EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Multi-Document Summarization using CS-ABC Optimization Algorithm

K. Chandra Kumar[1,2,*] and Sudhakar Nagalla[3]

[1]Research scholar in Acharya Nagarjuna University, India
[2]Lecturer at the Faculty of Computer Science, Kakinada Institute of Engineering and Technology, India
[3]Principal, Bapatla Engineering College, India

Abstract

In revolve handle to the information excess, the dramatic boost up documents, on the WWW, show the way of the accessibility of various credentials through the equal subject with conception. Within a limited time, a hard to inquire a suitable a particular document associated to a specific topic to fulfils user's compound data conditions. Hence, we have followed an effective document summarization system applying SVM classifier strategy by this paper. For choosing optimal sentence sets, the proposed technique applies the hybrid ABC-CS optimization algorithm. Further, established on few relevant features, SVM classifier approach is applied in finding the summary by ranking each of the optimal sentences. The operational proposal of JAVA and the results were examined for the methodology is implemented.

Keywords: Artificial Bee Colony based Cuckoo Search optimization technique, Support Vector Machine Classifier, Term Frequency, Inverse Sentence Frequency, Aggregate Cross Sentence Frequency.

Received on 03 February 2020, accepted on 10 March 2020, published on 19 March 2020

doi: 10.4108/eai.13-7-2018.163835

*Corresponding author. Email:chandrakumark2381@gmail.com

1. Introduction

To fascinate the crucial data controlled in huge quantity of manuscript, with inaugurate it in a concise, agent, with reliable review is the goal of summarization. On an afforded topic, a healthy printed review preserve crucially decrease the quantity of vocation essential to process of big quantities of manuscript. The presentation of synopses is as of now an undertaking best dealt with by people. Nonetheless, with the blast of accessible printed information, it is never again monetarily conceivable, or attainable, to make a wide range of outlines by dispense [1].

There are two main components in extractive summarization systems, namely sentence ranking and sentence selection at usually. To give a summary, the former affords an informative score to every sentence to evaluate its crucial while the latter, built on both the salience scores and redundancy between the sentences, chooses sentences. In extractive summarization, the sentence ranking has been extensively inquired [2]. For each pair of parallel sentences, the conventional word arrangements reproductions obtain a sentence-level orchestrated corpus as info with create a word-level arrangements. Consequently accumulated LCS information regularly contains no sentence-level arrangements, yet despite everything it has a few points of interest for preparing word arrangement models and machine translation (MT) frameworks which merit investigating [3].

Content outline tends to both the issue of picking the most noteworthy bits of content and the issue of giving sound rundowns. There are two sorts of a synopsis: extractive and abstractive. In the first record, extractive synopsis strategies disentangle the procedure by picking a delegate subset of the sentences. In the first source, an abstractive outline may assemble novel sentences that are

not initiated. Nonetheless, abstractive methodologies need deep natural language handling strategies [4].

The regulated techniques treat archive synopsis as a characterization assignment of discovering whether a sentence ought to be conceded in the outline or not. In any case, they ascertain upon the preparation tests, which are elusive. The unaided strategies for the most part use bunching calculations to score the sentences in the archives by comparing a lot of predefined highlights [5].

Another case of this sort of summarizer is initiated by Gupta and Siddiqui (2012). It compares single report synopses applying sentence bunching strategies to give multi-record outlines. It fills in as pursues: (i) First, it delivers a solitary archive rundown (utilizing sentence scoring technique); (ii) Then, it bunches the sentences applying both syntactic and semantic similitudes between sentences to speak to the pieces of the writings to be initiated in the outline; (iii) Finally, it gives the synopsis by removing a solitary sentence from each group [6].

Multi-document summarization (MDS) methods have been proposed to handle occasion summarization. A brief, on-point rundown from a lot of related documents, every now and again by extricating educational sentences from those documents is given by this strategy. Nonetheless, these methodologies are adjusted to reflectively outline an occasion managed a (top-notch) set of on-theme newswire articles about it and henceforth are not alluring to condense a progressing occasion after some time [7].

Rest of the paper is organized as follows. Section 2 surveys some previous literature which focused research on document summarization. Section 3 states the problem of existing system and proposes multi-document summarization using CS-ABC Optimization Algorithm. Results of this proposed approach are discussed in section 4. Finally, the conclusion of the paper is described in section 5.

2. Related survey

The possibility of applying the ontology in resolving multi-document summarization troubles in the field of failure organization has been developed by Lei Li and Tao Li [8]. They rendered an exact investigation of different methodologies wherein the metaphysics had been applied for summarization errands. Regarding the outline quality, broad trials on an assortment of official statements applicable to Hurricane Wilma in 2005 shows that philosophy based multi-document summarization strategies beat different baselines. However, quality of the summary is further to be improved.

A framework for abstractive summarization of multi-documents, which aids to choose contents of synopsis not from the first document sentences however from the semantic portrayal of the source documents, has been invented by Khan et al. [9]. The contents of the source documents were presented by predicate argument structures by utilizing semantic role labeling in that framework. Content chosen for summary was caused by ranking the predicate argument structures established on maximized features, and applying language generation for generating sentences from predicate argument structures. By presenting this scheme, the authors had achieved better precision and mean coverage score. Nevertheless, the performance of multi-document abstractive summarization is has to enhanced.

The various kind of matrix factorization technique, specifically Weighted Archetypal Analysis (WAA) to Q-MDS has been analyzed by Canhasia and Kononenko [10]. A chart portrayal of a lot of sentences weighted by comparability to the managed inquiry, emphatically and/or adversely striking sentences are values on the weighted informational collection a limit in query-focused summarization. In target documents set, and they decide to utilize WAA to ascertain these extraordinary qualities, archetypes, and hence to evaluate the importance of sentences. The authors had improved the performance of text summarization in terms of Recall Oriented Understudy for Gisting Evaluation (ROUGE). However, precision of the work is further to be enhanced.

The age of extractive synopses from a solitary document as a twofold improvement trouble where the quality (fitness) of the solutions was established on the weighting of individual statistical features of each sentences -, for example, location, punishment duration with the relationship of the rundown to the designation, equated with group features of similarity among candidate sentences in the summary and the original document, and between the candidate punishments of the summary has been implemented by Mendoza*et al.* [11]. The authors had improved the value of ROUGE. Nevertheless, mean coverage score of the text documentation is further to be improved.

The multi-document Summarization with Group Sparse learning (SGS) system, which can maximally revamp the first documents by means of lessening the estimation blunder and together pick rundown sentences with the scholarly group structure data between the sentences have been broke down by Ruifanget al. [12]. The rundown relatedness can be adjusted by limiting the reproduction models to be near one another and make multiple sentences share a typical basic structure to shape the outline content. Simulation results of the article showed that ROUGE of the proposed scheme had improved than the previous summarization systems. They had missed to improve the precision of the proposed summarization.

The making utilization of semantic job data to raise the diagram based positioning calculation for multi-document summarization has been concocted by Yan and Wan [13]. The first parse the sentences and locate the semantic jobs, and then recommend a newSRRank algorithm and two augmentations utilize the semantic job data. In a heterogeneous positioning procedure, their proposed algorithms can all the while ranking the sentences, semantic jobs, and words. Because of the proposed scheme, ROUGE recall scores were improved. However, they have to integrate SRRank-cluster and SRRank-span methods to get better results.

A using multi-view summarization to wireless video sensors to neglect redundant contents such that the density with communication influence can be decreased has been analyzed by Ouet al. [14]. A low-intricacy online multi-see video summarization strategy was suggested. While keeping significant events and experiments demonstrate that the proposed summarization method successfully decreases the video content. A power analysis of the system also demonstrated that an important amount of energy can be saved. By presenting this proposed scheme, the authors had reduced energy consumption and computational complexity. Nevertheless, F1 score of the proposed scheme is further to be improved.

A technique for unaided item highlight extraction for include situated assessment assurance has been discovered by Quan and Ren [15]. To determining the comparison detachment of field vectors, the field-specific features were extracted. A field vector was derived established on the organization principles among a quality with proportional field corpora. A new expression similarity evaluate (PMI–TFIDF) was inaugurated to estimate the association of candidate features and domain entities. Because of this proposed scheme, the authors had attained better distinction ability. However, computation complexity of the proposed scheme is further to be decreased.

3. Problem definition and Proposed Methodology

Since a collection of topic-based documents, the Multi-document summarization (MDS) has aim to extract the interior data. Basically a kind of information compression technique is MDS. Most of the being nonexclusive extractive summarization strategies can be generally isolated into two groups - unaided techniques and administered strategies. Solo techniques frequently utilize the positioning models to pick the sentences from a candidate set and concede the centroid-based strategy. In the document set, and even though there are many sentences, there may be a small some of the striking and enlightening. In the event that we see sentences as a kind of signals, in the original documents, MDS preserve the conceived to remove a subset of sentences that can best reproduce the full arrangement of sentences.

Another multi-document Summarization with Group Sparse learning (SGS) system, which can maximally remake the first documents by means of lessening the estimate blunder and together pick rundown sentences with the educated group structure data between the sentences, has been developed by Ruifanget al. [12]. The above mentioned article is sound in technical manner and the strategy towards the trouble was satisfactory but its lags in the results, the results found are good but doesn't achieved satisfaction level. This trouble aims us to work on the same work with some advanced approaches to attain better results.

The activated to a limited extent by the advancement of the WWW, with the mushrooming of the amount of online content data, and it is particularly valuable to have devices that can assist clients with processing data content. Content summarization endeavors to address this issue by considering an in part organized source content, extricating data content from it, and introducing the most vital substance to the client in a way delicate to the client necessities. The few present-day data recovery applications require summarization frameworks which scale up to huge volumes of unlimited content in misusing summarization. The normal issue which emerges is the presence of multiple documents finding comparable data, as on account of multiple news tales about an occasion or a grouping of occasions in such applications. A specific test for content summarization is to be competent to outline the likenesses and contrasts in data content between these documents such that it is touchy to the necessities of the client. Multi-document summarization applying cuckoo search based artificial bee colony (CS-ABC) streamlining algorithm has been proposed by this paper.

The relatedness can be adjusted by limiting the reproduction models to be near one another and produce multiple sentences to share a typical hidden structure to shape the synopsis content. We consider the worldwide data in gauge the centrality of sentences and further diminishing the repetition by this model. The simple block diagram of proposed SVM based summarizer is afforded in the below figure 1.

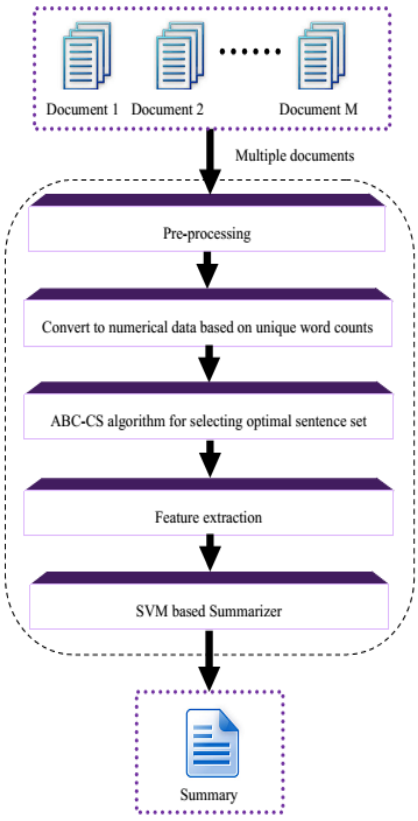

Figure 1. Overall Block diagram of proposed document summarization technique

The suggested model is aligned to process faster by neglecting unwanted HTML/XML tags, stopwords and stem words at firstly. Moreover, for choosing optimal sentence sets, ABC-CS approach is applied. Artificial Bee Colony is one of the proficient streamlining algorithms, demand departure from cuckoo search algorithm will handle to a lot quicker intermingling rate than some conventional strategies. To extracting important text features, SVM will be applied for summarizing those documents.

3.1 Document Pre-Processing

The HTML/XML Tag Images Removal, Sentence Tokenization, Stop word Removal and the Stemming processes takes place in Document Pre-Processing step.

HTML/XML Tag Images Removal

From any online origins, the process of removal of HTML/XML Tag Images is done when the input documents are gathered. Moreover, the extra white space, figures, equations and the special characters like '{[(<!@#$%^&*~`:+;>)]}?' are also removed.

Sentence segmentation

After the removal of unnecessary elements introduce on the documents, in summary creation, each sentence is to be segmented from the documents for the ease of processing. Let us conceiving document database, $N = \{n_1, n_2, \ldots n_T\}$, where, n_k represents the k^{th} document in the database, N. For each text documents 'n_k', the number of sentences are segmented separately as, $A = [A_1, A_2, \ldots, A_r]$. Where, A_q depicts q^{th} the sentence in the document for simple extraction of the short sentence and n is the number of sentences in the document.

Moreover, the Schematic Diagram of suggested document summarization approach is afforded in the below figure 2.

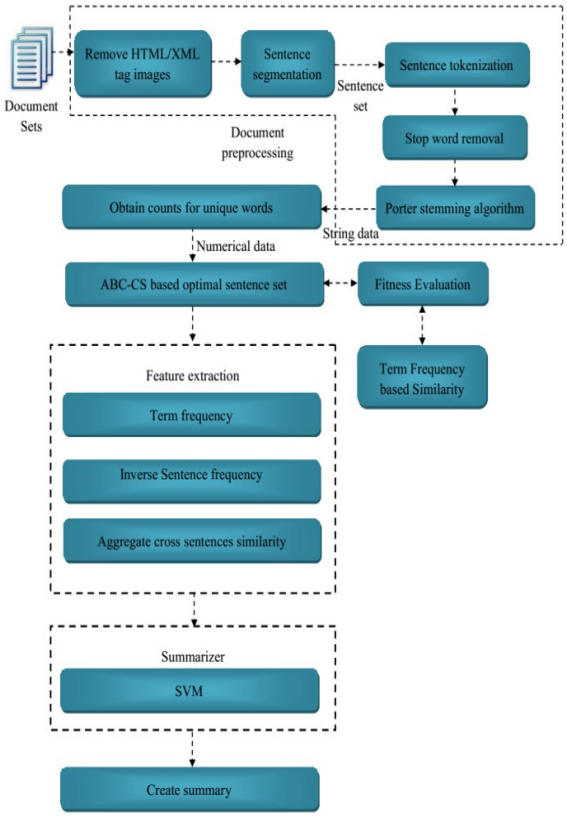

Figure 2. Schematic Diagram of proposed document summarization strategy

Tokenization

Nextly, the Sentence Tokenization is executed to divide the set of sentences from each document. Hence, the words/terms of each sentence are tokenized as, $X = [X_1, X_2, \ldots, X_m]$, where $X_p \in X$ for $p = 1,2,\ldots m$ represents all the distinct terms founding in document 'n_k' and 'm' is the number of unique terms.

Stop word Removal

After tokenizing the Sentences, the stop words are dispatched; where, the most usually words applied in English language like 'a', 'an', 'the' etc. which has less significant centrality as for the document are taken out.

Stemming by Porter Stemming Algorithm

Stemming is a procedure of slashing off the parts of the bargains a typical base structure. Here, the stemming is done set up on the Porter Stemming Algorithm. The algorithm is extremely straightforward in idea, with around 60 additions, two recoding rules and a solitary kind of setting the touchy standard to decide if a postfix ought to be taken out.

After pre-processing the documents, a set of optimal sentences are chosen from the multiple documents.

3.2. Data Conversion

The string data is exchanged to numerical data established on the unique word counts by this stage. The conversion of string data is afforded as in the below figure 3.

The Term Frequency Similarity Computation matrix can be found in the below figure. Where, (U_{Term} 1, U_{Term} 2, … U_{Term} m) are the unique terms from the sentences (Sentence1, Sentence 2, …, Sentence n). Thus, the optimal sentence set can be found by setting the objective function established on the above calculated matrix; so that the sentence chosen may comprise little useful information.

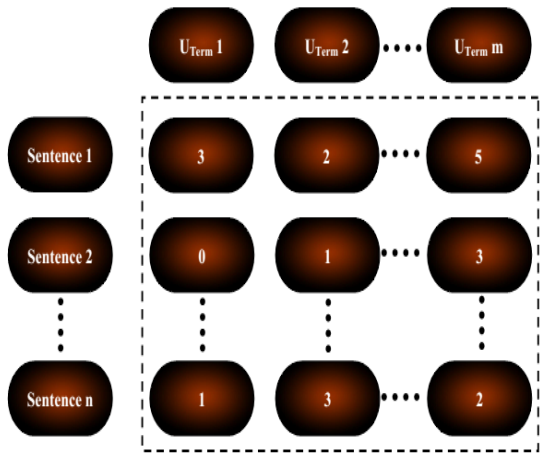

Figure 3. Term Frequency similarity Computation Matrix

3.3. CS-ABC based optimal Sentence Selection

The align of proposed Multi Document Summarizer can be modeled by means of simply maximizing the input sentence set by means of cuckoo search based artificial bee colony (CS-ABC) optimization algorithm before inserting the summarizer.

Artificial Bee Colony Algorithm

An optimization algorithm is Artificial Bee Colony algorithm, where the optimal solution is found and established on reproducing the ordinary foraging performance of a bee colony. Here, the input signal will be applied to give the initial population of the ABC algorithm.

The bees employed in ABC algorithm may be comprised of three categories:

- ◊ Employee Bee
- ◊ Onlooker Bee
- ◊ Scout Bees

Role of Employee Bee

To find out new food origins, the employee bees go for local search through the search space to determine the fittest solution (i.e., food source) from the neighborhood solution. If the better fitness value is found from the neighbor solution, then the employee bee replaces the new food source with the existing food source. The low food source with lower fitness value will be disposed.

Role of Onlooker Bee

Normally, from the neighbor, the onlooker bee also departs in search of the better food. So that, the probability of selecting food origin from the by onlooker will be computed and then selectedand established on the probability values. The neighbor food origin will be then chosen randomly and fitness estimated. For lower fitness values, the food source will be replaced with the food source with lower fitness values.

Role of Scout Bee

The scout bees were utilized to give novel food origin for the discarded food foundation by the employee bees with the onlooker bees at finally.

Cuckoo search optimization algorithm:

Cuckoo search algorithm is a metaheuristic algorithm proposed by Xin-She Yang and Suash Deb in 2009 by inspiration since environment. The algorithm is established on obligate brood parasitism of some cuckoo species and random walks named asLévy flights.

Cuckoo request computation depends on the accompanying three regular values.

1. Every cuckoo selects arbitrarily a nest and lays a single egg in the nest at once.
2. The greatest nests (solutions) controlled by the objective function characterized by the trouble are continued to the following creations.
3. The available home number is fixed and the probability of finding the cuckoo eggs by having flying winged creatures is in the extension of [0, 1].

The steps admitted in the hybrid ABC-CS optimization algorithm is afforded as of,

STEP 1: Solution Initialization

The ABC gives a accidentally circulated primary populace of solutions (S) in the first step, anywhere S represents the amount of utilized bees.

Each solution $S_b\,(b=1,2,...,S)$ is a D-dimensional vector wherever D is the amount of optimization limitations (i.e. the number of words inaugurating in a sentence). The b^{th} solution at v^{th} dimension can be generated applying the below equation,

$$S_{bv} = S_v^{\min} + rand[0,1]\left(S_v^{\max} - S_v^{\min}\right) \quad (1)$$

Where, $rand[0,1]$ is the regularly scattered accidental numeral, $S_v^{\min}$ and $S_v^{\max}$ are the bounds of S_b at v^{th} dimension.

STEP 2: Fitness Evaluation

The primary fitness of the population is estimated for after initialization. Here, the objective function is found from the maximum value of the term frequency similarity. The fitness of a sentence (S_x) can be estimated by means of the below equation.

$$fitness(S_x) = \max\left(\frac{1}{I-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(S_x(j) - S_i(j))^2\right) \quad (2)$$

Where, I represents the total number of sentences from multiple documents; J depicts the number of unique words. Also, $S_x(j)$ represents the j^{th} unique term in sentence 'S_x' and $S_i(j)$ is the j^{th} unique term in the other remaining sentences.

The number of inhabitants in the arrangements is then exposed to rehash cycles, for example, used bees, the passerby bees, and the scout bees.

STEP 3: Employee Bee Phase

For each applied bee, the new solutions (F_{bv}) is developed by applying the solution search equation.

$$F_{bv} = S_{bv} + \phi_{bv}(S_{bv} - S_{cv}) \quad (3)$$

Where, $c \in \{1,2,...,S\}$ and $v \in \{1,2,...,D\}$; ϕ_{bv} is uniformly distributed between the interval [0, 1]

The solution satisfying the purpose occupation is selected with is preceded to the next phase (onlooker bee phase).

STEP 4: Onlooker Bee Phase

In this step, the prospect morals (P_b) for the applied bee solutions (F_{bv}) are computed by the following eq. (13):

$$P_b = \frac{fitness_b}{\sum_{b=1}^{S} fitness_b} \quad (4)$$

In the event that a position can't be grown further through a foreordained number of cycles, the nourishment starting point ought to be abandoned (i.e., the nourishment source is abandoned when a similar nourishment source is found as the fittest one for the three preliminary checks).

STEP 5: Levy Flight Updation

In the onlooker bee phase, and if the food origins are not developed, the current solutions ($(F_{bv}^{g})'$) are aligned with a novel randomly created solution (S_{bv}^{g+1}).

In nature, the Lévy flight is one of the approaches which some animals utilize to investigate for food. Moreover, the Lévy flight can be determined as the random walk in, in which the step lengths are scattered and established on the Lévy distribution which has an unbounded difference with an interminable mean.

The Lévy flight distribution can be given as follows,

$$Levy \sim \delta = g^{-\delta}; \quad \text{Where, } 1 < \delta \le 3 \quad (5)$$

An established on the Lévy flight distribution, the arrangement can be refreshed by methods for the accompanying connection,

$$S_{bv}^{g+1} = (F_{bv}^{g})' + \beta \oplus Levy(\delta) \quad (6)$$

Where, β is the parameter to control step size ($\beta > 0$)

STEP 6:

Memorize the finest resolution that is obtained so extreme.

STEP 7:

Repeat the cycle awaiting the extinction condition is fulfilled.

Once the extinction criterions are attained, the best possible set of sentences is attained. Nextly, the SVM based summarizer is utilized so as to produce summary applying the optimal sentences.

3.4. Feature Extraction

The feature extraction method admits abridge the quantity of origins needed to depict huge locate of information accurately. Established on the estimated features, the summary of the sentence sets are produced andapplying SVM classification technique. Let the set optimal sentences be, $\hat{S} = \{\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_z\}$. From the set of maximized sentences, certain crucial features are extracted so as to rank

the importance of the sentences. The features admit, term frequency feature, Inverse Sentence Frequency feature and Aggregate Cross sentence similarity features.

Term Frequency Feature

The term frequency of a word (w_l) is depicted as the numeral of periods that term 'w_l' founds in the document (i.e., the chosen optimal sentence set). The term frequency feature can be afforded as,

$$F_2 = tf(w_l) = \frac{count\ (w_l)\ in\ \hat{S}}{M} \tag{7}$$

Where, M represents the sum of occurrences of all words in total number of unique word and $\hat{S}$ refers the optimal sentence set.

Inverse Sentence Frequency Feature

The inverse sentence frequency is an evaluation of how much data the word gives, that is, regardless of whether the term is regular or uncommon over all sentences.

The inverse sentence frequency feature can be written as follows,

$$F_2 = isf(w_l) = \log\left(\frac{M}{1+M(w_l)}\right) \tag{8}$$

Aggregate Cross Sentence Similarity Feature

The Aggregate Cross Sentence Similarity evaluate can be applied to find out the suitable applicant for the synopsis. In the input optimal sentence sets, and this estimate choose the sentence having most extreme comparability with every single other sentence. Henceforth the Aggregate Cross Sentence Similarity of a sentence, $\hat{S}_a$ can be calculated as,

$$F_2 = acs\_sim(\hat{S}_a) = \sum_{b=1}^{z-1} Sim\left(\hat{S}_a, \hat{S}_b\right) \tag{9}$$

Where, $\hat{S}_b$ represents the b^{th} optimal sentence; also, $\left(\hat{S}_a, \hat{S}_b\right) \in \hat{S}$.

The above features are extracted for all the sentences. Established on the range of feature vectors, the ranking of sentences is done in order to produce the summary.

3.5.SVM based Summarizer

The SVM represents a machine learning technique aligned in accordance with the statistical learning theory and is fruitfully utilized for classification and regression with high-dimensional space. A SVM algorithm was aligned to locate the optimal hyper- plane separating two classes with insufficient data. In addition, it can isolate inactive or not unmistakably perceptible examples accurately. In contrast to different classifiers, SVM lessens auxiliary hazard as opposed to observational hazard. During the preparation of SVM, it boosts the good ways from structures to the class isolating hyper-plane. The examples are not straightly divisible at as a rule; along these lines, nonlinear piece change is executed.

Let us conceiving the training samples be, ($P = \{p_1, p_2, .., p_y\}$) of feature vectors in D-dimensional space. Each feature vector is afforded as,

$$P_k = \{p_{k1}, p_{k2}, .., p_{kD}\}^T \in R^D \tag{10}$$

Also, the corresponding labels are afforded as, L. Where, value of L is established on the rank of each sentence.

The classification can be given as,

$$f(q, c^*, d^*) = \sum_{k=1}^{y} L_k c^* H(p_k, q) + d^* \tag{11}$$

In the above equation, c^*, d^* are the optimal parameters discovered from the training samples p and q is the new sample to anticipate. The kernel function $H(p_k, q)$ that estimates the distance among two features plays an crucial role in classification accuracy.

Usually applied kernel functions are of linear kernel, Polynomial kernel, Quadratic kernel, Sigmoid and Radial Basis function. The expressions for kernel functions are represented as below,

For Linear Kernel: $H_{lin}(P,Q) = p^T q + c$ (12)

Where u, v denotes the internal manufactures in linear kernel and c is a constant.

For Quadratic Kernel:

$$H_{quad}(P,Q) = 1 - \frac{\|p-q\|^2}{\|p-q\|^2 + c} \tag{13}$$

Where, u, v -are the vectors of the polynomial kernel purpose in the input gap

For Polynomial Kernel: $H_{poly}(P,Q) = (\sigma p^T q + c)^e, \sigma > 0$

$$H_{poly}(P,Q) = (\sigma p^T q + c)^e, \sigma > 0 \tag{14}$$

For Sigmoid Kernel: $H_{sig}(P,Q) = \tanh(\sigma p^T q + c), \sigma > 0$ (15)

Here, the Radial Basis Function kernelis used. The Radial Basis Function kernel is given as:

$$H_{RBF}(P,Q) = \exp\left(-\frac{\|p-q\|^2}{2\sigma^2}\right) \quad (16)$$

For each optimally chosen sentence, the classification of sentences affords the appropriate rank. As, a result a summary is produced from the optimal sentence sets established on the SVM ranks.

4. Result and Discussion

The consequence and conversation is instanced regarding the Multi-Document Summarization applying CS-ABC optimization algorithm by this section. In the working platform of JAVA, the proposed algorithm is accomplished and the experimentation is done in a framework which contains 4 GB RAM and 2.10 GHz Intel i-3 processor.

The documents were gathered from different databases for analysis. An established on word count, this gained information was pre-handled and traded into numerical information. For selecting optimal sentence sets, then the proposed CS-ABC algorithm is utilized. Features of each document were extracted and then summarized by applying SVM based summarizer.

4.1. Evaluation Metrics

The evaluation measurements, for example, Precision, Recall, and F-measure are depicted beneath:

Precision

Precision affords information about effectiveness of the proposed system. Precision is determined as the proportion of the number of important sentences recovered to the number of sentences recovered.

$$P = \frac{n(R_{\text{Re}})}{n(R)} \quad (17)$$

Where,

$n(R_{\text{Re}})$ - Number of relevant sentences retrieved

$n(R)$ - Number of sentences retrieved

Recall

Recall affords information about the accuracy of the proposed system. The proportion of the number of important sentences recovered to the all outnumber of significant sentences in the database is resolved as recall.

$$R = \frac{n(R_{\text{Re}})}{n(D_{\text{Re}})} \quad (18)$$

Where, D_{Re} is the full amount numeral of relevant sentences in the database.

F- Measure

The F-measure is an assessment of a test's exactness and is resolved as the weighted symphonious mean of the accuracy and review of the test. It is depicted as,

$$F - Measure = 2 * \frac{(\text{Pr}\,ecision * \text{Re}\,call)}{(\text{Pr}\,ecision + \text{Re}\,call)} \quad (19)$$

4.2. Performance analysis

This segment is instanced about the performance evaluated of the proposed CS-ABC optimization algorithm. In the below tables, the Precision, Recall and F-Measure values for the suggested and existing methods are tabulated.

Table 1. Precision, Recall and F-Measure values for the proposed CS-ABC technique

Iterations	Precision	Recall	F-measure
10	63.147	61.228	62.1727
20	63.249	62.049	62.64325
30	64.117	63.481	63.79741
40	65.387	63.854	64.61141

. Table 2. Precision, Recall and F-Measure values for existing ABC technique

Iterations	Precision	Recall	F-measure
10	62.98429	60.3895	61.65961
20	62.66616	61.42257	62.03813
30	63.92894	63.47462	63.70097

40	65.21471	63.24945	64.21705

Table 3. Precision, Recall and F-Measure values for existing Cuckoo technique.

Iterations	Precision	Recall	F-measure
10	62.32958	61.0372	61.67662
20	62.59287	61.41001	61.9958
30	63.13714	62.7669	62.95147
40	64.69874	63.0977	63.88819

From the analysis, it is clear that the precision value is reliable for the proposed CS-ABC for all the iterations on comparing with the values of the existing ABC and Cuckoo techniques. Table 1 & Figure 4 show the performance analysis of the proposed CS-ABC algorithm. As shown in the table and figure, precision, recall and F-measure of CS-ABC are analyzed by varying number of iterations. The precision of CS-ABC is increased when the number of iterations increases. Similarly, recall of CS-ABC is 61% at 10 number of iterations while it is 64% at 40 number of iterations. Besides, F-measure of CS-ABC is 65% at 40 number of iterations. From these analyses, we inferred that the performance of multi-document summarization is enhanced due to the hybridization of ABC and CS algorithm.

The performance analysis of existing ABC algorithm is shown in table 2 & figure 5. From the table and figure, precision of ABC algorithm is reduced to 0.26% than the proposed CS-ABC at 40 number of iterations. Also, recall and F-measure of ABC algorithm are decreased to 0.96% and 0.62% respectively that of CS-ABC at 40 number of iterations. Table 3 and figure 6 show the performance analysis of existing cuckoo search algorithm. Compared to the proposed CS-ABC algorithm, precision, recall and F-measure of the cuckoo search algorithm are reduced to 1.1.%, 1.2% and 1.14% respectively.

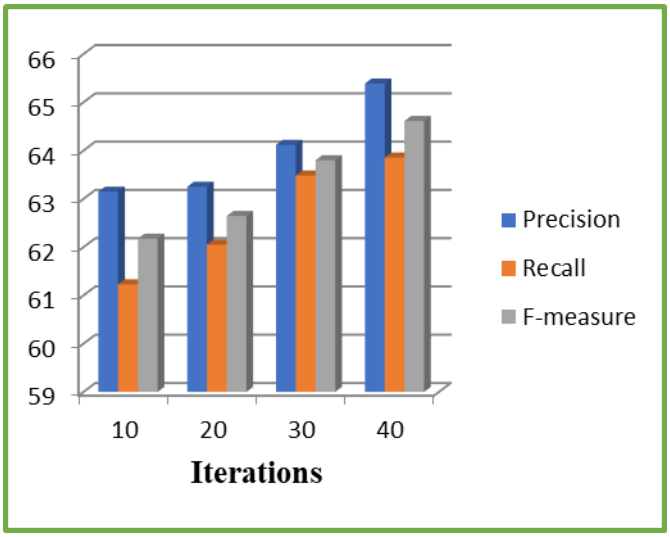

Figure 4. Precision, Recall and F-Measure values for the proposed CS-ABC technique.

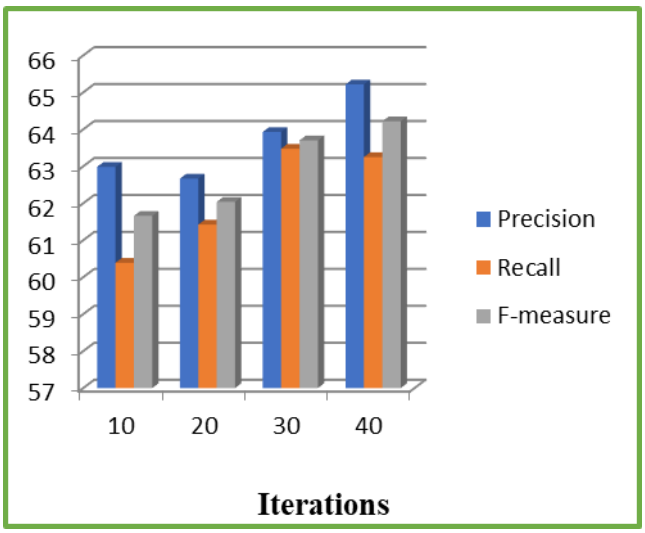

Figure 5. Precision, Recall and F-Measure values for the existing ABC technique.

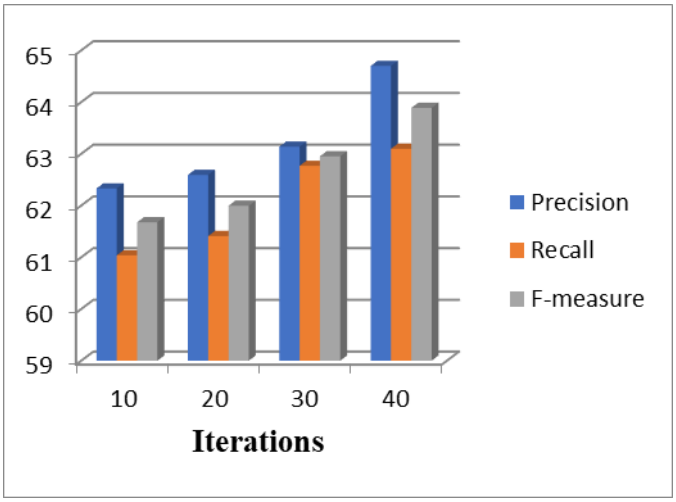

Figure 6. Precision, Recall and F-Measure values for the existing Cuckoo technique.

From the above comparison plot, the proposed CS-ABC technique is additional dependable than several extra existing technique. While equating Precision, Recall with F-Measure, the planned technique demonstrates 100% than the existing ABC and Cuckoo approaches.

5. Conclusion

In this paper, we have structured a multi-document summarization utilizing CS – ABC optimization algorithm for selecting optimal sentence sets. The proposed algorithm is compared with various existing advances to show the efficacy of the proposed method. From results got, it is particularly noticed that the planned multi – document summarization using CS –ABC optimization algorithm performs to be effective to summarize the documents.

References

[1] M.Kagebäck, O.Mogren, N.Tahmasebi, D.Dubhashi, "Extractive Summarization using Continuous Vector Space Models",Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), pp.31--39. Gothenburg, Sweden, Association for Computational Linguistics, (April), 2014.

[2] C.Ziqiang, W. Furu, D.Li, L.Sujian, Z. Ming, "Ranking with Recursive Neural Networks and Its Application to Multi-document Summarization", AAAI,. 2015.

[3] F.Lea, T.Ivan, P.Manfred, "A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge", 14th Conference of the European Chapter of the Association for Computational Linguistics,. 2014.

[4]F.Mohamed Abdel, "A hybrid machine learning model for multi-document summarization", springer, vol.40, no.4, pp.592–600, June. 2014.

[5] C.Ercan, K.Igor, "Multi-document summarization via Archetypal Analysis of the content-graph joint model", Springer, vol.41, no.3, pp.821–842, December. 2014.

[6] F.Rafael, C.Luciano de Souza, F. Frederico, L.Rafael Dueire, S.Gabriel de França, J. Steven Simske, F. Luciano, A multi-document summarization system based on statistics and linguistic treatment, Elsevier/Expert Systems with Applications vol.41, pp.5780–5787,. 2014.

[7] C.Richard Mc, M.Craig, , O.Iadh, "Incremental Update Summarization: Adaptive Sentence Selection based on Prevalence and Novelty", CIKM, Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp.301-310,. 2014.

[8]Lei, Li., Tao, Li., "An Empirical Study of Ontology-Based Multi-Document Summarization in Disaster Management", IEEE Transactions on Systems, Man And Cybernetics: Systems, 44, 2,.

[9] K.Atif, , Naomie, Salim., Yogan Jaya, Kumar., 2015, "A framework for multi-document abstractive summarization basedon semantic role labelling", Elsevier AppliedSoftComputing vol. 30, pp. 737–747,. 2014.

[10] C.Ercan, K. Igor, , "Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization", Elsevier/Expert Systems with Applications, vol.41, no.2, pp.535–543,. 2014.

[11] M.Martha, B.Susana, N.Clara, C. Carlos, L. Elizabeth, "Extractive single-document summarization based on genetic operators and guided local search", Elsevier/Expert Systems with Applications, vol.41, no.9, pp.4158–4169,. 2014.

[12] H.Ruifang, T.Jiliang, G.Pinghua, H.Qinghua, W.Bo, "Multi-Document Summarization via Group Sparse Learning", Elsevier Science Inc, vol.349, C, pp.12-24,. 2016.

[13] Y.Su, W.Xiaojun, "SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.22, no.12, 2014.

[14] O.Shun-Hsing, L. Chia-Han, V. Srinivasa Somayazulu, "On-Line Multi-View Video Summarization for Wireless Video Sensor Network", IEEE Journal of Selected Topics in Signal Processing, vol.9, no.1, Feb. 2015.

[15] Q.Changqin, Fuji, Ren., "Unsupervised product feature extraction for feature-oriented opinion determination", Elsevier, vol.272, pp.16–28,. 2014.