Original Research Article

PERFORMANCE EVALUATION OF LSA, NMF and ILSA IN ELECTRONIC ASSESSMENT OF FREE TEXT DOCUMENT

ABSTRACT

Aims: To evaluate the performance of an Improved Latent Semantic Analysis (ILSA), Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF) algorithms in an Electronic Assessment Application using metrics, Semantic Adequacy, Term-Term Similarity, Precision, Recall and F-measure function, Mean divergence and Assessment Accuracy.
Methodology: The three algorithms were separately applied in developing an Electronic Assessment application. One hundred students' responses to a test question in an introductory artificial intelligence course were used. Their performance was measured based on the following metrics, assessment accuracy, divergence from manual, adequacy of semantic representation and retrieval quality.
Results: ILSA outperformed the LSA and NMF with an assessment accuracy of 96.64, mean divergence from manual score of 0.03, and recall, precision and f-measure value of 0.83, 0.85 and 0.87 respectively.
Conclusion: The research observed the performance of an improved algorithm ILSA for electronic Assessment of free text document using Adequacy in Semantic Representation, Retrieval Quality and Assessment Accuracy as performance metrics. The results obtained from the experimental designs shows the adequacy of the improved algorithm in semantic representation, better retrieval quality and improved assessment accuracy.

Keywords: Electronic Assessment, Latent Semantic Analysis, Ant Colony Optimization

1. INTRODUCTION

Electronic Assessment is the use of information technology for any assessment-related activity (Jordan 2013). An E-assessment application is adjudged worthwhile if it generates nearly human grade assessment. There are two basic approaches to Electronic Assessment. These are Information Retrieval (IR) based and Linguistics-based. LSA, NMF and ILSA belong to the information retrieval approach, where keywords and their co-occurrence statistics are used to reveal hidden semantic links between a gold standard (lecturer's marking scheme) and the essay to be assessed. ILSA is a hybrid algorithm that integrates LSA, NMF and ACO to address the inadequacy associated with LSA. LSA has dimensionality approximation and noise reduction problem while NMF has the problem of convergence at local minimal (Hoenkamp, 2011; Ayesha et al., 2020). The performance of LSA or any of its hybrid can be improved by adjusting any of its operational parameters which are Document pre-processing, weighting, dimensionality reduction and similarity measurement. The ILSA algorithm improved the existing LSA in the area of dimension reduction by substituting the SVD in LSA with the ILSA. The LSA in ILSA is used for initialization of the factors of NMF while the objective function of NMF is optimally minimized using an ACO algorithm.

The rest of the paper is divided into 5 sections. Section introduced the topic. Section 2 reviewed related works revealing the strategies used by researchers, the limitation of work and the results achieved. Section 3 discuss the methodology used in terms of stages involved the electronic assessment system and the developed algorithm. Section 4 presents the observed comparative

experimental results while section 5 concludes the paper and other areas of application of the developed algorithm.

2. RELATED WORKS

Open ended question is a question whose response is not a single word but requires the composition of free-text document where students give their answers in different pattern and yet may be expressing the same thing. The fact that answer in an open-ended question is free text makes its analysis and comparison with the lecturer marking scheme difficult because we are not looking for a word match but a semantically similar text. Several algorithms have been developed to represent free-text document and check for semantic similarity in an electronic assessment context. A review of these algorithms is presented along with their performance value.

Amalia, Gunawan, Fithri and Aulia, (2019) applied LSA in automating assessment in an essay written in Bahasa Indonesia. The essays were initially subjected to pre-processing step which ensured the removal of punctuations and irrelevant symbols, conversion of the entire text to lower case, breaking the sequences of strings to minimal meaningful units, stop word removal and reducing the word content to their root form. A document-term matrix which has the document label as its rows and pre-processed term as the columns. Its entries are the frequency of occurrence of the term in the document. The entries of the matrix were weighted, after which Single Value Decomposition was used to decompose the matrix. The resultant matrix was compared with the lecturer row vector using cosine similarity rule to obtain the score of the students. The LSA technique was evaluated against manual assessment by Amalia et al (2019) and the result was 83.3%. This work demonstrates the portability of LSA across languages. However, the accuracy level needs to be improved.

The work of (Mokhtari-Fard, 2013) is a practical demonstration of the application of Natural Language Processing methods. In this algorithm, the questions and correct answers are separately received by the system in the form of natural language. Then, the accurate answer for each question is converted into objects which represent the object-based representations of the input texts. In the subsequent step, the user inputs the relevant answer for each question through Graphics Users Interface (GUI). The system converts the input text for each question into distinctive objects. For analysing the accuracy level of the answers, the created objects are compared with each other and the answering grades are calculated. On the other hand, it is hard to accomplish and very difficult to port across languages. Other systems that use the Natural Language approach are: C-rater and Paperless School free text Marking Engine (PS-ME). Another approach used in NLP is to grade the student essay by summarizing it so that only the relevant information is taken into account and this minimizes the presence of noise (Burstein and Marcu, 2000).

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

3. MATERIAL AND METHODS

ILSA is a hybrid algorithm that incorporates LSA, NMF and ACO to address specific problem area of the existing LSA for Electronic Assessment of Free-Text Document. One of the conspicuous problems associated to LSA is the interpretability problem which manifest as the presence of negative values in the resultant Document Term matrix. This problem was resolved by initialising the factors of NMF with the absolute value of the factors of LSA in order to form a bound optimisation problem of the form

Min f(W,H), W≥0, H≥0 1

And the objective function is

$$f(W,H) = \frac{1}{2}\|Z - WH\|_F^2 \quad 2$$

Where Z is a non-negative Document Term Matrix and W and H are the factors of NMF. The ACO optimization algorithm seek the values of W and H that minimizes the objective function in Equation 2. Figure 1 below shows the block diagram for the ILSA algorithm.

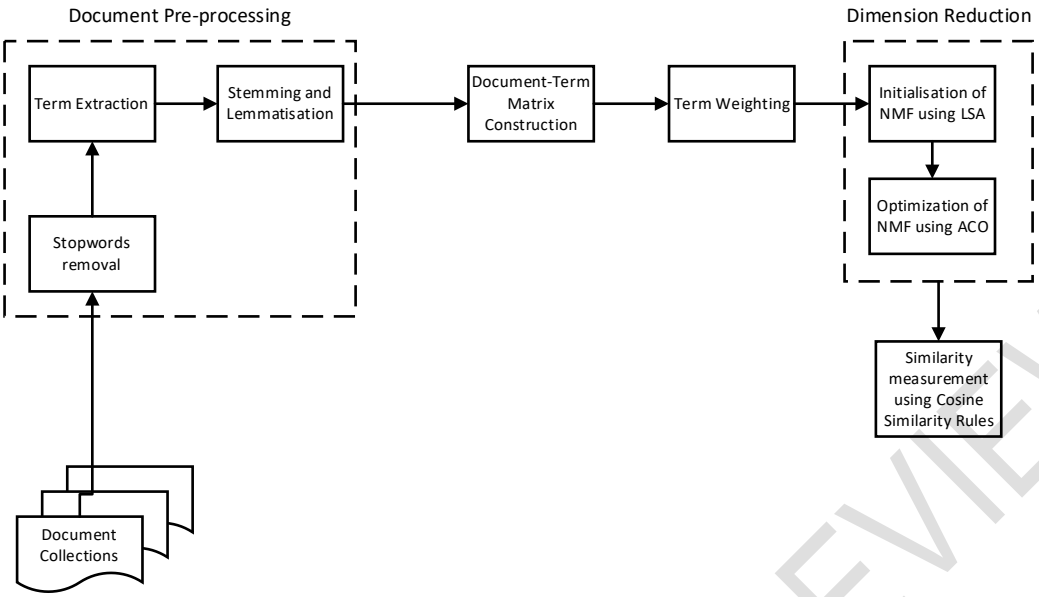

Figure 1: Block Diagram of The ILSA Algorithm

Relevant documents which comprise of the scripts to be graded and the lecturer marking scheme that will be used for the grading were collected. These documents were used for training and extraction of terms using Syntatic Analyser. The extracted terms and the documents from which they were extracted, were used to create a document-term matrix, where documents tagged as student1, student2, student3 (based on the number of students) serve as the matrix row headings and the terms as the column headings. The entries to the matrix were the frequencies of occurrence of a term in a particular document.

The entries were weighted using Term Frequency - Inverse Document Frequency (TF-IDF) weighting scheme in order to give emphasis to terms with higher semantic value. The weighted matrix was subjected to dimension reduction using a combination of Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization and Ant Colony Optimization Techniques in order to filter out noise and words with less semantic contribution. (Rufai et.al., 2021)

The similarity value between the lecturer Marking scheme and the students' responses was determined by evaluating similarity of their vectors using cosine similarity rule.

Algorithm 1 shows the steps/procedure of the ILSA techniques

Algorithm 1: The ILSA Algorithm that Integrates LSA,NMF and ACO

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

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

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

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

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=$\|z_i^r - w_i^r H\|_F$ using ACO

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

Step 9: Multiply the minimized rows of W with the minimized column of H to obtain $\hat{Z}$ which is the reduced dimension of Z

Step 10	Compute the similarity value between the first column of $\hat{Z}$ and its other columns using the cosine similarity rule expressed as: $$cosSIM(\vec{A}.\vec{B}) == \frac{\vec{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_{i=1}^2}}$$ Where A is the first column of $\hat{Z}$ and B is used interchangeably for any other column and it represent the students answer.

4. RESULTS AND DISCUSSION

The developed ILSA algorithm is compared with its component algorithms (i.e. LSA and NMF) to reveal its better performance using the following metrics:

b. Semantic Adequacy

It is a measure of how adequately the semantic space capture the semantic content of the documents involved. In this research we assessed the semantic representational adequacy by using two methods which are the Term-Term Cosine Similarity measure and The Precision, Recall and F-Measure Functions.

i. Term Similarity

Term-Similarity is a measure of how semantically close a term is to another term. Our approach uses Term similarity to confirm two naturally similar terms and two naturally dis-similar terms. The cosine angle was computed on the Document Term Matrix also known as the semantic space of LSA, NMF and ILSA to show their respective semantic adequacy. The similarity between two terms is given as

$$cosSIM(\overrightarrow{t_1}.\overrightarrow{t_2}) = cos(\theta) = \frac{\overrightarrow{t_1}.\overrightarrow{t_2}}{\|\overrightarrow{t_1}\|\|\overrightarrow{t_2}\|} \quad 3$$

Table 1: Similarity Value of Synonyms

Term1	LSA	NMF	ILSA	Term2
Computer	0.80	0.79	0.84	Hardware
Create	1.00	0.95	0.70	Make
Decision	1.00	1.00	0.96	Logic
Internet	0.80	0.79	0.84	Web
Laptop	0.80	0.79	0.84	Computer
Learn	0.78	0.77	0.74	Study
Learn	1.01	1.00	0.98	Knowledge
Learn	1.00	1.00	0.99	Acquire
Behaviour	1.00	1.00	0.88	Manner
Machine	0.89	0.86	0.89	Device
Develop	0.73	0.73	0.69	Design
Task	0.78	0.76	0.54	Work
Theory	1.00	0.99	0.92	Study
Think	1.00	1.00	0.93	Logic
Man	0.93	0.93	0.91	Human

The result of Table 1 shows the similarity values between Term1 and Term2 in the semantic space of LSA, NMF and ILSA. A similarity value of >=0.5 confirms similarity while <0.5 shows dissimilarity. The words compared in Table 1 are similar terms that can be used interchangeably in a sentence without altering the sentence meaning which is the characteristics of synonyms.

Table 2: Word Pairs with low Similarity values

Term1	LSA	NMF	ILSA	Term2

Human	0.37	0.33	0.15	Machine
Hardware	0.42	0.42	0.41	Software
Natural	0.75	0.7	0.33	Artificial
Solution	0.46	0.43	0.38	Problem
Input	0.29	0.29	0.29	Output

Table 2 shows the low similarity recorded for naturally dis-similar terms by the tree algorithms. However, LSA and NMF erred in their similarity values between "natural" and "artificial" which may be the consequence of there poor noise handling mechanism.

ii. Precision, Recall and F-Measure

The Precision, Recall and F-Measure measures the performance of the ILSA, LSA, NMF in terms of retrieval quality which indirectly reflects performance in terms of semantic capturing. Retrieval using any of these aforementioned approach compares the query text and documents using their semantic content to rank the documents in order of their similarity. Documents having a similarity of 0.5 portrays similarity and hence will be retrieved as relevant document to the query.

In this research we used 102 documents as the search space and 11 query texts. Relevant documents were sought for in the search space.

The procedures followed were:

1. Represent the query text as a vector of the terms in the Term Document matrix.
2. Convert the vector to a scaled, weighted sum of component term vectors using

$$q = q^T V_k S_k^{-1} \quad (4)$$

3. Compute the cosine similarity between the query vector and each document in the collection to determent relevant documents and level of relevance
4. Compute Precision, Recall and F-Measure using Equation 4, 5 and 6 respectively

Precision (P) is the fraction of retrieved documents that are relevant given as:

$$P = \frac{number\ of\ relevant\ documents\ retrieved}{number\ of\ retrieved\ documents} \quad (5)$$

Recall is given as:

$$R = \frac{number\ of\ relevant\ documents\ retrieved}{number\ of\ relevant\ documents} \quad (6)$$

F measure, which is the weighted harmonic mean of precision and recall is given as:

$$F = 2 \times \frac{P \times R}{P + R} \quad (7)$$

Table 3 shows the result. Precision shows the proportion of relevant documents that are retrieved why Recall shows the proportion of retrieved document that are relevant. It is expected that technique with better semantic capturing shows a better Precision, Recall and F-Measure result which can be observed in the result of ILSA. ILSA outperforms LSA, NMF in terms of retrieval quality. Figure 2 shows the graphical representation of this performance.

c. Mean divergence and Measurement of Accuracy

Divergence measures the difference between the machine generated score and the manual score at ± value. The machine score will be acceptable if the difference between it and the human score is minimal (Adesiji *et al*. 2016). The divergence variance V of result of a question q for n students is given in equation 4.7 and 4.8 as:

$$DF_q = |S - M|_q \quad (8)$$

$$V_q = \frac{\sum_i^n DF_q}{n}$$

where DF is set of score differences, M is score obtained from Human score, S is score obtained from machine i represents distinct student in set n.
$\sum_i^n DF_q$ is the sum of the differences between the machine score and the human score

Accuracy = 100-(100*Vq)

The result of the divergence and accuracy of the various algorithm is given in Table 3

Table 3: Mean Divergence and Assessment Accuracy of the Various Algorithm

	Mean Divergence	Assessment Accuracy
LSA	0.10	89.52
NMF	0.09	91.13
ILSA	0.03	96.63

Table 4: Precision, Recall and F-Measure Evaluation for LSA, NMF and LSA-NMF-ACO

		No of Document Correctly Retrieved(n)			No of Document Retrieved(N)			Number of documents that should be retrieved as related document(R)			PRECISION			RECALL			F-MEASURE		
QUERY	QUERY TEXT	LSA	NMF	LSA-NMF-ACO	LSA	NMF	LSA-NMF-ACO	LSA	NMF	LSA-NMF-ACO	LSA	NMF	LSA-NMF-ACO	LSA	NMF	LSA-NMF-ACO	LSA	NMF	LSA-NMF-ACO
Q1	Task World	68	80	97	102	102	102	88	100	100	0.67	0.78	0.95	0.86	0.98	0.98	0.75	0.87	0.97
Q2	Dependent Human	75	89	99	102	102	102	88	102	102	0.74	0.87	0.97	0.86	1.00	1.00	0.79	0.93	0.99
Q3	Computer	83	63	59	102	102	102	102	86	100	0.81	0.62	0.58	1.00	0.84	0.98	0.90	0.71	0.73
Q4	Visual World	62	65	93	102	102	102	88	98	100	0.61	0.64	0.91	0.86	0.96	0.98	0.71	0.77	0.94
Q5	Human Intelligence	82	73	87	102	102	102	102	100	100	0.80	0.72	0.85	1.00	0.98	0.98	0.89	0.83	0.91
Q6	Watch Theory	85	90	77	102	102	102	100	100	100	0.83	0.88	0.75	0.98	0.98	0.98	0.90	0.93	0.85
Q7	SCIENCE REQUIRE	76	97	66	102	102	102	102	100	100	0.75	0.95	0.65	1.00	0.98	0.98	0.85	0.97	0.78
Q8	WEB WORK	67	83	89	102	102	102	100	100	100	0.66	0.81	0.87	0.98	0.98	0.98	0.79	0.89	0.92
Q9	THINK VIRUS	96	60	80	102	102	102	102	99	100	0.94	0.59	0.78	1.00	0.97	0.98	0.97	0.73	0.87
Q9	ASSUME BASE AREA	59	70	80	102	102	102	88	100	100	0.58	0.69	0.78	0.86	0.98	0.98	0.69	0.81	0.87
Q10	ABLE LAPTOP	75	83	60	102	102	102	102	100	100	0.74	0.81	0.59	1.00	0.98	0.98	0.85	0.89	0.74
Q11	VISUAL RECOGNITION	84	91	81	102	102	102	99	100	100	0.82	0.89	0.79	0.97	0.98	0.98	0.89	0.93	0.88
										AVERAGE	0.75	0.77	0.79	0.95	0.97	0.98	0.83	0.85	0.87

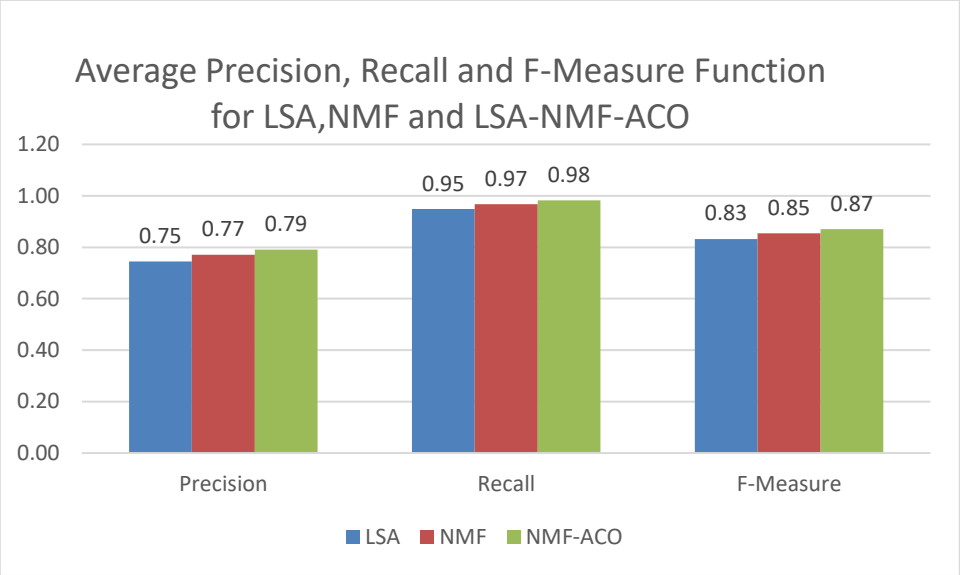

Figure 2: The Relative Performance of LSA-NMF-ACO for Precision, Recall and F-Measure function

CONCLUSION

The research observed the performance of an improved algorithm ILSA for electronic Assessment of free text document using Adequacy in Semantic Representation, Retrieval Quality and Assessment Accuracy as performance metrics. The results obtained from the experimental designs shows the adequacy of the improved algorithm in semantic representation, better retrieval quality and improved assessment accuracy. An accuracy of 96.63% was observed and an average precision, recall and F-Measure value of 0,79,0.98 and 0.87 were recorded respectively. The results show that the introduction of optimization to the LSA process improve the assessment results.

The work can be adopted by Examination conducting bodies and educational institutions for mass marking of theoretical questions. However, future work can be geared towards investigating the performance of other optimization techniques on noise reduction and its effect on assessment accuracy.

References

1. Amalia, A., Gunawan, D., Fithri, Y., & Aulia, I. Automated Bahasa Indonesia essay evaluation with latent semantic analysis. In Journal of Physics: Conference Series IOP Publishing.2019; 1235(1):012100.

2. Ayesha, S., Hanif, M. K., & Talib, R.. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 2020;59:44-58.

3. Burstein, J., and Marcu, D. Towards using text summarization for essay-based feedback. In La 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN. 2000

4. 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.

5. Hoenkamp, E. Trading spaces: on the lore and limitations of latent semantic analysis. In Conference on the Theory of Information Retrieval. Springer, Berlin, Heidelberg. 2011; 40-51

6. Jordan, S. E-assessment: Past, present and future. New Directions in the Teaching of Physical Sciences, 2013;9:87-106.

7. Mokhtari-Fard, I. Natural Language Understanding for Grading Essay Questions in Persian Language. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Berlin, Heidelberg. 2013; 144-153

8. Rufai M. M, Afolabi A, Fenwa O. D, & Ajala F.A. (February,). An Improved LSA Model for Electronic Assessment of Free Text Document. International Journal of Innovative Technology and Exploring Engineering Blue Eyes Intelligence Engineering and Sciences Publication. 2021;10(4):152-159