Journal of Informatics and Mathematical Sciences
Vol. 10, No. 4, pp. 703–714, 2018
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications
DOI: 10.26713/jims.v10i4.891

http://www.rgnpublications.com

Research Article

A Constraints Driven PSO Based Approach for Text Summarization

Shrabanti Mandal[1], Girish Kumar Singh[1,*] and Anita Pal[2]

[1] *Department of Computer Science & Applications, Dr. Harisingh Gour Central University, Sagar 470003, India*
[2] *Department of Mathematics, National Institute of Technology, Durgapur, West Bengal, India*
*__Corresponding author:__shrabmandal@gmail.com

Abstract. In the present scenario we are living in a digital media and virtual world. To conveniently communicate in digital world electronic data have to gradually increase. So it is a serious challenge to manage the huge digital and electronic resources efficiently and accurately. One of the important solutions of the above problem is text summarization i.e. an application of text mining. Representing the gist of a text document is called summary. A rich summary always covers the maximum coverage, high level of diversity and with user defined size. This paper proposes an approach for summarizing the text documents by extractive way using Particle Swarm Optimization (PSO) that is known as population based stochastic optimization technique and it has many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The huge volume and dimensions of terms have managed by the concepts of term document matrix followed by K-Means clustering with PSO for acquiring optimal number of concepts clusters. Then apply constraint-driven concept for selecting the best one. These key concepts were used to identify the significant gist in documents for text summarization.

Keywords. Text summarization; Particle swarm optimization; K-means; Fitness function; Cosine measure and ROUGE

MSC. 68P20 (Information storage and retrieval)

Received: April 13, 2018 **Accepted:** October 6, 2018

1. Introduction

Now-a-days World Wide Web and digital library are flourished by creating and storing electronic information and in parallel researchers take interest in text mining [3, 26]. As the amount of electronic information increases on the Internet, it becomes difficult to extract the information quickly and most efficiently. The conventional Information Retrieval (IR) technologies have been used in extracting the information but they are insufficient to retrieve required information effectively. The volume of text information is gradually increasing, so documents summarization becomes one of the essential tasks for information retrieval. The summarized information represents the great volumes of information in a concise or compressed way. One of the best ways of searching information effectively and efficiently is automatic summarization. The text summarizer extracts the gist from the original documents and presents it in a condensed form to the user [14, 41].

Text summarization is a process to concise a huge amount of information into a small form by selecting or extracting the highly important sentences and discarding the redundant one. Abstractive and extractive are two types of summarization techniques are available. Abstractive methods are described as expressing the same concepts by generating the new sentences after reading and understanding the source document while extractive method selects or extracts the sentences according to their weight and then concatenates them into concise form. The linguistic quality control and content selection method in both summarization techniques are the common challenging task. Linguistic quality ensures the coherent of content and very significant in fluency. Content selection includes the selection of important content as well as removal of redundant information and keeping of high level of content coverage.

Text summarization may be applied on single document or on set of multiple documents [24, 42] . Mani and Maybury [21] proposed a text summarization method based on their importance of sentences according to user's needs. In this method a partial structure is generated from source document which cover almost same topic.

The search engines like Google, Yahoo!, AltaVista, and others do not unnecessarily browse the large volume of documents instead they produce the group of documents as per user's interest with a brief of each document to facilitate the process of finding required information efficiently [9, 32, 33, 41].

A text summarization model based on extractive content selection also proposed by Raj Kumar *et at*. [26]. Text summarization method can also be classified in two categories namely generic and query-focused [24, 35, 36]. A generic summary presents a sense of overall document's content while a query-focused summary focuses on the information that is relevant to the said queries. The goal of the generic summary is to fetch the core information which is central to the source documents and aim of query-focused summary is to build up a summary that can fulfill the need of query generator.

2. Related Works

Various abstractive based and extractive based techniques have been proposed for generic multi-document summarization. The heart of extractive method is to select highly weighted sentences.

Radev el al. has proposed one of the popular extractive methods MEAD [25]. MEAD is a centroid based method; it uses the centroid information of the clusters of sentences for selecting sentences. Latent Semantic Analysis (LSA) was proposed by Gong *et at*. [13]. Higher priorities sentences are chosen for summary by LSA. Some other examples of text summarization methods include the Non-negative Matrix Factorization (NMF) based topic specification [19, 40] and Conditional Random Fields (CRF) based summarization [32]. The CRF method evaluates importance of sentences of original documents and arranges them according to their importance [40] and NMF model was based on semantic and symmetric of sentences.

Weighted Feature Subset Non-negative Matrix Factorization (WFSNMF) is an unsupervised approach for clustering. WFSNMF method clusters the data by using the important features and then assigns a numeric value as weights to data points [39]. This model has been applied to document clustering, summarization, and visualization. Wang *et at*. [37] also proposed weighted consensus summarization method which work by combining the result of different summarization techniques for determining the relative contribution of an individual method to the consensus.

Redundancy and diversity should be minimized to have effective document summarization. Redundancy calculates the number of repeated terms in the document while diversity indicates the number of terms which are different in the document. Redundancy can be overcome by firstly selecting the top most sentences and then compare the similarity of rest of the sentences to that of already selected one and then choose those sentences that have maximum dissimilarity [31]. Summarization task is a global inference problem and it is an attempt to optimize relevance, redundancy, and length jointly [22]. One of the many approaches Maximal Marginal Relevance (MMR) [10] is to minimize the redundancy. For improving the performance of multi-document summarization system Sarkar has proposed a sentence compression technique for extractive summarization by using various local and global sentence-trimming rules [31]. Hybrid model for text summarization problem are in use which is based on diversity selection, fuzzy concepts and swarm intelligence [8]. The concept of swarm intelligence has used to suggest the weights of the text features and to adjust the text feature scores. Important and unimportant features can be differentiated easily using feature score. The trapezoidal membership function of fuzzy logic is used to fuzzify the numerical values (crisp) of text features.

Selecting the best sentences by pursing the local greedy approximation is easier than the selecting the best summary globally (global optimization problem) [4, 5, 15].

An extractive summarization based model has been proposed by Fillaova *et at*. [12]. The resultant summary is considered as a unit which must conceptually cover maximum aspects of source documents. A method for Text summarization, Maximum Coverage Problem with Knapsack (MCKP) proposed in [34]. MCKP can easily identify whether all the concepts are covered by the resultant summary or not. Baysian Sentence-based Topic Model (BSTM) works on term-document and term-sentence associations used for multi-document summarization, proposed by Wang *et at*. [38]. This model provides a principled way for text summarization by doing probability distribution of selecting sentence of given topics. Huang *et at*. have focused on the four objectives like coverage, significance and redundancy of information and cohesion of text in their proposed technique [16]. Optimization of distortion of information based approach has been proposed by Ma and Wan in [20].

Maximum Coverage and Less Redundancy (MCLR) was proposed in [2] for summarization of multiple documents based on Quadratic Boolean Programming (QBP) problem, which includes two steps first according to the coverage of information or features score is calculated for every sentence and then choose those sentences which are being considered to be added to the final summary,if the sentence covers the maximum important aspects and minimum similarity with other sentences. Besides these algorithms a lots of techniques have been proposed for the same purpose based on clustering algorithms like Suffix Tree Clustering [43], Scatter [23], and Bisecting K-Means Clustering [17]. A graph based model for information has been proposed in [30]. This model has a remarkable impact on the real world industrial problem.

Kovaleva *et at*. [17] has focused on fact of enhancing the Principal Direction Division (PDD) clustering method with the stopping criteria. Then this approach has been extended to Bisecting K-Means and it is able to project the data onto random directions. This result has shown that performance of this approach is better than others like hierarchical clustering algorithms, with respect to the quality of clustering. Besides that, this algorithm is much more efficient.

A generic summarization technique has been presented by Rautray and Balabantaray which [28] tracking the key feature as content coverage and redundancy aspects of single document summary by using Particle Swarm Optimization (PSO) algorithm. The objective function is designed for solving such problem by taking weighted average of content coverage and redundancy features. The objective function used in [28] is also used in another single document summarizer in [29] which takes the features of text as an input arguments instead of sentence weights as input arguments as in [28]. One of the PSO based extractive summarizers [7] uses expression of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) as fitness function. Asgari *et at*. [6] presented a summarization model based on PSO by considering summary features such as content coverage, readability and length. In [1] a summarization technique based on PSO and clustering has been proposed. The input of the technique is taken from multiple documents. After taking input similarity score has been calculated between sentences. The output is ensured to achieve the maximum content coverage and reach to threshold diversity. Similarity metric also used by Alguliev *et at*. [4] for achieving the goals to fulfill content coverage, diversity and length constant applicable for multi-documents summarization. Another Cat Swarm Optimization (CSO) based multi-documents summarization technique has been proposed by Rautray *et at*. [27]. The proposed method focuses on the some key aspects like content coverage, readability and cohesion. This method is found better than the Particle Swarm Optimization (PSO) and Harmony Search Algorithm (HSA) after evaluating the output of the method over DUC dataset.

3. Proposed Model

The summarization is based on k-means clustering algorithm. To measure the similarity or dissimilarity cosine measure is used. After clustering summary is generated by using different constraints like diversity, content coverage and length.

3.1 The Cosine Measure

The similarity measure of text mining is one of the important issues of Natural Language Processing, Information Retrieval and Text Mining. Vector Space Model (VSM) technique is used to represent the textual units in vector form [ ]. This vector form is used to measure similarity and dissimilarity between textual units. Textual units are represented in vector form which is used to calculate the similarity between them.

Let the source document has m number terms, say $t_1,t_2,\dots,t_k$ then a s_i number is given by $s_i=[w_{i1},\dots,w_{im}]$.

In this vector representation $w_{i1},\dots,w_{im}$ represents the weights of term $t_1,t_2,\dots,t_m$ in sentence s_i respectively. The weight or importance w_{ik} of term t_k in sentence s_i is calculated using the concept of Term Frequency (TF) and Inverse Sentence Frequency (ISF)

$$w_{ik}=\mathrm{TF}_{ik}\times\log(n/n_k), \tag{3.1}$$

where TF_{ik} is TF that indicates total number of term t_k present in sentence s_i, n is the total number of sentences in documents and n_k indicates how many total number of sentences in which term t_k appears. The term $\log(n/n_k)$ is commonly called ISF represents for global weight of term t_k.

The traditional IR system initially used the ISF concept to enhance the power of discrimination of the terms. Nowadays, cosine similarity measure is used to calculate similarity between two sentence vectors [ ]. Cosine similarity between two vectors s_i and s_j is defined as below:

$$sim(s_i,s_j)=\frac{\int_{k=1}^{m}w_{ik}w_{jk}}{\sqrt{\int_{k=1}^{m}w_{ik}^2\cdot\int_{k=1}^{m}w_{jk}^2}},\quad i,j=1,\dots,n. \tag{3.2}$$

3.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population based technique, mainly used for discovering favourable regions within search space. In this content two terms are defined one is particle another is swarm. Particle is a member of population and swarm indicates all particles in a group. Flying velocity of particle is adjusted dynamically by using flying experience of its own and also its companions' to get the best position in search space which it ever encountered. The best position obtained by each particle is broadcasted for swarm. In the local variant topology, each particle can be assigned to its neighbours group, which comprises a predefined number of particles.

The PSO algorithm starts by considering solutions that is mainly a group of particles (swarm) randomly. Then in every cycle swarm will update its best velocity and position value by using eq. (3.3) and eq. (3.4) respectively. The optimal solution will come after multiple iterations.

$$Vid(t+1)=Vid(t)+c1*r1d*(Pid-Xid)+c2*r2d*(Pgd-Xid) \tag{3.3}$$

$$Xid(t+1)=Xid(t)+Vid(t+1) \tag{3.4}$$

where

the velocity of particle i is $Vid(t)$ at iteration t,

the position of particle i is $Xid(t)$ at iteration t,

the velocity of particle $iVi(t+1)$ at iteration $t+1$,
the position of particle i is $Xi(t+1)$ at iteration $t+1$,
$r1d$, $r2d$ are random number between $(0,1)$,
$c1$ cognitive acceleration coefficient,
$c2$ social acceleration coefficient

The PSO algorithm works in following steps.

1. Initialize the particle of population with random position and velocity vector.
2. For each particle's position (p) fitness value is calculated.
3. If fitness $p > pBest$ then $pBest = p$.
4. Go to step 2 utile all particle exhausted.
5. $gBest = max(pBest)$.
6. update particle velocity and position using the equation
7. go to step 2 until optimal solution.

The many version of PSO have been proposed such as continuous particle swarm optimization and binary particle swarm optimization. Esmin in [19] it has been proved that PSO covered search space easily without being trapped in local minima or maxima, so it performed faster in comparison to others. In PSO, swarm (cluster) contains a number of data points (particle). It always starts with a seed value. A particle is used to represents the *cluster* centroid Nc.

3.3 The Proposed Algorithm

1. Construct Term-Document Matrix A from unstructured text document.
2. Define maximum number of cluster k as suitable number.
3. Initialize each clusters with the cluster centroid $\{O_1, O_2, \dots, O_k\}$ $\forall$ $k > 1$.
4. Select seed particle with random position and velocity.
5. Do
6. for all particles(sentence) do

   6.1. Assign each sentence to the nearest cluster centroids.

   $$d(O_i, x_i) < d(O_i, x_j), \quad i \neq j,\ i = 1, 2, \dots, n,$$

   where $d(O_i, x_i) = v(O_i - x_i)^2$.

   6.2. Recalculate each cluster center to be equal to the mean of all vector points within that cluster.

   $$O_i \leftarrow \frac{1}{n} \sum_{j=1}^{n} x_j\,.$$

   6.3. Calculate similarity of a particle as fitness f with its cluster centroid by eq. (3.2).

   $$f \leftarrow \min(sim(s_i, s_j))\,.$$

   6.4. if $f(O_i) < pBest$ then

   $$pBest \leftarrow f(O_i)$$

```
    6.5. if $pBest < gBest$ then
             $gBest \leftarrow pBest$
    6.6. Update particle velocity and position using eq. (3.3) and eq. (3.4).
 7. Save the all clusters centroids with smallest fitness value.
 8. End for
 9. Until (maximum iteration >maxIteration or noChange(gBest)).
10. Return all the $k$ clusters with fitness value in ascending order.
11. Final cluster is selected by using eq. (3.5) and eq. (3.6) and eq. (3.8).
```

3.4 Cluster Selection for Summary

After getting all the clusters the following constraints [24] like coverage, diversity and length are calculated. The diversity constraints are given by

$$\Theta_{\text{driver}} = \sum_{i=1}^{n-1} \sum_{j=j+1}^{n} sim(s_i,s_j)x_i \cdot x_j\,, \tag{3.5}$$

$$\sum_{s_i,s_j \in S} sim(s_i,s_j) = \Theta_{\text{driver}}\,, \tag{3.6}$$

where $i \neq j$ and Θ_{driver} is the threshold of diversity (the diversity constraint). The higher value of Θ_{driver} achieves low level of diversity.

$$\Theta_{\text{count}} = sim(o,o^s) \cdot \sum_{i=1}^{n} sim(o,s)x_i\,. \tag{3.7}$$

The constraints of content coverage may be written as

$$sim(S,D) \geq \Theta_{\text{count}}\,, \tag{3.8}$$

where Θ_{count} is content coverage (the content coverage constraint) and o and o^s indicate the mean of original document and summary respectively. The higher value of Θ_{count} corresponds to higher level of similarity between generated summary and source document:

$$\sum_{i=1}^{n} l_i x_i = L\,, \tag{3.9}$$

$$x_i \in \{0,1\}\,, \tag{3.10}$$

where L and l_i indicate the length of summary and sentence s_i respectively. The eq. (3.4) shows that length constraints never be violated and eq. (3.5) implies the integrity constraints.

Each swarm consists of most dissimilar particles. Then content coverage and diversity constraints are estimated for each swarm and select the most desirable one. Then numbers of sentences are placed into the final summary by eq. (3.9).

4. Result Analysis

The proposed algorithm has been evaluated on ten different documents collected from legal case report data set contains Australian legal cases from the Federal Court of Australia (FCA). This is an open source dataset. For evaluating the result Recall-Oriented Understudy for Gisting

Evaluation (ROUGE) [42] is used to compare the system generating summary with the human generated summary to analyze the quality. Here we use ROUGE-N and $N = 1$ and 2 and these are compatible for evaluating signal document summarization technique [18].

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{xsumm}_{\text{ref}}} \sum_{\text{N-gram} \in s} \text{Count}_{\text{match}}^{(\text{N-gram})}}{\sum_{s \in \text{xsumm}_{\text{ref}}} \sum_{\text{N-gram} \in s} \text{Count(N-gram)}},$$

where N is length of the N-gram, the highest number of N-grams is represented by $\text{Count}_{\text{match}}^{\text{N-gram}}$ among candidate summary and reference-summaries. Count(N-gram) indicates number of N-grams present in the reference summaries.

Table 1. Average Precision, Recall and F-Measure using ROUGE-1

Dataset	Precision	Recall	F-Measure
Document1	42	40.67	41.32
Document 2	46	42.67	44.27
Document 3	47.1	45.5	46.29
Document 4	49.02	43.9	46.32
Document 5	46.9	44.4	45.62
Document 6	50.7	41.2	45.46
Document 7	48.4	42.63	45.33
Document 8	48.07	43	45.39
Document 9	50.4	43.9	46.93
Document 10	53.01	43.7	47.91
Average	48.16	43.15	45.48

Table 2. Average Precision, Recall and F-Measure using ROUGE-2

Dataset	Precision	Recall	F-Measure
Document 1	22	20.3	21.12
Document 2	24.2	19	21.29
Document 3	21	21.2	21.10
Document 4	20.9	16	18.12
Document 5	23.01	17	19.55
Document 6	21.9	19.4	20.57
Document 7	19.5	13	15.6
Document 8	22.1	15.9	18.5
Document 9	23	15	18.16
Document 10	20.8	14.7	17.22
Average	21.841	17.12	19.12

The average precision, recall and F-measure of swarm model are 47.741, 43.028 and 44.669 for ROUGE-1, respectively. By using ROUGE-2 average precision, recall and F-measure of swarm model are 21.622, 18.828 and 19.776, respectively.

Table 3. Result comparison

	Proposed Model (Average)			Swarm Model (Average)		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
ROUGE-1	48.16	43.15	45.48	47.741	43.028	44.669
ROUGE-2	21.841	17.12	19.12	21.622	18.828	19.776

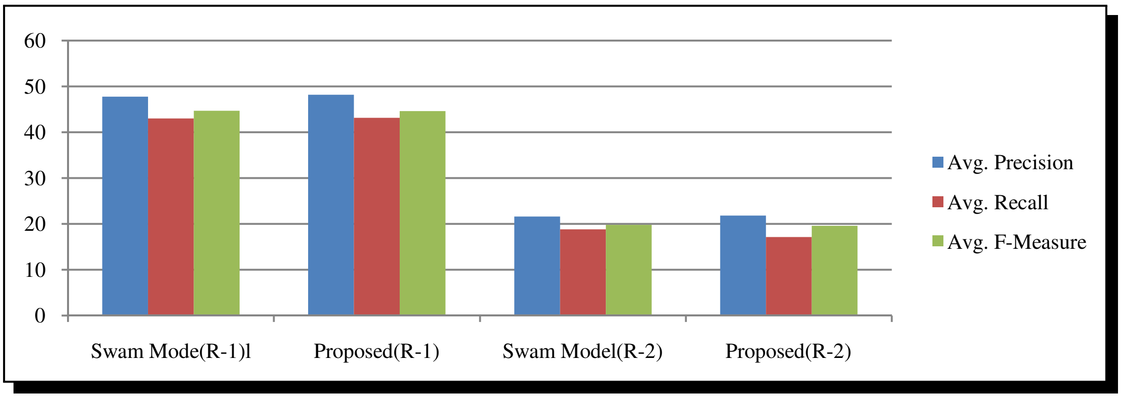

Figure 1. Evaluation graph of Proposed Method

5. Conclusion

In this paper the proposed method is based on extractive summarization. In extractive summarization to select the relevant sentence is a very complex task. So to generate the better result we have applied the PSO algorithm with similarity measurement as a fitness function with the three well defined constraints. To analysis the result we have improved recall, precision, f-factor by ROUGE-1 and precision by ROUGE-2. In future we can enhance the other attributes by ROUGE-2.Throughout the model we never consider the complexity. So we can focus and enhance the model to achieve the low complexity. In future the proposed algorithm may be used to get the optimal solution from the multi-documents source.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] R. M. Alguliev, R. M. Aliguliyev and C. A. Mehdiyev, Sentence selection for genericdocument summarization using an adaptive differential evolution algorithm, *Swarm Evolutionary Comput.* **1**(4) (2011), 213 – 222, DOI: 10.1016/j.swevo.2011.06.006.

[2] R. M. Alguliev, R. M. Aliguliyev and M. S. Hajirahimova, GenDocSum + MCLR: Generic document summarization based on maximum coverage and less redundancy, *Expert Systems with Applications* **39**(16) (2012), 12460 – 12473, DOI: 10.1016/j.eswa.2012.04.067.

[3] R. M. Alguliev, R. M. Aliguliyev and N. R. Isazade, CDDS: Constraint-driven document summarization models, *Expert Systems with Applications* **40** (2013), 458 – 465, DOI: 10.1016/j.eswa.2012.07.049.

[4] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova and C. A. Mehdiyev, MCMR: Maximum coverage and minimum redundant text summarization model, *Expert Systems with Applications* **38**(12) (2011), 14514 – 14522, DOI: 10.1016/j.eswa.2011.05.033.

[5] R. M. Aliguliyev, Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization, *Computational Intelligence* **26**(4) (2010), 420–448, DOI: 10.1111/j.1467-8640.2010.00365.x.

[6] H. Asgari, B. Masoumi and O. S. Sheijani, Automatic text summarization based onmulti-agent particle swarm optimization, in *Intelligent Systems (ICIS), 2014 Iranian Conference on, IEEE*, February, pp. 1 – 5 (2014), DOI: 10.1109/IranianCIS.2014.6802592.

[7] M. S. Binwahlan, N. Salim and L. Suanmali, Swarm based text summarization, in *Computer Science and Information Technology-Spring Conference, 2009, IACSITSC'09, International Association of, IEEE*, 2009, April, pp. 145 – 150, DOI: 10.1109/IACSIT-SC.2009.61.

[8] M. S. Binwahlan, N. Salim L. Suanmali, Fuzzy swarm diversity hybrid model for text summarization, *Information Processing & Management* **46**(5) (2010), 571 – 588, DOI: 10.1016/j.ipm.2010.03.004.

[9] O. Boydell and B. Smyth, Social summarization in collaborative web search, *Information Processing & Management* **46**(6) (2010), 782 – 798, DOI: 10.1016/j.ipm.2009.10.011.

[10] J. G. Carbonell and J. Goldstein, The use of MMR, diversity-based re-ranking for reordering documents and producing summaries, in *Proceedings of the 21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia pp. 335 – 336 (1998), DOI: 10.1145/290941.291025.

[11] M. A. Fattah and F. Ren, GA, MR, FFNN, PNN and GMM based models for automatic text summarization, *Computer Speech and Language* **23**(1) (2009), 126 – 144, DOI: 10.1016/j.csl.2008.04.002.

[12] E. Filatova and V. Hatzivassiloglou, A formal model for information selection in multi-sentence text extraction, in *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 397 – 403 (2004), DOI: 10.3115/1220355.1220412.

[13] Y. Gong and X. Liu, Generic text summarization using relevance measure and latent semantic analysis, in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, USA, pp. 19 – 25 (2001), DOI: 10.1145/383952.383955.

[14] S. Harabagiu and F. Lacatusu, Using topic themes for multi-documen summarization, *ACM Transactions on Information Systems* **28**(3) (2010), 13:1 – 13:47, DOI: 10.1145/1777432.1777436.

[15] L. Huang, Y. He, F. Wei and W. Li, Modeling document summarization as multi-objective optimization, in *Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics*, Jinggangshan, China, pp. 382 – 386 (2010), DOI: 10.1109/IITSI.2010.80.

[16] L. Huang, Y. He, F. Wei and W. Li, Modeling document summarization as multi-objective optimization, in *Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics*, Jinggangshan, China, pp. 382 – 386 (2010), DOI: 10.1109/IITSI.2010.80.

[17] E. V. Kovaleva and B. G. Mirkin, Bisecting K-means and 1D projection divisive clustering: A unified framework and experimental comparison, *Journal of Classification* **32**(3) (2015), 414 – 442, DOI: 10.1007/s00357-015-9186-y.

[18] J. S. Lee, H. H. Hah, S. C. Park, Less-redundant text summarization using ensemble clustering algorithm based on GA and PSO, *Wseas Transactions On Computers* **16** (2017), `http://www.wseas.org/multimedia/journals/computers/2017/a085805-082.php`.

[19] J.-H. Lee, S. Park, C.-M. Ahn and D. Kim, Automatic generic document summarization based on non-negative matrix factorization, *Information Processing & Management* **45**(1) (2009), 20 – 34, DOI: 10.1016/j.ipm.2008.06.002.

[20] T. Ma and X. Wan, Multi-document summarization using minimum distortion, in *Proceedings of the 10th IEEE International Conference on Data Mining*, Sydney, Australia, pp. 354 – 363, (2010), DOI: 10.1109/ICDM.2010.106.

[21] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization*, p. 442, MIT Press, Cambridge (1999).

[22] R. McDonald, A study of global inference algorithms in multi-document summarization, in *Proceedings of the 29th European Conference on IR Research*, Rome, Italy, No. 4425, pp. 557 – 564 (2007), DOI: 10.1007/978-3-540-71496-5_51.

[23] A. Notsu and S. Eguchi, Robust clustering method in the presence of scattered observations, *Neural Computation* **28**(6) (2016), 1141 – 1162, DOI: 10.1162/NECO_a_00833.

[24] Y. Ouyang, W. Li, S. Li and Q. Lu, Applying regression models to query focused multi-document summarization, *Information Processing & Management* **47**(2) (2011), 227 – 237, DOI: 10.1016/j.ipm.2010.03.005.

[25] D. Radev, H. Jing, M. Stys and D. Tam, Centroid-based summarization of multiple documents, *Information Processing & Management* b(6) (2004), 919 – 938, DOI: 10.1016/j.ipm.2003.10.006.

[26] V. S. Raj Kumar and D. Chandrakala, An effective generic summary creation of multi and single documents using genetic algorithm, *International Conference on Breakthrough in Engineering, Science & Technology*, Vol. **3** (3), 154 – 158, March 2016, `http://ijetst.in/article/si/35%20ijetst.pdf`.

[27] R. Rautray and R. C. Balabantaray, Cat swarm optimization based evolutionaryframework for multi document summarization, *Phys. A: Stat. Mech. Appl.* **477**(2017), 174 – 186, DOI: 10.1016/j.physa.2017.02.056.

[28] R. Rautray and R. C. Balabantaray, Comparative study of DE and PSO over document summarization, in *Intelligent Computing, Communication and Devices*, Springer, India, pp. 371 – 377 (2015), DOI: 10.1007/978-81-322-2012-1_38.

[29] R. Rautray, R. C. Balabantaray and A. Bhardwaj, Document summarization usingsentence features, *Int. J. Inf. Retrieval Res.* **5** (1) (2015), 36 – 47, DOI: 10.4018/IJIRR.2015010103.

[30] K. C. Santosh, g-DICE: graph mining-based document information content exploitation, *International Journal on Document Analysis and Recognition* **18**(4), 9 September 2015, DOI: 10.1007/s10032-015-0253-z.

[31] K. Sarkar, Syntactic trimming of extracted sentences for improving extractive multi-document summarization, *Journal of Computing* **2**(7) (2010), 177 – 184.

[32] D. Shen, J.-T. Sun, H. Li, Q. Yang and Z. Chen, Document summarization using conditional random fields, in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, pp. 2862 – 2867 (2007).

[33] W. Song, L. C. Choi, S. C. Park and X. F. Ding, Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization, *Expert Systems with Applications* **38**(8) (2011), 9112 – 9121, DOI: 10.1016/j.eswa.2010.12.102.

[34] H. Takamura and M. Okumura, Text summarization model based on maximum coverage problem and its variant, in *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece, pp. 781 – 789, (2009), `https://dl.acm.org/citation.cfm?id=1609154`.

[35] C. Teng, N. Xiong, Y. He, L. T. Yang and D. Liu, A behavioural mode research on user-focus summarization, *Mathematical and Computer Modelling* 51(7-8) (2010), 985 – 994, DOI: 10.1016/j.mcm.2009.08.015.

[36] X. Wan, Using only cross-document relationships for both generic and topic-focused multi-document summarizations, *Information Retrieval* **11**(1) (2008), 25 – 49, DOI: 10.1007/s10791-007-9037-5.

[37] D. Wang and T. Li, Weighted consensus multi-document summarization, *Information Processing & Management* **48**(3) (2012), 513 – 523, DOI: 10.1016/j.ipm.2011.07.003.

[38] D. Wang, S. Zhu, T. Li and Y. Gong, Multi-document summarization using sentence-based topic models, in *Proceedings of the ACL-IJCNLP Conference*, Singapore, pp. 297 – 300 (2009), DOI: 10.3115/1667583.1667675.

[39] D. Wang, T. Li and C. Ding, Weighted feature subset non-negative matrix factorization and its applications to document understanding, in *Proceedings of the 2010 IEEE International Conference on Data Mining*, Sydney, Australia, pp. 541 – 550 (2010), DOI: 10.1109/ICDM.2010.47.

[40] D. Wang, T. Li, S. Zhu and C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore pp. 307 – 314 (2008), DOI: 10.1145/1390334.1390387.

[41] C. C. Yang and F. L. Wang, Hierarchical summarization of large documents, *Journal of the American Society for Information Science and Technology* **59**(6) (2008), 887 – 902, DOI: 10.1002/asi.20781.

[42] 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* **44**(4) (2008), 1600 – 1610, DOI: 10.1016/j.ipm.2007.09.007.

[43] J. Zhang, X. Ma, W. Li and W. Jin, Social Network Recommendation Based on Hybrid Suffix Tree Clustering, in *Computer Science and its Applications*, Springer, Berlin — Heidelberg, pp. 47 – 53 (2015), DOI: 10.1007/978-3-662-45402-2_8.