Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 11, November 2018, pg.142 – 149

A Hybrid Approach to Single Document Extractive Summarization

Som Gupta[1], S.K Gupta[2]

[1]PhD Computer Science Engineering Department, AKTU Lucknow, India

[2]BIET Jhansi, Computer Science Department, Jhansi, India

[1] somi.11ce@gmail.com; [2] guptask_biet@rediffmail.com

Abstract— Extractive summarization is a technique to compress the text in such a manner that only the important original sentences representing the text are considered for the summary. Creating a concise and non-redundant summary is not an easy task. There are a number of techniques which are already available for creating the summaries like graph-based techniques, feature-based techniques, fuzzy logic, etc.
In this paper we propose a hybrid approach using graph-based technique (PageRank), MMR along with K-means clustering and feature-based extraction technique along with the fuzzy logic to create the extractive summaries. The experiments show that our hybrid approach gives either better results or comparable results than the individual techniques.

Keywords— summarization, extractive summarization, pageRank, MMR, Feature Based Extraction

I. INTRODUCTION

This document is a template. The exponential growth of information has posed a big challenge of representing the information in a useful manner. Task of manually creating summaries of lengthy documents is not an easy task. Automatic text summarization is one of the most growing fields of research in the field of natural language processing which reduces the content of text in such a manner that the main thought of the documents remains preserved [1]. It is one of the best way to help reduce the information overload. Text summarization techniques are categorized into extractive and abstractive on the basis of how the sentences are represented. Extractive summarization [2] [3] is to extract the important sentences from a document in the same manner in which they are present in the original document. Extractive summarization mainly involves calculating how important the sentence or a paragraph is in order to extract the important sentences. Abstractive summarization [4] deals with linguistic techniques to interpret the meaning of document and generates a human like summary. Abstractive summarization techniques build internal semantic relationships and then applies natural language processing techniques.

There are various approaches which have been used to perform extractive summarization like TF-IDF [5] approach where the importance of word is calculated on the basis of frequency of the appearance of word in the document but because of the fact that many times, the longer sentence gets the high score due to the large number of words, TF-IDF alone is not a very effective approach to perform summarization. Cluster based approach is also used for extractive text summarization where the clusters are generated from high TF-IDF

terms which represents the main theme to filter out the sentences. Neural based techniques, Genetic and fuzzy based approaches have also been used to create summaries.

Text summarization is not an easy task as it is very subjective and there does not exist a best summary. Not only this but issues like coverage, consistency, redundancy, coherence, sentence ordering, anaphora resolution, dealing with references make it even more difficult [6]. For example, while arranging the sentences during text summarization, if the relationships between sentences are not considered then salience and coverage cannot be achieved properly. Salience means including the sentences which are most relevant to the main topics of the text and coverage means covering more and more information in the summary while maintaining the less redundancy among the sentences included for the summary. Even though statistical features methods, help increase the salience property of summaries but they don't help achieve good coverage. Coverage can be achieved by applying techniques like MMR. Reducing redundancy is not much possible in case of extractive summarizations but we have tried reducing the redundancy to an extent by using MMR technique.

In this paper, we have proposed a technique for creating single document extractive summaries. We have used a hybrid approach utilizing the power of feature based extraction, fuzzy logic, MMR and PageRank for creating summaries. Feature based extraction helps capture the word-level relationships in the sentences or sentence level relationships by considering the features. MMR helps achieve non- redundancy. PageRank helps capture the sentence-level relationships without having any domain-knowledge. Fuzzy logic helps deal with the uncertainties in our approach. For feature based extraction, we have used unique terms feature(UniqueTerm), title similarity feature(Title Similarity), number of numeric tokens feature(Numeric Tokens), TF-ISF(TF ISF) for a sentence feature, cue-phrase(Cue Phrase) feature, sentence Length ,and number of positive and negative keywords feature (PosNeg). We have taken the individual summaries and then created a summary by taking all the common sentences or high TF-ISF sentences from different summaries produced by individual approaches. We have used ROUGE-n Score for evaluating the summaries. Organization of paper is as follows : Section 2 discusses the related word done in this field, Section 3 discusses about the methodology used for our approach. Section 4 is about experiments performed and Results obtained and finally the conclusion.

II. RELATED WORK

The work in the field of text summarization is not a new topic. It started in 1958 when Luhn [5] studied how the weight of a sentence depends on the high frequency words. Most of the initial studies in the field of text summarization are based on the features like sentence location, title similarity and unique words, etc. Then in 1969 Edmundson[7] along with the term frequency added 3 more components namely pragmatic words or Cue-Phrases, similarity to title and sentence location for the summarization. In 2001, Radev et al. [8] proposed an approach called MEAD, a cluster based approach for extractive summarization by using centroid score, similarity to first sentence and sentence location. In 2005 Jagadeesh et al. [9] used presence of verbs, parts of speech tag, named entities, font, familiarity of word and occurence of headings and subheadings along with the previously used features.

Ladda Suanmali et al.[3] in 2009 used fuzzy logic along with the feature extraction to improve the quality of summaries obtained. They have used sentence length, sentence position, title similarity, tf-idf, number of pronouns, number of numerical data , thematic data and sentence-to-sentence similarity to find the sentence score using features. Ferreira et al. [10] in 2013 performed the qualitative and quantitative assessment of various features on sentence scoring and has emphasized on the need of considering morphological transformation, synonyms, co-references, ambiguity and redundancy to improve the sentence ranking. Babar et al. [11] in 2014 used feature based extraction along with the Latent Semantic Analysis to obtain the extractive summaries. PadmaLahari et al. [12] in 2014 used the feature based extraction and created the summary of size equal to number of paragraphs by using successive threshold method. Kurmi et al. [13] in 2014 used enhanced MMR approach for summarization where the maximal marginal relevance was calculated between the sentences. Jafari et al.[14] in 2016 used syntactic parameters like TF-ISF, Sentence Length, Sentence Location, Similarity to Title, Similarity to Keywords, number of nouns and semantic parameters like semantic similarity between sentences and order to words in a sentence to calculate the feature score for a sentence and then applied fuzzy logic to find the degree of importance and correlation to improve the results. Liu et al. [15] in 2017 used variation of PageRank by personalizing it, based approach for multi-document summarization.

III.METHODOLOGY

The Fig 1 illustrates the complete flow of our approach. The First step of our approach is pre-processing of the text.

• Pre-Processing of Text: pre-processing of text involves segmenting the text into sentences by identifying the sentence boundaries, stop word removal , punctuations removal and stemming. Stop word removal is to

remove the words without meanings from the sentences like conjunctions and articles. Stemming is to find the stems of the words.

The first step in the process of text summarization involves:

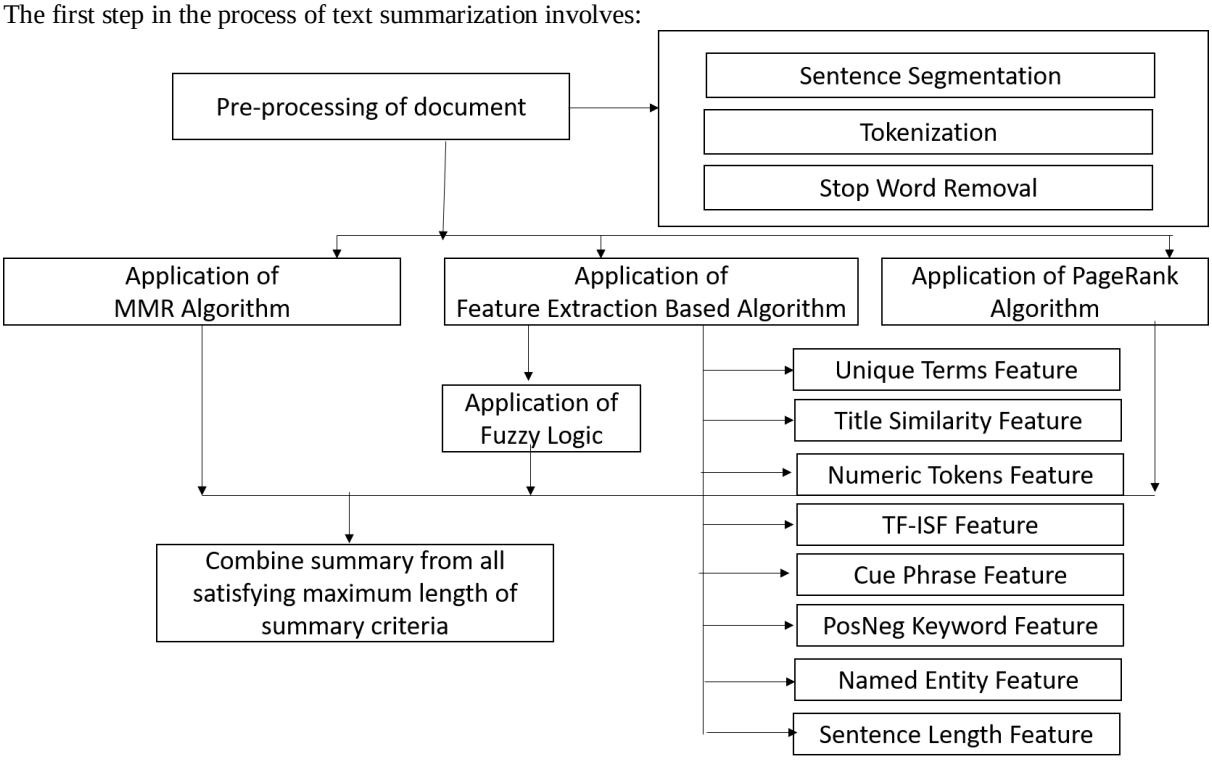

Fig 1: Our Method

After pre-processing of the text, processing of the text is done.

• Processing of Text: For the processing of text, we have used MMR approach, Feature-Extraction approach and PageRank approach individually and obtained the summaries..

A) Features-Based Extraction Method : In this method, the score of sentence is calculated on the basis of various features like how similar is the sentence from the title, how many unique terms, named-entities, numeric-terms, cue-phrases, positive-negative keywords a sentence has, TF-ISF score of a sentence and sentence length. In order to find out the important sentences, we have calculated the scores for the features and then normalized them to range between 0 to 1. For each sentence, we have created a feature vector, which is used by fuzzy logic module fuzzifier for performing further calculations. Even though there are a number of features which can be considered but for our approach[10] we have used the following features based on the results of Ferreira et al. [10] where they have found the features which performed best with the datasets.

Features-Extraction:

a) Unique-Terms Feature: It represents the number of unique terms in the sentence. More the number of unique

words in a sentence, more important the sentence is.

UniqueTermsValue=No of unique terms in a sentence / No of terms in a sentence---------- (1)

b) Title-Similarity Feature: It represents the similarity of words in a sentence with the title. We have also used the synonyms to capture the similarity of words in case different word is used which is similar to title. More the similarity is with the title, more important the sentence is as it represents the theme of the document.

Title_SimilarityScore=No of common terms between title and sentence / No of terms in a sentence----- (2)

c) Numeric Tokens Feature: It represents the number of numeric terms in a sentence. If the number of numeric tokens are more, it represents either the statistics of the topic or some event. So more the number of tokens, more is the importance of sentence.

Numeric_TokensScore= No of numeric tokens in sentence / No of tokens in a sentence---------- (3)

d) TF-ISF Feature:TF-ISF[14] is a single document version of TF-IDF. Term-frequency assigns the weights to the words of a sentence on the basis of the frequency count of words in the document whereas inverse-sentence frequency decreases the weights of most frequently occurring words indicating that more frequently occuring words are not very important.TF-IDF is used for multiple documents. TF represents the terms-frequency for a document and represents the number of times a particular word has occurred in the sentence. ISF represents inverse sentence frequency and represents the number of times the word has occurred in the document. If the ISF value is more, it means it is used many times in the sentence and most probably, these words are not of much importance.

$$TF[word] = 0.5 + 0.5 * f1/totalwords \quad (4)$$

where f1 is the frequency of word in the text

totalwords is the number of words in the text

ISF[word] = math.log(N/f2) ----------(5)

where N is the frequency of word in the text f2 is the number of sentences in which the word has appeared

$$TF - ISF = tf[word] * isf[word] \quad (6)$$

e) Cue-Phrase Feature: Cue-Phrases are the words which describes the representation of sentences[1] like “in the summary”, ”because”, ”our survey”., etc and they are considered to be useful for inclusion in the summary.

f) Positive Negative Keywords Feature: Here we have calculated the number of positive and negative keywords in the sentence and then normalized them in between 0 to 1.

g) Named-Entities Count Feature: Here we have normalized the number of entities found in the sentence to the maximum number of entities found in any sentence of text. Finding the named entities means finding what the real word entity is like whether it is a person or a geographical location or an event, etc. Named entities help finding the relationships between the entities. More the number of named entities, more important the sentence is.

h) Sentence Length Feature: Here we have normalized the length of sentence and found the sentence length. The sentences which are smaller than a fixed threshold are considered not useful.

Sentence Length=No of words in a sentence / No of words in the longest sentence---------- (7)

Along with the statistical calculation of sentence scores on the basis of feature values and to deal with the uncertainties we have used fuzzy logic to improve the quality of sentences chosen. Fuzzy logic is a reasoning similar to human reasoning and instead of giving answers in "true" or "false", it gives result in terms of degree of truth. Degree of membership for the feature vector given as an input to fuzzifier is calculated using trapezoidal membership function. If-Then rules were used for rule based engine to find the important and unimportant features [16]. Final score for the sentence on the basis of importance is calculated using centroid method. Fuzzy engine of the system is shown in Fig 2.

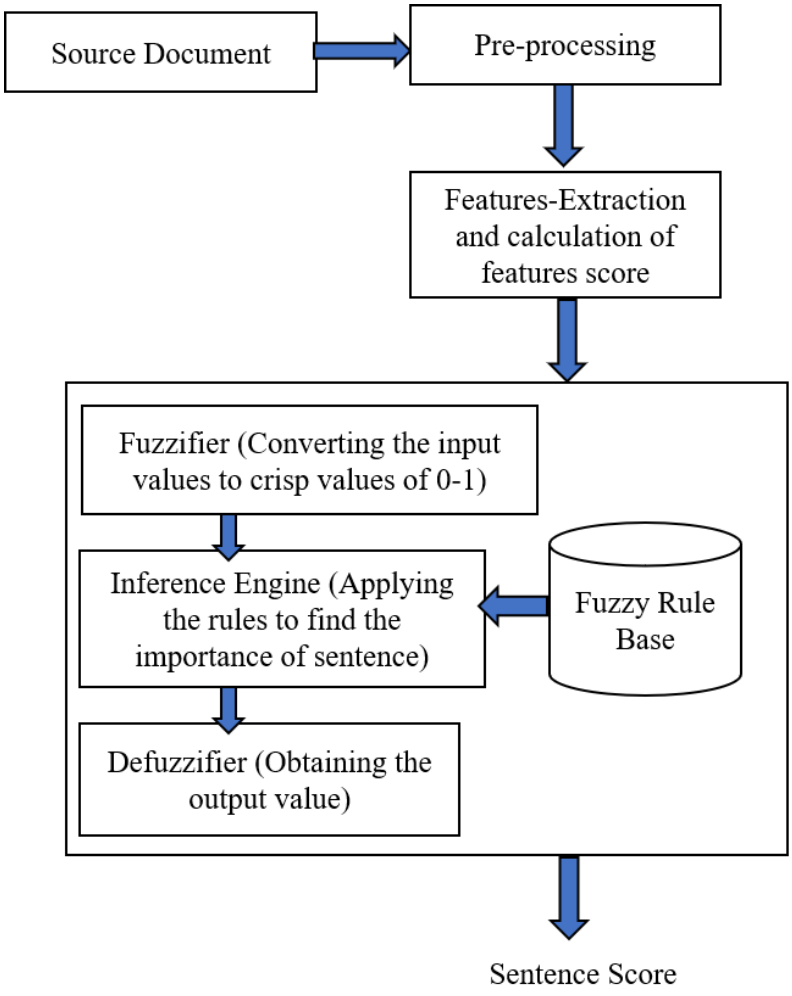

Fig. 2: Fuzzy Engine

B) PageRank : It is a graph based ranking algorithm where sentences represent vertices and the edges represent the relationships between the sentences. Edges represent the semantic similarity between the sentences. PageRank score is calculated for each sentence. PageRank helps find the ranking of sentences in a document by their importance [17]. Top-k sentences are extracted from the text according to the PageRank score. They are based on the centrality. As PageRank approach does not required linguistic knowledge and is independent of the domain, thus it is very portable and which makes it a good candidate to capture the sentence level relationships [18].

C) MMR : It is a semi-supervised incremental greedy approach for performing text summarization. A sentence is said to have high marginal relevance if it is similar to the query and has minimum similarity to the previously selected sentences. It is a query based summarization technique [19]. For creating a query for the technique, K-means algorithm is used to find the keywords which represent the theme of the text. It is an algorithm to quantify the degree of dissimilarity among the current sentence and the previously selected sentences. MMR helps in removing the redundancy among the sentences while maintaining the query-relevance. It also helps get the informative sentences. But MMR alone does not give very good results [20]. Clustering is one of the widely used technique for multi-document summarization to find the clusters of similar documents. In this paper, we have found the cluster of words by considering each sentence as an individual document. Here in our approach we have used K-means for creating the query. The reason for choosing the K-Means clustering is its simplicity and its efficiency on working with large datasets[1].

IV. EXPERIMENTS AND ANALYSIS

In this section, we present the dataset used for experiments, evaluation measures used for comparing the results and implementation detail in terms of experimental set up.

A. Dataset

We have used Opinosis Opinion Dataset 1.0 [21] for our experiments purpose. There are 51 articles in this dataset, each comprising of approximately 100 sentences and which are mostly about opinions about mobile phones, Cars, etc which are collected from sites like TripAdvisor, Edmunds.com, Amazon.com, etc. Dataset also contains gold summaries for the articles. For each article, there are 4 gold summaries which are in natural form.

B. Evaluation Measures

There are 2 ways to evaluate the summaries- intrinsic and extrinsic. Intrinsic evaluation is basically human evaluation while extrinsic evaluation is task based evaluation. We have used intrinsic evaluation to evaluate the summaries [1]. ROUGE which is Recall-Oriented Understudy for Gisting Measures, is one of the automatic way to calculate the effectiveness of summaries. Precision, Recall and F-Score are mainly used for evaluating whether the summary obtained is good or not.

Precision is the ratio of number of sentences occurring in both the gold summary and the system generated summary to the number of sentences in the system generated summary. Recall is the ratio of number of sentences occurring in both the gold summary and the system generated summary to the number of sentences in the gold summary. F-Score combines both the precision and recall and is the harmonic mean of both precision and recall. Precision and Recall individually cannot determine whether the summary obtained is good or not so F-Score is what that we look to determine the quality of summary.

To find out the precision and recall, we have used ROUGE-N calculations where similarity of n-grams is found between the system generated summary and gold summary [22].

C. Implementation Details

In our experiment we have randomly chosen 9 articles and have applied Feature Based approach, MMR algorithm, PageRank algorithm and then applied hybrid approach using all feature based with fuzzy logic, MMR and PageRank.

We have calculated precision, recall and F-score for evaluating the algorithm. For comparison we have calculated the precision, recall and F-Score for each manual summary individually and then calculated the average of all the individual results. We have created the summaries of maximum 250 characters length. From the results, the proposed method gives better results than the individual approaches.

V. CONCLUSIONS

The version of this template is V2. In this paper, we proposed a hybrid approach to automatic text summarization which utilizes the power of feature based extraction, fuzzy logic, semi-supervised approach MMR, unsupervised approach PageRank. Feature based extraction helps capture the word level relationships, fuzzy logic helps deal with the uncertainties of summaries whereas PageRank helps capture the sentence level relationships. MMR helps achieve non-redundancy in our summarization system. The experimental results obtained by using ROUGE framework depicts that the combined approach gives better or comparable results than the individual approaches.

Although the hybrid approach gives better results, but in some cases the precision is not even 1 percent. Also the dangling anaphora problem still remains. In future we will be using abstractive techniques along with the extractive techniques to create the summaries.

Dataset	Average Precision	Average Recall	Average F-Score
Article1	0.143	0.791	0.211
Article2	0.194	0.795	0.309
Article3	1	0.013	0.033
Article4	1	0.023	0.048
Article5	0.054	0.894	0.096
Article6	0.177	0.723	0.261
Article7	0.178	0.768	0.269
Article8	0.007	1	0.014
Article9	0.133	0.804	0.210

TABLE 1: Evaluation Measures from PageRank Algorithm

Dataset	Average Precision	Average Recall	Average F-Score
Article1	0.140	0.815	0.211
Article2	0.19	0.730	0.456
Article3	0.223	0.703	0.331
Article4	0.155	0.862	0.255
Article5	0.0642	0.944	0.112
Article6	0.137	0.783	0.220
Article7	0.190	0.709	0.274
Article8	0.004	1	0.002
Article9	0.174	0.757	0.251

TABLE 2: Evaluation Measures from MMR Algorithm

Dataset	Average Precision	Average Recall	Average F-Score
Article1	0.084	0.863	0.143
Article2	0.252	0.711	0.371
Article3	0.200	0.777	0.310
Article4	0.131	0.848	0.227
Article5	0.023	0.964	0.0444
Article6	0.103	0.826	0.176
Article7	0.113	0.819	0.191
Article8	0.00	1	0.005
Article9	0.198	0.742	0.273

TABLE 3: Evaluation Measures from Feature Based Extraction Approach

Dataset	Average Precision	Average Recall	Average F-Score
Article1	0.230	0.743	0.276
Article2	0.239	0.707	0.354
Article3	0.239	0.646	0.344
Article4	0.131	0.848	0.227
Article5	0.053	0.924	0.095
Article6	0.178	0.725	0.263
Article7	0.212	0.724	0.300
Article8	0.008	1	0.015
Article9	0.198	0.742	0.270

TABLE 4: Evaluation Measures from our Hybrid Approach

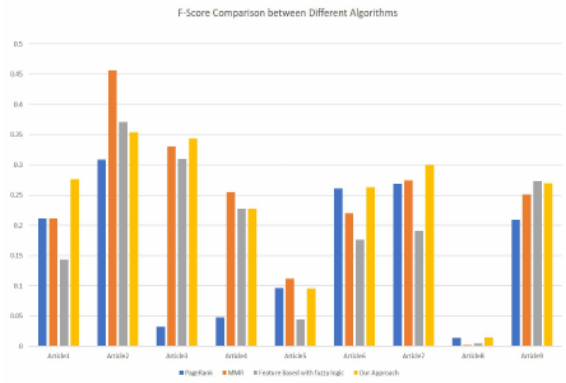

Fig. 3: Comparison F-Score between different algorithms

REFERENCES

[1] S. alZahir, Q. Fatima, and M. Cenek, "New graph-based text summarization method," in IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Aug 2015.
[2] S. Shazia Begum and P. S. Sajja, "Literature review on extractive text summarization approaches," International Journal of Computer Applications, vol. 156, no. 12, Dec 2016.
[3] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy logic based method for improving text summarization," International Journal of Computer Science and Information Security, vol. 2, Jan 2009.
[4] A. KHAN and N. SALIM, "A review on abstractive summarization methods," Journal of Theoretical and Applied Information Technology, vol. 59, pp. 64–72, Jan 2014.
[5] H. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research and Development, vol. 2, pp. 159–165, 1958. [6] N. Nazar, Y. Hu, and H. Jiang, "Summarizing software artifacts: A literature review," Journal of Computer Science and Technology, vol. 31, no. 5, Sep 2016.
[7] H. P. Edmundson, "New methods in automatic extracting," Journal of the ACM (JACM), vol. 16, no. 2, pp. 264–285, Apr 1969.
[8] D. R. Radev, "Experiments in single and multi-document summarization using mead," First Document Understanding Conference, 2001.
[9] V. V. Jagadeesh J, Prasad Pingali, "Sentence extraction based single document summarization," Workshop on Document Summarization, 2005.
[10] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. de Frana Silva, and F. Freitas, "Assessing sentence scoring techniques for extractive text summarization," Elsevier Expert Systems with Applications, May 2013.
[11] S.A.Babar and P. D.Patil, "Improving performance of text summarization," in International Conference on Information and Communication Technologies, vol. 46, 2014.
[12] E. Padma Lahari, D. S. Kumar, and S. S. Prasad, "Automatic text summarization with statistical and linguistic features using successive thresholds," in IEEE Conference on Advanced Communication Control and Computing Technologies, 2014, pp. 1519–1524.
[13] R. Kurmi and P. Jain, "Text summarization using enhanced mmr technique," in International Conference on Computer Communication and Informatics, Jan 2014.
[14] M. Jafari, A. S. Shahabi, Y. Q. Jing Wang, X. Tao, and M. Gheisari, "Automatic text summarization using fuzzy inference," in 22nd International Conference on Automation and Computing, Sep 2016.
[15] Y. Liu, X. Wang, J. Zhang, and H. Hu, "Personalized pagerank based multi-document summarization," in IEEE Workshop on Semantic Computing and Systems, 2017, pp. 169–173.
[16] J. Yadav and Y. K. Meena, "Use of fuzzy logic and wordnet for improving performance of extractive automatic text summarization," in Intl.Conference on Advances in Computing, Communications and Informatics, 2016, pp. 2071–2077.
[17] P. Gustavsson and A. Jonsson, "Text summarization using random indexing and pagerank."
[18] K. S. Thakkar, R. Dharaskar, and M. Chandak, "Graph-based algorithms for text summarization," in Third International Conference on Emerging Trends in Engineering and Technology, 2010, pp. 516–519.
[19] N. Rahman and B. Borah, "A survey on existing extractive techniques for query-based text summarization," in International Symposium on Advanced Computing and Communication, 2015.
[20] C. R. Chowdary and P. S. Kumar, "Update summarizer using mmr approach."
[21] K. Ganesan, C. Zhai, and J. Han, "Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions," in Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010, pp. 340–348.
[22] J. Steinberger and K. Jezek, "Evaluation measures for text summarization," Computing and Informatics, vol. 28, pp. 1001–1026, Mar 2009.

