RESEARCH ARTICLE OPEN ACCESS

MULTI DOCUMENT SUMMARIZATION BY USING GRAPH BASED TEXT MINING TECHNIQUES

*1Mr. Venkatasubramaniyan T, Msc, M.Phil, *2 Mr. Sivabalan M., MCA, M.Phil,

*1Research Scholar, PG and Research Department of Computer Science, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India.

*2Guest Lecturer, PG and Research Department of Computer Science, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India

--***--

Abstract: The World Wide Web has become one of the largest information and knowledge repositories in the world. In spite of its easy access, it is virtually impossible for any user to browse or read a large number of such individual documents available online. Text summarization fulfils such information-seeking goals by providing a method for the user to quickly view the highlights or relevant portions of document collection. With tons of information uploaded on the web on a daily basis, the task of summarizing becomes a necessity. Also, locating and browsing information quickly from a collection of documents within a short span of time becomes possible with the help of summarization. This has led to large-scale research efforts in text summarization. The issues discussed above necessitate the need for an automated summarization system. The summarization techniques are broadly categorized into two schemes, extraction and abstraction. Extraction involves picking up the most important sentences from a document using statistical approaches. Abstraction, on the other hand, involves the reformulation of content depending upon the type of summary. This technique involves more adoptable linguistic processing tools. Though abstraction leads to better summaries, extraction is the preferred approach and is widely adopted by the research community. The process of text summarization using either a single document or multiple documents is quite tricky and challenging, with multi-document summarization facing additional challenges. As this thesis focuses on multi-document summarization, the first task is to cluster the documents based on their contents. To measure the similarity among the documents, several choices are available like cosine, dice, and Jaccard**.**

Keywords: Text Mining, Summarization Techniques, Clustering data Analysis, Similarity Measuring.

INTRODUCTION

Automated text summarization has drawn a lot of interest among the Natural Language Processing and Information Retrieval communities in recent years. The initial interest for automated text summarization started during the late 1960s in American research libraries, where a large number of scientific papers and books were to be digitally stored and made searchable. Before the invention of personal computers and the emergence of the World Wide Web (WWW) as a global digital library, locating text materials of relevance was a strenuous task. After the advent of WWW the form and function has been altered, where in people, academicians, researchers or lay end users get huge benefits by browsing the contents online. Though this has reduced the burden of information gathering, the task of acquiring the relevant information in a concise manner is still a challenge. Text summarization is the solution to address this issue. Summarization is a technique in which a computer automatically creates an abstract or summary of one or more documents. Automated text summarization is the process of automatically constructing summaries for a text depending on the user‘s needs. A summary is a precise representation of information depending on the specified target compression ratio. Systems summarizing single documents are called single document summarization systems, while systems which perform the same task with multiple related sets of documents are called multi-document summarization systems.

Document summarization, in general, is a difficult task by itself with multi-document summarization facing additional difficulties compared to single document aspects, like removal of redundancy among the document sets, handling large number of documents, time stamping etc. Thus, multi-document summarization becomes not only an issue to be focused on at the research level, but is quite a challenging task. There are several ways by which the process of summarization can be defined.

RELATED WORK

two techniques for both single and multi-document text summarization. One is adding a new feature Sim With First (Similarity With First Sentence) with MEAD (Combination of Centroid, Position, and Length Features) called CPSL and another is the combination of LEAD and CPSL called LESM. Finally simulate and compare the results of new techniques with conventional ones called MEAD with respect to some evaluation techniques. Simulation results demonstrate that CPSL shows better performance for short summarization than MEAD and for remaining cases it is almost similar to MEAD. Simulation results demonstrate that LESM also shows better performance for short summarization than MEAD but for remaining cases it does not show better performance than MEAD.[1] Md. Mohsin Ali, Monotosh Kumar Ghosh, an approach from the literature based on atomic events is compared to a novel approach based on generic relation extraction (GRE), which aims to build systems for relation identification and characterization that can be transferred across domains and tasks without modification of model parameters. The various representations are substituted in the interpretation phase of a multi-document summarization task and used as the basis for extracting sentences to be placed in the summary. System summaries are compared by calculating term overlap with reference summaries created by human analysts [2]. Ben Hachey, The summarizer does not use any expensive linguistic data. This Summarizer uses Vector Space Model for finding similar sentences to the query and Sum Focus to find word frequency; they achieved high Recall and Precision scores. The accuracy achieved using the proposed method is comparable to the best systems presented in recent academic competitions i.e., TAC (Text Analysis Conference). Text summarization is a data reduction process. As summary is concise, accurate and explicit, it has great significance. Sentence similarity means then calculate how much they are similar. Sentence similarity is calculated by most widely used Vector Space Model (VSM)[3], : A. P. Siva kumar,

The content-based approach has its roots in the research field of information retrieval which has been studied since the late fifties. In an information retrieval system a user enters a request for information and the system responds by identifying information sources that are relevant to the query. Many techniques that are incorporated in an information retrieval system can also be employed by a content-based filtering system[4] Vasudevan, V.Sharmila. This report provides a description of the methods applied in Center for Intelligence Science and Technology (CIST) system participating ACL MultiLing 2013. Summarization is based on sentence extraction. Hierarchical Latent Dirichlet Allocation (hLDA) topic model is adopted for multilingual multi-document modeling. Various features are combined to evaluate and extract candidate summary sentences. Sentences are clustered into sub-topics in a hierarchical tree. A sub-topic is more important if it contains more sentences[5] Lei Li, Wei Heng, Jia Yu, Yu Liu, Shuhong Wan. The Pattern-based Topic Model (PBTM) and Structural Pattern-based Topic Model (StPBTM). The main distinctive features of the proposed models include, (1) user information needs are generated in terms of multiple topics; (2) document relevance ranking is determined based on topic distribution and topic related semantic patterns; (3) patterns are organized structurally based on the patterns‘ statistical and taxonomic features for representing user interests for each topic. (4) Significant matched patterns and maximum matched patterns are proposed based on the patterns‘ statistical and taxonomic features to enhance the pattern representations and document ranking[6] Yang Gao, Yue Xu, and Yuefeng Li.

PERVIOUS IMPLEMENTATION

There are several parameters by which similarity can be evaluated. The first category of similarity evaluation is based on the document size and structure. The length of the document, the number of paragraphs, number of sentences, average number of characters per word, average number of words per sentence etc. The second category is based on ―style‖, whether the contents have been written in the first person conversational style or in the third person and so on. Thirdly, similarity can be based on the set of words used in the document. For example the original text of the novel ―A Tale of two cities‖ written by Charles Dickens may contain 20,000 distinct words, whereas the same novel rewritten for seventh standard students may contain only a set of 1000 words. The fourth category of similarity is ―content similarity‖ which reflects to what extent the contents of the two documents are alike. This category is adopted throughout this thesis wherever similarity is talked of hereafter.

$$Cosine(ti,tj) = \sum_{h=1}^{k} (t_{ih} * t_{jh}) * (idf_h) \Big/ \sqrt{\sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2}$$

In these formulas it is assumed that the similarity is being evaluated between two vectors *ti ={ ti1,…….. tik}* and *tj ={ tj1,…….. tjk*}, and the vector entries usually are assumed to be nonnegative numeric values. IDF denotes the Inverse Document Frequency of the document, which is discussed further in the next section.

$$Dice(ti,tj) = 2 \sum_{h=1}^{k} (t_{ih} t_{jh}) * (idf_h) \Big/ \sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2$$

A collection of _N‘ documents can thus be viewed as a collection of vectors, leading to the natural view of a collection as a term-document matrix; this is an M× N matrix whose rows represent the M terms (dimensions) of the N columns, each of which corresponds to a document.

$$Jaccard(ti,tj) = \sum_{h=1}^{k} t_{ih} t_{jh} * (idf_h) \Big/ \sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2 - \sum_{h=1}^{k} t_{ih} t_{jh} * (idf_h)$$

concluded that the maximum term resolving power of significant words lies somewhere in the middle between The upper and lower cutoff terms.

PROPOSED METHODOLOGY:

To find the enhancements to existing graph-based methods for summarizing single documents and multi-document clusters. The objective of automated text summarization is to condense the given text to its essential contents, based upon the user‘s choice of brevity. The summarization techniques are broadly categorized into two schemes, extraction and abstraction. Extraction involves picking up the most important sentences from a document using statistical approaches. Abstraction, on the other hand, involves the reformulation of content depending upon the type of summary. This technique involves more adoptable linguistic processing tools. Though abstraction leads to better summaries, extraction is the preferred approach and is widely adopted by the research community. The process of text summarization using either a single document or multiple documents is quite tricky and challenging, with multi-document summarization facing additional challenges. As this thesis focuses on multi-document summarization, the first task is to cluster the documents based on their contents.

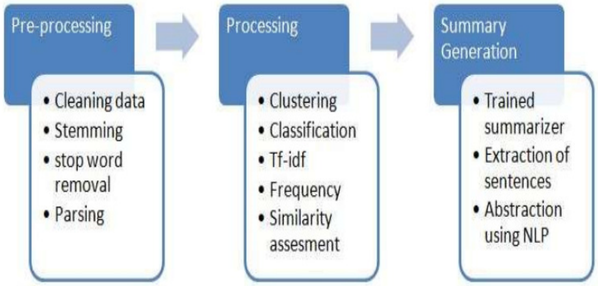
PHASES OF DOCUMENT SUMMARIZATION

There are basic three phases of document summarization as follows:

1) Pre-processing,

2)Processing,

3) Summary Generation.



Phases of Document Summarization

Pre-processing

Pre-processing is defined here as cleaning the data. For cleaning unwanted characters, symbols, extra spaces, hyperlinks etc are removed. The stop words like ‘a’, ‘the’, ’and’. Stemming is done where the words are reduced to their word root for example ‘playing’ would be reduced to ‘play’. Moreover the parsing of the above data is done where the words are bifurcated into nouns, adjectives, verbs etc. The pre-processing is done as per the requirement using one or combination of the above defined techniques.

Processing

This is the next phase of document summarization. After the Text data is cleaned and pre-processed, the processing techniques are applied in which the tf-idf scores, frequency of words are calculated. The data is grouped on the basis of similarity, dissimilarity so that the summary generation can be made efficiently. For this different clustering techniques are used.

Summary Generation

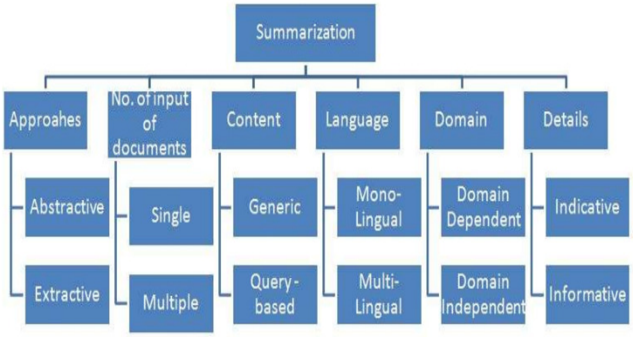
After the data is processed, the summary is to be generated based on the requirement. Extraction of sentences is done from the processed data and the summary is processed. The words to added to sentences, reducing the sentences based on score etc is done in this step to produce summary. The representation of the summary can be in the form of words, sentences, paragraphs, graphs etc. for the summary to be generated different summarizers are used.

Text Mining and Information Extraction

“Text mining” is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured text. Several techniques have been proposed for text mining including conceptual structure, association rule mining, episode rule mining, decision trees, and rule induction methods. In addition, Information Retrieval (IR) techniques have widely used the “bag-of-words” model for tasks such as document matching, ranking, and clustering. The related task of information extraction aims to find specific data in natural-language text.

TYPES OF DOCUMENT SUMMARIZATION

Document summarization is mainly divided into following types:



Types of Document summarization

Based on types of Approaches

There are two type of summarization based on their appraoches.

1) **Abstractive method**

Abstractive method builds an internal semantic representation and then uses Natural Language Processing (NLP) to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

2) **Extractive method**

In extractive summarization selection of a subset of existing words, phrases or sentences in the original text to form a summary is done.

Based on Number of Input Documents

1)Single document. The summary of a data from a single document is generated giving the brief idea about the document.

2) Multi document. The summary from two or more documents is given based on similarity or dissimilarity of the content in the documents.

Based on Content

Generic. The generalised summary of the document is given based on the frequency and the importance of the sentence.

Query-based. The summary is generated based on the query from the user about a single topic, word etc.

Based on Language

Mono-lingual . The summary only for a single language can be generated and it may not give results for other languages.

Multi-lingual. The summarizer can produce summary for multiple languages .

Based on Domain

Domain Dependent. In this the summary to be produced is dependent on a specific domain and cannot be applied for different domains

Domain Independent. The summaries formed under this category does not depend on any specific domain and the result can be obtained for all the domains.

Based on Details

Indicative . In the indicative summaries the it mainly gives the main idea about the topic and does not give any other related information available in the document.

Informative. This type of summary gives the details related to the topic along with the important data hence giving more coverage.

DOCUMENT SUMMARIZATIONTECHNIQUES

Term frequency-Inverse document frequency.

This technique is used for extraction of word to form the summary.

Tf-Idf of each document is generated using below formula:

Tf-Idft,d= Tft,d∗Idft

Where Tf is term frequency of term t in document d and Idft=log(N/dft)

Where N is the number of documents in corpus and df is the document frequency of t.

1)Cluster based:

In this technique the sentences or words are clustered on the basis of similarity or dissimilarity and the data is extracted from each cluster to form the summary of that cluster. Different clustering techniques like k-means, hierarchical clustering etc are used

2)Graph theoretic :

In this technique the sentences or words are represented in the form of nodes of a unidirectional graph. This method gives the idea about the coverage about a specific sentence or word and also the most important once can be extracted on the basis of high cardinality.

3) Classification:

The data is classified into multiple classes as per the requirement. The data from the specific class can be extracted. Again this is a extraction technique for document summarization. Decision tree, rule based, support vector machine, k-nearest neighbour are various classification techniques used.

4) Neural network:

In neural networks, the information is processed in a similar way to te human brain. Initially it learns about all the features which must be present in the sentence, and then extracts them on the basis of it ,removes dissimilar features and merges the similar ones.

5) Ontology based:

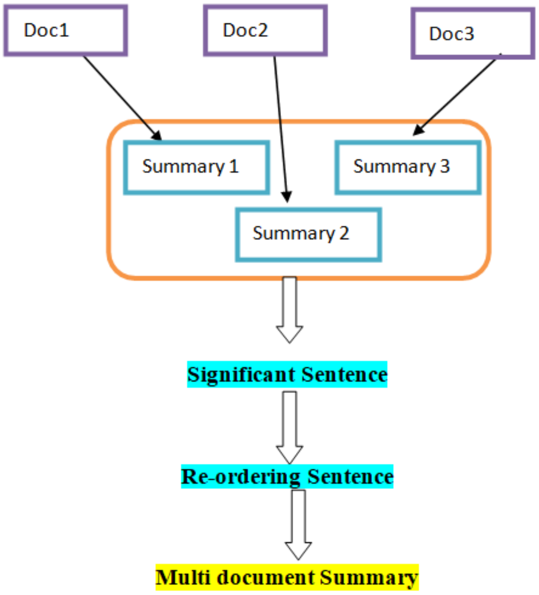
This is a abstraction based technique where the dictionary of words is used for the summarization process.

6) Fuzzy logic based:

In this summarization technique, instead of binary output the output is obtained in n-ary form.

8) Sematicbased:

This summarization technique uses the semantic analysis abstraction method to generate summaries which are more accurate like the human generated summaries.



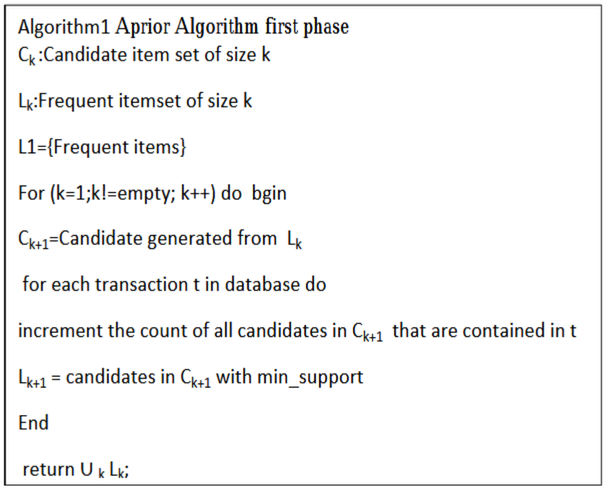
Multi Document Summary

Generation of Association Rule Using Apriori Algorithm for MDS

The goal of association rule is to explore the relationship between a group of items in a database. In 1993 Association rule was first introduced . There are two parts of an association rule, an antecedent (if) is an item appear in the data and a consequent (then) an item that appear in combination with the antecedent.

Generation of association rules can be done by analyzing the data for frequent if/then patterns and applying the criteria support and confidence to specify the most significant relationships. Support refers to how the item appears frequently in the database.

Confidence refers to how many the if/then statements have been appearing to be true. There are two modes in our proposed automatic MDS training mode and testing mode.

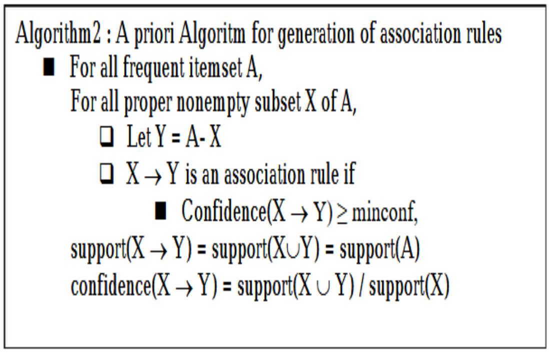


```
Algorithm1 Aprior Algorithm first phase
$C_k$ :Candidate item set of size k
$L_k$:Frequent itemset of size k
L1={Frequent items}
For (k=1;k!=empty; k++) do  bgin
$C_{k+1}$=Candidate generated from  $L_k$
 for each transaction t in database do
increment the count of all candidates in $C_{k+1}$  that are contained in t
$L_{k+1}$ = candidates in $C_{k+1}$ with min_support
End
 return $\cup_k L_k$;
```

1)Training mode:

Where the extracted features from (70) manually summarized English documents fed to the system to obtain the constructed model. As illustrated in figure (1) Apriori Algorithm applied to the fuzzified features. Apriori algorithm is an important algorithm for mining frequent itemsets for Boolean association rules. It was first introduced in 1994 by Agrawal and R.Srikant [20] and is the first algorithm that control the exponential increase of candidate itemset. The support is used for pruning candidate itemset [21]. Apriori algorithm is composed of two main phases, the creating frequent itemsets and discovering the association rules from the created frequent itemsets. At the first step, it finds the frequent itemsets. The frequent itemset is an itemset whose support is larger than some specified minimum support. The following algorithms illustrate the first phase of Aprior algorithm.

At the second phase of Apriori algorithm the association rules are chosen by applying the minimum confidence support. The following algorithm illustrates the generation of association rules.



```
Algorithm2 : A priori Algoritm for generation of association rules
■ For all frequent itemset A,
  For all proper nonempty subset X of A,
    ❑ Let Y = A- X
    ❑ $X \rightarrow Y$ is an association rule if
        ■ $\text{Confidence}(X \rightarrow Y) \geq \text{minconf}$,
  $\text{support}(X \rightarrow Y) = \text{support}(X \cup Y) = \text{support}(A)$
  $\text{confidence}(X \rightarrow Y) = \text{support}(X \cup Y) / \text{support}(X)$
```

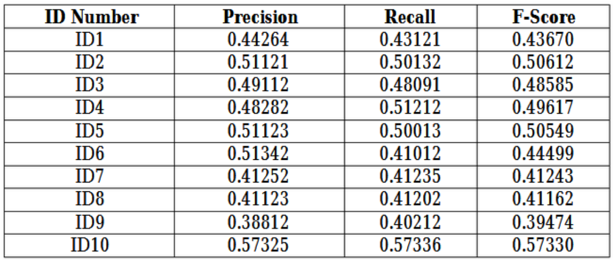
Where min conf is the minimum confidence in our proposed automatic text MDS we use min conf equal to 75%.All rules with confidence is less than 75%will be ignored. The final results of this phase will be rules which have the following form:

2) Testing Mode: in this mode only the documents that produced from the fuzzy logic is fed to the constructed rules that generated from the previous mode. There is a (30) documents in this mode.

PERFORMANCE ANALYSIS

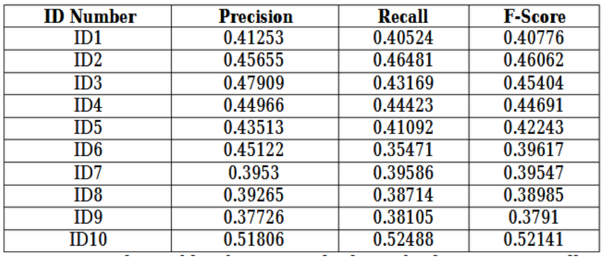
In our proposed system for multi-document text summarization. As illustrated in the previous section the dataset consists of 10 topics, each of the10 documents. Here we divided the dataset into two sets, the first set for training and a second set for testing. Each of them consists of (70)documents where (7) documents from each topic is selected as a train data and all the (100) documents as the test data. The results show a good performance of the proposed systems as appear in the table (1) which represents the output of ROUGE-1.

Table (1) The scores of ROUGE-1 produced by the proposed system



ID Number	Precision	Recall	F-Score
ID1	0.44264	0.43121	0.43670
ID2	0.51121	0.50132	0.50612
ID3	0.49112	0.48091	0.48585
ID4	0.48282	0.51212	0.49617
ID5	0.51123	0.50013	0.50549
ID6	0.51342	0.41012	0.44499
ID7	0.41252	0.41235	0.41243
ID8	0.41123	0.41202	0.41162
ID9	0.38812	0.40212	0.39474
ID10	0.57325	0.57336	0.57330

Table (2) the scores of ROUGE-1 for the system summary



ID Number	Precision	Recall	F-Score
ID1	0.41253	0.40524	0.40776
ID2	0.45655	0.46481	0.46062
ID3	0.47909	0.43169	0.45404
ID4	0.44966	0.44423	0.44691
ID5	0.43513	0.41092	0.42243
ID6	0.45122	0.35471	0.39617
ID7	0.3953	0.39586	0.39547
ID8	0.39265	0.38714	0.38985
ID9	0.37726	0.38105	0.3791
ID10	0.51806	0.52488	0.52141

We can see from table1 that our results for each of Precision, Recall and F-Score is better than the table2 which represent the result of peer summary for the TAC dataset. These results show the effect of the selected features and the effect of the rule representation method to get good results.

CONCLUSION

The extraction of information from a multi-document is very necessary. MDS is the choosing of important sentences from the original text with keeping the main ideas for that summarized documents. In this paper a new method for MDS is introduced which depends on linguistic and statistical features of the sentences. Apriori algorithm applied for extraction of association rules. Two important criteria play main role in getting good results, confidence and the number of training data. Where many confidences applied until reaching 75%, which gives balance between the number of generated rules and the importance of these rules. The number of training dataset is also very important such that, less efficient results obtained when number of trained documents less than(60).

FUTURE ENHANCEMENT

Generating a summary is a tricky and challenging task, especially for multi document cases. Only two aspects, namely, discounting and position weight have been considered for the study. The results obtained are not only promising but provide a good scope for further improvement using some additional features. Thus, summary generation techniques in this work do not take temporal information into account. If such a feature is considered in future, then the summary of each document could be generated and merging can be done based on a time sequence. Linguistic processing tools may be used to analyze the semantics of the documents to improve the quality of summaries.

BIBILIOGRAPHY:

[1] R. Kumar and D. Chandrakal" A survey on text summarization using optimization algorithm," ELK Asia Pacific Journals vol. 2, no. 1, 2016.

[2] Y. K. Meena and D. Gopalani, "Evolutionary Algorithms for Extractive Automatic Text Summarization," Procedia Comput. Sci., vol. 48, no. Iccc, pp. 244–249, 2015.

[3] K. Duraiswamy and G. Padma Priya, "An Approach for Text Summarization Using Deep Learning Algorithm," J. Comput. Sci., vol. 10, no. 1, pp. 1–9, 2014.

[4] R. He, J. Tang, P. Gong, Q. Hu, and B. Wang, "Multi-document summarization via group sparse learning," Inf. Sci. (Ny)., vol. 349–350, pp. 12–24, 2016.

[5] A. John and D. M. Wilscy, "Random Forest Classifier Based Multi-Document Summarization System," IEEE Recent Adv. Intell. Comput. Syst. RANDOM, pp. 31–36, 2013.

[6] S.A. Babar and P. D. Patil,"Improving Performance of Text Summarization ,"Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 354–363, 2015.

[7] R. M. Aliguliyev, "Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization," Comput. Intell., vol. 26, no. 4, pp. 420–448, 2010.

[8] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm diversity hybrid model for text summarization," Inf. Process. Manag., vol. 46, no. 5, pp. 571–588, 2010.

[9] R. Aliguliyev, "A new similarity measure and mathematical model for text summarization," no. January 2015.

[10] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, "Text summarization features selection method using pseudo genetic-based model," Proc. - 2012 Int. Conf. Inf. Retr. Knowl. Manag. CAMP'12, pp. 193–197, 2012.

[11] Porter stemming algorithm:
http://www.tartarus.org/martin/PorterStemmer

[12] ANSAMMA JOHN, "Multi-Document Summarization System: Using Fuzzy Logic and Genetic Algorithm," Int. J. Adv. Res. Eng. Technol., vol. 7, no. 1, pp. 30 – 40 , 2016.

[13] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," Comput. Speech Lang., vol. 23, no. 1, pp. 126–144, 2009.

[14] C. N. Satoshi, S. Satoshi, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, "Sentence Extraction System Assembling Multiple Evidence," Proc. 2nd NTCIR Work., pp. 319–324, 2001