical methods in natural language processing. p. 1949–54.
14. Salton G, Singhal A, Mitra M, Buckley C. “Automatic Text Structuring and Summarization”. Information Process Management 1997; 33(2):193–207.
15. Medelyan O. “Computing Lexical Chains with Graph Clustering”. In: Proceedings of the ACL 2007 Student Research Workshop. p. 85–90.
16. Nandhini K, Balasundaram SR. “Improving Readability through Extractive Summarization for Learners with Reading Difficulties”. Egyptian Information Journal 2013; 14 (3):195–204.
17. G. Vaitheeswaran, L. Arockiam, “Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets”, International Journal of Advance Research in Computer Science and Management Studies, Volume 4, Issue 5, May2016.
18. Eisha Nathan, Geoffrey Sanders, James Fairbanks, Van Emden Henson, and David A. Bader, “Graph Ranking Guarantees for Numerical Approximations to Katz Centrality”, International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland, pp 68–69.

AUTHORS PROFILE

Anish Mathew Kuriakose is a Ph.D Research Scholar at Jairams Arts and Science College Karur affiliated to Bharathidasan University . He completed MCA and MBA from Bharathidasan University Tiruchirappalli. His areas of interest are Data mining, Mobile Communication and Big Data.

Dr. V.Umadevi M.Sc(CS &IT). M.Tech (IT)., M.Phil., PhD., D.Lit.Currently working as Director, Department of Computer Science, Jairams Arts and Science College, Karur.She is a reputed author of few journal publications and wrote three books.