ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Indian Journal of Science and Technology, *Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/92307, July 2016*

MODC: Multi-Objective Distance based Optimal Document Clustering by GA

Annaluri Sreenivasa Rao[1*], S. Ramakrishna[2] and P. Chitti Babu[3]

[1]Department of Computer Science and Engineering, MRCE, Hyderabad - 500100, Telangana State, India; annaluri.rao@gmail.com
[2]Department of Computer Science,Sri Venkateswara University,Tirupathi - 517502, Andhra Pradesh, India; drsramakrishna@yahoo.com
[3]Annamacharya PG College of Computer Studies, Rajampet - 516126, Andhra Pradesh, India; drpcbit@gmail.com

Abstract

Background/Objective: Unsupervised learning of text documents is an essential and significant process of knowledge discovery and data mining. The concept, context and semantic relevancy are the important and exclusive factors in text mining, where as in the case of unsupervised learning of record structured data, these factors are not in scope. **Methods/ Statistical Analysis**: The current majority of benchmarking document clustering models is keen and relies on term frequency, and all these models are not considering the concept, context and semantic relations during document clustering. In regard to this, our earlier works introduced a novel document clustering approaches and one of that named as Document Clustering by Conceptual, Contextual and Semantic Relevance (DC3SR). The lessons learned from the empirical study of this contribution motivated us to propose aMulti-Objective Distance based optimal document Clustering (MODC) approach that optimizes resultant clusters using the well-known evolutionary computation technique called Genetic Algorithm. **Findings**: The significant contribution of this proposal is feature formation by concept, context and semantic relevance and optimizing resultant clusters by genetic algorithm. An unsupervised learning approach to form the initial clusters that estimates similarity between any two documents by concept, context and semantic relevance score and further optimizes by genetic algorithm is proposed. This novel method represents the concept as correlation between arguments and activities in given documents, context as correlation between meta-text of the documents and the semantic relevance is assessed by estimating the similarity between documents through the hyponyms of the arguments. The meta-text of the documents considered for context assessment contains the authors list, keywords list and list of document versioning time schedules. **Application/Improvements:**The experiments were conducted to assess the significance of the proposed model.The results obtained from experiments concluding that the MODC is performing exceptionally well under divergent document count and evincing the cluster formation accuracy as 97%. The dimensionality reduction by concept, context and semantic relevance is left for future enhancement of the proposed model.

Keywords: Concept Distance,Context Distance, Document Clustering, Meta-text,MODC, Multi Objective Distance Function, Text Mining, Unsupervised Learning

1. Introduction

The objective of the act of text mining is to retrieve critical patterns and associations of the text or documents, which are often unable or overlooked by domain experts due to dense in size and dimensionality. The outcomes of the text mining are more probabilistic due the explicit factors such as semantic, concept and context of the data that compared to other mining models[1,2].Hence the text mining is challenging in order to retrieve patterns, associations, classes and groups with high sensitivity.

The documents are usually grouped according to the similarity scope by supervised or unsupervised learning[3]. The supervised learning is the process of grouping documents in to known classes and the process of grouping the given documents without knowing the possible

*Author for correspondence

groups to be formed is known as unsupervised learning. The learning process that groups the given documents in to partially known classes and new classes formed according to the document similarity factors is known as semi-supervised learning strategy. Among these unsupervised learning is more critical and challenging[3]. In this strategy, documents are grouped according to their relevance scope assessed dynamically. The mining process called clustering is one among the unsupervised learning strategies.

Thereare several benchmarking text mining model in contemporary literature. Majority of the existing models are using the word frequency as a feature to perform mining, which often fails to delivers target patterns, associations, classes or groups with high sensitivity under semantic, context and concept factors. Hence the current research is potentially contributing to deliver optimal text mining models under semantic, concept and context factors. In this regard most of the contributions considerably justifyingthe semantic relevance, but the concept and context relevance is still a significant research objective since the ambiguity observed in text data framing that differs in concept and context[4–8]. This evincing the research scope to contribute optimal mining models that identifies the distance between the given documents by concept and context along with semantic relevance. The unsupervised learning strategy called clustering is a mining process that groups the documents by the similarity measure adopted. Segregating the given documents or text corpus in to optimal clusters is one critical research objective[9] that adopted here in this manuscript. In this line of research, the existing models are questionable for optimality of the clusters derived, number of clusters, process overhead and resource usage overhead[10].

The evolutionary computational techniques are evincing vital role in handling optimization issues[11–14].One of that is genetic algorithm[15], which is used to optimize the clusters delivered by DC3SR[16].

The proposed "Multi-objective distance based Optimal Document Clustering (MODC) by GA" is forming the clusters by using our earlier contribution called DC3SR[16].

The contemporarydata centric clustering algorithms[17–21] are not optimal to define labeled clusters. Hence these approaches are least significant for clustering documents[22].

The retrieval of optimal clusters and cluster count is the objective of the model called "variable string length genetic algorithm"[23]. The distance between documents is estimated by their semantic relevance, which is done by Davis-Bouldin index[24]. A hybrid strategy that combines GA and PSO was used[25] to cluster the given documents. The objective of this model that achieved successfully is to reduce the search space by using PSO and optimizing the clusters using GA.The other hybrid models[26,27]are the combination of Particle Swarm Optimization and Latent Semantic Index that are successfully reduced the search space and dimensionality. The other evolutionary models are KPSO[28] that combines PSO with K-Means[21], and FCPSO[28], which is the combination of PSO and Fuzzy C Means[21]. The FCPSO evinced as best that compared to KPSO about the optimal clusters defined. The other model[29], which is optimizing clusters by Bees Algorithm. Optimizing the document clusters using ACO[30] is another benchmarking cluster optimization model that randomized the Ant path track to discover optimal clusters.

Few other significant contributions[31,32]found in recent literature also significant towards text clustering. Nagaraj. R et al.,[31] are used the directed ridge regression to estimate the correlation between documents, which further used as distance metric to cluster the documents. Devi, S. S et al.,[32] devised a novel harmony search strategy that clusters text documents using constraint based approach. These contributions[31,32] also not considering the context, concept and semantic similarities in order to cluster the documents that found to be significant constraint in the context of this manuscript.

The constraints observed for all of these models is the least significant membership, cost or objective functions adopted to estimate the optimality of the resultant clusters. This is since, these adopted functions mainly relied on term frequency, combination of these frequent terms and their occurrence, and the optimality assessment is not considering the semantic, concept and context factors, which are more specific to optimize the clusters formed from the text corpus or the text content of the documents given. Hence the optimality of the clusters formed from these models is questionable. The other constraint observed is computational complexity, which is exponential to the given document countand the number of clustersformed. The other limit of these models is least significant to define optimal cluster count for document set with lessdiscrepancy.

In order to this, here we devised a document clustering techniquecalledMulti-objective distance based Optimal

Document Clustering (MODC) by GA, which is aimed to achieve optimality in cluster formation and cluster count, also target to achieve linearity in process overhead and resource utilization. The MODC is an extension to our earlier clustering technique called DC3SR. The critical factors of the text content such as semantic, concept, context are considered by the proposal to define optimal clusters. Further optimizing these clusters by GA that estimates the cluster fitness by the distance measure proposed (see sec 2.1.4). The progressive evolution strategy[33] is adopted to simplify the process overhead of the genetic algorithm.

2. Multi-Objective Distance based Optimal Document Clustering

Here in our proposed model, the given input documents will be clustered into minimum k clusters. Since the proposed model is an unsupervised learning approach, the documents is grouped according to the distance by concept, context and semantic relevance. The distance between any two documents is measured according to the heuristics called (i) concept distance, which is the inverse of number of common concepts between any two documents, (ii) the context distance is the sum of inverse of common authors, inverse of common topics and inverse of common versioning time frames and (iii) the semantic distance is the inverse common hyponyms between any given two documents. The detailed exploration of the distance measuring objective is explored in sections 2.1.4 and 2.2.4.

a. Multi-objective distance based Optimal Document Clustering (MODC):

Let $DS = \{d_1, d_2, ...d_{|DS|}\}$ is the document set and $MTS = \{(mt(d_1), mt(d_2), ..., mt(d_{|DS|})\}$ is the related meta-text set to be used to perform unsupervised learning using DC^{3}SR further optimizing the clusters by GA. The meta-text $mt(d_i) = \{\{\mathrm{a}\,l(d_i)\}, \{kw(d_i)\}, \{\mathrm{vt}(d_i)\}\}$ of each document d_i contains the authorslist $\{al(d_i)\}$, list of keywords $\{kw(d_i)\}$ and the list of document versioning times $\{vt(d_i)\}$.

2.1.1 Preprocessing

The preprocessing will be applied on each input document in sequence. The initial step of preprocessing is extracting the text content of the given document, then tokenizes, removes stop words, symbols and non-English words, characters from the characters. Further bipartite the words as arguments and activities and includes all hyponyms of the arguments. And then stemming up all these arguments and their hyponyms. Once these steps applied on all documents given, the process continues as follows.

The argument and activity pairs will be extracted from each document and prepares a two dimensional vector VDS with variable row length, such that each row represents an input document and the columns of each row is the activity and argument pairs found in the document that represented by respective row.

2.1.2 Finding Centroids

In regard to find the initial centroids Order the VDS documents by their column size and select top K rows as initial centroids represented as a set acs .

2.1.3 Clustering by K-Means

1. i. Each row $\{dv \forall dv \in VDS\}$ in sequence find the distance (see sec 2.1.4) between each centroid $\{c \forall c \in scs\}$ in sequence and move the row dv to the cluster $clstr(c_i)$ labeled by the centroid c_i such that the distance $dst(dv, c_i)$ between dv and c_i is minimal that compared to the distance found between dv and other centroids.
   ii. Find new optimal centroids for each cluster and add to set $tacs$
   iii. If acs and $tacs$ are identical then stop clustering process, else empty acs move all entries from $tacs$ to acs repeat above two steps.

The resultant clusters are said to be initial clusters and those will be used as input to the process of cluster optimization by GA (see sec 2.2).

2.1.4 Distance Function

- The semantic distance between document dv and centroid c can be measured as follows
- ○ Intersect the all terms and hyponyms of dv and c as $i_{sd}(dv, c)$
  ○ Find the ratio of number of hyponyms and terms in $i(dv, c)$ per number of hyponyms and terms in c as $\frac{|i_{sd}(dv, c)|}{|c|}$
  ○ Find the semantic distance $dst_{sd}(dv, c)$ as $1 - \frac{|i_{sd}(dv, c)|}{|c|}$
  ○ The concept distance between document dv and centroid c can be measured as follows

  - Intersect the argument and activity pairs of dv and c as $i_{cd}(dv,c)$
  - Find the ratio of number of pairs in $i_{cd}(dv,c)$ per number of pairs in c as $\frac{|i_{cd}(dv,c)|}{|c|}$
  - Find the distance $dst_{cd}(dv,c)$ as $1-\frac{|i_{cd}(dv,c)|}{|c|}$

- The context distance between document dv and centroid c can be measured as follows

  - Intersect the keywords $\{kw(dv)\}$ of document dv and keywords $\{kw(c)\}$ of document represented by centroid c as $i_{kw}(dv,c)$
  - Find the ratio of number of keywords in $i_{kw}(dv,c)$ per number of keywords in $\{kw(c)\}$ as $\frac{|i_{kw}(dv,c)|}{|\{kw(c)\}|}$
  - Find the distance $dst_{kw}(dv,c)$ as $1-\frac{|i_{kw}(dv,c)|}{|\{kw(c)\}|}$
  - Intersect the version update time frames of document dv and document represented by centroid c as $i_{vt}(dv,c)$
  - Find the ratio of number of time frames in $i_{vt}(dv,c)$ per number of time frames in $\{vt(c)\}$ as $\frac{|i_{vt}(dv,c)|}{|\{vt(c)\}|}$
  - Find the distance $dst_{vt}(dv,c)$ as $1-\frac{|i_{vt}(dv,c)|}{|\{vt(c)\}|}$
  - Intersect the authors list $\{al(dv)\}$ of document dv and authors list $\{al(c)\}$ of document represented by centroid c as $i_{al}(dv,c)$
  - Find the ratio of number of authors in $i_{al}(dv,c)$ per number of authors in $\{al(c)\}$ as $\frac{|i_{al}(dv,c)|}{|\{al(c)\}|}$
  - Find the distance $dst_{al}(dv,c)$ as $1-\frac{|i_{al}(dv,c)|}{|\{al(c)\}|}$
  - Then the context distance can be measured as $dst_{cod}(dv,c)=1-\left(dst_{kw}(dv,c)+dst_{al}(dv,c)+dst_{vt}(dv,c)\right)^{-1}$ //context role is defined by keyword relation, author relation and versioning update relation

- The overall distance between document dv and centroid c can be assessed as follows
  - $dst(dv,c)=1-(dst_{cd}(dv,c)+dst_{cod}(dv,c)+dst_{vt}(dv,c))^{-1}$ //The overall distance is defined by the ratio of concept, context and semantic distances respectively.

2.2 Cluster Optimization by Genetic Algorithm

2.2.1 Fitness function

The fitness of the newly formed clusters from crossover process is assessed as follows:

➢ For each document in the given input cluster, find distance with other documents in that cluster, which is done by the distance function devised (see sec 2.1.4)

```
1. Begin
2. $\forall_{i=1}^{|cl|}\{d_i \exists d_i \in cl\}$ begin //For-each document entry $d$ in given input cluster $cl$
     $dst_{cd}(d_i) \leftarrow \phi$,
     $dst_{sd}(d_i) \leftarrow \phi$,
     $dst_{kw}(d_i) \leftarrow \phi$,
     $dst_{vt}(d_i) \leftarrow \phi$,
     $dst_{al}(d_i) \leftarrow \phi$,
     $dst_{ctx}(d_i) \leftarrow \phi$
     // sets to store concept, semantic, topic, versioning, author list and context distances between document $d_i$ andother documents of the respective cluster $cl$
3.   $\forall_{j=1}^{|cl|}\{d_j \exists d_j \in cl \forall i \neq j\}$ begin //For-each document entry $d_j$ in given input cluster $cl$
4.     $dst_{cd}(d_i) \leftarrow 1-\frac{|(d_i \cap d_j)|}{|d_i|}$ //concept distance between document and centroid
5.     $dst_{sd}(d_i) \leftarrow 1-\frac{|(hn(d_i) \cap hn(d_j))|}{|hn(d_j)|}$ //semantic distance between document and centroid
6.     $dst_{kw}(d_i) \leftarrow 1-\frac{|\{kw(d_i)\} \cap \{kw(d_j)\}|}{|\{kw(d_i)\}|}$ //Topic distance between the document and centroid
7.     $dst_{vt}(d_i) \leftarrow 1-\frac{|\{vt(d_i)\} \cap \{vt(d_i)\}|}{|\{vt(d_i)\}|}$ //Versioning time frames Distance between the document and centroid
8.     $dst_{al}(d_i) \leftarrow 1-\frac{|\{al(d_i)\} \cap \{al(d_j)\}|}{|\{al(d_i)\}|}$ //Authors list distance between the document and centroid
9.     $dst_{ctx}(d_i) = 1-\left(dst_{kw}(d_i)+dst_{al}(d_i)+dst_{vt}(d_i)\right)^{-1}$ //context role is defined by keyword relation, author relation and versioning update relation
10.    $dst_{d_i} \leftarrow 1-(dst_{cd}(d_i)+dst_{ctx}(d_i)+dst_{vt}(d_i))^{-1}$ //The overall distance is defined by the ratio of concept, context and semantic distances respectively
11.  End For //of step 3
```

12. $\left\langle dst_{d_i}\right\rangle = \dfrac{\sum_{j=1}^{|dst_{d_i}|} dst_{d_i}(j)}{|dst_{d_i}|}$ // mean of the distances found between d_i and other documents of the cluster cl

13. $dst_{cl}\{d_i\} \leftarrow \left\langle dst_{d_i}\right\rangle$

14. End For //of step 2

15. Find the average of inverse of distances (similarities) $\overline{\left\langle dst_{cl}\right\rangle}$ observed for all documents in the given cluster cl as follow.

$$\overline{\left\langle dst_{cl}\right\rangle} = \frac{\sum_{j=1}^{|cl|} dst_{cl}(d_j)^{-1}}{|cl|}$$

16. Find mean absolute variance $\overline{\left\langle dst_{cl}\right\rangle}_{mad}$ of the inverse of distances observed for all documents in the cluster.

$$\overline{\left\langle dst_{cl}\right\rangle}_{mad} = \frac{\sqrt{\sum_{j=1}^{|cl|}\left(\overline{\left\langle dst_{cl}\right\rangle} - dst_{cl}(d_j)^{-1}\right)^2}}{|cl|}$$

17. If mean absolute variance is approximately 0, then finalize the cluster cl, else If $\overline{\left\langle dst_{cl}\right\rangle}$ is greater than the any of the parent cluster, then consider the new cluster.

2.2.2 Genetic Algorithm with Progressive Evolutions

Let CL be the set of clusters formed,which is further used in progressive evolutions of the GA as follows:

```
    $ls \leftarrow true$ //loop state initialized with Boolean value true
    GA-Main ( $CL$ ) Begin
        $tCL \leftarrow CL$ // create the copy of clusters $CL$ as $tCL$
        $\overline{CL} \leftarrow \phi$ //A set to store newly formed clusters
                //Find cross over points (common entries in given both clusters) that should not be the first entry in both clusters.

1. $\forall_{i=1}^{|tCL|}\{cl_i \exists cl_i \in tCL\}$ Begin
2. $\forall_{j=1}^{|tCL|}\{cl_j \exists cl_j \in tCL \wedge i \neq j\}$ Begin
3. $\forall_{p=1}^{|cl_i|}\{d_p \exists d_p \in cl_i\}$ Begin
4. $\forall_{q=1}^{|cl_j|}\{d_q \exists d_q \in cl_j\}$ Begin
5. If $(d_p \equiv d_q)$ Begin
        $cl_p \leftarrow \overleftarrow{cl_i}$    // new cluster $cl_p$ contains predecessor documents $\overleftarrow{cl_i}$ of $d_p$ in cluster $cl_i$
        $cl_p \leftarrow \overrightarrow{cl_j}$   followed by successor documents $\overrightarrow{cl_j}$ of $d_q$ in cluster $cl_j$
        $cl_q \leftarrow \overleftarrow{cl_j}$    // new cluster $cl_q$ contains predecessor documents $\overleftarrow{cl_j}$ of $d_q$ in cluster $cl_j$
        $cl_q \leftarrow \overrightarrow{cl_i}$   followed by successor documents $\overrightarrow{cl_i}$ of $d_p$ in cluster $cl_i$
        Afterwards estimate the fitness of these $cl_p$ and $cl_q$ using fitness function that devised in sec 3.2.1
        if$\left(\overline{\left\langle dst_{cl_p}\right\rangle}_{mad} \cong 0\right)$then $cl_p$ be the stable cluster
        else if$\left(\left(\overline{\left\langle dst_{cl_p}\right\rangle} > \overline{\left\langle dst_{cl_i}\right\rangle}\right) \| \left(\overline{\left\langle dst_{cl_p}\right\rangle} > \overline{\left\langle dst_{cl_j}\right\rangle}\right)\right)$then $\overline{CL} \leftarrow cl_p$
        If$\left(\overline{\left\langle dst_{cl_q}\right\rangle}_{mad} \cong 0\right)$then $cl_q$ be the stable cluster
        Else if $\left(\left(\overline{\left\langle dst_{cl_q}\right\rangle} > \overline{\left\langle dst_{cl_i}\right\rangle}\right) \| \left(\overline{\left\langle dst_{cl_q}\right\rangle} > \overline{\left\langle dst_{cl_j}\right\rangle}\right)\right)$ $\overline{CL} \leftarrow cl_q$
        End If // of step 5
        End For //of step 4
        End For //of step 3
        End For//of step 2
        End For//of step 1
        $CL \leftarrow CL \cup \overline{CL}$
        Then redefine clusters set $CL$ by removing clusters those are subset of other clusters if any, combine the clusters if they approximately equal, which is explore following.

1. $\forall_{i=1}^{|CL|}\{cl_i \exists c_i \in CL\}$ Begin
2. $\forall_{j=1}^{|CL|}\{cl_j \exists c_j \in CL \wedge i \neq j\}$ Begin
        If $(cl_i \subseteq cl_j)$ then
            $CL \leftarrow CL \setminus cl_i$ // delete $cl_i$ from $CL$
3. Else if $(cl_i \cong cl_j)$ then Begin // $c_i$ and $c_j$ approximately equal on threshold $\Delta$
            $cl_i \leftarrow cl_i \cup cl_j$ // new cluster that contains the all of $c_i$ and $c_j$
            $CL \leftarrow CL \setminus cl_j$ // delete $cl_j$ from $CL$
    End If//of step 9
    End For //of step 8
    End For //of step 7
    If $(CL \neq tCL)$ then GA-Main ( $CL$ )
    End Function GA-Main
```

The CL represents all stable and optimal clusters

3. Experimental Setup and Performance Analysis

The significance of the model MODC is explored through the experiments done on divergent datasets formed using research articles and their meta-text collected from different publishers.

3.1 Data set Exploration

In order to assess the scalability and clustering accuracy of the proposed multi-objective distance based optimal document clustering using genetic algorithm, we adopt the scientific research articles with preassigned cluster labels from divergent domains. The words used in content formation arewidely similar in majority of the domains opted for selected research articles but all of these input articles are differentiate by concept and context. The clusters formed by DC3SR are used as input to the cluster optimization done by progressive genetic algorithm.

3.2 Performance Analysis

The statistical metrics[34]such as precision, sensitivity, specificity and accuracy wereassessed to estimate the performance of the MODC and DC3SR. In regard to this, the true-positives (clusters that contains documents with similar pre assigned labels), false-positives (The clusters contain documents which are divergent according to the preassigned labels), true negatives (clusters contains documents those not related to any of the preassigned labeled documents) and false negatives (The clusters contain preassigned labeled documents as unrelated documents). To attain the true negatives and false negatives, we included set of deferential documents to the input document set, which are of a distinct cluster. Since the Genetic Algorithm is often influenced by process overhead and resource overhead, the process and resource overhead also assessed.

Since the assessment metrics computational and resource complexity also included in performance analysis, a computer with i5 processor and 4GB ram is used. The implementation is done using java and Statistical metrics analysis was done using explorative language R[35]. The input and obtained results were explored in Table 1.The observations indicating that MODC delivered significant clustering accuracy that compared to DC3SR (see Figure 1). In the available benchmarking approaches[4–8], the clusters are either influenced by the distance measurement by any one of the objective like term frequency, but the devised DC3SR and its extension MODC is assessing the distance between centroid and document through multiple objectives and that are concept, semantic and context. The distance by context is again a multifold of distance between topics, authors, and versioning time frames.The metric values to scale the DC3SR and MODC are explored in table 1.The MODC is robust under divergent count of input documents (see Figures 2 and 3).

The results (see Table 1, Figures1, 2 and 3) obtained, more particular, the cluster formation accuracy observed from experiments indicating that the proposed cluster

Table 1. Results obtained from DC3SR and MODC

	MODC	DC3SR
Total Number of Documents	4500	4500
Total Number of clusters formed	27	31
True Positives	3021	2901
False Positives	84	134
True Negatives	1353	1303
False Negatives	42	162
Precision	0.97	0.96
Sensitivity	0.99	0.95
Specificity	0.94	0.91
Accuracy	0.97	0.93

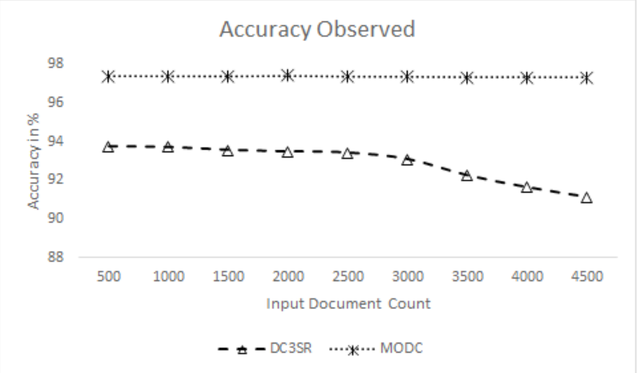

Figure 1. Cluster accuracy ratio observed for DC3SR and MODC for divergent count of input documents.

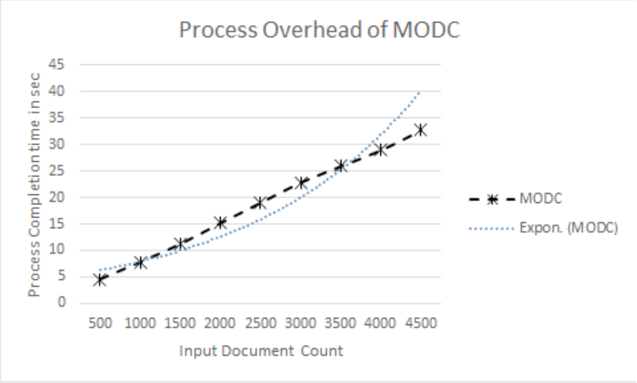

Figure 2. Process completion time observed for divergent count of documents.

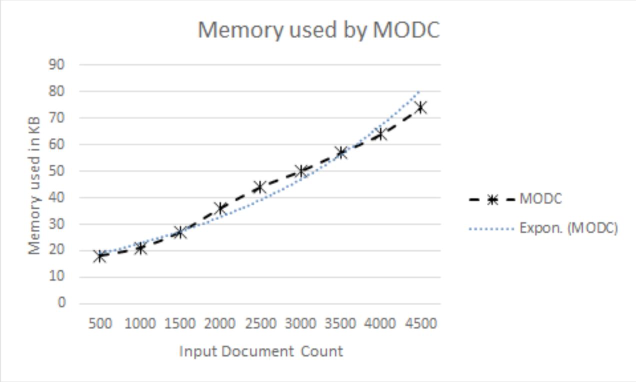

Figure 3. Memory usage ratio observed for divergent count of input documents.

optimization process by GA that applied on the clusters derived by DC3SR is significant that compared to DC3SR.

4. Conclusion and Future Work

The proposedmodel is an extension to our earlier unsupervised learning strategy called DC3SR[16]for document clustering under concept, context and semantic relevance as key factors.The proposed model is titled as Multi-objective distance based Optimal Document Clustering (MODC) by GA. The objective of the proposal is to improve the cluster formation accuracy by applying genetic algorithm on the initial clusters formed by DC3SR.The concept, context and semantic relevance of the documents to be clustered are considered as multi-objectives to estimate the distance between the documents. The results obtained from experiments concluding thatthe MODC is performingexceptionally well under divergent document count and evincing the cluster formation accuracy as 97%, which is substantially high that compared to the cluster formation accuracy of the DC3SR, which is 93%. The dimensionality reduction by concept, context and semantic relevance is left for future enhancement of the proposed model. The experimental results extremely motivating us to extend this MODC by considering fuzzy reasoning as fitness function of the Genetic Algorithm, which would also be the substantial future work.

5. References

1. Berkhin P. A survey of clustering data mining techniques. Grouping Multidimensional Data. Springer-Verlag;2006. p. 25–71.
2. Huang A. Similarity measures for text document clustering; 2008.
3. Hastie TT.Unsupervised learning. New York: Springer; 2009.
4. FungBCM, Wan K, Ester M. Hierarchical document clustering using frequent itemsets. SDM. 2003.
5. Sedding J, Kazakov D.Wordnet-based text document clustering. 3rd Workshop on Robust Methods in Analysis of Natural Language Data; 2004. p. 104–13.
6. LI Y, Chung SM. Text document clustering based on frequent word sequences. Proceedings of the. CIKM. Bremen, Germany; 2005 Oct 31 – Nov 5.
7. Zheng H-T, Kang B-Y, Kim H-G. Exploiting noun phrases and semantic relationships for text document clustering. Information Science. 2009; 179(13):2249–62.
8. Rao AS, Ramakrishna S. DCCR: Document Clustering by Conceptual Relevance as a factor of unsupervised learning. International Journal of Scientific and Engineering Research. 2014 Oct; 5(10):2229–5518.
9. Cui XG.A flocking based algorithm for document clustering analysis. Journal of Systems Architecture. 2006:505–15.
10. Narayanan NJ.Enhanced distributed document clustering algorithm using different similarity measures. IEEE Conference on Information and Communication Technologies (ICT); 2013. p. 545–50.
11. Castillo O, Martínez-Marroquín R, Melin P, Valdez F, Soria J. Comparative study of bio-inspired algorithms applied to the optimization of type-1 andtype-2 fuzzy controllers for an autonomous mobile robot. Information Sciences.2012 Jun; 192:19–38.
12. Kang F, Li J, Ma Z. Rosenbrock artificial bee colony algorithm for accurate global optimization of numerical functions. Information Sciences. 2011;181(16):3508–31.
13. Kundu D, Suresh K, Ghosh S, Das S, Panigrahi BK, Das S. Multi-objective optimization with artificial weed colonies. Information Sciences.2011; 181(12):2441–54.

14. Yang X. Nature-inspired metaheuristic algorithms. Luniver Press; 2008.
15. Haupt RL, Haupt SE. Practical genetic algorithms, second ed., John Wiley and Sons; 2008.
16. Rao AS, Ramakrishna S. DC3SR: Document Clustering by Concept, Context and Semantic Relevance as factors of unsupervised learning. International Journal of Applied Engineering Research.2015; 10(21):42213–18.
17. Carpineto C, Osiński S, Romano G, Weiss D. A survey of Web clustering engines. ACM Computing Surveys. 2009; 41(3):1–38.
18. Hammouda K. Web mining: clustering web documents a preliminary review;2001. p. 1–13.
19. Jain AK, Dubes RC. Algorithms for clustering data; 1988.
20. Jain AK, Murthy MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys. 1999; 31(3):264–323.
21. Steinbach M, Karypis, Kumar V. A comparison of document clustering techniques. ACM Boston;2000. p. 1–20.
22. Cobos CMV.Clustering of web search results based on the cuckoo search algorithm and balanced Bayesian information criterion. Information Sciences.2014:248–64.
23. Park WS. Genetic algorithm for text clustering based on latent semantic indexing. Computers and Mathematics with Applications.2009:1901–7.
24. Bolshakova N. Cluster validation techniques for genome expression data. Signal Processing.2003:825–33.
25. Natarajan KP. Hybrid PSO and GA models for document clustering. International Journal of Advanced Soft Computing Applications. 2010:2074–8523.
26. Hasanpour EH. PSO algorithm for text clustering based on latent semantic indexing. The Fourth Iran Data Mining Conference. Tehran, Iran; 2010.
27. Hasanzadeh ME. Text clustering on latent semantic indexing with Particle Swarm Optimization (PSO) algorithm. International Journal of the Physical Sciences.2012:116–20.
28. Karol VS. Evaluation of a text document clustering approach based on Particle Swarm Optimization. CSI Journal of Computing. 2012.
29. Nihal M,AbdelHamid MB. Bees algorithm-based document clustering. ICIT 2013 The 6th International Conference on Information Technology; 2013.
30. KayvanAzaryuon BF. A novel document clustering algorithm based on ant colony optimization algorithm. Journal of Mathematics and Computer Science. 2013:171–80.
31. Nagaraj R, Thiagarasu V. Correlation similarity measure based document clustering with directed ridge regression. Indian Journal of Science and Technology. 2014; 7(5):692–97.DOI: 10.17485/ijst/2014/v7i5/50135.
32. Devi SS, Shanmugam A. An integrated harmony search method for text clustering using a constraint based approach. Indian Journal of Science and Technology. 2015; 8(29).DOI: 10.17485/ijst/2015/v8i29/73986.
33. Layzer D. Genetic variation and progressive evolution. American Naturalist. 1980:809–26.
34. Tabachnick BG,Fidell LS, Osterlind SJ. Using multivariate statistics. Pearson; 2001.
35. Ihaka R. R: Alanguage for data analysis and graphics. Journal of Computational and Graphical Statistics. 1996:299–314.
36. Sajana T,Rani CMS, Narayana KV. A survey on clustering techniques for big data mining. Indian Journal of Science and Technology. 2016 Jan; 9(3). DOI: 10.17485/ijst/2016/v9i3/75971.
37. Hariharan R,Mahesh C, Prasenna P, Kumar RV. Enhancingprivacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journalof Science and Technology. 2016 Jan; 9(3). DOI:10.17485/ijst/2016/v9i3/86386.