Proceedings of 13th International Conference on Computer and Information Technology (ICCIT 2010)
23-25 December, 2010, Dhaka, Bangladesh

Automated Essay Scoring Using Generalized Latent Semantic Analysis

Md. Monjurul Islam [#1], **A. S. M. Latiful Hoque** [#2]

[#]Department of CSE, Bangladesh University of Engineering & Technology, Dhaka, Bangladesh (BUET)
[1]mdmonjurul@gmail.com , [2]asmlatifulhoque@cse.buet.ac.bd

Abstract

Automated Essay Grading (AEG) is a very important research area in educational technology. Latent Semantic Analysis (LSA) is an information retrieval technique used for automated essay grading. LSA forms a word by document matrix and then the matrix is decomposed using Singular V alue Decomposition (SVD) technique. Existing AEG systems based on LSA cannot achieve higher level of performance to be a replica of human grader. We have developed an AEG system using Generalized Latent Semantic Analysis (GLSA) which makes n-gram by document matrix instead of word by document matrix. We have evaluated this system using details representation and showed the performance of the system. Experimental results show that our system outperforms the existing system.

Keywords: Automatic Essay Grading, Latent Semantic Analysis, Singular Value Decomposition, N-gram.

I. INTRODUCTION

Automated essay scoring is an essential part of educational process. Scoring students' writing is one of the most expensive and time consuming activity for educational assessment. The interest in the development and in use of automated assessment system has grown exponentially in the last few years [1], [5], [7], [10]. Most of the automated assessment tools are based on objective-type questions: i.e.multiple choice questions, short answer, selection/association, hot spot, true/false and visual identification [1], [10]. Most researchers in this field agree that some aspects of complex achievement are difficult to measure using objective-type questions. The assessment of essays written by students is more fruitful for measurement of complex achievement. Essay grading is a time consuming activity. It is found that about 30% of teachers' time is devoted to marking [1], [2]. This issue may be faced through the adoption of automated essay grading (AEG) system.

Several AEG systems have been developed under academic and commercial initiative using statistical [17], Natural Language Processing (NLP) [18], Bayesian text classification [5], Information Retrieval (IR) technique [3], amongst many others.

Latent Semantic Analysis (LSA) is a powerful IR technique that uses statistics and linear algebra to discover underlying "latent" meaning of text and has been successfully used in English language text evaluation and retrieval [3], [8], [9]. LSA applies Singular Value Decomposition (SVD) to a large term by context matrix created from a corpus, and uses the results to construct a semantic space representing topics contained in the corpus. Vectors representing text passages can then be transformed and placed within the semantic space where their semantic similarity can be determined by measuring how close they are from one another.

The main dimension of measuring performance of AEG is how much the system accurate with human grade. The existing AEG techniques which are using LSA do not consider the word sequence of sentences in the documents.

In existing LSA methods the creation of word by document matrix is somewhat arbitrary [3]. Automated essay grading by using these methods are not a replica of human grader.

In our research we have focused on IR based LSA technique for essay grading. We have developed an AEG system by using Generalized Latent semantic Analysis (GLSA). The GLSA consider word sequence of sentences in the documents. The new AEG system grades essay with more accuracy than the existing techniques.

The rest of the paper is organized as follows: In section II we have presented existing approaches to the automated assessment of essays. In section III we have discussed system architecture of our model. In section IV we have analyzed our developed model. Our proposed technique is compared with the existing in the section V.

II. EXISTING ESSAY GRADING SYSTEMS

Automatic E ssay Grading (AEG) system is a very important research area for using technology in education. Researchers have been doing this job since the 1960's and several models have been developed for AEG. As early as 1966, Page developed Project Essay Grader (PEG) which uses multiple regression technique which grades essays on the basis of writing quality, taking no account of content [1]. Vector–space model was developed by Sparck Jones at 1972 which starts with co-occurrence term-document matrix formed from the essay. TF-IDF is used for weighting the elements of matrix and Cosine correlation is used for scoring. It is less successful at judging overall quality essay [3].

E-rater is an essay-scoring system developed in 1990 by Educational Testing Service. E-rater uses multiple regressions with NLP to extract writing features of essays. By comparing human and E-Rater grades essays with 87% accuracy [1], [18].

Hearst et al. developed Intelligent Essay Assessor (IEA) which is based on the Latent Semantic Analysis (LSA) technique that was originally designed for indexing documents and text retrieval. Whittington et al. defined that LSA represents documents and their word content in a large two-dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationship are modified to more accurately represent their true significance. The correlation from 0.59 to 0.89 has been achieved between the IEA and human graders [10], [13].

Bayesian Essay Test Scoring System (BETSY) is a program that classifies text based on trained material and is being developed by Lawrence M. Rudner. BETSY uses Multivariate Bernoulli Model (MBM) and the Bernoulli Model (BM). Using BESTSY an accuracy of over 80% was achieved.

IntelliMetric uses a blend of Artificial Intelligence (AI), Natural Language Processing (NLP), and statistical technique. IntelliMetric process assists to examine the essay according to the main characteristics of standard written English. Shermis et al. is claimed that the correlation between human garder and IntelliMetric is poor [1], [10], [18].

Bin L. et al. designed an essay grading technique that used text categorization model which incorporates K-Nearest Neighbor (KNN) algorithm. Transforming the essays into the vector space model (VSM), TF-IDF and IG are applied for feature selection from the feature pool of words, phrases and arguments. After training for the KNN algorithm, a precision over 76% is achieved on the small corpus of text [6].

Nahar K. M. et al. is developed an Automatic Grading for Online exams in Arabic with Essay Questions using statistical and computational linguistics techniques. By this method 60% of accuracy is gained [7]. In the next section we have discussed our new architecture of essay grading system.

III. AEG WITH GLSA: SYSTEM RCHITECTURE

We have developed a system for essay grading using Generalized Latent Semantic Analysis (GLSA). Normally LSA represents documents and their word content in a large two-dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationship are modified to more accurately represent their true significance. Each word represents a row in the matrix, while each column represents the sentences, paragraphs and other subdivisions of the context in which the word occurs.

The traditional word by document matrix creation of LSA does not consider word sequence in a document. Here the formation of word by document matrix the word pair “carbon dioxide” makes the same result of “dioxide carbon”.

We have proposed a system for essay grading by using Generalized Latent Semantic Analysis (GLSA).

In GLSA n-gram by document matrix is created instead of a word by document matrix of LSA. According to GLSA, a bi-gram vector for “carbon dioxide” is atomic, rather than the combination of “carbon” and “dioxide”. The GLSA preserve the proximity of word in a sentence. We have used GLSA because it generates clearer concept than LSA.

Our whole system architecture has been spited into two main parts: generation of training essay set and essay evaluation of submitted essays using training essay sets.

A. Training essay set generation

The training essay set generation is shown by Fig. 1. We can select essays of a particular subject of any levels. The essays are graded first by more than one human experts of that subject. The number of human graders may increased for the non-biased system. The average value of the human grades has been treated as training score of a particular training essay. The preprocessing has been done on training essay set. In preprocessing steps the stopwords have been removed from the essay and words have been stemmed to their roots. For stemming we have used M. F. Porter’s stemming algorithm [12].

N-grams i.e. unigrams, bigrams, trigram, …, ngrams index terms have been selected for making the n-gram by documents matrix. At first we have selected some important words as unigram and then their neighbors’. For bigram the first and second neighbors’ of the selected words have been selected, for trigram first, second and third neighbors’ have been selected. In this way we have generated n-grams. The n-gram by document matrix has been created by using the frequency the n-gram present in an essay. Each cell of the matrix has been filled by the frequency of n-grams in the document.

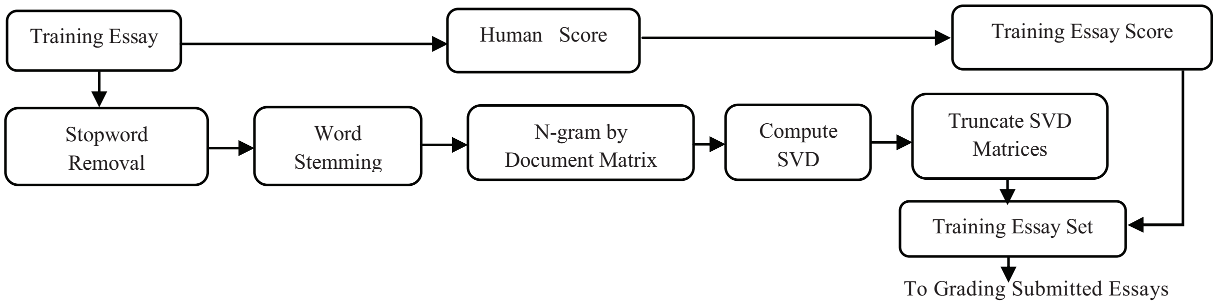

Fig. 1 Training set generation

Fig. 2 shows the decomposition of n-gram by document matrix using SVD of matrix. According to SVD a matrix A_{txd} has been decomposed as follows:

$$A_{txn} = U_{txn} * S_{nxn} * V_{dxn}^{T} \qquad (1)$$

Here, A is n-gram by documents matrix, U is an orthogonal matrix, S is a diagonal matrix and V^T is the transpose of an orthogonal matrix V.

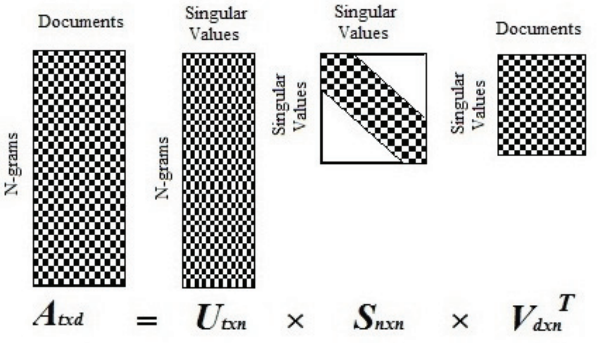

Fig. 2 Singular Value Decomposition of matrix

The columns of U are orthogonal eigenvectors of AA^T. The columns of V are orthogonal eigenvectors of A^TA and S is a diagonal matrix containing the square roots of eigenvalues of V in descending order.

Fig. 3 shows the dimension reduction of SVD matrices.

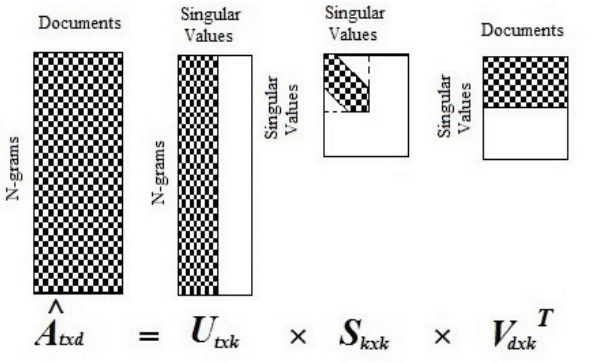

Fig. 3 The dimensionality reduction of SVD matrices

As in fig. 3 the dimensionality reduction operation has been done by removing one or more smallest singular values from singular matrix S and also deleted the same number of columns and rows from U and V, respectively.

The purpose of the dimensionality reduction is to reduce the noise and unimportant details in the data so that the underlying semantic structure can be used to compare the content of essays. The truncated SVD matrices have been used for making the training essay vectors.

Training essay vectors d'_j have been created for each document vector d_j from the truncated SVD matrices as in

$$d'_j = d_j^{T} * U_{txk} * S_{kxk}^{-1} \qquad (2)$$

Here, d_j^T is the transpose of document vector d_j, U_{txk} is truncated left orthogonal matrix and S_{kxk} is truncated singular matrix of truncated SVD.The document vectors d'_j along with human grades of training essays make the training essay set.

B. The evaluation of submitted essay

Fig. 4 shows the evaluation part of our architecture. The submitted essays have been graded first by the human grader. The pregarded essays have been checked for lingual errors. Positive or negative marking has been given on the basis of lingual error. Stopwords have been removed from the essays and the words have been stemmed to their roots. Query matrix (q) has been formed by the submitted essay according to rules of making n-gram by documents matrix. Query vector q' has been created from the submitted essay as in

$$q' = q * U_{txk} * S_{kxk}^{-1} \qquad (3)$$

Here, q is query matrix U_{txk} is truncated left orthogonal matrix and S_{kxk} is truncated singular matrix of SVD.

Similarity between query vector (q') and training essay set vectors d'_j has been calculated by Cosine similarity as follows

$$Sim(q', d'_j) = \frac{\sum_{j=1}^{t} w_{qj} * d_{ij}}{\sqrt{\sum_{j=1}^{t} (d_{ij})^2 * \sum_{j=1} (w_{qj})^2}} \qquad (4)$$

Here, w_{qj} is the j^{th} weight of query vector (q')and d_{ij} is the i^{th} weight of training essay set vectors d'_j. The highest correlation value from the Cosine of query vector and the training essay vector has been used for grading the submitted essay.

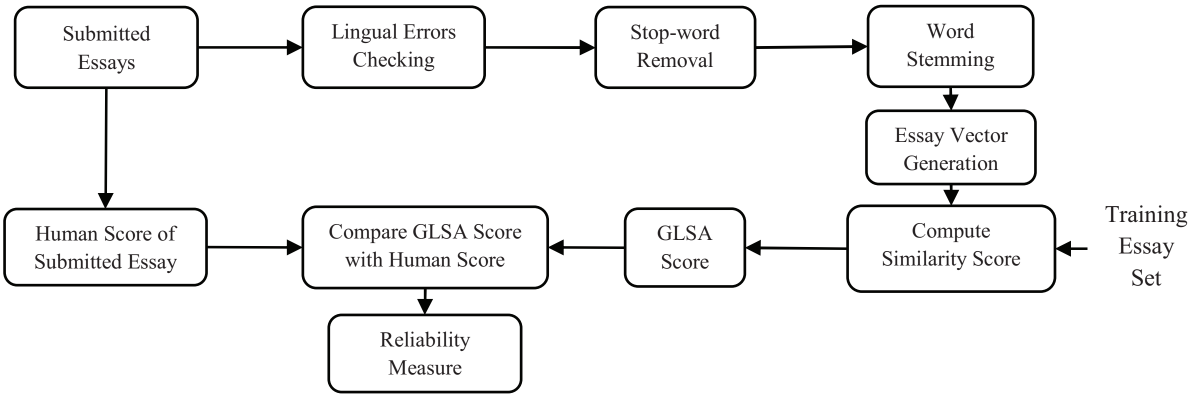

Fig. 4 Evaluation of submitted essay

The grade of submitted essay has been assigned by the grade of training which made a maximum similarity. The grade has been treated as LSA score.

IV. ANALYSIS OF AEG WITH GLSA

The preprocessing has been done by removing stopwords and word stemming. The M. F. Porter stemming algorithm has been used for stopword removal and stemming the words to their roots. The preprocessing steps increase the performance of our AEG system.

The n-gram by document matrix has been created by using the frequency of n grams in a document. For each cell n-gram by document matrix has been filled by a_{ij}= tf_{ij}. The n-gram by documents matrix has been decomposed by singular value decomposition (SVD) of matrix. The SVD of matrix has been done by the following algorithm.

```
Input: Matrix A of order mxn
Output: The $U_{mxp}$, $S_{pxp}$, $V_{pxn}$ Matrices when,
        $A=U*S*V^T$
Multiplicate A by the transpose of A and put it to T
Compute $\lambda_1, \lambda_2, \lambda_3,\ldots, \lambda_n$ the eigenvalues of T
FOR i=1 to n DO
        $\mu_i$=sqrt($\lambda_i$)
ENDFOR
Sort $\mu_1, \mu_2,\ldots, \mu_n$ in descending order
FOR i =1 to m DO
        FOR j =1 to n DO
                IF (i = j) THEN
                        set $S_{ij} = \mu_i$
                ELSE
                        set $S_{ij}$ =0
                ENDIF
        ENDFOR
ENDFOR
FOR i=1 to n DO
$u_i$ =eigenvector of $\lambda_i$
ENDFOR
Create a matrix V having the $u_i$ as columns
$V^T$ = the transpose of V
$U = A*V*S^{-1}$
```

The complexity of SVD algorithm is $O(mn^3)$for a matrix of order m x n.

The SVD matrices $\mathbf{U}_{txn}$, $\mathbf{S}_{nxn}$ and $\mathbf{V}_{dxn}^T$ have been truncated by removing one or more smallest singular values from singular matrix S and also deleted the same number of columns and rows from U and V, respectively. The dimension reduction algorithm for SVD matrices is as follows.

```
Input: $U_{txn}$, $S_{nxn}$, $V_{dxn}^T$
Output: $U_k$, $S_k$ and $V_k^T$
Set flag to 0
FOR i = 0 to n DO
        IF ($S_{i,i}$ < 0.5) THEN
                flag = i - 1
        ENDIF
        Increment i
ENDFOR
$S_k$  = The submatrix of S of order k x k
$U_k$  = The submatrix of U of order t x k
$V_k^T$ = The submatrix of V of order k x p
```

The SVD decomposition reduces errors for the AEG system. Moreover, reduction of SVD matrices makes clearer concepts. The essay vectors have been computed by the algorithm of training essay set generation. The algorithm of training essay set generation is as follows.

```
Input: A set training essays, E = {E1, E2, ……, En}
Output: A set of essay vectors, D = {d1, d2,…, dn}
Step 1: FOR i = 1 to n DO
            a. Remove stop-words from essay Ei
            b. Stemmed words of essay Ei to their root
        ENDFOR
Step 2: Selects ngarms for the index terms of n-garm
            bydocument matrix.
Step 3: Builds an n-gram by document matrix, $A_{mxn}$
            where each matrix cell $a_{ij}$ is the frequency of n-
            gram $N_i$ that appears in the document $d_j$.
Step 4: Decompose $A_{mxn}$ matrices using SVD of matrix,
            such that, $A_{mxn}=U_{mxp}*S_{pxp}*V_{pxn}^T$
Step 5: Truncate the U, S and $V^T$ and make
            $A_{kxk}= U_{mxk} * S_{kxk} * V_{kxn}$
Step 6: FOR j =1 to n DO
            Make the essay vector, $d_j=d_{jk}^T * U_{mxk}* S_{kxk}^{-1}$
        ENDFOR
```

In the evaluation part the query matrix (q) has been formed by the submitted essay according to rules of making n-gram by documents matrix. Query vector has been created by the following algorithm.

```
Input: A Submitted Essay for Grading, Q
Output: Query vector $\acute{Q}$
Step 1: Preprocess the Submitted Essays
            a) Remove stop-words from essay Q
            b) Stemmed words of essay Q to their root
Step2: Builds an one dimensional query matrix $q_{mx1}$
            same as the rule of the creating n-gram by
            document matrix.
Step 3: Makes the query vector $\acute{Q} = q_{mx1}^T * U_{mxk}* S_{kxk}^{-1}$
```

The above query vector along with the document vector grades the submitted essay by the following AEG algorithm.

```
Algorithm AEG
Input: Submitted Essay, Q
Output: Grade calculated by AEG of submitted essay, G
Step 1: Grades the submitted essay Q by human expert,
            the grade is H
Step 2: Compute query vector $\acute{Q}$ from the submitted
            essay
Step 3: Compute the Cosine similarity between the $\acute{Q}$
            and the each essay vector Di, such that,
```

Step 4: Finds the maximum value of cosine similarity M
Step 5: Assigns the grade of the training essay to G which creates M
Step 6: Compares G with H

In the evaluation phase the grades of submitted essays have been compared with the human grade for reliability measure of our method. For comparison, we have computed the mean of errors by averaging the magnitude each machine score deviated from its corresponding human score. In addition, we also computed the standard deviation of errors refer to "(4)" and "(5)" respectively.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

$$SD = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ...+ (x_n - \bar{x})^2}{n}} \tag{6}$$

When
$\bar{x}$ is the arithmetic mean from all errors
x_i is an absolute value of an error between human score and machine score
n is the number of data set where n =120

In the next section we have discussed our experimental results and compared our results with existing LSA based AEG techniques.

V. EXPERIMENTAL RESULTS

We have trained our system by 960 essays written by undergraduate students. The themes of the essays are "Digital Divide", "Tree Plantation" and "E-Governance". We have tested our model by 120 essays written by undergraduate students. The total mark is 100. The score of each essay ranged from 2 point to 4 points, where a higher point represented a higher quality. The score of an essay has obtained by averaging the scores from three teachers which has been treated as human grade. The numbers of essays corresponding to different scores in the range 0.00, 2.00, 2.5, 3.00, 3.5, 4.00 for obtained marks less than 40, 40-49, 50-59, 60-69, 70-79 and 80-100, respectively. Table I shows the summary of essay set used in this experiment.

Table I The essay sets used in the experiments

Set no.	Topic	Level	Training Essay	Test Essays
1	Digital Divide	Undergraduate	40	40
2	Tree Plantation	Undergraduate	40	40
3	E-governance	Undergraduate	40	40

The performance of a method for scoring essays can be evaluated by measuring how much the automated grade closer to the human grade. The more closely the automated grade to the human grade is more accurate. Table II shows the relationship between human scores and the scores graded by the proposed method.

Table II Performance of AEG using generalized latent semantic analysis

Human Score	No. of Test Essay	AEG with GLSA Score					
		4.00	3.50	3.00	2.50	2.00	0.00
4.00	20	10	5	5	0	0	0
3.50	22	5	10	5	2	0	0
3.00	20	0	8	7	5	0	0
2.50	20	0	0	4	9	5	2
2.00	20	0	0	1	5	10	4
0.00	18	0	0	0	1	2	15

From the table II we have found that our system's results are closer to human grades. From the results of above table II we have calculated the Standard Deviation (SD) error which is only 0.22.
We have tested our data both on Traditional LSA and generalized LSA. The comparison between traditional LSA and Generalized LSA are shown in the table III.

Table III Comparison of LSA with GLSA

Experiment	Essay grading with LSA	Essay grading with GLSA
Mean of Errors	.80	.33
Standard Deviation of Errors	.82	.22

We have been compared AEG grades with human grades which are represented by fig. 5. We have found that most of machine scores by AEG with GLSA are equal to the human scores.

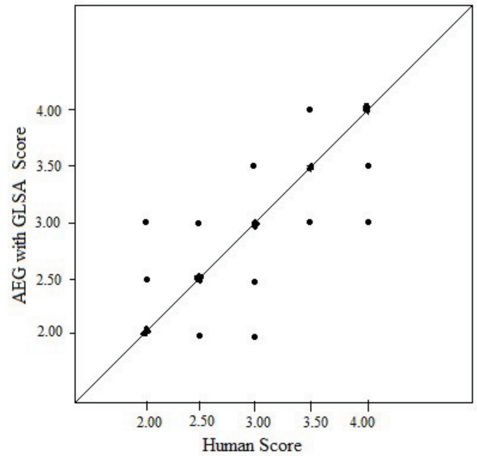

Fig. 5 Human score versus AEG with GLSA score

We have compared our system with the previous performance of the previous systems which are based on LSA. Table IV contrasts the performance comparison of new technique to that of previous method. Valenti et al. indicate that the accurate rate of LSA based IEA is from 0.85 to 0.91. Kakkonen et al. indicate that Automatic Essay Assessor (AEA) is 0.75 accurate with human grade. Lemaire B. et al. indicate the Apex (for an Assistant for Preparing EXams), a tool for evaluating student essays based on their content using LSA gives 0.59 accurate with human grade. The Table IV shows Comparison between the Performances of Three AEG approaches. Table IV shows the performance of the proposed method for scoring essays is very close to human grades.

Table IV Comparison between the performances of three AEG approaches

AEG Technique	Accuracy
IEA usig LSA	0.85-0.91
AEA using LSA	0.75
Apex using LSA	0.59
Our Syatem using GLSA	0.89-0.95

VI. CONCLUSION

We have trained our system by 960 training essays and tested by 120 submitted essays. We have gained 89%-96% of accuracy which show that our system is very closer to human grader. We hope this system will be designed as a replica of human grader.

REFERENCES

[1] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," Journal of Information Technology Education, vol. 2, 2003, pp. 319-330.

[2] T. Miller, "Essay assessment with latent semantic analysis," Department of Computer Science, University of Toronto, 2002. IEEE Conferences, 2009, pp. 333–338.

[3] A. M. Olney, "Generalizing latent semantic analysis," In Proceedings of 2009 IEEE International Conference on Semantic Computing, IEEE Conferences, 2009, pp. 40–46.

[4] M. M. Hasan, "Can information retrieval techniques meet automatic assessment challenges?," In Proceedings of the 12th International Conference on Computer and Information Technology (ICCIT 2009), Dhaka, Bangladesh, 2009, pp. 333–338.

[5] L. M. Rudner, and T. Liang, "Automated essay scoring using Bayes' theorem," The Journal of Technology, Learning, and Assessment , vol. 1, No. 2, 2002.

[6] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, "Automated essay scoring using the KNN algorithm," In Proceedings of the International Conference on Computer Science and Software Engineering (CSSE 2008), IEEE Conferences, 2008, pp. 735–738.

[7] K. M. Nahar, and I. M. Alsmadi, "The automatic grading for online exams in Arabic with essay questions using statistical and computational linguistics techniques," MASAUM Journal of Computing, vol. 1, no. 2, 2009.

[8] C. Loraksa, and R. Peachavanish, "Automatic Thai-language essay scoring using neural network and latent semantic analysis," In Proceedings of the First Asia International Conference on Modeling & Simulation (AMS'07),+ 2007, pp. 400–402.

[9] D. T. Haley, P. Thomas, A. D. Roeck, and M. Petre, "Measuring improvement in latent semantic analysis based marking systems: using a computer to mark questions about HTML," In Proceedings of the Ninth Australasian Computing Education Conference (ACE), 2007.

[10] S. Ghosh, and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context," In Proceedings of TENCON-2008, IEEE Region 10 Conference, 2008, pp. 16-21.

[11] T. Kakkonen, N. Myller, J. Timonen and E. Sutinen, "Automatic essay grading with probabilistic latent semantic analysis," In Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, June 2005, pp. 29–36.

[12] Porter Stemming [Online] Available: http://www.comp.lancs.ac.uk/computing/research/stemming/generall/porter.fitm

[13] T. Kakkonen, N. Myller, J. Timonen, and E. Sutinen, "Comparison of dimension reduction methods for automated essay grading," International Forum of Educational Technology & Society (IFETS), 2008, pp. 275–288.

[14] L. S. Larkey, "Automatic essay grading using text categorization techniques," In Proceedings of the 21st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, 1998, Melbourne, Australia, pp. 90-95.

[15] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, 1990, pp. 391-407.

[16] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," In Proc. of 11th annual int'l ACM SIGIR conference on Research and development in information retrieval, 1988, pp. 465-480.

[17] E. B. Page, "Statistical and linguistic strategies in the computer grading of essays," In Proceedings of the International Conference on Computational Linguistics, 1967, pp. 1-13.

[18] Y. Attali, and J. Burstein, "Automated essay scoring with e-rater® V.2," The Journal of Technology, Learning and Assessment, vol. 4, no. 3, 2006.