Multi-Document Summarization using Fuzzy Logic and Firefly Algorithm

Assist. Prof. Dr. Suhad Malallah
Dept. Computer Science
University of Technology
Baghdad,Iraq
suhad_malalla@yahoo.com

Zuhair Hussein Ali
Dept. Computer Science
Al- Mustansiriya university
Baghdad,Iraq
zuhair72h@yahoo.com

Abstract

Due to the huge amount of documents in the internet made it difficult to get useful information. Automatic text summarization is a good solution for such problem, which is based on a selection of important sentences from one or multi-document without losing the main ideas of the original text. In this paper a new method was proposed which depend upon selection of seven features for every sentence in the documents. These features fed into the fuzzy logic system to give scores to these sentences. Firefly algorithm applied as association rule mining to minimize the set of rules generated by the fuzzy logic system and finally redundancy reduce performed to remove redundant sentences. The proposed model is performed using dataset supplied by the Text Analysis Conference (TAC-2011) for English documents. The results were measured by using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The obtained results support the effectiveness of the proposed model.

Keywords: Firefly, Fuzzy logic, confidence, Frequent itemset, Association Rule Mining.

تلخيص النصوص المتعددة باستخدام المنطق الضبابي وخوارزميه ذبابة النار

أ.م.د سهاد مال الله م.زهير حسين علي

الخلاصة

بسبب كثرة المعلومات الموجودة في الانترنيت ولاهميتها أصبح موضوع تلخيص النصوص من المواضيع المهمة حيث يعتمد على أختيار الجمل المهمه من النصوص متعدده مع المحافظة على الفكره الاساسيه للنصوص الملخصة. في هذا البحث تم أستخلاص النصوص بلاعتماد على أستخراج سبعة خصائص لكل جمله من جمل النصوص الملخصة. يتم تقديم هذة الخصائص الى المنطق الضبابي لاعطاءها تصنيفات بعدها تم أستخدام خوارزميه ذبابه النار لاستخراج القوانين المهمة الخاصه تصنيف ROUGE لاختبار النظام وأحتسبت النتائج باستخدام برنامج (TAC–2011)الجمل . تم أختيار قاعده بيانات

I. Introduction

Due to the rapid growth of information on the World Wide Web a huge amount of documents has been produced. This burst of documents made it hard to get useful information from them [1]. A lot of pertinent and motivating information is discarded by the user due to the huge amount of documents. To deal with such problem of information overload, an Automatic Text Summarization (ATS) has been used as a solution [2]. The main goal of the ATS is to create a summary from multi document or single document that express a complete meaning of the document with the least number of words. The main purpose of text summarization is to help users in discovering information from source documents by gathering the indispensable information and giving its shortened form. In such manner, text summarization can be considered as an arbiter between information included in many documents and users [3]. Text summarization methods are categorized as an abstractive summarization and extractive summarization. Abstractive summarizatio***n depends on Natural*** Language Processing (NLP) strategies for parsing, finding and creating content. Currently, NLP machinery is computationally cost-effective but it has less precision. Conversely, extractive summarization can be defined as the technique for verbatim extraction of the literary components like passages, sentences and so on from

the source content. Abstractive summarization is noticed to be complex and consumes more time as compared to extractive summarization [4]. The fundamental objective of extraction technique is the picking of suitable and pertinent sentence from the input documents. . A technique to acquire the suitable sentences is assigned a weight for each sentence which indicates the salience of a sentence for choosing to be included in the summary [5].

Depending on the quantity of documents to be summarized, the summarization can be classified to a single document summarization (SDS) and multi-document summarization (MDS) . Only one document can be condensed into a shorter one in sing SDS, whereas many documents is condensed into one summary in MDS [6]. Also summarization can be classified as either generic based or query based, on generic based a summary is created for all documents independent on specific subjects, whereas in query based a summary is created depending on user who want specific information [7]. In this paper, we have proposed a new method for automatic MDS based on computing seven different features for each sentence, then a fuzzy logic used to assign scores for each of these features. Firefly algorithms applied to get most confidence rules which identify the most important sentences to be included in the summary. The proposed method applied to the Text Analysis Conference (TAC-2011) data set for English documents only.

II. Related Works

Identifying the most important sentences to be included in the summary is the main goal of the ATS. Understanding the main content of the document and producing a short summary, consider a hard natural language processing problem. There are many methods to perform such process. Most of these methods focused on extraction summarization [8]. In this section we investigate some of these methods. In [9] a set of desired features for each sentence be extracted. These features include whether it's first or last sentence in the paragraph, the similarity between the sentence the title, the length of the sentence, the number of thematic words and finally the number of emphasizing words, these extracted features used as input to the fuzzy inference system where a Bell function used to give a membership for each feature. The output of the membership is one of three {important, average, unimportant} which identify the importance of each rule. There are too many rules generated by the fuzzy inference system so a combination of genetic algorithm and genetic programming performed to optimized the number of rules. In [10] suggested fuzzy-swarm hybrid diversity A method that merges three models depend on swarm, diversity and fuzzy-swarm. The diversity-based model forms, sentence groups arranged in a binary tree according to their scores. It then executes Maximal Marginal Importance (MMI) to choose

the sentences for embedding in the summary. The model based on binary Particle Swarm Optimization (PSO) is applied to optimize the weight related to every feature of the objective function. The location of the particle is a string of bits, where one means that the related feature is chosen, otherwise it has a zero. On obtaining the weights, the score is given for every sentence and the sentences that have higher score are chosen to be incorporated in the candidate summary. In the model based on fuzzy logic and swarm, the sentence score is calculated by a system of inference in the fuzzy algorithm, starting with the weights creates with PSO. The sentences are sorted depending on the score and the summary is acquired. In [11] a new method for generic multi document summarization based on optimization model was proposed. The approach creates a summary by extracting the most important sentences from the documents by computing sentence to sentence relation, summary to document collection and sentence to document collection. An improved differential evolution algorithm was used to solve the optimization problem that reduces the redundancy in the created summary and get the most salient sentences to be included in the summary. In paper [12] a set of features extracted for each sentence. These features used as input to the combination model which consist of Cellular Learning Automata (CLA), PSO and fuzzy logic. The CLA was used to calculate the similarity between sentences to reduce the redundancy. While the PSO was used to set a weight for each feature. The fuzzy logic used to give scores to the sentences, then the sentences were arranged in descending order, the sentence with higher score was selected to be included in the created summary.

III. Problem Statement and Formulation

To produce a good summary for any MDS system two issues must be considered. These issues are

1-Relevance: can be defined as the goodness of information included in the created summary. A summary considered as relevant if it includes many information relevant to the main topic of the documents.

2-Redundancy: The generated summary must include less redundant information. Since want to cover most of the relevant information in the documents thus by reducing redundancy can cover most of the main topics in the original documents.

Then, to formulate the problem suppose have a corpus which consist of many clusters, each cluster contain a set of documents called D with the same topic. The set D can be defined as D= {d1, d2,…, dn} where n is the number of distinct document in D. Each D can be represented by a set of sentences called Si, i.e D= {Si | 1<=i<=M} M represents the total number of sentences in

the set D. Our goal is to find a subset of set D called A i.e. A □ D that satisfies both relevant content and reduce redundancy.

IV. The Proposed Method

In this paper a new method for MDS was proposed which depend on computing seven features for each sentence, then these features are introduced to fuzzy logic to give scores to the sentences. The outputs of fuzzy logic are a lot of sets of rules. A firefly algorithms used to as Association Rules Mining (ARM) by getting the most confidence rules finally reduce redundancy performed. There are five main steps in our MDS systems.

1-Preposessing

2-Feature Extraction

3-Fuzzy logic, scoring and generations of rules

4- Association Rule mining using firefly

5- Reduce redundancy

1. Preprocessing

There are four steps for preparing the data these steps are

A- Sentence segmentation: is an important approach in text processing such as machine translation, information extraction, text summarization and syntactic parsing. Sentence segmentation is done by splitting the source text into sentences according to the period "." between sentences.

B-Tokenization: Is the process of splitting sentence into word.

C-Stop word removal: Is the third step in preprocessing steps, where words which don't give the necessary information for identifying significant meaning of the document content and appear frequently are removed. There are a variety of methods used for specifying of such stop words list. Presently, a number of English stop word list is usually used to help text summarization process. Regardless of its repetition and having no effect to the meaning, these words contribute an important percentage of the overall documents. Removing of such words can increase the efficiency and effectiveness of information retrieval process. The document size can be minimized without affecting its meaning, less memory and time consumed.

D-Word Stemming: is the process of producing root of the word, in This paper word stemming is performed by removing suffixes proposed by Porter's stemming algorithm [13].

2. Features Extraction

It's an important part of Text summary, which include compute of features score for every sentence. These features include sentence position, sentence length, numerical data, Thematic word, title word, proper noun and centroid value

A-Sentence positions: Where the higher score will give to the first sentence, and the score decreases according to the sentence position in the document. This feature can be computed according to equation (1).

$$F1 = \frac{N-P+1}{N} \qquad (1)$$

Where N represents the total number of sentences in the document

P current position of the sentence

B-Sentence length: This feature is helpful for deleting the short sentences, the short sentences are the new article which includes writer name, datelines which is not necessary to be included in the summary. This feature is calculated by dividing the sentence length by the length of longest sentence in the document as in equation (2).

$$F2 = \frac{sentence\ length}{longest\ sentence\ length\ in\ the\ document} \qquad \ldots. \quad (2)$$

C-Numerical data: The numerical information appearing in the document has important information and it would more probably incorporate into the summary [13]. This feature is calculated according to the following equation(3)

$$\mathbf{F3} = \frac{\textbf{Number of numerical data in the sentence}}{\textbf{Total sentence length}} \qquad \ldots (3)$$

D-Thematic Words: is the term that appears most frequently in the document. This feature can be calculated by computing the frequencies of all terms in the document, then top (n) terms with the highest frequency is selected, in this research, we used top (5). This feature is calculated by dividing the number of thematic words in the sentence by the maximum thematic words in the document as explained in equation (4).

$$\mathbf{F4} = \frac{\textbf{Number of thematic words in the sentence}}{\textbf{Max number of Thematic}} \quad \ldots (4)$$

E-Title word: This feature is important when summarizing the document. The score was calculated by dividing the number of title words in the sentence and total number of words in the title as follows (5).

$$\mathbf{F5} = \frac{\textbf{Number of title word in the sentence}}{\textbf{NO. Of word in the title}} \quad \ldots (5)$$

F-Proper noun: A higher score is given to the sentence if it contains the maximum number of proper nouns [14]. And it's calculated As in (6)

$$\mathbf{F6} = \frac{\textbf{Number of proper nouns in the sentence}}{\textbf{Snetence length}} \quad \ldots (6)$$

G-Centroid value: Is a feature used to specify salient sentences in the multiple documents [15]. This feature can be calculated as follows

$$F7 = \sum_{i=1}^{n} C\mathbf{wi} \qquad (7)$$

$$\mathbf{C_{wi}=TF\ *IDF} \qquad (8)$$

$$\mathbf{DF=log[}\frac{Total\ NO.\ of\ documents}{NO.\ of\ documents\ contaning\ the\ given\ word}\mathbf{]} \quad \ldots \quad (9)$$

Where Cw is the centroid of the words

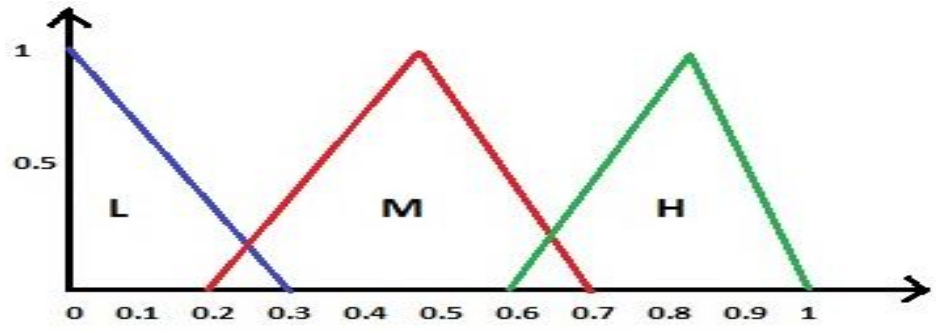
TF is the term frequency which represents the frequency of a given term in the document. IDF is the inverse term frequency computed by division of the total number of documents and the number of documents including the given

3. Fuzzy Logic Scoring

The obtained features from the previous section are introduced as inputs to the fuzzy logic. The fuzzy logic system uses the triangle membership function to partition the score of each feature into one of three values that are high, medium and low [16]. The triangle membership function is defined in equation (10)

$$f(X: a, b, c) = \begin{cases} 0 \text{ if } x < a \text{ or } x > c \\ \frac{x-a}{b-a} \text{ if } a \leq x \leq b \\ \frac{a-x}{b-a} \text{ if } b \leq x \leq c \end{cases} \qquad (10)$$

At the end of this step many rules will be generated. Figure(1) shows how triangle membership function assigns a value to the input variable sentence length. Where an input variable (X) assigned a value according to eq. (10) then as shown in figure (1) it take on of the values (low, medium or high)



Fig(1) triangle membership function.

4. Firefly Algorithm

In 2008 Xin she Yanq proposed a firefly algorithm (FA) which identified by their flashing light. The main goal of flashing light is to attract other fireflies. FA formulated by assuming the following

A-All fireflies can consider as a unisex.

B-The brighter firefly attracts the less bright firefly.

C-The move of firefly will be randomly if the given firefly is brighter than all others fireflies[17]. Figure (2) shows the main FA operations

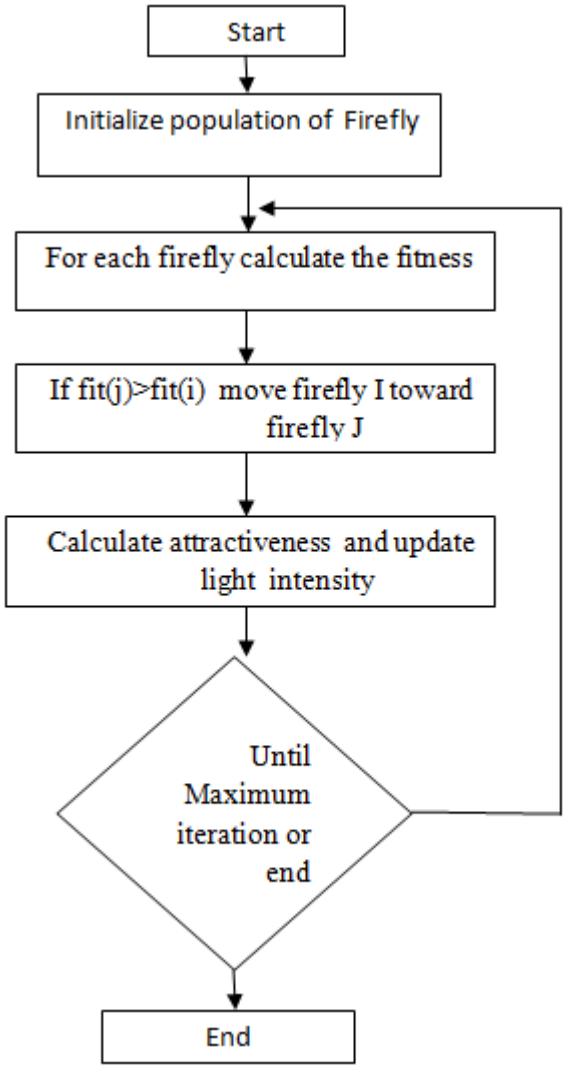


Figure (2) Firefly operation

5. Association Rule Mining using Firefly Algorithm

As we have shown there are too many rules generated by the fuzzy logic so our goal is to optimized these rules and find the most confidence rules. ARM applied to reach this goal, ARM is the most important technique in data mining that can be used to extract necessary information from large databases. The main goal of ARM is to find the relationship between frequent itemsets, and it has the form A →B in which A and B represent the antecedent and consequent part of the rule respectively [18]. There are two important parameters in association rule which called support and confidence.

Support: the support can be defined as the number of transactions in the database that include both A and B

$$support(A \rightarrow B) = \frac{A \cup B}{|D|} \quad (11)$$ where |D| represent the number of records.

Confidence: represent the strength of the rule

$$confidence(A \rightarrow B) = \frac{\sup(A \cup B)}{\sup(A)} \qquad (12)$$

There are many algorithms that can be used for ARM such as Aproiri and Fp-growth algorithm, but most of these algorithms required to specify minimum support and minimum confidence which set by the user [19]. In recent years there are many algorithms that used in rule mining such as genetic algorithm, evolutionary algorithm and swarm algorithm. These algorithms can be used without the need to define a minimum support and minimum confidence also can get the required rules in one stage. In our proposed method we used firefly algorithm that described in [20] with some modification to the fitness function.

We can explain the two basic steps of the algorithms as follows

Each rule represented by one firefly this technique used to compute association rule that has high frequency.

Encode and compute the fitness values for every firefly. The fitness value for our proposed method is as follows.

Fitfit(i)=[sup(A ∪ B)/sup(A)].[sup(A ∪B)/sup(B)].[1-sup(A∪B)B)/ |D|]+max(cos_sim (sent(i),Ref_summary)) (13)

Where

A- $\left[\frac{\sup(A \cup B)}{\sup(A)}\right]$ Shows how the antecedent part creates the probability of the rule.

B- $\left[\frac{\sup(A \cup B)}{\sup(B)}\right]$ Shows how the consequent part creates the probability of the rule.

C-$\left[1 - \frac{\sup(A \cup B)}{|D|}\right]$Shows how the whole data, create the probability of the rule

max(cos_sim (sent(i),Ref_summary))

Shows the cosine similarity between the sentence and the reference summary.

This the first three terms acts to compute the rule mining in the database while the fourth term acts as a measurement of how the sentence belong to the reference summary.

V. Reduce Redundancy

One of the most problems in the MDS is the redundancy, because there are many documents on the same topic. Some sentences may be repeated. Therefore, reduce redundancy is very important to allow the generated summary express the main ideas of the documents that we want to be summarized. The equation used to compute redundancy between two sentences as in [21].

$$R = \frac{2*Ms}{M1+M2} \qquad (14)$$

Where R number of redundancy between two sentences
Ms number of similar words between two sentences
M1 number of words in sentence one
M2 number of words in sentence two.

Where every sentence selected to be added to the candidate summary will be compared to all previous selected sentences in the summary. If R is greater than the specified threshold, then the selected sentence will be ignored. There are two modes in our proposed model training mode and testing mode. In the training features that extracted from 50 documents are used as trainers. In the testing mode the remaining 50 documents were used.

VI. Dataset and Evaluation metrics

The dataset used in our proposed method is the Text Analysis Conference (TAC-2011) which consist of seven languages (English, Arabic, Greek, Czech, French, Hindi, Hebrew). There are10 topics, each of 10 documents for each language [22]. Our proposed method deal with English language only.

Evaluation ROUGE [23] was used to evaluate the proposed system the output of rouge package is three numbers which represent, precision (P), Recall (R) and F−score. They formulated as follows.

$$\mathbf{P} = \frac{\mathbf{S\epsilon human\ summary} \cap \mathbf{S\epsilon system\ summary}}{\mathbf{S\epsilon system\ summary}} \quad \ldots \quad (\mathbf{15})$$

$$\boldsymbol{R} = \frac{\boldsymbol{S\epsilon human\ summary} \cap \boldsymbol{S\epsilon system\ summary}}{\boldsymbol{S\epsilon human\ summary}} \quad \ldots \ (\mathbf{16})$$

$$\mathbf{F} = \frac{(\mathbf{1} + \boldsymbol{\beta}^{\mathbf{2}})\mathbf{R} * \mathbf{P}}{\mathbf{R} + \boldsymbol{\beta}^{\mathbf{2}}\,\mathbf{P}} \qquad (\mathbf{17})$$

Where

$$\boldsymbol{\beta} = \frac{\mathbf{P}}{\mathbf{R}}$$

VII. Experimental Results

The TAC-20111 dataset was used in our proposed system for MDS for English document only. Created summary evaluated using the measures F-score, Recall and precision. The following figure shows the results of our proposed method and the peer summary of the system

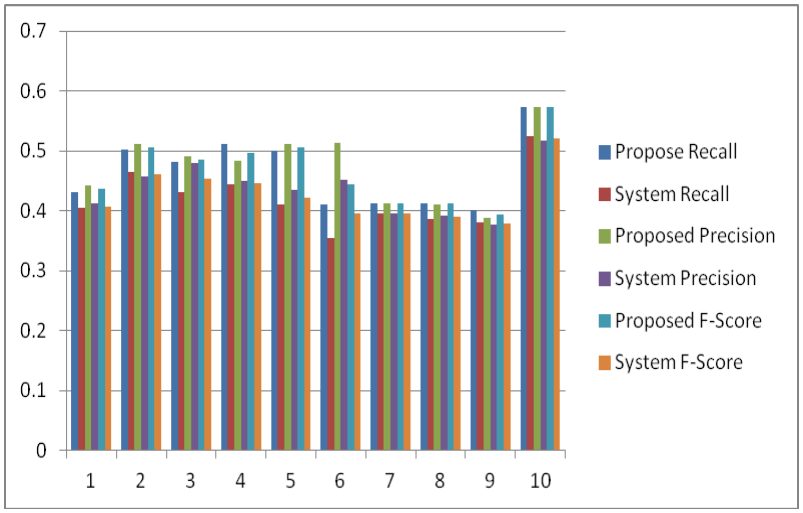


Figure (3) Evaluation Measures.

Figure (3) shows that our proposed method gives better results than system results. Where X-axis represents the document set id1,id2… id10 for each of the topics in the data set and y-axis represent the values of Recall, Precision and F-score of our proposed system compared to the system values.

VIII. Conclusion

In our proposed method for MDS we have extracted seven features from the dataset TAC-2011 for every sentence. The seven features fed to the fuzzy logic to score the sentences and generate a set of rules then a firefly algorithm applied as an ARM to optimize the set of rules. Appling firefly algorithm for ARM avoided the system the problem of setting support and confidence parameters by the user. The incorporation of fuzzy logic and firefly algorithm showed good results for MDS.

IX. References

[1] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, "Expert Systems with Applications MCMR : Maximum coverage and minimum redundant text summarization model," Expert Syst. Appl., vol. 38, no. 12, pp. 14514–14522, 2011.

[2] C. Li, Y. Liu, F. Liu, L. Zhao, F. Weng, and P. Alto, "Improving Multi-documents Summarization by Sentence Compression based on Expanded Constituent Parse Trees," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ,pp. 691–701, 2014.

[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] K. Patil and P. Brazdil, "Sumgraph: text summarization using centrality in the pathfinder network," IADIS Int. J. Comput. Sci. Inf. Syst., vol. 2, no. 1, pp. 18–32, 2007.

[5] F. Kyoomarsi, H. Khosravi, E. Eslami, and P. Khosravyan, "Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization," International Journal of Hybrid Information Technology,vol. 2, no. 2, pp. 105–116, 2009.

[6] D. M. Zajic, B. J. Dorr, and J. Lin, "Single-document and multi-document summarization techniques for email threads using sentence compression," Information Processing & Management vol. 44, pp. 1600–1610, 2008.

[7] H. Dave, "Multiple Text Document Summarization System using Hybrid Summarization Technique," International Conference on Next Generation Computing Technologies, pp. 4–5, 2015.

[8]B. . Samei, M. . Eshtiagh, F. . Keshtkar, and S. . Hashemi, "Multi-document summarization using graph-based iterative ranking algorithms and information theoretical distortion measures," Proc. 27th Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2014, pp. 214–218, 2014.

[9] . K. Kiani and M. R. Akbarzadeh-T, " Automatic Text Summarization Using Hybrid Fuzzy GA-GP," IEEE Int. Conf. Fuzzy Syst., pp. 977–983, 2006.

[10] 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.

[11] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "Expert Systems with Applications Multiple documents summarization based on evolutionary

optimization algorithm," Expert Syst. Appl., vol. 40, no. 5, pp. 1675–1689, 2013.

[12] R. A. Ghalehtaki, "A combinational method of fuzzy , particle swarm optimization and cellular learning automata for text summarization,",IEEE conference, vol. 15, no. 1, 2014.

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

[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.

[15] 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.

[16]] 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.

[17] Yang, X.S.: Nature-Inspired Metaheuristic Algorithms. Luniver Press 2008.

[18] E. Duneja, "A Survey on Frequent Itemset Mining with Association," International Journal of Computer Applications vol. 46, no . 23, pp. 18–24, 2012.

[19] J. Han and M. Kamber, Data Mining: Concepts and Techniques, vol. 54, no.1 Second Edition. 2006.

[20] P. Sehrawat, "Association Rule Mining Using Firefly Algorithm," IJLTET, vol. 3, no. 2, pp. 263–270, 2013.

[21] D. R. Radev, H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," Information Processing and Management , vol. 40, pp. 919–938, 2004.

[22] G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma, "TAC 2011 MultiLing Pilot Overview," no. November, 2011.

[23] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proc. Work. text Summ. branches out (WAS 2004), no. 1, pp. 25–26,2004.

.