**CHAPTER ONE**

**INTRODUCTION**

* 1. **BACKGROUND OF THE STUDY**

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics and other sciences. Optimization forms an important part of our day-to-day life. The task of optimization is that of determining the values of a set of parameters so that some measure of optimality is satisfied, subject to certain constraints. This task is of great importance to many professionals (Van den, 2006).

Optimization problems arise in a variety of fields, including engineering design, operational research, information science and related areas. Effective and efficient optimization algorithms are always needed to tackle increasingly complex real world optimization problems. Stochastic optimization algorithms, such as genetic algorithm (GA) particle swarm optimization (PSO) ant colony optimization (ACO), biogeography-based optimization (BBO), harmony search (HS), and artificial bee colony (ABC) algorithm, have been shown to be successful in dealing with many optimization problems.

Artificial Bee Colony algorithm, developed by Karaboga based on simulating the foraging behavior of honey bee swarm. Numerical comparisons demonstrated that the performance of ABC algorithm is competitive to other population-based algorithms with an advantage of employing fewer control parameters (D. Karaboga, B. Basturk, 2005). Due to its simplicity and ease of implementation, ABC algorithm has captured much attention and has been applied to solve many practical optimization problems since its invention in 2005.

In practice, the exploration and exploitation contradicts to each other. In order to achieve good performances on problem optimizations, the two abilities should be well balanced. While, we it is good that the solution search equation of ABC algorithm which is used to generate new candidate solutions based on the information of previous solutions, is good at exploration but poor at exploitation, which results in the above two insufficiencies.

Therefore, accelerating convergence speed and avoiding the local optima have become two most important and appealing goals in ABC research. A number of variant ABC algorithms have, hence, been proposed to achieve these two goals (B. Alatas, 2014). However, so far, it is seen to be difficult to simultaneously achieve both goals. For example, the chaotic ABC algorithm (CABC3) in (S. Rahnamayan, et al., 2008) focuses on avoiding the local optima, but brings in a more extra function evaluations in chaotic search as a result.

To achieve both goals, inspired by DE, ABC algorithm called ABC/best, which is based on that each bee searches only around the best solution of the previous iteration to improve the exploitation. In addition, to enhance the global convergence, when producing the initial population and scout bees, both chaotic systems and opposition-based learning method are employed. The rest of this paper is organized as follows.

Non-negative matrix factorization (NMF or NNMF), also non-negative matrix approximation on the other hand is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. Also, in applications such as processing of audio spectrograms or muscular activity, non-negativity is inherent to the data being considered. Since the problem is not exactly solvable in general, it is commonly approximated numerically (Inderjit. Dhillon & Suvrit, 2015)

NMF finds applications in such fields as computer vision, document clustering, chemometrics, audio signal processing and recommender systems (Yannis, 2011). Many standard NMF algorithms analyze all the data together; i.e., the whole matrix is available from the start. This may be unsatisfactory in applications where there are too many data to fit into memory or where the data are provided in streaming fashion. One such use is for collaborative filtering in recommendation systems, where there may be many users and many items to recommend, and it would be inefficient to recalculate everything when one user or one item is added to the system. The cost function for optimization in these cases may or may not be the same as for standard NMF, but the algorithms need to be rather different.

The propose project Hybrid Text summarization using Artificial Bee Colony (ABC) and Non negativity matrix factorization (NMF) is a software that is aim at helping to optimizing text to a better but minimum useful context provision it carries the sense convey to manipulated the text and still give same result, without diminishing the grammar and lexical quality nor the semantic relationship between each terms in a giving string.

**1.2 OBJECTIVES OF THE STUDY**

The Text summarization software is a desktop based application has been targeted to adopt the following features.

1. To develop an application that seeks to provide quality and accurate summary of text or sentences.
2. Aims to provide a better means of optimizing sentence that can be further use.
3. Query local and global data in other to know the minimum point laps to maximum point of repetitions.
4. To develop an interactive application for easy access.
5. To develop an application that can be further integrated or developed for further research.
   1. **STATEMENT OF THE PROBLEMS**

The Text summarization using ABC and NMF describes students who must give answers between N answers. Each answers having it unique difference. Each of those unique differences between the answers has one or more weights attached. Most important, it has applications in science and engineering. An efficient solution to this problem reduces production costs for the manufacturer.

The basic problems of the existing system are listed as follow:

1. The inconsistency of the previous system leading to an in appropriate text summarization and dissimilarity in distances.
2. Inconsistency in summarization provide by the same application on processing the same document.
3. Non flexibility of processed documents that can be better utilize at appropriate result.

**1.4 AIM OF THE STUDY**

The core aim of this system: Text summarization using Artificial Bee Colony (ABC) and Non negativity matrix factorization (NMF) is to help in optimizing text to a better but minimum useful context provision it carries the sense convey to manipulated the text and still give same result, without diminishing the grammar, lexical quality nor the semantic relationship between each terms in a giving document.

**1.5 SIGNIFICANCES OF STUDY**

Text summarization will require being very precise on cost effectiveness to produce an efficient optimized text combinatorial. Hence, the crucial points that this research Text summarization using ABC and NMF.

* The project will provide to users the total time taken to summarize document.
* The project will also provide a breakdown of how each step in summarizing document are achieved.
* This project will provide several alternative summarizations to document base on combination of algorithm specified (in this case ABC and NMF).
* This application will be a lot easier to use i.e. more interactive compare to the previous existing ones.

**1.6 SCOPE AND LIMITATIONS OF THE STUDY**

The current system can be implemented with some other related system by providing appropriated data to the system for further process, analysis or decision making. This system may be integrated with third party application either for application development or research work that required a working system like these.

1. This project is a Text Summarization using ABC and NMF i.e. this project does not automatically address any error from the used or referenced libraries), plug-ins sources.
2. This project is not the solution to every problem on text optimization, and it does not promise total security except frequent improvement on it.
3. This project is limited to the information gathered during the development of the project.
4. The project is a good tool for continuity and maintainability.
   1. **DEFINITION OF TERMS**

* **Document summarization:** is the process of shortening a text document with software, in order to create a summary with the major points of the original document.
* **Artificial Bee Colony techniques:** In computer science and operations research, the **Artificial Bee Colony algorithm** (ABC) is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs.
* **E-Assessment:** Electronic assessment is the use of technology to manage and deliver assessment
* **Optimization:** This is the process of minimizing or maximizing opportunities as best suit to the situation.
* **Pheromones** : are chemical substances which attract other bee searching for food.
* **Computer system:** These is an electronic device that is capable of accepting data, process it following a preset logic and generate output as require by the user.
* **Program, software or application:** A set of logical instruction combined together to get data and perform a specific task on a given data.
* **Input:** Data supplied to the computer for processing.
* **Output:** The result of a processed data.
* **Data:** A raw fact yet to be processed.
* **Information:** Data that has been proceed.

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.1 THEORITICAL BACKGROUND**

The theory of optimization has its roots in the isoperimetric problem faced by Queen Dido in 1000 BC. She procured for the founding of Carthage the largest area of land that could be surrounded by the hide of a bull. From the hide she made a rope, which she arranged in a semicircle with the ends against the sea. Queen Dido’s intuitive solution was correct. But it was many centuries before a formal proof was presented, and the mathematical and systematic solution to this problem proved to be a very difficult problem in the calculus of variations. The calculus of variations essentially handles problems where the decision variable is a vector.

sThe calculus of variations, or the first systematic theory of optimization, was born on June 4, 1694, when John Bernoulli posed the Brachistochrone (Greek for “shortest time”) problem, and publicly challenged the mathematical world to solve it. The problem posed was, “What is a slide path down which a frictionless object would slip in the least possible time?” Earlier attempts to solve this problem were made by many well-known scientists including Galileo who proposed the solution to be a circular arc, an incorrect solution, and Leibnitz who presented ordinary differential equations without solving them. Then John Bernoulli proved the optimal path to be a cycloid. From that point, efforts continued in the area of the calculus of variations, leading to the study of multiple integrals, differential equations, control theory, problem transformation and so on. Although this research mainly involved theory and analytical solutions, it formed the basis for numerical optimization developed during and after World War II. World War II made scientists aware of numerical optimization and solutions to physics and engineering problems. In 1947 Dantzig proposed the simplex algorithm for linear programming problems. Necessary conditions were presented by Kuhn and Tucker in the early 1950s, which formed a focal point for nonlinear programming research. Now, numerical optimization techniques constitute a fundamental part of theoretical and practical science and engineering (Diwekar, 1995).

Optimization has pervaded all spheres of human endeavor. Although optimization has been practiced in some form or other from the early prehistoric era, this area has seen progressive growth during the last five decades. Modern society lives not only in an environment of intense competition but is also constrained to plan its growth in a sustainable manner with due concern for conservation of resources. Thus, it has become imperative to plan, design, operate, and manage resources and assets in an optimal manner. Early approaches have been to optimize individual activities in a standalone manner, however, the current trend is towards an integrated approach: integrating synthesis and design, design and control, production planning, scheduling, and control. The functioning of a system may be governed by multiple performance objectives. Optimization of such systems will call for special strategies for handling the multiple objectives to provide solutions closer to the systems requirement. Uncertainty and variability are two issues which render optimal decision making difficult. Optimization under uncertainty would become increasingly important if one is to get the best out of a system plagued by uncertain components. These issues have thrown up a large number of challenging optimization problems which need to be resolved with a set of existing and newly evolving optimization tools (Daniel D. Lee & H. Sebastian Seung, 2011).

Optimization theory had evolved initially to provide generic solutions to optimization problems in linear, nonlinear, unconstrained, and constrained domains. These optimization problems were often called mathematical programming problems with two distinctive classifications, namely linear and nonlinear programming problems. Although the early generation of programming problems were based on continuous variables, various classes of assignment and design problems required handling of both integer and continuous variables leading to mixed integer linear and nonlinear programming problems (MILP and MINLP). The quest to seek global optima has prompted researchers to develop new optimization approaches which do not get stuck at a local optimum, a failing of many of the mathematical programming methods. Genetic algorithms derived from biology and simulated annealing inspired by optimality of the annealing process are two such potent methods which have emerged in recent years. The developments in computing technology have placed at the disposal of the user a wide array of optimization codes with varying degrees of rigor and sophistication. The challenges to the user are many fold. How to set up an optimization problem? What is the most suitable optimization method to use? How to perform sensitivity analysis? An intrepid user may also want to extend the capabilities of an existing optimization method or integrate the features of two or more optimization methods to come up with more efficient optimization methodologies.

**2.2 REVIEW OF RELEVANT LITERATUES**

Various works has been done by different researchers on a variety of research topics using NMF. ABC is also a favorite choice for researchers for solving different kind of optimization problems. These two topics are relatively new than the CREDIT scoring problem. For a long time people are dealing with credit scoring. High investor risk and low consumer satisfaction was the root cause for required improvement in this area. Now researchers are using different techniques for reducing risk and improving satisfaction and they are also successful up to some extent.

**2.2.1 Image Processing**

Nonnegative Matrix Factorization can trace its history back to 1970’s, but has attracted lots of attention due to the research of (Lee & Seung, 2012). In their works, the model was applied to image processing successfully. In image processing, the data can be represented as an n×m nonnegative matrix X, each column of which is an image described by n nonnegative pixel values. Then NMF model can find two factor matrices F and G such that X=FG. F is the so-called basis matrix since each column can be regarded as a part of the whole such as nose, ear or eye, etc. for facial image data. G is the coding matrix and each row is the weights by which the corresponding image is reconstructed as the linear combination of the columns of F.

**2.2.2 Clustering**

One of the most interesting and successful applications of NMF is to cluster data such as text, image or biology data, i.e. discovering patterns automatically from data. Given a nonnegative n×m matrix X, each column of which is a sample and described by n features, NMF can be applied to find two factor matrices F and G such that X = FG, where F is n×r and G is m×r, and r is the cluster number. Columns of F can be regarded as the cluster centroids while G is the cluster membership indicator matrix. In other words, the sample i is of cluster k if Gik is the largest value of the row Gi. The good performance of NMF in clustering has been validated in several different fields including bioinformatics (tumor sample clustering based on microarray data, [3]), community structure detection of the complex network ([4]) and text clustering ([5, 6, 7]).

**2.2.3 Semi Supervised Clustering**

In many cases, some background information concerning the pair wise relations of some samples are known and we can add them into the clustering model in order to guide the clustering process. The resulting constrained problem is called semi-supervised clustering. The incorporated user provided constraints in data clustering (Chen and Rege, 2015).

**2.2.4 Bi-Clustering (Co-Clustering)**

Bi-clustering was recently introduced by (Cheng & Church, 2014) for gene expression data analysis. In practice, many genes are only active in some conditions or classes and remain silent under other cases. Such gene-class structures, which are very important to understand the pathology, cannot be discovered using the traditional clustering algorithms. Hence it is very necessary to develop bi-clustering models/algorithms to identify the local structures. Bi-clustering models/algorithms are different from the traditional clustering methodologies which assign the samples into specific classes based on the genes’ expression levels across ALL the samples, they try to cluster the rows (features) and the columns (samples) of a matrix simultaneously. In other words, the idea of bi-clustering is to characterize each sample by a subset of genes and to define each gene in a similar way. As a consequence, bi-clustering algorithms can select the groups of genes that show similar expression behaviours in a subset of samples that belong to some specific classes such as some tumour types, thus identify the local structures of the microarray matrix data. Binary Matrix Factorization (BMF) was presented for solving bi-clustering problem: the input binary gene-sample matrix X is decomposed into two binary matrices F and G such that X=FGT. The binary matrices F and G can explicitly designate the cluster memberships for genes and samples. Hence BMF offers a framework for simultaneously clustering the genes and samples.

**2.2.5 Underlying Trends in Stock Market**

In the stock market, it has been observed that the stock price fluctuations does not behave independently of each other but are mainly dominated by several underlying and unobserved factors. Hence try to identify the underlying trends from the stock market data is an interesting problem, which can be solved by NMF. Given an n×m nonnegative matrix X, columns of which is the records of the stock prices during n time points, NMF can be applied to find two nonnegative factors F and G such that X = FGT , where columns of F are the underlying components. Note that identifying the common factors that drive the prices is somewhat similar to blind source separation (BSS) in signal processing. Furthermore, G can be used to identify the cluster labels of the stocks and the most interesting result is that the stocks of the same sector is not necessarily assigned into the same cluster and vice versa, which is of potential use to guide diversified portfolio, in other words, investors should diversify their money into not only different sectors, but also different clusters (Konstantinos, 2008).

**2.2.6 Discriminant Features Extraction in Financial Distress Data**

Building appropriate financial distress prediction model based on the extracted discriminative features is more and more important under the background of financial crisis. Ref presents a new prediction model which is indeed a combination of K-means, NMF and support vector machine (SVM). The basic idea is to train a SVM classifier in the reduced dimensional space which is spanned by the discriminative features extracted by NMF, the algorithm of which is initialized by K-means (Bernardete, Catarina, Armando, & João, 2009).

NMF can discover the common basis hidden behind the observations and the way how the images are reconstructed by the basis. But further researches have also shown that the standard NMF model does not necessarily give the correct part of whole representations, hence many efforts have been done to improve the sparseness of NMF in order to identify more localized features that are building parts for the whole representation.

**2.3 ARTIFICIAL BEE COLONY (ABC)**

ABC is an optimization algorithm, which imitates the real acts of honey bees (Karaboga & Dervis, 2005). The most important components of ABC algorithm are its food source, employed and unemployed bees. The main theme of this algorithm is to arrive at the best food source. The standard pseudo code of the ABC algorithm is presented in algorithm (Das, Swagatam, and Amit Konar, 2009).

The most basic ABC algorithm consists of three phases. They are initialization, employed, onlooker and scout bees phase. Each phase is replayed until the maximum count of iterations is reached. In the initial phase, the count of solutions and the control parameters are fixed. The employed bees phase deals with the search of new high quality food sources in the nearby locality of old food source. The new food source is then evaluated for its fitness, which is then followed by the comparison of the old and the new food source by means of greedy selection. The collected knowledge about the food source is distributed among the onlooker bees present in the beehive.

In the next phase, the onlooker bees follow a probabilistic approach to select the food sources with respect to the information provided by the employed bees. This is followed by the calculation of the fitness function of the food source, which is located nearby the selected food source. Finally, the old and the new food sources are compared by the greedy selection In the final phase, the employed bees turn to scout bees, when their solutions cannot be enhanced within a predefined count of iterations. The solutions so found by the bees are dropped out. At this point, the scout bees search for new food source again. Using this functionality, the poor solutions are dropped out. These three phases continue its process untl the stopping point is reached (Karaboga, Dervis and Celal Ozturk, 2014 ).

The below is the ABC algorithm

*1: Input: Training data;*

*2: Produce initial population*

*3: Calculate the fitness function of the population*

*4: Fix counter=1*

*5: Do*

*// Employed bees phase*

*6: Search for the food source;*

*7: Calculate the fitness function;*

*8: Employ greedy selection process;*

*9: Compute the probability for the food source;*

*// Onlooker bees phase*

*10: Select food source based on the probability values;*

*11: Generate new food source;*

*12: Calculate the fitness function;*

*13: Apply greedy selection process;*

*// Scout bees phase*

*14: If food source drops out then swap it with new food source;*

*15: Save the best food source;*

*16: Counter + =1;*

*17: While counter=MC;*

**

Figure.1 ABC Food Source Diagram

**2.4 NON NEGATIVE MATRIX FACTORIZATION (NMF)**

Nonnegative Matrix Factorization (NMF) is recent development for document clustering. Initial work on NMF (Lee & Seung 1999; 2001) emphasizes contain coherent parts of the original data (images). Later work (Xu, Liu, & Gong 2003; Pauca et al. 2004) show the usefulness of NMF for clustering with in experiments on documents collections, and a recent theoretical analysis (Ding, He, & Simon 2005) shows the equivalence between NMF and Kmeans / spectral clustering.

In many settings in science and engineering the observed data are admixtures of multiple latent sources. We would typically want to infer the latent sources as well as the admixture distribution given the observations. Nonnegative matrix factorization (NMF) is a natural mathematical framework to model many admixture problems. In NMF we are given an observation matrix M 2 Rnm, where each row of M corresponds to a data-point in R m.

It is assume that there are r latent sources, modeled by the unobserved matrix W 2 Rrm, where each row of M characterizes one source. Each observed data-point is a linear combination of the r sources and the combination weights are encoded in a matrix A 2 Rnr. Moreover, in many natural settings, the sources are non-negative and the combinations are additive. The computational problem is then is to factor a given matrix M as M = AW, where all the entries of M;A and W are non-negative. We call r the inner-dimension of the factorization, and the smallest possible r is usually called the nonnegative rank of M. NMF was first purposed by (Lee & Seung, 2009), and has been widely applied in computer vision (Lee & Seung, 2000), document clustering (Xu et al., 2003), hyperspectral unmixing( Nascimento & Dias, 2004; Gomez et al., 2007), computational biology (Devarajan, 2009), etc.

Nonnegative Matrix Factorization has been proved to be valuable in many fields of data mining, especially in unsupervised learning. The special point on NMF is its ability to recover the hidden patterns or trends behind the observed data automatically, which makes it suitable for image processing, feature extraction, dimensional reduction and unsupervised learning.

Intuitively there are three ideas on disguising sensitive data. One is to transform original data into protected, publishable data by data perturbation. An alternative to data perturbation is to generate a new dataset (synthetic dataset), not from the original data, but from random values that are adjusted in order to have the same feature pattern as the original data. A third possibility is to build a hybrid dataset as a mixture of a distorted one and a synthetic one. The idea of positive matrix factorization is developed by P. Paatero at the University of Helsinki, and to be popular in the computational science community. Interest in positive matrix factorization increased when a fast algorithm for Non-negative Matrix Factorization (NNMF), based on iterative update, was developed by (Lee and Seung, 2010), particularly as they were able to show that it produced intuitively reasonable factorizations for a face recognition problem. NNMF has recently been shown to be very useful technique in approximating high dimensional data where the data are comprised of non-negative components. NNMF is a vector space method to obtain a representation of data using non-negative constraints. These constraints can lead to a parts-based representation because they allow only additive, not subtractive, combinations of the original data. This is in contrast to techniques for finding a reduced dimensional representation based on SVD.

**2.5. NMF AND ABC FOR TEXT SUMMARIZATION**

The transparent decision support systems in the sector have an important role in the analysis and the decision process. This goal can be achieved only by having a good technical tool to identify creditworthy candidate. The main motivation to apply this approach is to obtain a credit scoring model from a data set that not only have the required performance, but it is relatively interpretable. This objective is achieved through two steps and using complementary soft computing methods. In the first step the original data will be factorize into two non-negative parts using NMF algorithm where ABC will be used to increase the convergence rate and accuracy. And in the second step the reduced data set will be used by different classification algorithms to create a good credit scoring model.

The factorization of matrices representing complex multidimensional datasets is the basis of several commonly applied techniques for pattern recognition and unsupervised learning. Similarly to principal components analysis (PCA) or independent component analysis (ICA), the objective of non-negative matrix factorization (NMF) is to explain the observed data using a limited number of basis components, which when combined together approximate the original data as accurately as possible (Desai, Crook, & Overstreet, 2007). The distinguishing features of NMF are that both the matrix representing the basis components as well as the matrix of mixture coefficients are constrained to have non-negative entries and that no orthogonality or independence constraints are imposed on the basis components. This leads to a simple and intuitive interpretation of the factors in NMF, and allows the basis components to overlap. The Non-Negative Matrix Factorization (NMF) is an algorithm able to learn a parts-based representation by imposing non negativity constraints that allow only non-subtractive combinations. The idea of using Artificial Bee Colony (ABC) in finding the non- negative matrix factors came because of fast convergence of ABC over other optimisation techniques. ABC is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems. The detailed description of ABC is available in subsequent section.

Document clustering has been widely used as a fundamental and effective tool for efficient document organization, summarization, navigation and retrieval of large amount of documents.

Generally document clustering problems are determined by the three basic tightly-coupled components: a) the (physical) representation of the given data set; b) The criterion/objective function which the clustering solutions should aim to optimize; c) The optimization procedure (Li 2005).

Among clustering methods, the K-means algorithm has been the most popularly used. A recent development is the Probabilistic Latent Semantic Indexing (PLSI). PLSI is a unsupervised learning method based on statistical latent class models and has been successfully applied to document clustering (Hofmann 1999). (PLSI is further developed into a more comprehensive Latent Dirichlet Allocation model (Blei, Ng, & Jordan 2003).

Documents can be represented by various models. However, the most frequently employed model for document representation is Vector Space Model (VSM). Vector Space Model represents the text documents as vectors.

Let Doc=(Doc1,Doc2,Doc3,....Doci) be the accumulation of documents and TS be the term set of the documents Doc, where TS=(T1,T2,....TK) In this work, every document is defined as a point in a k dimensional vector space. Thus, the dimension is based on term set of the document. Every dimension denotes a different term Docm= (Wtm1, Wtm2, Wtm3 ,...... Wtmk). Where, the value of m ranges between 1 to I . Each element of mDoc is the weight of the terms available in the document. Thus, mDoc is based on the level of association between the term and the document. Measurement of the level of association between the term and the document is referred to as term weighting process. Term frequency-inverse document frequency model is the famous model to determine the level of association between the term and the document. This model focuses on the local and the global occurrence frequency of the term in the documents and is presented in equation below.





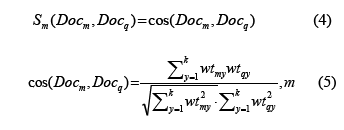
where i my is the occurrence frequency of the term y in mth document and i.d.f is given by equation below



Inverse document frequency determines the occurrence frequency of term y in a group of documents. In case of the presence of a term in all the documents, the term cannot play the role of a keyword. For instance, when the term frequency of a word equals the total number of documents, then the weight assigned to that term becomes 0. Mostly, such terms can be articles, connectors, prepositions and pronouns. These terms are not considered to be significant in this work and so are eliminated.

**2.5.1 Similarity Measure**

The process of clustering completely depends on the similarity of terms or documents, as similar entities can alone be grouped or clustered. This step is the predecessor of the clustering process. The similarity measure determines the level of association between the documents. There are several performance similarity measures such as Euclidean distance, cosine similarity, Jaccard coefficient and Pearson correlation coefficient (T Li, C Ding, 2008). This work proposes to incorporate cosine similarity measure because of its wider range of applications in text mining. The main objective of a similarity measure is to obtain the degree of association between two documents. This is achieved by the calculation of cosine angle between two different vectors and can be calculated by Equation below



**2.5.2 Clustering algorithm**

A combination of Artificial Bee Colony and k-means algorithm is proposed for clustering the web documents. ABC colony algorithm is an efficient population based optimization algorithm and it imitates the behaviour of real bees. The k-means algorithm is efficient and fast, however the problem is on finding initial cluster point. This work proposes to locate the initial cluster point with the help of bees and these clusters are refined by the k-means algorithm. We propose to combine both ABC and k-means algorithm, so as to inherit the merits of both the algorithms. ABC is efficient but consumes more time for convergence. The k-means algorithm is also known for its faster convergence but struggles in locating the initial cluster point. Thus, a new algorithm is presented for improving the efficiency and reducing the execution time. The steps involved in the proposed algorithm are explained below.

**Step 1:** Parameters initialization

The parameters that need to be initialized are the maximum count of iterations or the maximum time can be given in milliseconds, position of the food source (cluster center) and occurrence frequency threshold of the terms for labeling clusters. The fitness value for this algorithm is the degree of relevance between two documents. Thus, the initial populations of food sources are distributed randomly.

**Step 2:** Document pre-processing

This step is concerned with the removal of articles, connectors, prepositions and pronouns. This step is to weed out the unwanted terms, so as to make the clustering process effective. The stop words are eliminated. Some of the sample stop words which were removed from the documents are presented in table 1. #

|  |  |  |  |
| --- | --- | --- | --- |
| **Sample stop words** | | | |
| a | an | the | about |
| above | According | despite | along |
| again | among | apart | around |
| before | after | between | but |
| by | because | with | without |
| over | under | near | on |
| of | till | until | in |
| below | behind | beside | beyond |
| you | me | I | we |

Table 1. Sample stop words

Let Doc = (Doc1, Doc2, Doc3 … Doc i) be the accumulation of documents and be the term set of the documents Doc, where TS= ( T1,T2, …TK) In this work, every document is defined as a point in a dimensional vector space. Thus, the dimension is based on the term set of the document. Every dimension denotes a different term and it is DocM =(wtM1, wtM2, wtM3, … wtMk)

where the value of *m* ranges between 1 to*i* . Each element of  *DocM* is the weight of the terms present in the document. Thus, *DocM* is based on the level of association between the term and the document. Similarity between the documents or the similarity between the document and the cluster centre is calculated by equation (2).

**Step 3:** The execution of the proposed algorithm depends on the source of food, which are the solutions. Let food source *Fs= 1,2..i* is a k dimensional vector, where *k* is the multiplicative result of documents and the cluster size. The initial count of documents in a cluster is 2 and the maximum number of documents to be in a cluster is 8.

**Step 4:** In this step, the fitness of the population is calculated by the Equation (6) and is given below.



Where, c is the cluster, CC is the cluster center and doc is the document. The above equation determines the distance between the document and the cluster center. The main objective is to have minimal fitness value.

**Step 5:** After finding the fitness of the population, the employed bees search for the new food source in its neighbourhood and provides a new food source from its locality. This new food source is tested for its fitness by the K-means algorithm and the greedy selection is applied. In case, if the degree of similarity (fitness) of the new document with the cluster center is more than the similarity between the old document and the cluster centre, then its memory is loaded with the new document and is computed by Equation (7). By this way, the employed bee search for the better documents with respect to the cluster centre. This is followed by the calculation of probability of the food source and is calculated by Equation (8).



The employed bees search for the new documents with better degree of relevance with respect to the cluster centre, in its neighbourhood.

In (7), *ai,j* is the location of the initial document and is stored in memory of the employed bee, *ai,j*

is the new document and *d* is in the range of [1, -1]. The new document is found by changing a dimension over *a* . By this way, the employed bee moves in its neighbourhood and the location bound is reset.



Where, *fi* is the fitness of the *ith* document and n is the total number of cluster centres. Thus, the probability is calculated.

**Step 6:** The onlooker bee, which is waiting on the bee-hive, chooses the document based on the so calculated probability value and tries to find a document in the neighbourhood. If a document is found in its neighbourhood, then the degree of relevance is found. This is followed by the application of K-means and the greedy selection, as in step 6. This notifies that any number of onlooker bees can probabilistically select a single document with high fitness value. Finally, the obtained best solution is saved. This process continues till the stopping point for execution is reached. The overall structure of proposed document clustering algorithm is presented in algorithm 2.

|  |
| --- |
| **Algorithm 2** |
| *1. Initialize the algorithm parameters*  *2. Pre-process the document*  *3. Randomly assign the population of food*  sources  *4. Determine the fitness of the population by (6)*  *5. While*  *6. For each employed bee*  *7. Produce new food source;*  *8. Calculate fitness of the food source;*  *9. Employ K-means and greedy selection;*  *10. Calculate the probability of food*  source by (8);  *11. For each onlooker bee*  *12. Choose the food source with respect to step 1.*  *13. Produce new food source and compute its*  fitness by (6);  *14. Apply K-means and greedy selection*  *15. Compare and swap the solutions if*  new source is better;  *16. Save the best food source;*  *17. end while (termination condition not met);* |

**CHAPTER THREE**

**METHODOLODY**

**3.1 RESEARCH APPROACH**

This research adopts a Text summarization using ABC and NMF approach to optimize documents for proper processing and accurate mining of text within documents. The project will take different document as input process the document in other to extract stop words from it and the output of the project will be important word space that quantifies and qualifies the meaning of the words in each documents

The various activities carried out and different modules implemented to ensure application feature behaviors are intercepted for the use of Text summarization are

1. Input Document
2. Extract Text
3. Extract Text stop words
4. Check for Repeated Text
5. Intercept on a table to generate output
6. Display summarized text

User wants to summarize text

Load Document from specified location

Hybrid Algorithm For Text Summarization system

Display the summarized and compressed document output

Load summarized output for further analysis

Remove repeated word using stemming

Manipulate loaded document to produce summarized output

***Figure 3.1: Architecture of Hybrid Text summarization using ABC and NMF***

**3.1 Materials**

The application of the Hybrid Text summarization using ABC and NMF will be test on a laptop machine with an intel-centrino-dual2-pro processor, 4GB of available memory and 500GB Hard Disk Drive (HDD). This machine runs windows 8 Operating System while NetBeans 7.2 Integrated Development Environment (IDE) will be used as the Java Developer Kit (JDK). The server used to test runs the project is Xampp that contains apache for rendering the application.

**3.1.2 Methods**

The approach toward which this project will takes as an desktop application, that will mainly be develop to address the issue of text summarization. This project which will be develop using java technology such as word processor, text extractor or abstractor and some other related tools for text summarization.

The application will be tailor toward desktop text summarization in other to provide Automatic summarization by processing shortening a text document in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax.

**3.2 PROCEDURAL MODEL OF NMF AND ABC**

Stop words removal

Lexical analysis of text

Obtaining optimal dimension using NMF

Term weighting

Term-document matrix construction

Stemming

Dimension reduction using NMF and ABC

Ranking/similarity measurement using cosine rules

Document collection

***Figure 3.2: System Procedural Diagram***

**3.2.1 Document Collection**

The source of the document shall be a document containing text which can either be course material and the students’ free text response to each question. The answer to each question by the lecturer and the student will constitute a document of its own. That is each student will generate five documents for a set of five questions. The responses shall be supplied as a soft copy through an interface designed to capture the data. These documents will serve as the input data to the system. Obviously, the responses are not expected to be the same in spelling but some will be closer in relational concept. This perfectly fit into the description of non-negative matrix factorization which discovers the semantic relationship between keywords and documents in the document set in order to achieve the goal of concept based assessment by eliminating the influence of different word usage.

**3.2.2 Document processing and Term Extraction**

Document processing comprises of three stages which are as follows:

**3.2.2.1 Lexical Analysis of text**: this is the process of converting a sequence of characters into a sequence of tokens (strings with an assigned and thus identified meaning).

It breaks down streams of text up into words, phrases, symbols or other meaningful element called tokens. The list of tokens becomes input for further processing.

**3.2.2.2 Stop words removal:** This is the elimination of the most common words such as: the, as, a, in, to etc. in a language. These words are usually excluded from the analysis as they do not contribute much (if any) meaning to a context. The filtered words become input for further processing.

**3.2.2.3 Stemming:** This is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form. The stem needs not to be identical to the morphological root of the word. A stemming algorithm reduces the word “fishing”, “fished” and “fisher” to the root word “fish”.

The purpose of term extraction is to generate list of terms that are relevant to the input domain. A well generated or extracted term will facilitate the assessment process. It is made up of two stages which are the training stage and the extraction stage. At the training stage, a model is created for identifying terms using training documents. The extraction stage chooses term from a test document using the model that was created at the training stage.

The training stage makes use of the following procedure:

1. Read the input document which is the training document
2. Extract noun phrase from each sentence in the training document using syntactic parser. The parser will analyze each sentence and generate a list of syntactic information such as Noun, Noun-Phrase etc.
3. The extracted noun and noun phrase are preprocessed and stem to remove stopwords and produce a list of clean noun phrase as term.
4. A set of five features(domain relevance, domain consensus, term cohesion, first occurrence and length of noun phrase) are calculated for each candidate term which are subsequently used to calculate the score and rank the term based on their score.

The extraction stage consists of the following procedures:

1. Read the test documents
2. Perform preprocessing operation
3. Perform feature generation
4. Conduct term ranking
5. Generate list of terms

**3.2.3 Term document matrix construction:** In this stage, each row represents documents in the collection and each column a term, and respective cells of the matrix contain the frequencies with which the term occurs in the document (Each cell contains the number of times that index word occurs in the document). In general, the matrices built during NMF tend to be very large, but also very sparse (most cells contain 0). That is because each document usually contains only a small number of all the possible words. This sparseness can be taken advantage of in both memory and time by more sophisticated NMF implementations.

**3.2.4 Term weighting generation:** In Latent Semantic Analysis systems, the raw matrix counts are usually modified so that rare words are weighted more heavily than common words. For example, a word that occurs in only 5% of the documents should probably be weighted more heavily than a word that occurs in 90% of the documents. The reason for this because rare words reveal better similarity features among documents. The most popular weighting is TFIDF (Term Frequency - Inverse Document Frequency). Under this method, the count in each cell is replaced by the following formula.

For the purpose of term generation, a set of five features are used to characterize the noun phrases in the document. These features are calculated for each term and are used at both stages (i.e. training and extraction stages). The features are:

1. **Domain Relevance:** It is a measure of the amount of information captured in the target document with respect to contrastive documents. If Di is a set of relevant document in a domain of interest Di and [Di…Dn] is the sets of documents in another domain.
2. **Domain Consensus**:-It measures the distributed use of a term in a Domain Dk.
3. **Term Cohesion:** It is used to calculate the cohesion of the multi-word terms. This measure is proportional to the co-occurrence frequency and the length of the term.
4. **First Occurrence**: It is calculated as the number of words that precede the phrase’s first appearance, divided by the number of words in the document. The result is a number between 0 and 1 that represents how much of the document precedes the phrase’s first appearance.
5. **Length of noun phrase**: candidate length is also a useful feature in extraction as well as in candidate selection, because the majority of terms are one or two words in length. Length of noun phrase score is calculated as its frequency times its length (in words).

**3.2.5 Dimension reduction using NMF**

Dimension reduction using NMF has been observed in literatures to have the following setback: It does not lead to proper storage management and conservation, the presence of negative value in the cell of term-document matrix makes it un-interpretable. This research intends to solve these problems by hybridizing ABC with NMF a particle swarm optimization techniques.

**3.2.6 The Mechanism for Application Design**

Application Designed Interface(s) with Java codes

ABC NMF document processing system

Xampp (MySQL) server database design holding the application data as a repository

***Figure 3.3: Application Design and Program***

**3.3 HYBRID TEXT SUMMARIZATION USING ABC AND NMF ACTIVITIES**

Document location

Summarized Text

ABC processor

Text Extraction

NMF Optimizer

Document Term weight output

Get Provided document

***Figure 3.4: Data flow Diagram***

The following are the systems activities for the Hypermedia Optimization Using Adaptive Compressor

Step 1: Supply document location.

Step 2: Extract Text from each documents.

Step 3: Load the extracted text for optimization using ABC.

Step 4: Load the optimization output for factorization using NMF.

Step 5: Generate the output of the summarized text.

**CHAPTER FOUR**

**SYSTEM IMPLEMENTATION AND DOCUMENTATION**

**4.1 SYSTEM IMPLEMENTATION**

The implementation phase of this project is more concerned with the integrating of NMF and ABC optimization techniques on free text document for text summarization service with the use of formatted free text document, by installing the Text Mining Language (TML) and it applications from the appropriate respective online/offline repository. The application required Java Virtual Machine (JVM) which is available on most recent Operating Systems. In case the JVM is not available on the machine, then, one has install Java Developer Kit (JDK), the application also required the use of TML. It invariably means that the coding perspective of building the system provides the blue print for the system and helps provide the platform for the user.

**4.2 CHOICE OF PROGRAMMING LANGUAGE**

The technologies used in building the code are Java, sequential query language (SQL) server and TML.

1. Java is an object-oriented programming language. It provides support for software engineering principles such as checking array sounds checking, detection of attempts to use uninitialized variables and automatic garbage collection. It also provides software justness, durability, and programmers' productivity.

The language is used in developing software component suitable for deployment in distributed environments.

1. SQL (Sequential language) server is a relational database management system first developed by Microsoft in 1989. As a database, it is a software product whose primary function is to store and retrieve data as requested before other software application be it those on the same computer or those running on another computer across a network including the internet.
2. TML is a standard plugging for text factorization and optimization having various kind factorization techniques such as NMF, LSA etc. this technology is built on java framework to mine data of huge.

**4.3 HARDWARE AND SOFTWARE SPECIFICATION**

**This is the Section where we discuss the two-basic software division of computer.**

**4.3.1 SOFTWARE SPECIFICATION**

The following are the software requirements for this application

* NetBeans 7.2 Integrated Development Environment (IDE)
* Sequential query language (SQL) server 2005
* Windows operating system (at least windows vista installed)
* Anti-virus package to prevent that application from virus attack

**4.3.2 HARDWARE SPECIFICATION**

The above listed software will work perfectly with the under listed specification, as a computer is not complete without either the software or the hardware

* 1 gigabyte RAM
* 1.5GH2 processor
* 200GB Hard Disk or higher (recommended)
* Uninterruptible Power Supply (UPS)
* Mouse and enhanced keyboard

**4.4 TESTING**

After the implementation, the following tests are going to take place:

**i Alpha Test:** This means self or in-house test for the application for any error or exceptions.

**ii Beta Test:** I am going to release a virtual data for the application user to test the application.

**4.5 IMPLEMENTATION PHASE**

The implementation phase of this project is divided into two which are the NMF factorized result and the ABC summary. These phases are explained below:

**i.** The NMF factorization result is generated using a plugin called Text mining language (TML), TML implementation analysis each paragraph, sentence and word. Nonnegative Matrix Factorization has been proved to be valuable in many fields of data mining, especially in unsupervised learning. The special point on NMF is its ability to recover the hidden patterns or trends behind the observed data automatically, which makes it suitable for image processing, feature extraction, dimensional reduction and unsupervised learning.

**ii** The ABC summary technique as ABC algorithm consists of three phases. They are initialization, employed, onlooker and scout bees phase. Each phase is replayed until the maximum count of iterations is reached. In the initial phase, the count of solutions and the control parameters are fixed. The employed bees phase deals with the search of new high quality food sources in the nearby locality of old food source. The new food source is then evaluated for its fitness, which is then followed by the comparison of the old and the new food source by means of greedy selection. The collected knowledge about the food source is distributed among the onlooker bees present in the beehive.

**4.6 SYSTEM DOCUMENTATION**

**Installation Procedure**

This program is already packaged to some extence having all requirement to function with the database and tml mining. some computer programs can be executed by simply copying them into a [folder](http://en.wikipedia.org/wiki/Folder_%28computing%29) stored on a computer and executing but this is quit advanced in nature because of the advancement in technology Other programs are supplied in a form not suitable for immediate execution and therefore need an installation procedure. Once installed, the program can be executed again and again, without the need to reinstall before each execution.

But it is important to note that users must have an SQL server and TML package in the C: directory installed and running before execution as the program is based on local server.

The following are the step involve in installing this Application

1. Install Java on your system by visiting oracle.com
2. Download and install TML from online
3. Move TML to the C dir
4. Copy the Free text summary to a dir on your system and launch the jar
5. You are set to Go.

**4.6.1 System Maintenance**

The program may be maintained on the ground that the system requires an upgrade. Though it is compiled as a standalone software the database can be tempered with but it’s advisable that the admin put a password on the file to secure the database from intrusion.

**The following precaution should be done**

* Ensure that the computer is kept in clean areas.
* System should be kept in cool places.
* Air conditioner is important to reduce room temperature and keep it constant.
* Backup of data is important

**4.6.2 SYSTEM EVALUATION**

This System is a high standard program that can weather the storm of technology advancement, it is most needed in all financial institutions, it will be very helpful to clients and customers in the marketing business. All it needs some computer-oriented personnel to make it work, it is stand alone and automated.

**CHAPTER FIVE**

**SUMMARY, CONCLUSION AND RECOMMENDATION**

**5.1 SUMMARY**

NMF finds applications in such fields as computer vision, document clustering, chemometrics, audio signal processing and recommender systems. Many standard NMF algorithms analyze all the data together; i.e., the whole matrix is available from the start. This may be unsatisfactory in applications where there are too many data to fit into memory or where the data are provided in streaming fashion. One such use is for collaborative filtering in recommendation systems, where there may be many users and many items to recommend, and it would be inefficient to recalculate everything when one user or one item is added to the system. The cost function for optimization in these cases may or may not be the same as for standard NMF, but the algorithms need to be rather different.

A combination of Artificial Bee Colony and k-means algorithm is proposed for clustering the web documents. ABC colony algorithm is an efficient population based optimization algorithm and it imitates the behaviour of real bees. The k-means algorithm is efficient and fast, however the problem is on finding initial cluster point. This work proposes to locate the initial cluster point with the help of bees and these clusters are refined by the k-means algorithm. We propose to combine both ABC and k-means algorithm, so as to inherit the merits of both the algorithms. ABC is efficient but consumes more time for convergence. The k-means algorithm is also known for its faster convergence but struggles in locating the initial cluster point. Thus, a new algorithm is presented for improving the efficiency and reducing the execution time. The steps involved in the proposed algorithm are explained below.

**5.2 CONCLUSION**

Optimization theory had evolved initially to provide generic solutions to optimization problems in linear, nonlinear, unconstrained, and constrained domains. These optimization problems were often called mathematical programming problems with two distinctive classifications, namely linear and nonlinear programming problems. The system uses NMF and Firefly algorithm for student assessment will be designed and developed not to replace the normal activities of lecturers in conducting end of semester assessments and marking of assessment scripts but to provide a better and more viable way of doing the aforementioned and seems to be a valuable tool for developing a framework for the students’ assessment.

Various works has been done by different researchers on a variety of research topics using NMF. ABC is also a favorite choice for researchers for solving different kind of optimization problems. These two topics are relatively new than the CREDIT scoring problem. For a long time people are dealing with credit scoring. High investor risk and low consumer satisfaction was the root cause for required improvement in this area. Now researchers are using different techniques for reducing risk and improving satisfaction and they are also successful up to some extent

**5.3 RECOMMENDATION**

The following recommendations are suggested in order to and efficiently increase the use of Summary Electronic free text documents of free text document using NMF and ABC.

1. Using NMF and ABC to summarize free text document using should not be limited to a student environment alone.
2. This techniques should be adopt by every assessment organization because it is considered accurate efficient.
3. This project can be enhanced for future works to authenticate in a dark place.

In future works, it could be also implemented on a special device.