International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-4, February 2021

An Improved LSA Model for Electronic Assessment of Free Text Document

Rufai M. M, Afolabi Adeolu, Fenwa O. D, Ajala F.A

Abstract*: *Latent Semantic Analysis (LSA) is a statistical approach designed to capture the semantic content of a document which form the basis for its application in electronic assessment of free-text document in an examination context. The students submitted answers are transformed into a Document Term Matrix (DTM) and approximated using SVD-LSA for noise reduction. However, it has been shown that LSA still has remnant of noise in its semantic representation which ultimately affects the assessment result accuracy when compared to human grading. In this work, the LSA Model is formulated as an optimization problem using Non-negative Matrix Factorization(NMF)-Ant Colony Optimization (ACO). The factors of LSA are used to initialize NMF factors for quick convergence. ACO iteratively searches for the value of the decision variables in NMF that minimizes the objective function and use these values to construct a reduced DTM. The results obtained shows a better approximation of the DTM representation and improved assessment result of 91.35% accuracy, mean divergence of 0.0865 from human grading and a Pearson correlation coefficient of 0.632 which proved to be a better result than the existing ones.

Keywords: Ant Colony Optimization, Electronic Assessment, Latent Semantic Analysis, Non-Negative Matrix Factorization

I. INTRODUCTION

Latent Semantic Analysis (LSA) is a statistical-based method for inferring meaning from a text. The technique was initially designed for indexing and information retrieval. Further exploration reveals that the technique can be applied in the semantic representation of free text document. LSA is considered a suitable tool for electronic assessment of free text document because it focuses on the content of the essay and not on the surface features or keyword-based content analysis. It can be applied on relatively low amount of human graded essays and yield good result (Kakkonen et al. 2005). The assessment is done by using the technique to extract the conceptual similarity between the student's candidate text and the teacher's reference text by looking for repeated patterns between them. However, LSA is characterized by random noise which manifest in its objective function that is used to determine its optimal decomposition. This becomes so manifest with large document collections. The presence of noise in LSA has negative effects on its performance such as inadequate semantic representation and poor assessment result (Hoenkamp, 2011). Selection of a dimension reduction method is an important factor in achieving efficient noise reduction and adequate semantics capturing (Anandarajan, Hill and Nolan, 2019)(Ayesha et. Al., 2020). This work addressed the issue of minimizing noise presence in LSA and ultimately improve its assessment accuracy by integrating Non-Negative Matrix Factorization (NMF) and Ant Colony Optimization (ACO) in its dimension reduction process. Dimension Reduction is a pre-processing step that identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis. Mathematically the problem of dimension reduction can be defined as: given a r-dimensional random vector $X = (x_1, x_2, \dots x_r)T$, the objective is to find a representation of lower dimension $S = (s_1, s_2, \dots s_k)T$ where $k < r$, which preserves the content of the original data, as much as possible according to some criterion.

Manuscript received on February 05, 2021.
Revised Manuscript received on February 13, 2021.
Manuscript published on February 28, 2021.
***** Correspondence Authors
Rufai Mohammed Mutiu*, Computer Technology Department, Yaba College of Technology, Yaba, Lagos, Nigeria. Email: mohammed.rufai@yabatech.edu.ng
Prof. A. O. Afolabi, Computer Science Department, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria. Email: aoafolabi@lautech.edu.ng
Dr. (Mrs.) O. D. Fenwa, Computer Science Department, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria. Email: odfenwa@lautech.edu.ng
Dr. (Mrs.) F. A. Ajala, Computer Science Department, Ladoke Akintola University of Technology, Ogbomoso, Oyo State, Nigeria. Email: faajala@lautech.edu.ng

II. RELATED WORKS

The performance of LSA in electronic assessment is influenced by tunning of its operational parameters which are Document preprocessing, weighting, dimensionality and similarity measurement (Wild et al, 2005). This assertion by Wild et. al forms the basis for improvement of LSA by various researchers. Klein et al (2011) implemented the Latent Semantic Analysis on unstructured free text responses in computer science. His objective was to prove the efficiency of computer graded questions over manual human grading. The effect of varying some operational parameters of LSA was observed on its performance. Parameters whose effect were observed are the weight of the term frequency matrix, exclusion of the first dimension, the choice of the k (reduced dimension) and the choice of the similarity measures. The developed system was evaluated using the Pearson correlation coefficient to measure the similarity between the human graded score and the computer score which yielded a value above 0.8. Islam and Hoque (2010) worked on Automated Essay Scoring Using Generalized Latent Semantic Analysis. They used n-gram by document matrix for the document matrix representation as against the conventional word by document matrix in Latent Semantic Analysis.

Their contention with the traditional LSA approach is that LSA does not recognize word order of sentence in a document, consequently not seeing the difference between a word like "Carbon dioxide" and "dioxide carbon" in its document matrix representation. The system's reliability was tested by comparing the score generated by the system against that of the human grader. An accuracy level of 0.89-0.90 was achieved while a standard deviation of 0.22 was observed. Zhang *et al* (2014) developed a system that solves the problem associated with large dataset representation in a way that will maximize the use of storage and minimize the computing time in automated essay scoring. He introduced the concept of Incremental Latent Semantic Analysis for Automated Essay Scoring. In this method essays to be graded are segmented into words and constructed into essay vectors. This resulted in to a term-document matrix. A weighting formula $w_{ij} = TF_{ij} \times IDF_{ij}$ was applied on each entry in the matrix. A set of first M essay vectors (initialized vector) are derived from the general vector, the conventional SVD is applied on this set which produces an intermediate result of decomposition. This result is updated with a new set of N columns (batch size) taken from the rest part of the general essay vector. The process of updating with subsequent batch size continued until all the essays have been added to the mediate result. The next step was to re-project, which can be described as re-projecting an essay d_j to the semantic space of the training using the formula $\hat{d}_j = \Sigma^{-1}.U^T.d^j$. The Support Vector Machine was used to automatically score the essay. The ILSA was evaluated on computing time, memory usage and scoring and was observed to be an improvement on the conventional LSA. However, more still need to be done to get the optimum batch size that will bring about better performance and improve the correlation between human and predicted scoring. Darwish *et al* (2019) worked on automated essay evaluation by applying Latent Semantic Analysis and Fuzzy Ontology. The LSA was used in checking the semantic of the essays involved while the Fuzzy Ontology was used to check the essays for consistency and coherence thereby resolving the problem of vagueness in language. The system scores the syntax of the essay, measures his semantic coherence and provides feedback to students about their mistakes. However, further work needs to be done to improve the semantic attributes representation and the feedback algorithm

III. REVIEW OF METHODS

The procedures used in LSA in assessing an essay can be summarized as follow:

i) Document Collection: Essays to be graded and the lecturer reference material are collected
ii) Term Extraction: Relevant terms in each of the documents are extracted
iii) Stopwords Removal: - Words with low discriminatory value are removed
iv) Document-Term-Matrix Construction: The Documents are used as the matrix row labels, the terms are the matrix columns labels. The entries are the frequency of occurrence of the term in each document
v) Term weighting: The entries are weighted using Term Frequency - Inverse Document Frequency(TF-IDF) weighting formular expressed as

$$w_{t,d} = \left(1 + log tf_{t,d}\right).log\frac{N}{df_t} \quad (1)$$

vi) Dimension Reduction using SVD
vii)Ranking
viii) Cosine Similarity Measurement using the cosine similarity rule expressed as

$$cosSIM\left(\vec{A}.\vec{B}\right) == \frac{\vec{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_{i=1}^2}}$$

IV. THE IMPROVED LSA ALGORITHM USING NMF-ACO

Dimension reduction using LSA has been observed in literatures to have the following setback:

1. It does not lead to proper approximation for the original matrix,
2. The presence of negative value in the cell of term-document matrix makes it un-interpretable.
3. It does not adequately capture the document semantic content

This paper solves these problems by integrating Latent Semantic Analysis with Non-Negative Matrix Factorization and Ant Colony Optimization.

The LSA system is replaced with LSA-NMF-ACO to build a low rank approximation to the term-document matrix in order to solve the problem of poor approximation and inadequate semantic representation.

The primary algorithm used for the dimension reduction is the Non-Negative Matrix Factorization (NMF). Other algorithms such as Ant Colony Optimization technique iteratively searches for the value of the decision variables in NMF that minimizes the objective function without compromising its semantic content. Non-Negative Matrix Factorization (NMF) is a low rank approximation technique with reduced storage and run-time requirements and reduced redundancy and noise. It allows for additive parts-based, interpretable representation of the data. NMF approximates a matrix V by

$$V_{n\times m} \approx W_{n\times r} H_{r\times m}$$

where W and H are NMF factors and all entries in V, W and H are to be non-negative. r,m,n represent the rank of the matrix r which is chosen to satisfy $(n+m)r < nm$
The goal of NMF is to minimize the original matrix V. The objective function used is the Frobenious Norm shown in equation 4.

$$min\|V - WH\|_F^2 = min\sum\sum\left(V_{ij} - WHij2\right) \quad (4)$$

NMF perfectly fits in as a better alternative to SVD in dimension reduction in LSA because of its scarcity and non-negativity; reduction in storage and its interpretability. However, its major challenge is its convergence issue because different NMF algorithm can converge to different local minima.
This challenge is addressed by choosing the right initialization and update strategy.

In this research the problem of dimension reduction is tackled by using two strategies: the first strategy is to use SVD-LSA to initialize the factors for minimizing NMF objective function prior to factorization while the second strategy seeks to iteratively improve the quality of the dimension reduction accuracy using Ant Colony Optimization Technique. The proposed improved algorithm is divided into three phases which are:

a. Initialization
b. Dimension Reduction
c. Optimization

These three phases reflect in the improved algorithm below

Table 1: The Extracted Term/Document Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
1	0.50	0.17	0.21	0.50	0.23	0.25	0.56	0.16	0.08	0.12	0.36	0.50	0.34	0.03	0.00	0.30	0.32	0.14	0.21	0.17	0.13	0.14	0.03	0.20	0.01	0.06	0.02	0.04	0.03	0.06	0.01	0.01	0.00	0.00	0.00
2	0.37	0.16	0.57	0.84	0.36	0.09	0.04	0.38	0.00	0.03	0.20	0.06	0.37	0.17	0.25	0.15	0.31	0.08	0.09	0.13	0.02	0.04	0.05	0.06	0.06	0.11	0.00	0.03	0.01	0.01	0.02	0.02	0.00	0.00	0.00
3	0.07	0.01	0.02	0.02	0.33	0.07	0.18	0.05	0.08	0.13	0.27	0.20	0.18	0.07	0.26	0.14	0.04	0.15	0.00	0.01	0.03	0.02	0.00	0.09	0.05	0.08	0.09	0.01	0.06	0.09	0.02	0.01	0.00	0.00	0.00
4	0.11	0.01	0.04	0.02	0.23	0.22	0.19	0.19	0.14	0.25	0.09	0.24	0.10	0.11	0.06	0.03	0.10	0.19	0.07	0.15	0.13	0.05	0.08	0.03	0.01	0.11	0.07	0.04	0.04	0.03	0.03	0.00	0.00	0.00	0.00
5	0.14	0.07	0.19	0.02	0.11	0.08	0.04	0.17	0.11	0.04	0.05	0.31	0.02	0.35	0.04	0.18	0.01	0.25	0.18	0.02	0.03	0.03	0.02	0.05	0.02	0.01	0.00	0.00	0.04	0.02	0.02	0.00	0.00	0.00	0.00
6	0.21	0.65	0.07	0.05	0.06	0.01	0.01	0.04	0.08	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
7	0.21	0.65	0.07	0.05	0.06	0.01	0.01	0.04	0.08	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
8	4.25	0.49	0.90	0.47	0.22	0.10	0.01	0.13	0.12	0.14	0.02	0.25	0.11	0.13	0.07	0.32	0.07	0.04	0.09	0.11	0.13	0.02	0.13	0.03	0.07	0.03	0.05	0.07	0.02	0.01	0.03	0.04	0.00	0.00	0.00
9	0.11	0.01	0.19	0.24	0.01	0.06	0.01	0.05	0.04	0.03	0.00	0.00	0.11	0.05	0.10	0.09	0.10	0.04	0.12	0.21	0.20	0.33	0.07	0.20	0.18	0.08	0.10	0.04	0.03	0.10	0.09	0.03	0.00	0.00	0.00
10	0.12	0.03	0.03	0.07	0.05	0.11	0.13	0.46	0.10	0.33	0.10	0.03	0.06	0.24	0.04	0.02	0.07	0.06	0.08	0.03	0.04	0.03	0.05	0.00	0.01	0.01	0.02	0.05	0.03	0.01	0.01	0.01	0.00	0.00	0.00
11	0.11	0.01	0.04	0.02	0.23	0.22	0.19	0.19	0.14	0.25	0.09	0.24	0.10	0.11	0.06	0.03	0.10	0.19	0.07	0.15	0.13	0.05	0.08	0.03	0.01	0.11	0.07	0.04	0.04	0.03	0.03	0.00	0.00	0.00	0.00
12	0.12	0.07	0.08	0.01	0.01	0.02	0.05	0.07	0.12	0.11	0.01	0.06	0.02	0.01	0.01	0.09	0.18	0.00	0.33	0.01	0.28	0.25	0.05	0.05	0.12	0.06	0.02	0.08	0.29	0.10	0.03	0.01	0.00	0.00	0.00
13	0.21	0.65	0.07	0.05	0.06	0.01	0.01	0.04	0.08	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
14	0.14	0.07	0.19	0.02	0.11	0.08	0.04	0.17	0.11	0.04	0.05	0.31	0.02	0.35	0.04	0.18	0.01	0.25	0.18	0.02	0.03	0.03	0.02	0.05	0.02	0.01	0.00	0.00	0.04	0.02	0.02	0.00	0.00	0.00	0.00
15	3.89	0.13	0.54	0.42	0.12	0.09	0.39	0.36	0.28	0.65	0.37	0.28	0.27	0.12	0.12	0.23	0.05	0.07	0.07	0.08	0.09	0.06	0.02	0.03	0.09	0.06	0.06	0.10	0.05	0.05	0.11	0.00	0.00	0.00	0.00
16	0.23	0.13	0.14	0.02	0.02	0.14	0.21	0.05	0.31	0.05	0.02	0.00	0.20	0.30	0.16	0.05	0.32	0.03	0.05	0.05	0.16	0.29	0.36	0.03	0.40	0.02	0.04	0.17	0.10	0.02	0.02	0.02	0.00	0.00	0.00
17	0.21	0.65	0.07	0.05	0.06	0.01	0.01	0.04	0.08	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
18	0.21	0.65	0.07	0.05	0.06	0.01	0.01	0.04	0.08	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00

The Improved Dimension Reduction Algorithm that Integrates LSA, NMF and ACO

Step 1: Compute the rank P of factorization such that P<$\frac{mn}{m+n}$

Step 2: Decompose Z using SVD-LSA in order to obtain Z=$U\Sigma V^T$ with a rank of P

Step 3: Initialise NMF factors with SVD-LSA factors as

W=$|U|$ and H=$|\Sigma V^T|$

Step 4: Update W and H using the multiplicative Update equation

$$H = H \times \frac{(WT\,Z)}{(WTWH + \varepsilon)}$$

$$W = W \times \frac{(ZHT)}{(WHHT + \varepsilon)}$$

Step 5: Compute the Distance Matrix (D) as D=Z – WH

Step 6: Compute the row-wise Frobenious Norm of D as

$$\|D\|_F^{RW} = (\sum_{i=1}^{m} |d_i^r|^2)^{1/2}$$

Step 7: Identify the rows of D with the highest norm and look for the corresponding rows of W that minimizes D=$\|z_i^r - wirHF$ using ACO

Step 8: Identify the columns of D with the highest norm and look for the corresponding columns of H that minimizes D=$\|z_j^c - Wh_j^c\|$ using ACO

Step 9: Multiply the minimized rows of W with the minimized column of H to obtain the reduced dimension of Z

Formulating the Dimension Reduction Problem as an Optimization Problem

To minimize the objective function D of the Dimension reduction problem, the problem must be modelled as a continuous optimization problem. A continuous optimization problem has all its optimization variables as continuous variables. A Continuous Optimization problem(Q) can be modelled as a 3 turple relationship given as:

Q = (S,X, f)

Where

i) S is the search space defined over a finite set of decision variable. It represents a set of unknowns or variables which affect the value of the objective functions. In this case we are looking for the set of W that minimizes the objective function $f = \|Z - WH\|$
ii) X is a set of constrain. The problem is considered as unconstrained problem, hence no constrain is assigned
iii) D is the objective function which represents the quantity to be minimized. D=F is the objective function $f = \|Z - WH\|$

The ACOR Minimization Algorithm

The algorithm has three parts which are:

a) Pheromone Representation
b) Solution Construction
c) Pheromone Update

Table 2: Reduced Dimension LSA-NMF-ACO

	abil	accomplish	acquir	act	activ	actual	aid	aim	algorithm	allow	analysi	appl	area	artifici	assum	attribut	base	behav	branch	call	capac	carri	code	cognit	complet	complex	compon	comput	concern	contol	contrast	correct	creat
Lecturer	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
student001	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
student002	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student003	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student004	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student005	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
student006	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
student007	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student008	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student009	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student010	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student011	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student012	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student013	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
student014	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student015	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
student016	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student017	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student018	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student019	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
student020	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0
student021	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student022	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student023	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
student024	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student025	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
student026	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
student027	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

a) Pheromone Representation and Initialization

1. Initialize decision variable W i.e. W_ACO(i,k)=W(i,k)
2. Compute Solution Archive as a function of W i.e., $D_k = f(W_{i,k})$ which implies there are i number of decision variables for each solution archive D_k
3. Sort Solution Archive D_k in descending order of the rank

$$\begin{matrix}1\\2\\\vdots\\j\\\vdots\\k\end{matrix}\begin{bmatrix}D_1\\D_2\\\vdots\\D_j\\\vdots\\D_k\end{bmatrix}\begin{bmatrix}d_1^1 & d_1^2 & \cdots & d_1^i & \cdots & d_1^n\\ d_2^1 & d_2^2 & \cdots & d_2^i & \cdots & d_2^n\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ d_j^1 & d_j^2 & \cdots & d_j^i & \cdots & d_j^n\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\ d_k^1 & d_k^2 & \cdots & d_k^j & \cdots & d_k^n\end{bmatrix}\begin{bmatrix}\omega_1\\\omega_2\\\vdots\\\omega_j\\\vdots\\\omega_k\end{bmatrix}$$

Figure 1: Solution Archive for ACOR algorithm

b) Ant Solution Construction

Solution is constructed on variable-by-variable basis

1. Choose a solution D_k from the set of Solution Archive D_i based on a probability of selection function p_j expressed as $p_j = \frac{\omega_j}{\sum_{r=1}^{k}\omega_r}$ where w_j is the weight associated with Solution j, is calculated using the formular

$$\omega_j = \frac{1}{q^k\sqrt{2\pi}}e^{\frac{-(rank(j)-1)^2}{2q^2k^2}}$$

q is a parameter of the algorithm known as intensification factor and is set to 0.5 and k is the size of the solution archive set to 72

2. For variable 1 to n, compute new solution using a Probability Density Function(P) expressed as

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu = S_j^i$ and $\sigma = \xi\sum_{r=1}^{k}\frac{\left|S_r^i - S_j^i\right|}{k-1}$

$\xi = Pheromone\ Evaporation\ rate > 0$

c) Perform Pheromone update

1. Repeat the whole step b m number of times to generate m number of new constructed solutions
2. Append these m solutions to the initial k solutions
3. Order the k+m solutions in ascending order of rank
4. Remove the m number of worst solutions to retain the k number of good solutions

V. EXPERIMENTAL EVALUATION

The algorithm was implemented using MATLAB. Matlab has some built-in statistical and mathematical functions that make matrix manipulation easy hence its choice. Results obtained include the following:

- The Weighted Term Document matrix which is a product of Term-Extraction

- The reduced Matrix using LSA-NMF-ACO
- The Electronic Assessment Result using LSA and The Modified Algorithm
- The Mean Divergence and Assessment Accuracy
- The Pearson Correlation Coefficient

Table 3: The E-Assessment Result using LSA and LSA-NMF-ACO

SN	Documents	Manual	LSA	LSA-NMF-ACO	DIFF LSA	DIFF NMF ACO
1	Lecturer	1.00	1.00	1.00	0.00	0.00
2	Student1	1.00	1.00	1.00	0.00	0.00
3	Student2	0.90	1.00	0.92	0.10	0.02
4	Student3	0.62	0.87	0.56	0.25	0.06
5	Student4	0.75	0.99	0.81	0.24	0.06
6	Student5	0.86	1.00	0.97	0.14	0.11
7	Student6	0.80	0.91	0.79	0.11	0.01
8	Student7	0.80	0.86	0.83	0.06	0.03
9	Student8	0.91	0.36	0.66	0.56	0.25
10	Student9	0.91	0.89	0.67	0.02	0.24
11	Student10	0.72	0.86	0.56	0.14	0.16
12	Student11	0.95	1.00	0.95	0.05	0.00
13	Student12	0.71	0.64	0.68	0.07	0.03
14	Student13	0.75	0.62	0.63	0.13	0.12
15	Student14	0.94	0.93	0.88	0.00	0.06
16	Student15	0.94	0.93	0.88	0.00	0.06
17	Student16	0.85	1.00	0.94	0.15	0.09
18	Student17	0.89	0.98	0.72	0.09	0.17
19	Student18	0.96	0.99	0.89	0.03	0.07
20	Student19	0.90	1.00	0.92	0.10	0.02
21	Student20	0.90	0.80	0.67	0.10	0.23
22	Student21	0.42	0.36	0.38	0.07	0.05
23	Student22	0.87	0.99	0.94	0.12	0.07
24	Student23	0.71	1.02	0.67	0.31	0.04
25	Student24	0.72	0.86	0.74	0.14	0.02
26	Student25	1.00	1.00	0.95	0.00	0.05
27	Student26	0.88	0.99	0.94	0.12	0.06
28	Student27	0.77	0.86	0.74	0.09	0.03
29	Student28	0.78	0.77	0.59	0.01	0.19
30	Student29	0.79	0.80	0.51	0.01	0.27
31	Student30	0.78	0.99	0.69	0.21	0.09
32	Student31	0.85	0.99	0.84	0.14	0.01
33	Student32	0.75	1.00	0.89	0.25	0.14
34	Student33	0.89	1.00	0.99	0.11	0.10
35	Student34	0.90	0.99	0.73	0.09	0.17
36	Student35	0.99	0.99	0.90	0.00	0.09
37	Student36	0.93	1.00	0.96	0.07	0.03
38	Student37	0.85	1.00	0.86	0.15	0.01
39	Student38	0.72	1.00	0.68	0.28	0.04
40	Student39	0.53	0.98	0.60	0.45	0.07
41	Student40	0.72	1.00	0.95	0.28	0.23
42	Student41	0.95	0.93	0.88	0.02	0.07
43	Student42	0.92	1.00	0.97	0.08	0.05
44	Student43	0.64	0.99	0.58	0.35	0.06
45	Student44	0.52	1.00	0.88	0.48	0.36
46	Student45	0.90	1.00	0.89	0.10	0.01
47	Student46	0.40	1.02	0.35	0.62	0.05
48	Student47	0.72	0.44	0.68	0.28	0.04
49	Student48	0.86	0.73	0.50	0.13	0.36
50	Student49	0.82	0.73	0.50	0.09	0.32
51	Student50	0.83	0.90	0.82	0.07	0.01
52	Student51	0.95	0.99	0.98	0.04	0.03
53	Student52	0.93	1.00	0.97	0.07	0.04

SN	Documents	Manual	LSA	LSA-NMF-ACO	DIFF LSA	DIFF NMF ACO
54	Student53	0.92	1.00	0.97	0.08	0.05
55	Student54	0.79	1.00	0.76	0.21	0.03
56	Student55	0.92	0.96	0.92	0.04	0.00
57	Student56	0.95	0.99	0.92	0.04	0.03
58	Student57	0.88	0.93	0.85	0.05	0.03
59	Student58	0.85	1.00	0.91	0.15	0.06
60	Student59	0.82	0.99	0.85	0.17	0.03
61	Student60	0.78	0.93	0.70	0.15	0.08
62	Student61	0.92	0.91	0.84	0.00	0.07
63	Student62	0.95	0.99	0.94	0.04	0.01
64	Student63	0.78	1.00	0.82	0.22	0.04
65	Student64	0.94	0.96	0.93	0.02	0.01
66	Student65	0.81	1.00	0.80	0.19	0.01
67	Student66	0.93	0.99	0.94	0.06	0.01
68	Student67	0.61	0.87	0.64	0.26	0.03
69	Student68	0.90	0.99	0.92	0.10	0.02
70	Student69	0.72	0.84	0.74	0.12	0.02
71	Student70	0.82	1.00	0.85	0.18	0.03
72	Student71	0.83	0.99	0.82	0.16	0.01
73	Student72	0.83	0.82	0.71	0.01	0.12
74	Student73	0.93	1.00	0.96	0.07	0.03
75	Student74	0.91	0.99	0.93	0.08	0.02
76	Student75	0.85	0.90	0.84	0.05	0.01
77	Student76	0.87	0.99	0.91	0.12	0.04
78	Student77	0.93	0.64	0.54	0.29	0.39
79	Student78	0.75	1.00	0.90	0.25	0.16
80	Student79	0.86	1.00	0.80	0.14	0.06
81	Student80	0.87	0.86	0.68	0.00	0.19
82	Student81	0.85	0.84	0.74	0.01	0.11
83	Student82	0.88	1.01	0.56	0.13	0.32
84	Student83	0.89	0.99	0.86	0.10	0.03
85	Student84	0.87	0.99	0.86	0.12	0.01
86	Student85	0.95	0.84	0.74	0.11	0.21
87	Student86	0.61	0.61	0.35	0.00	0.26
88	Student87	0.80	1.00	0.88	0.20	0.08
89	Student88	0.91	0.99	0.62	0.08	0.29
90	Student89	0.95	0.99	0.95	0.04	0.00
91	Student90	0.89	0.99	0.89	0.10	0.00
92	Student91	0.89	0.99	0.89	0.10	0.00
93	Student92	0.86	1.00	0.85	0.14	0.01
94	Student93	0.82	0.99	0.81	0.17	0.01
95	Student94	0.80	1.00	0.89	0.20	0.09
96	Student95	0.64	0.62	0.25	0.01	0.38
97	Student96	0.90	0.93	0.91	0.04	0.01
98	Student97	0.81	1.00	0.83	0.19	0.02
99	Student98	0.85	1.00	0.87	0.15	0.02
100	Student99	0.82	1.00	0.95	0.18	0.13
101	Student100	0.95	0.87	0.70	0.08	0.25
102	Student100	0.95	0.98	0.89	0.03	0.06
		Mean Divergence	0.127	0.0865		
		Assessment Accuracy	87.29	91.35		

A. Approximation Error

The approximation error is measured by calculating frobenious norm of the distance between the Weighted Term-Document Matrix matrix and the LSA-NMF-ACO dimensionally reduced matrix. The idea is that LSA-NMF-ACO matrix approximates the weighted matrix without compromising its vital semantic content. The computed error reveals the level of reduction. The approximation error consists of the absolute error and the relative error. The absolute error(ε) is given as

$$\varepsilon = \|\hat{x} - x\|$$

While the relative error rε=

$$r\varepsilon = \frac{\|\hat{x} - x\|}{\|x\|}$$

Where x is the weighted term document matrix that is being approximated and $\hat{x}$ is the dimensionally reduced matrix using LSA-NMF-ACO. A larger ε and $r\varepsilon$ indicates a better reduction.

The approximation error was computed for LSA and LSA-NMF-ACO. Table 3 shows the result. The higher the absolute and relative error the better and higher the reduction which implies that NMF-ACO has a better reduction than LSA.

Table 3: Comparative Absolute and Relative Approximation Error

TECHNIQUES	ε	$r\varepsilon$
LSA	2.190541	0.065943
NMF_ACO	10.580527	0.318513

B. Electronic Assessment Result

The improved modified algorithm was applied in assessing student performance on an introductory course in Artificial Intelligence for 101 students. Table 4 shows the comparative assessment using Manual, LSA and LSA-NMF-ACO.

C. Mean divergence and Measurement of Accuracy

Mean Divergence shows the ratio by which the automated system score differs (i.e. LSA-NMF-ACO) from the manual score at ± value. The difference is as a result of human emotional and cognitive scoring attribute associated to the manual system. Therefore, the divergence variance V of result of a question number q for n students is written as:

$$DF_{i,q} = |S_I - M_I|_q \quad (5)$$

$$V_q = \frac{\sum_i^n DF_{q,i}}{n} \quad (6)$$

where DF is set of score differences, M is score obtained from manual process, S is score
obtained from automated system respectively and i represents distinct student in set n.

The sum of the differences between the manual score and the system score

$\sum_i^n DF_{q,i}$= 8.82

The Mean Divergence $V_q = \frac{8.82}{102} = 0.087$

Accuracy of the system is calculated as:
Accuracy = 100 - (0.087× 100) = 91.35

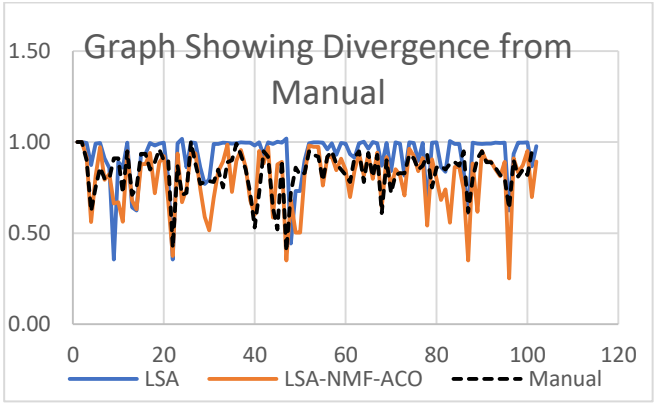

Figure 1: Divergence of Machine Graded Essay from Manual Grading

D. Pearson Correlation Analysis

The performance of the new system compared to the old manual system was measured by carrying out Pearson correlation analysis. Pearson's correlation determines the degree to which two linearly dependent variables are related. This was applied by computing the Pearson Correlation Coefficient (PCC) between the human grade scores and the machine grade, using the obtained values from Table 4. A correlation of 0.6321 exists between Human (HG) and Machine Grade (MG) which indicates that the two variables are closely related. The result is shown in Table 5. Figure 1 shows the graphical representation of this correlation.

Table 5: Pearson Correlations between Machine Grade and LSA-NMF-ACO Grade

	Pearson Correlation Coefficient (PCC)
MANUAL	1
LSA-NMF-ACO	0.6321

VI. CONCLUSION

This paper presented a modified LSA Algorithm using NMF-ACO. It demonstrates how the new algorithm improves the assessment result which is the consequence of further noise reduction in the semantic space. The LSA model was formulated as an optimization problem with the objective of minimizing the Frobenious Norm function of the NMF. The factors of NMF were initialized by the factors of the LSA for the purpose of quick convergence and iteratively refined using the ACO. The result achieved shows the new algorithm has a better result in terms of distant error, assessment accuracy and divergence from manual grading when compared to the existing LSA

REFERENCES

1. Anandarajan, M., & Nolan, T. (2019). Practical Text Analytics. Maximizing the Value of Text Data.(Advances in Analytics and Data Science. Vol. 2.) Springer.

2. Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44-58.
3. Darwish, S. M., & Mohamed, S. K. (2019, March). Automated Essay Evaluation Based on Fusion of Fuzzy Ontology and Latent Semantic Analysis. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 566-575). Springer, Cham.
4. Hoenkamp, E. (2011, September). Trading spaces: on the lore and limitations of latent semantic analysis. In Conference on the Theory of Information Retrieval (pp. 40-51). Springer, Berlin, Heidelberg.
5. Islam, M. M., and Hoque, A. L. (2010). Automated essay scoring using generalized latent semantic analysis. In *2010 13th International Conference on Computer and Information Technology (ICCIT)* (pp. 358-363). IEEE.
6. Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Journal of Educational Technology and Society*, *11*(3), 275-288.
7. Klein, R., Kyrilov, A., and Tokman, M. (2011, June). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education* (pp. 158-162). ACM.
8. Socha, K. (2009). Ant colony optimisation for continuous and mixed-variable domains. Saarbrücken: VDM Publishing.
9. Wild, F., Stahl, C., Stermsek, G., and Neumann, G. (2005). Parameters driving effectiveness of automated essay scoring with LSA.
10. Zhang, M., Hao, S., Xu, Y., Ke, D., and Peng, H. (2014). Automated essay scoring using incremental latent semantic analysis. *Journal of Software*, *9*(2), 429-437.

AUTHORS PROFILE

Rufai Mohammed Mutiu, A PhD student of Ladoke Akintola University of Technology**.** Did his first degree at Ogun State University, Ago-Iwoye and His masters degree at University of Lagos, Akoka. He is a member of Nigeria Computer Society, International Association of Engineers, International Association of Computer Science and Information Technology. He is currently a principal lecturer in the Department of Computer Technology, Yaba College of Technology, Yaba

A. O. Afolabi is a Professor of Computer Science in Department of Computer Science and Engineering, Ladoke Akintola University of Technology Ogbomoso, Nigeria.

Dr. (Mrs.) O. D. Fenwa, is a Senior Lecturer in the Department of Computer Science and Engineering, Ladoke Akintola University of Technology
Ogbomoso, Nigeria

Dr. (Mrs.) F. A. Ajala, **is an** Associate Professor of Computer Science in the Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria