Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INCREMENTAL EVOLUTIONARY GENETIC ALGORITHM BASED OPTIMAL DOCUMENT CLUSTERING (ODC)

A.KOUSAR NIKHATH[1], K.SUBRAHMANYAM[2]
[1]Research Scholar, Department of Computer Science and Engineering,
Koneru Lakshmaiah University, Guntur-522502, AP, India
[1]Asst .Professor, Department of Computer Science and Engineering,
VNR Vignana Jyothi Institute of Engineering and Technology, Bachupally, Hyderabad.
[2]Professor, Assoc Dean, Department of Computer Science and Engineering,
Koneru Lakshmaiah University, Guntur-522502, AP, India

Email: [1]kousarnikhath@vnrvjiet.in,[2]smdkodukula@kluniversity.in

ABSTRACT

Clustering is one of the phenomenal process towards information retrieval and knowledge discovery. Cluster optimality is still a questionable factor for current benchmarking clustering strategies. In particular document clustering is most sensible towards information retrieval and knowledge discovery, which is due to the curse of high volume and high dimensionality observed in recent times. In order to this many of document clustering models have been devised in recent times, but all of these models are questionable either the case of cluster optimality, process time complexity or adoptability. Henceforth, here we devised a deep machine learning approach called incremental evolutionary genetic algorithm based optimal document clustering (ODC) process. The experiments were done on documents dataset with curse of high dimensionality and volume. The results obtained from the experiments observed to be remarkably optimistic towards document clustering and also evincing the linearity in time complexity and memory usage.

Keywords: *Text Mining, Unsupervised Learning, Document Clustering, Cluster Optimization, Evolutionary Computation, ODC*

1 INTRODUCTION

Due to the exponential raise in internet usage towards document collection and storage such as journal and news archives, information retrieval and knowledge discovery from document corpus become monotonous task, which due to the curse of high volume in number of documents and high dimensionality of the document concepts[21,8] In order to this the documents need to be segregated into groups according to their similarity scope. This can be done by supervised or unsupervised learning[17]. Prior knowledge of the group identity labels helps to assess and group the relative documents is known as supervised learning, which often is not possible since the most of the times these labels are obsolete or unknown[17] In such situations documents should be classified by their relevance scope assessed dynamically, which is known as the process called unsupervised learning. Document clustering is one of such unsupervised learning strategy. The significant research objective in document clustering is the optimality of the clusters and cluster count[7]. Many of existing algorithms are questionable for either the case of cluster optimality or optimal cluster count or both[28].

Bio-inspired are playing phenomenal role to handle optimization issues [34],[13],[9],[42] One of that bio-inspired approach is genetic algorithm that can be used to resolve the optimization issues [37].The other few significant bio-inspired strategies are simulated annealing (SA)[10],the ant colony optimization (ACO)[25]and the particle swarm optimization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(PSO)[20],[3],[14],[19],[44]andmany more. This paper proposed an optimal document clustering approach that uses incremental evolutionary genetic algorithm[23] to optimize the clusters those initially formed. The ODC is assessing the hamming distance [31] of the documents to form the initial clusters. Further Incrementalevolutionary genetic algorithm [23] is used to optimize these clusters. The fitness function proposed is using Jaccard index [18]to estimate the optimality of the cluster.

The rest of the article is organized as follows. Section 2 is explored the associated models in document clustering. Section 3 is elaborating the proposed approach that followed by the section 4, which is discussing of experimental setup and performance analysis. The section 5 summarizing the contributions of the article.

2 RELATED WORK

Traditional clustering algorithms [35],[5],[15] [2],[27],[1] are data centric, which are not optimal in in need of identifying labeled clusters. These algorithms are not adoptable for document clustering, since they generally grouped into labeled clusters [6].
The variable string length genetic algorithm [36] is aimed to identify both the optimal clusters and cluster count. The traditional Genetic Algorithm is used to identify the semantic structure in order to define the optimal clusters. The fitness function is used to identify the semantic similarity of the documents in a given cluster, which is done by Davis-Bouldin index [4].

The combination of GA and PSO is proposed [29]for document clustering, which is using PSO to search in large spaces and the GA is used to define the optimal clusters for given document set. This hybrid model is evincing optimal performance to identify optimal cluster count under high diversity observed between given documents.

The document clustering algorithms [16], [12] are the combination of Particle Swarm Optimization and Latent Semantic Index. These are aimed to achieve optimality in search and reduce the dimensionality. The experiments indicating the advantage of these models to reduce dimensionality and search complexity.

The PSO based document clustering algorithms KPSO and FCPSO[40] are hybridizing the PSO with K-Means[39] Fuzzy C Means[39](Steinbach, 2000)[39]respectively. The clusters obtained from FCPSO are optimal than KPSO, K-means [40] and Fuzzy C Means (Steinbach, 2000). [35]

The other document clustering algorithm(Nihal M. AbdelHamid, 2013),[30] which is using Bees Algorithm to optimize the discovered clusters. The objective of the model is to discover the optimal cluster and the same is claimed by comparing with GA based document clustering(Park, 2009), [36]K-Means (Steinbach, 2000).[35],An ACO based document clustering algorithm[22]is another benchmarking evolutionary model. The Ant movement is completely randomized in order to span the search towards optimal cluster discovery. The theme of the warm intelligence (paraffin based search) taken from the ants is discarded, hence the model is least significant to claim as Ant colony approach and it is not much contradict to claim the search process is resembles the CUCKOO Search [6].

The observed computational complexities of all of these benchmarking models are nonlinear and cluster count and cluster optimality is questionable due to the curse dimensionality reduction by semantic relevance. All of these models are least significant to define optimal cluster count for document set with fewer divergence. Since the complexity of the traditional evolutionary strategies like GA, the process complexity is observed as nonlinear.

In order to this here we devised a novel optimal document clustering by incremental evolutionary genetic algorithm, which is considering the constraints called computational complexity and optimal cluster count and optimal clusters of the benchmarking models as objective of optimality.

Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In order to simplify the initial cluster formation, we adopted computationally much simplified approach called hamming distance [31] to identify the document similarity. Since the adopted genetic algorithm is based on incremental evolutions, the computational complexity expected to be linear. The dimensionality reduction by concept, context and semantic relevance is left for future enhancement of the proposed model.

3 INCREMENTAL EVOLUTIONARY GENETIC ALGORITHM BASED OPTIMAL DOCUMENT CLUSTERING.

The GA based optimal document clustering proposed here in this article is explored in this section. The Overall process is done in 4 stages and those are 1: dataset preprocessing, 2: initial cluster formation, 3: defining fitness function and optimizing clusters by Incremental Evolutionary Genetic Algorithm. The exploration of the process is done following subsections.

3.1 Dataset Preprocessing

For each document $\{d_i \exists d_i \in DS \wedge i = 1, 2, ..., |DS|\}$ Begin

Form a word vector $W(d_i) = \{w_1, w_2, ..., w_{|W(d_i)|}\}$

Remove noise (special symbols) and stop-words from the vector $W(d_i)$

End

3.2 Initial Cluster formation

Initial clusters will be formed for each document d_i and the other documents having hamming distance with d_i less than the given threshold hdt. The model of initial cluster formation is explored below:

1. For each word vector $\{W(d_i) \wedge i = 1, 2, 3, ... |DS|\}$ document Begin

   $c_i \leftarrow i$ // c_i is the cluster initialized with index i of the document d_i

   Find the hamming distance with other all documents as follows

2. For each word vector $\{W(d_j) \exists i \neq j \wedge j = 1, 2, 3, ... |DS|\}$
   Begin
   For a given two vectors $W(d_i) = \{wi_1, wi_2, ..., wi_{|W(d_i)|}\}$
   and
   $W(d_j) = \{wj_1, wj_2, ..., wj_{|W(d_j)|}\}$
   of size $|W(d_i)|$ and $|W(d_j)|$ respectively. Hamming Distance can be measured as follows
   Let $W \leftarrow \phi$ // is a vector of size 0
3. $foreach\ \{k \exists k = 1, 2, 3, ... \max(|W(d_i)|, |W(d_j)|)\}$
   Begin

   $$if\ (\{wi_k \exists wi_k \in W(d_i)\} - \{wj_k \exists wj_k \in W(d_j)\}) \equiv 0\ then$$

   $$W \leftarrow \{wi_k \exists Wi_k \in W(d_i)\} - \{wj_k \exists wj_k \in W(d_j)\}$$

   Else

   $$W \leftarrow 1$$

   End // end of loop in step 3

$$hd_{W(d_i) \leftrightarrow W(d_j)} = \frac{\sum_{l=1}^{|W|} W\{l\}}{\max(|W(d_i)|, |W(d_j)|)}$$

// $hd_{W(d_i) \leftrightarrow W(d_j)}$ is the hamming distance between $W(d_i)$ and $W(d_j)$, $W\{l\}$ is the l^{th} element of the vector W and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$|W|$ is the size of the vector W

If($hd_{W(d_i)\leftrightarrow W(d_j)} < hdt$) then

$c_i \leftarrow j$ // since the hamming distance between d_i and d_j is less than the threshold hdt index j of document d_j moved to the cluster c_i

End // end of loop in step 2

$C \leftarrow c_i$ // C be the set, contains the clusters formed

End //end of loop in step 1

Discard the clusters from C those are subset or equal to any of other cluster, merge the clusters those are approximately equal under given threshold. This will be done as follows

4. For each $\{c_i \exists c_i \in C \wedge i = 1,2,....|C|\}$ Begin
5. For each $\{c_j \exists c_j \in C \wedge i \neq j \forall j = 1,2,....|C|\}$ Begin

   If $(c_i \subseteq c_j)$ then

   $C \leftarrow \{C\} - \{c_i\}$ // discarding c_i from C

6. Else if $(c_i \square c_j)$ then Begin // c_i and c_j approximately equal on threshold Δ

   $c_k \leftarrow c_i \bigcup c_j$ // new cluster that contains the all of c_i and c_j

   $C \leftarrow c_k$ // adding new cluster c_k to C

   $C \leftarrow \{C\} - \{c_i\}$ //Discarding cluster c_i

   $C \leftarrow \{C\} - \{c_j\}$ // discarding cluster c_j

   End // of condition in step 6

   End// end of loop in step 5

   End//end of loop in step 4

3.3 Fitness function

The cluster fitness can be assessed as follows:

- Find Jaccard similarity of each document with all other documents of the cluster as follows.

  For a given cluster c_i

  $wv \leftarrow \phi$ //word vector that contains all words of the documents of the cluster c_i

  For each index $\{j \exists j \in c_i\}$ Begin

  $wv \leftarrow wv \bigcup W(d_j)$

  End

  For each index $\{j \exists j \in c_i\}$ Begin

  $$js_{c_i \leftrightarrow d_j} = \frac{|(W(d_j) \bigcap wv)|}{|(W(d_j) \bigcup wv|}$$

  End

- Find the average of Jaccard similarity $\langle js(c_i) \rangle$ observed for all documents in the given cluster c_i as follow.

$$\langle js(c_i) \rangle = \frac{\sum_{j=1}^{|c_i|} js_{c_i \leftrightarrow d_{c_i\{j\}}}}{|c_i|}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Find mean absolute distance $\langle js(c_i) \rangle_{mad}$ of the Jaccard similarity observed for all documents in the cluster.

$$\langle js(c_i) \rangle_{mad} = \frac{\sqrt{\sum_{j=1}^{|c_i|} \left(\langle js(c_i) \rangle - js_{c_i \leftrightarrow d_{c_i\{j\}}} \right)^2}}{|c_i|}$$

If mean absolute distance is approximately 0, then finalize the cluster c_i, else If $\langle js(c_i) \rangle$ is greater than the any of the parent chromosome, then consider the new cluster.

3.4 Incremental Evolutionary Genetic Algorithm

Each pair of clusters from C are considered as input to the incremental evolution process of the genetic algorithm. The strategy of incremental evolutions on the clusters applied as follows:

$ls \leftarrow true$ //loop state initialized with Boolean value true

While (ls) Begin

$tC \leftarrow C$ // clone the set of clusters C as tC

$\overline{C} \leftarrow \phi$ //An empty set of clusters

//Find the common documents as cross over points follows, such that the number of documents as predecessor and successor are not zero.

1. For each cluster $\{c_i \forall c_i \in C\}$ Begin
2. For each cluster $\{c_j \exists (c_j \in C \wedge j \neq i\}$ Begin
3. For each $\{k \exists k \in c_i\}$ Begin
4. For each $\{l \exists l \in c_j\}$ Begin

   //Split each cluster of the pair on cross-over point and form new cluster from the left part of the one cluster and right part of the other cluster as follows
5. If $(k \equiv l)$ Begin

Partite cluster c_i in to two at cross point k , and label the left part as $\overleftarrow{c_i}$ and right part as $\overrightarrow{c_i}$

Partite cluster c_j in to two at cross point l , and label the left part as $\overleftarrow{c_j}$ and right part as $\overrightarrow{c_j}$

Form cluster c_p by connecting left part of c_i and right part of c_j

Form cluster c_q by connecting left part of c_j and right part of c_i

//Find fitness of each new cluster as explored in sec 3.3
Assess fitness of the clusters c_p and c_q (see sec 3.3)

if$\left(\langle js(c_p) \rangle_{mad} \cong 0 \right)$ finalize the cluster c_q

else if
$$\left(\langle js(c_p) \rangle > \langle js(c_i) \rangle \right) || \left(\langle js(c_p) \rangle > \langle js(c_j) \rangle \right)$$
$\overline{C} \leftarrow c_p$

if$\left(\langle js(c_q) \rangle_{mad} \cong 0 \right)$ finalize the cluster c_p

else if
$$\left(\langle js(c_q) \rangle > \langle js(c_i) \rangle \right) || \left(\langle js(c_q) \rangle > \langle js(c_j) \rangle \right)$$
$\overline{C} \leftarrow c_q$

End //end of condition in step 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

End //end of loop in step 4

End //end of loop in step 3

End //end of loop in step 2

End //end of loop in step 1

$C \leftarrow C \bigcup \overline{C}$

Discard the clusters from C those are subset or equal to any of other cluster, merge the clusters those are approximately equal under given threshold. This will be done as follows

7. For each $\{c_i \exists c_i \in C \wedge i = 1,2,....|C|\}$ Begin
8. For each $\{c_j \exists c_j \in C \wedge i \neq j \forall j = 1,2,....|C|\}$ Begin

If $(c_i \subseteq c_j)$ then

$C \leftarrow \{C\} - \{c_i\}$ // discarding c_i from C

9. Else if $(c_i \square c_j)$ then Begin // c_i and c_j approximately equal on threshold Δ

$c_k \leftarrow c_i \bigcup c_j$ // new cluster that contains the all of c_i and c_j

$C \leftarrow c_k$ // adding new cluster c_k to C

$C \leftarrow \{C\} - \{c_i\}$ //Discarding cluster c_i

$C \leftarrow \{C\} - \{c_j\}$ // discarding cluster c_j

End // of condition in step 9

End// end of loop in step 8

End//end of loop in step 7

If $\left(C \cong tC\right)$ then $ls \leftarrow false$

End // end of the while loop (completion of the GA process)

The C contains set of all finalized clusters

4 EXPERIMENTAL STUDY AND PERFORMANCE ANALYSIS

4.1 The Dataset

The objective of the model is to perform the optimal document clustering using incremental evolutionary genetic algorithm (citation required). To assess the scalability and clustering accuracy, we adopt the manually labeled of scientific research articles from divergent domains. The terms mostly similar in most of these domains but the articles are divergent in terms of concepts like wired, wireless, communication and ad hoc networks, data mining, data science, knowledge discovery and information retrieval and same impact can observe even in distribute computing as terms used are similar but articles are divergent under concepts like cloud computing, grid computing and parallel computing. We initially cluster the documents by their concept relevance and obtained prior knowledge of the possible clusters and documents of those clusters.

4.2 Assessment metrics and strategy

The metrics that we considered to assess the accuracy of the clusters formed by ODCare precision, sensitivity, specificity and accuracy, which are estimated by using true-positives, false-positives, true negatives and false negatives.In order to obtain the true negatives and false negatives, we considered set of reverential documents, which can be grouped as separate cluster.

The adopted model is an evolutionary strategy, which is often complexed towards process and resource utilization. Hence the time complexity and process complexity of the proposed algorithm also being assessed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.3 Experimental setup and Results

Since the assessment metrics computational and resource complexity also included in performance analysis, a computer with i5 processor, 4GB Ram and Nvidia 4GB graphics card[33]used. The implementation is done in CUDA. [32]Statistical metrics analysis is done using explorative language R [18]. The input and obtained results are explored in Table 1.

Total Number of Documents	Labeled: 1021, unlabeled: 479
Total Number of actual clusters	14 from labeled documents
Total Number of Initial Clusters	67 from all labeled and unlabeled documents
Total Number of Predicted Clusters by ODC	20 from all labeled and unlabeled documents
True Positives	1007
False Positives	28
True Negatives	451
False Negatives	14
Precision	0.972947
Sensitivity	0.986288
Specificity	0.969892
Accuracy	0.972

Table 1: Input and observed metric values from the experiments

The performance of the model is assessed on a document set of size 1500. Among these documents 1021 documents already with known labels, which are notice to be fit into 14 clusters. In order to assess the accuracy, the documents of size 479 of divergent concepts, which are far different from the concepts of the labeled documents, are considered. The labeled documents are considered as positives and unlabeled documents are considered as negatives towards the actual clusters defined. Further the clusters predicted by ODC are assessed, which is based on the association of the documents given. The Metric values indicating that prediction of document associability under Jaccard index (document relevancy to the cluster) by the ODC is phenomenally significant (precision is 0.972947). The true positive Rate that indicates the true prediction of ratio of documents for relevant cluster is also considerably high (sensitivity is 0.986288) for ODC. The prediction rate of irrelevant documents to the defined clusters is also remarkably high (specificity is 0.969892). The overall document clustering optimality by ODC is observed as thebest, since the 97% of the documents grouped into relevant labels under the given input and experimental setup (accuracy is 0.972).

The computational complexity and resource cost is also assessed, which is done under divergent count of initial clusters as input. The time complexity observed to be linear for given initial clusters as input (see fig 1). The memory usage of Incremental evolutionary genetic algorithm is also being noticed as linear for given input clusters (see fig 2).

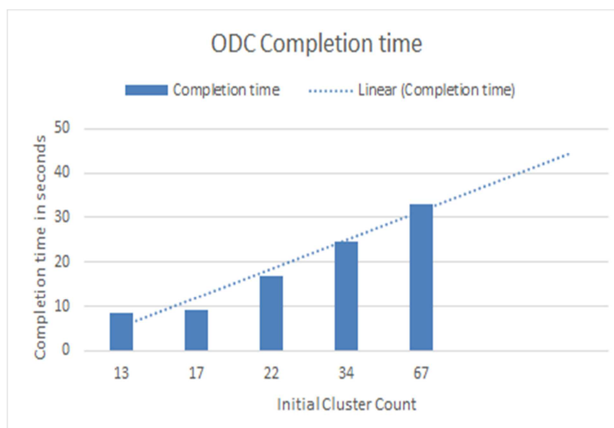

Figure 1: Incremental Evolutionary Genetic Algorithm Completion Time Observed For Divergent Count Of Input Clusters

Journal of Theoretical and Applied Information Technology
31st May 2016. Vol.87. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

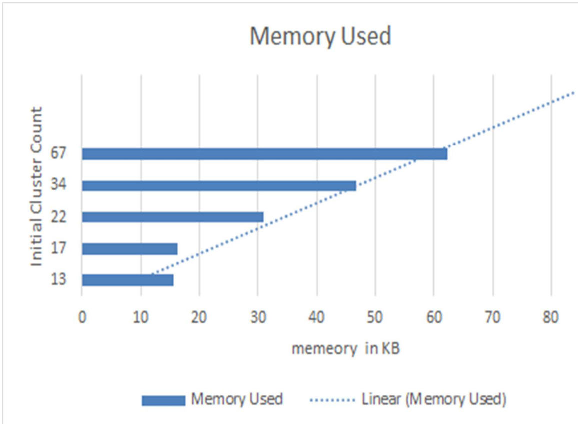

Figure 2: Memory Used For Incremental Evolutionary Genetic Algorithm

5 CONCLUSION

Optimal Document Clustering (ODC) by Incremental Evolutionary Genetic Algorithm is proposed in this article. The overall procedure is done in three level hierarchies. First level of the ODC is the formation of the initial clusters, in which hamming distance is used to identify the term based similarity between documents.

Further the fitness function is defined that estimates the fitness of the cluster using Jaccard index. The initial clusters further optimized using incremental evolutionary genetic algorithm, which is the third level of the ODC.

Experiments are done in the context of assessing the accuracy of the ODC by statistical metrics called precision, sensitivity, specificity and accuracy. The time complexity and memory usage also assessed in order to estimate the scalability of the incremental evolutionary genetic algorithm.

In order to this a set of documents that already labeled manually are taken as input. The accuracy, robustness and scalability of the ODC are phenomenally significant. Unlike traditional Genetic algorithm, the incremental evolutionary genetic algorithm is observed to be linear in time complexity and resource utilization. The performance analysis of the results obtained from the experimental setup motivates us to stretch the research further to perform the document clustering by concept, context and semantic relevance of the documents. Also our future contributions can be the optimized document clustering by deep machine learning through evolutionary computational strategies, whichreduces the dimensionality by concept, context and semantic relevance.

REFERENCES

[1] A.K. Jain, M. M. (1999). Data clustering: a review. *ACM Comput.*, 264-323.

[2] A.K. Jain, R. D. (1988). Algorithms for clustering data.

[3] B.Y. Qu, J. L. (2012). Niching particle swarm optimization with local search for multi-modal optimization. *Information Sciences*, 131–143.

[4] Bolshakova, N. &. (2003). Cluster validation techniques for genome expression data. *Signal processing*, 825-833.

[5] C. Carpineto, S. O. (2009). A survey of Web clustering engines. *ACM Comput*, 1-38.

[6] Cobos, C. M.-C.-M.-V. (2014). Clustering of web search results based on the cuckoo search algorithm and balanced Bayesian information criterion. *Information Sciences*, 248-264.

[7] Cui, X. G. (2006). A flocking based algorithm for document clustering analysis. *Journal of systems architecture*, 505-515.

[8] D. Hsek, J. P. (2009). Web data clustering. *Studies in Computational Intelligence, springer*, 325–353.

[9] D. Kundu, K. S. (2011). Multi-objective optimization with artificial weed colonies. *Information Sciences*, 2441–2454.

[10] D.S. Johnson, C. A. (1989). Optimization by simulated annealing. An experimental evaluation. Part I. Graph partitioning,. *Operations Research*, 865–872.

[11] E. Rashedi, H. N.-p. (2009). GSA: a gravitational search algorithm. *Information Sciences*, 2232–2248.

[12] EisaHasanzadeh, M. r. (2012). Text clustering on latent semantic indexing with

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

particle swarm optimization (PSO) algorithm. *International Journal of the Physical Sciences*, 116 – 120.

[13] F. Kang, J. L. (2011). Rosenbrock artificial bee colony algorithm for accurate global optimization of numerical functions. *Information Sciences*, 3508–3531.

[14] H. Shah_Hosseini. (2007). Problem solving by intelligent water drops. *IEEE Congress on Evolutionary Computation*, 3226–3231.

[15] Hammouda, K. (2001). Web Mining: Clustering Web Documents A Preliminary Review. 1-13.

[16] Hasanpour, E. H. (2010). PSO Algorithm for Text Clustering Based on Latent Semantic Indexing. *The Fourth Iran Data Mining Conference.* Tehran, Iran.

[17] Hastie, T. T. (2009). *Unsupervised learning.* New Yark: springer.

[18] Ihaka, R. &. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 299-314.

[19] J. Gonzalez, D. P.-S. (2011). A new metaheuristic bat-inspired algorithm, in: Nature Inspired Cooperative. *Springer*, 65–74 .

[20] J. Kennedy, R. E. (1995). Particle swarm optimization. *IEEE International Conference on Neural Networks*, (pp. 1942–1948).

[21] K. Cios, W. P. (n.d.). *Data Mining-Methods for Knowledge Discovery.* Kluwer Academic Publishers.

[22] Kayvan Azaryuon, B. F. (2013). A Novel Document Clustering Algorithm Based on Ant Colony Optimization Algorithm. *Journal of mathematics and computer Science*, 171-180.

[23] Layzer, D. (1980). Genetic variation and progressive evolution. *American Naturalist*, 809-826.

[24] M. Bramer, R. E.-S. (2010). Firefly algorithm, Lévy flights and global optimization, Research and Development in Intelligent Systems. *Springer*, 209–218.

[25] M. Dorigo, C. B. (2005). Ant colony optimization theory: a survey,. *Theoretical Computer Science*, 243–278.

[26] M. Fathian, B. A. (2007). Application of honey-bee mating optimization algorithm on clustering. *Applied Mathematics and Computation*, 1502–1513.

[27] M. Steinbach, G. K. (2000). A comparison of document clustering techniques,. *ACM Boston*, 1-20.

[28] Narayanan, N. J. (2013). Enhanced distributed document clustering algorithm using different similarity measures. *IEEE Conference on Information & Communication Technologies (ICT)*, (pp. 545-550).

[29] Natarajan, K. P. (2010). Hybrid PSO and GA Models for Document Clustering. *International Journal of Advanced Soft Computing Applications*, 2074-8523.

[30] Nihal M. AbdelHamid, M. B. (2013). BEES ALGORITHM-BASED DOCUMENT CLUSTERING. *ICIT 2013 The 6th International Conference on Information Technology.*

[31] Norouzi, M. F. (2012). Hamming distance metric learning. *Advances in neural information processing systems*, 1061-1069.

[32] Nvidia. (2008). C. U. D. A. Programming guide.

[33] NVIDIA. (2015). *PNY-NVIDIA-GeForce-GTX-960-4GB-XLR8.pdf.* Retrieved from pny.com: https://www.pny.com/File%20Library/Support/PNY%20Products/Resource%20Center/Graphics%20Cards/GTX%20900%20Series/PNY-NVIDIA-GeForce-GTX-960-4GB-XLR8.pdf

[34] O. Castillo, R. M.-M. (2012). Comparative study of bio-inspired algorithms applied to the optimization of type-1 and. *Information Sciences* , 19–38.

[35] P. Berkhin, J. K. (2006). A Survey of Clustering Data Mining Techniques, in: Grouping Multidimensional Data. *Springer-Verlag,*, 25-71.

[36] Park, W. S. (2009). Genetic Algorithm for text clustering based on latent semantic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

indexing. *Computers and Mathematics with applications*, 1901-1907.

[37] R.L. Haupt, S. H. (2004). *Practical Genetic Algorithms,*. second ed., John Wiley & Sons.

[38] .Real, R. &. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic biology*, 380-385.

[39] Steinbach, M. K. (2000, august). A comparison of document clustering techniques. *KDD workshop on text mining*, pp. 525-526.

[40] Stuti Karol, V. (2012). Evaluation of a Text Document Clustering Approach based on Particle Swarm Optimization. *CSI Journal of Computing*.

[41] Y. Marinakis, M. M. (2011). Honey bees mating optimization algorithm for the Euclidean traveling salesman problem. *Information Sciences*, 4684–4698.

[42] Yang, X. (2008). *Nature-Inspired Metaheuristic Algorithms.* Luniver Press,.

[43] Yang, X. S. (2009). Cuckoo search via Lévy flights. *World Congress on Nature & Biologically Inspired Computing*, (pp. 210-214).

[44] Z. Geem, X.-S. Y. (2009). Harmony search as a metaheuristic algorithm, in: Music-Inspired Harmony Search Algorithm. *Springer*, 1-14.