Journal of Engineering and Applied Sciences 13 (11): 4198-4204, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

A Comparative Analysis of Euclidean Distance and Cosine Similarity Measure for Automated Essay-Type Grading

[1]Odunayo. E. Oduntan, [2]Ibrahim A. Adeyanju, [3]Adeleye S. Falohun and [4]Olumide O. Obe
[1]Department of Computer Science, Federal Polytechnic, Ilaro, Nigeria
[2]Department of Computer Engineering, Federal University, Oye-Ekiti, Nigeria
[3]Department of Computer Science and Engineering, Ladoke Akintola University of Technology, Ogbomoso, Nigeria
[4]Department of Computer Science, Federal University of Technology, Akure, Nigeria

Abstract: Evaluation of student's performance is inevitable in any educational setting, allocating scores to student's response is a function of how close the answer supplied to the question is to expected answer. This study delves into analyzing the effectiveness of cosine similarity measure and Euclidean distance which are both used in similarity measures for Automated Essay Type Grading System (AETGS). AETGS involves transcription of the contents of the marking schemes into electronic form to derive a txt file extension using text editor while student's answers assumed txt format. The inherent stopwords and stemming in the txt document were pre-processed to address morphological variations using standard stopwords list and porters stemmer algorithm, respectively. N-gram terms were derived for each student's response and the Marking Schemes (MS) using the vector space model. A Document Term Matrix (DTM) was generated with N-gram terms of MS and students response representing columns and rows, respectively. Modified principal component analysis algorithm was used to reduce the sparseness of the DTM to obtain a vector representation of the student's answers and the marking scheme. The reduced vector representation of the student's answers was graded according to the mark assigned to each question in the marking scheme using cosine similarity measure and the Euclidean distance measure. The developed Automated Essay-Type Grading System (AETGS) was implemented in Matrix Laboratory 8.1 (R2013a). The effect of the similarity measures on the developed system was performed using Pearson Correlation coefficient of two courses: CMP401-Organization of Programming Languages and CMP205-Operating System I. The result showed that cosine similarity measure has a high positive correlation than the Euclidean distance.

Key words: Evaluation, cosine similarity measure, Euclidean distance measure, modified principal component algorithm, automated essay-type grading system, reduced vector

INTRODUCTION

Assessment is the process of gathering information using various methods to systematically gauge the effectiveness of the institution and academic programs to document student learning, knowledge, behaviors and skills as a result of their collegiate experiences (Allen, 2004; Valenti *et al*., 2003). In education, the term assessment refers to the wide variety of methods or tools that educators use to evaluate, measure and document the academic readiness, learning progress, skill acquisition or educational needs of students. Assessment is a crucial part of the learning process. It enables students to gauge their progress, tutors to judge the effectiveness of teaching and can also be used as a teaching tool, to give individuals or groups feedback designed to enable them to improve their performance in the future.

Evaluating of students performances can be done either manually or through automation. In the manual assessment which is involves the examiner setting questions and expecting the students to give answers to the question within the context of what has been earlier taught. The examiner then scores or grade the performance of the student by comparing the answers supplied by the students to the marking scheme prepared by the examiner in order to check the student understanding of the course content and to evaluate the similarity that exist between the marking scheme and the students answers.

Using an automated assessment tool, involves the examiner making available the questions to be attempted by the students available on an electronic device and the respective students supplies their answers in a real time system. A computer program is then used to grade the responses or answers supplied by the students by

Corresponding Author: Odunayo. E. Oduntan, Department of Computer Science, Federal Polytechnic, Ilaro, Nigeria

examining the content of what the student have written and matching it with the electronic marking scheme supplied by the examiner (Kaplan *et al*., 1998; Foltz *et al*., 1999; Kakkonen *et al*., 2005; Ade-Ibijola *et al*., 2012; Islam and Hogue, 2012).

This study focuses on comparing the similarity measure used in comparing the student's answers and the examiner marking scheme and its effect on automated essay-type grading system. Two similarity measures will be discussed: the Euclidean distance and the cosine similarity measure.

Literature review

Automated grading systems: Mason and Grove-Stephenson (2002) were of the opinion that teachers all over the world spend a great deal of time just marking student's. They have to cut down the time they can devote to their other duties. Even doing that, sometimes they do not have enough time to properly assess the big number of students they have. Therefore, many researchers believe that this situation has to be solved and some of them have presented the computer as a new assessing tool. These researchers do not attempt to substitute the teacher with the computer but to help the teachers with the computer software (Larkey, 1998; Foltz *et al*., 1999; Kakkonen and Sutinen, 2004).

Automated assessment of student's essays is regarded by many as the Holy Grail of Computer Assisted Assessment (Whittington and Hunt, 1999; Mitchell *et al*., 2002; Hanna and Dettmer, 2004; Oduntan and Adeyanju, 2017). The technical approaches that these tools are undertaking are very different but the goal and concepts underlying are just the same for all of them (Larkey, 1998). On the other hand, there have always been hard critics about the idea of a computer grading human essays. Nowadays, there are still some skeptical researchers that do not consider the automatic grading possible. However, the advances in natural language processing, machine learning and neural network techniques, the lack of time to give them appropriate feedback (despite the general assumption of its importance) and the conviction that multiple choice questions are a poor assessment method are favoring a change in this situation. Automatic assessment of student's texts could be seen as the higher level of a hierarchy in which two subcategories could be identified: the automatic assessment of short answers and the automatic assessment of essays. The focus of this research work is on automatic grading of essay-type questions.

In this modern culture, text is the most common vehicle for the formal exchange of information. Although, extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool (Feldman and Dagan, 1995; Guven *et al*., 2006). The tools used in text mining system has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information (Landauer *et al*., 1997; Jensen and Shen, 2008; Adeyanju *et al*., 2010).

The theoretical frameworks on which most automated essay-type grading systems are developed is the area of text mining. Text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents.

Similarities measures: In this study, similarities measures to be used include: Euclidean distance, cosine similarity, Pearson correlation coefficient and Jaccard coefficients.

Cosine similarity measure: Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1 and it is <1 for any other angle (Braga, 2009). It is thus, a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0 and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space where the outcome is neatly bounded in [0, 1].

Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors as only the non-zero dimensions need to be considered (Salton *et al*., 1975; Lehmann and Casella, 1998). The equation for cosine similarity is:

$$\text{CosSim}(\text{Dj},\text{Q}) = \frac{\sum_{i=1}^{t}(d_{ij} * q_i)}{\sqrt{\sum_{i=1}^{t} d_{ij}^2 * \sum_{i=1}^{t} q_i^2}} \qquad (1)$$

Where:

d_{ij} = The jth weight of query vector

D_j and q_{ij} = The ith weight of training essay vector Q

This method is useful when finding the similarity between two text documents whose attributes are word frequencies. A perfect correlation will have a score of 1 (or an angle of 0) and no correlation will have a score of 0 (or an angle of 90°).

Pearson coefficient: Pearson product moment correlation (r) signifies the degree of relationship that exists between dependent variables and independent variable. With Pearson correlation coefficient, the valid result for r lies between -1 and +1. If the result lies between 0 and 1, it shows there is a positive correlation that is X increases as Y increases. If r = 1, it shows that the result is perfect positive. If r is between 0.5 and 1, it shows a high positive correlation, when r is between 0 and 0.49, it exhibits a low positive correlation. When r = -1, it shows a perfect negative correlation that is the rate at which the dependent variable increases is exactly equal to the rate at which the independent variable decreases. When r is between -0.5 and 0, it shows a weak negative correlation, when r is between -0.49 and -1, it exhibits a strong negative correlation. The formula for deriving the Pearson correlation coefficient (Islam and Hogue, 2012) is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \qquad (2)$$

Euclidean distance: The basis of many measures of similarity and dissimilarity is Euclidean distance. The distance between vectors X and Y is defined as follows:

$$d(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \qquad (3)$$

In other words, Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Note that the formula treats the values of X and Y seriously: no adjustment is made for differences in scale. Euclidean distance is only appropriate for data measured on the same scale. The correlation coefficient is (inversely) related to the Euclidean distance between standardized versions of the data (Jolliffe, 2002).

Euclidean distance is most often used to compare profiles of respondents across variables. For example, suppose our data consist of demographic information on a sample of individuals, arranged as a respondent by variable matrix. Each row of the matrix is a vector of m numbers where m is the number of variables. We can evaluate the similarity (or in this case, the distance) between any pair of rows. Notice that for this kind of data, the variables are the columns. A variable records the results of a measurement. For our purposes, in fact, it is useful to think of the variable as the measuring device itself. This means that it has its own scale which determines the size and type of numbers it can have. For instance, the income measurer might yield numbers between 0 and 79 million while another variable, the education measurer, might yield numbers from 0-30. The fact that the income numbers are larger in general than the education numbers is not meaningful because the variables are measured on different scales. In order to compare columns we must adjust for or take account of differences in scale (Rencher and Christensen, 2002).

Jaccard coefficient: Jaccard coefficient measures the similarities between two finite sets to capture the number of elements common among the two sets, Its value belongs to the interval of (0, 1) (Rencher and Christensen, 2002). This coefficient is defined as:

$$SIM_J(\vec{t}_a, \vec{t}_b) - \frac{\vec{t}_a, \vec{t}_b}{|\vec{t}_a|^2 + |\vec{t}_b|^2 \;\; \vec{t}_a, \vec{t}_b} \qquad (4)$$

where, t_a and t_b are two document vectors to be compared.

MATERIALS AND METHODS

In order to improve on the existing automated essay grading techniques, this study carried out a comparative analysis of the impact of cosine similarity measure and Euclidean distance on the data set which comprises of the softcopy of student's answers and the softcopy of the examiner marking scheme captured in .txt file format. Modified Principal Component Analysis (MPCA) technique was used to reduce the sparseness of the document vectors and to address word order issues in automated essay type grading (Oduntan and Adeyanju, 2017). The development tool used was MATLAB 8.1 R2013a Version on Windows 7 Ultimate 32 bit operating system, Intel®Pentium® CPU B960@2.20GHZ Central Processing Unit, 4GB Random Access Memory and 500 GB hard disk drive.

Figure 1 gives a description of the framework of automated grading system using Modified Principal Component Analysis (MPCA) which involved the collection of data set comprising of the essay-type marking scheme and softcopy student's answers. The hardcopies of the marking scheme were transcripted into electronic form and the softcopies essay-type student's answers in .txt file format. The inherent stopwords and stemming in the .txt document were pre-processed to address morphological variations using

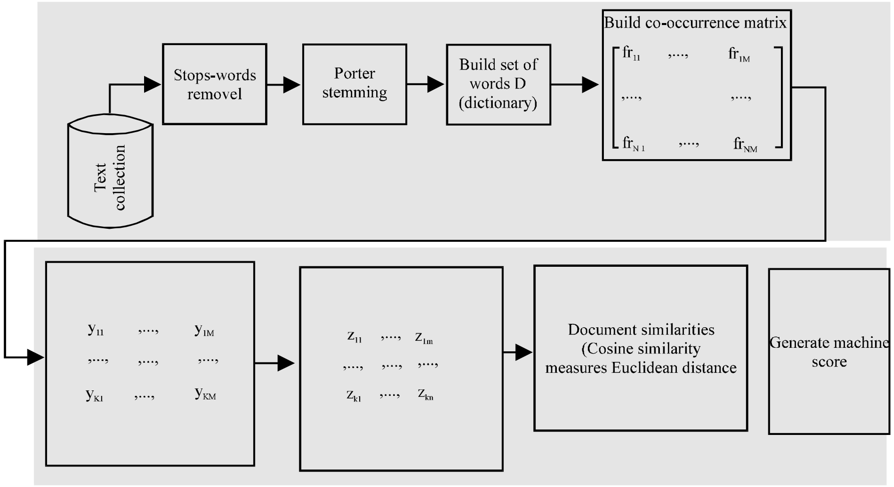

Fig. 1: Framework of the automated essay grading system

standard stopwords list and porters stemmer algorithm, respectively. N-gram terms were derived for each student's response and the Marking Schemes (MS) using the vector space model. A Document Term Matrix (DTM) was generated with N-gram terms of the marking scheme and student's response representing columns and rows, respectively. Principal Component Analysis (PCA) algorithm was modified by integrating n-gram terms as input into existing PCA to derive Modified Principal Component Analysis (MPCA) algorithm.

The MPCA was used to reduce the sparseness of the DTM to obtain a vector representation of the student's answers and the marking scheme. The reduced vector representation of the student's answers was graded according to the mark assigned to each question in the marking scheme using cosine similarity measure and Euclidean distance. The developed Automated Essay-Type Grading System (AETGS) was implemented in MATLAB 8.1 (R2013a). Performance of the effect of the similarity measures (Cosine and Euclidean distance) on students score was determined using the Pearson correlation coefficient (r) and coefficient of determination (R^2). Development of an automated essay type grading system involves.

Data acquisition: Data in the study comprises of softcopy of student's answer and the softcopy of examiner's marking scheme. A total of fifty students were examined with two different marking scheme on CMP401: Organization of programming languages and CMP205: Operating system I. Text were captured in .txt file format.

Text pre-processing: This is a process of carrying out stemming and stopword removal from captured .txt file. This is done to enhance the effectiveness of the text to be compared and to reduce the storage space required for the data.

Document representation: This refers to the pattern of extracting needed information from raw text. In this study, N-gram document representation and the vector space model was used. An N-gram is a subsequence of n items from a given sequence. Unique words were extracted to possess 'coordinates' for vector space model. N-gram terms were derived for each student's response and the Marking Schemes (MS) using the vector space model. Document term matrix is a representation of the certain text in the space which is built in according to Vector Space Model. Figure 2 shows a diagram explaining the text processing stage of the study. A Document Term Matrix (DTM) referred to as the co-occurrence matrix was generated with N-gram terms of marking scheme and student's response representing columns and rows, respectively. This was used to represent text in suitable form for further machine analysis.

Feature extraction using Modified Principal Component Analysis (MPCA): This is the transformation of input data into a set of features. It can be done by the process of dimensionality reduction. In text processing, feature extraction is performed by reducing sparseness of a document vector. Sparseness is the sequencing out of any zero elements in a matrix from a full matrix. The purpose of the feature extraction step 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 algorithm for the modified principal component analysis (MPCA) is:

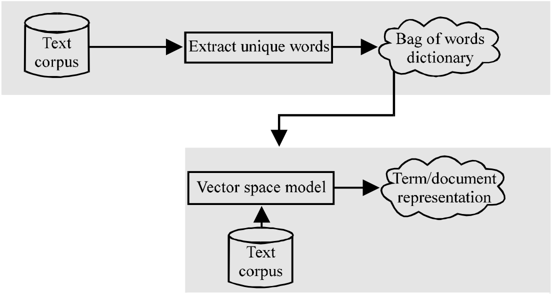

Fig. 2: Text processing with vector space model

1. Input N-gram from the document vector (x_i) (where i= 2, 3, ..., n)
2. Obtain the Mean of $x_{i::}$ using Eq. 5:

$$\overline{x}_i = \sum_{i=1}^{n} x_i \tag{5}$$

3. Subtract the mean from the data dimension $D_{[r, c]}$ using Eq. 6:

$$D_{[r, c]}-\sum_{i=1}^{n} x_i \tag{6}$$

4. Calculate the covariance matrix, using Eq. 7:

$$\text{Covariance-matrix } C = \frac{1}{b-a} A^{'}*AA^{T}*A \tag{7}$$

5. Calculate the eigenvalue and eigenvector using Eq. 8:

$$(A-\lambda I)X = 0 \tag{8}$$

where A is covariance matrix, λ is the eigenvalue and X is the eigenvector

6. Choose components to form a normalised document vector (nx_i)

Documents similarities: Documents similarities deals with the comparison of two separate documents to examine the level at which the items of one document matches the other. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1 and it is <1 for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0 and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space where the outcome is neatly bounded in [0, 1].

In this study, documents compared were the generalized document vector of the marking scheme and the student's script. The generalized document vector was compared using the cosine similarity to generate the similarity score. The similarity score was multiplied with the mark allocated by the examiner to derive the weighted score per question. A summation of the weighted score was used to determine the machine score. The cosine similarity measure and the Euclidean distance measure were used. Cosine similarity formula while the Euclidean distance formula is as stated in Eq. 9 and 10, respectively:

$$\text{CosSim}(L_j, M) = \frac{\sum_{i=1}^{t}(l_{ij} * m_i)}{\sqrt{\sum_{i=1}^{t} l_{ij}^2 * \sum_{i=1}^{t} m_i^2}} \tag{9}$$

Where:

l_{ij} = The weight of the ith term in the essay-type marking scheme document term matrix (L_j)

m_i = The weight of the ith term in the essay-type student script document term Matrix (M)

$$d(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \tag{10}$$

Performance evaluation: Pearson product moment correlation (r) was used. It signifies the degree of relationship that exists between dependent variables and independent variable. In this study, the dependent variable is the human score denoted as X while the independent variable is the machine score. The formula for deriving the Pearson correlation coefficient as stated by Islam and Hogue (2012) is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \tag{11}$$

RESULTS AND DISCUSSION

In Table 1, CMP401 dataset result shows that the Pearson correlation coefficient of the student score generated for cosine similarity measure gave 0.70 while the Pearson correlation coefficient of Euclidean distance have 0.45. Table 1 represents the results of CMP205 dataset indicating that the Pearson correlation coefficient of the student score generated for Euclidean distance measure gave 0.50 while Pearson correlation coefficient for cosine similarity measure gave 0.75.

Figure 3 is a chart describing the performance correlation of the result. The Pearson correlation coefficient shows a positive relation when compared with the human score. This has shown a better computational power of feature extraction performed on the data set. From the results evaluated, it was observed that performing similarity measure with the cosine similarity is more efficient for automated essay type grading system

Table 1: Pearson coefficient result

Dataset	Cosine similarity measure	Euclidean distance
CMP 401	0.70	0.45
CMP 205	0.75	0.50

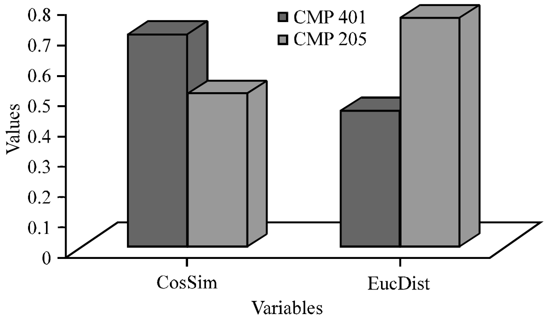

Fig. 3: Performance correlation of results

than the use of the Euclidean distance measure. A Pearson correlation of 0.70 and 0.75, respectively were derived with cosine similarity measure which is a high positive correlation. This study have been able to come up with the modified principal component analysis method of removing unwanted noise from document term matrix to improve computability.

CONCLUSION

This study has been able to perform a comparative analysis of two similarity measures used in data analysis. They are: cosine similarity and Euclidean distance measures, these measures have been applied in the document similarity of two datasets, namely the marking scheme and the student's responses in automated essay grading system. It has been established that cosine similarity measure has a high positive correlation when the machine score generated was compared with the human examiner than the Euclidean distance measure.

REFERENCES

Ade-Ibijola, A.O., I. Wakama and J.C. Amadi, 2012. An expert system for Automated Essay Scoring (AES) in computing using shallow NLP techniques for inferencing. Intl. J. Comput. Appl., 51: 37-45.

Adeyanju, I., N. Wiratunga, J.A. Recio-Garcia and R. Lothian, 2010. Learning to author text with textual CBR. Proceeding of the European Conference on Artificial Intelligence (ECAI'10), August 16-20, 2010, University of Lisbon, Portugal, Europe, pp: 777-782.

Allen, M.J., 2004. Assessing Academic Programs in Higher Education. Anker Publishing Company, Bolton, Massachusetts.

Braga, I.A., 2009. Evaluation of stopwords removal on the statistical approach for automatic term extraction. Proceedings of the 7th Brazilian Symposium on Information and Human Language Technology (STIL'09), September 8-11, 2009, IEEE, Sao Carlos, Sao Paulo, Brazil, ISBN:978-1-4244-6008-3, pp: 142-149.

Feldman, R. and I. Dagan, 1995. Knowledge discovery in textual databases (KDT). Proceedings of the Conference on Knowledge Discovery and Data Mining Vol. 95, August 20-21, 1995, AAAI, Montreal, Quebec, Canada, pp: 112-117.

Foltz, P.W., D. Laham and T.K. Landauer, 1999. Automated essay scoring: Applications to educational technology. Proceedings of the EdMedia: World Conference on Educational Media and Technology, June 19-24, 1999, Association for the Advancement of Computing in Education (AACE), Waynesville, North Carolina, ISBN:978-1-880094-35-8, pp: 939-944.

Guven, A., O.O. Bozkurt and O. Kalipsiz, 2006. Advanced information extraction with n-gram based LSI. World Acad. Sci. Eng. Technol., 17: 13-18.

Hanna, G.S. and P.A. Dettmer, 2004. Assessment for Effective Teaching using Context-Adaptive Planning. Pearson/Allyn and Bacon, Boston, Massachusetts, ISBN:9780205389414, Pages: 444.

Islam, M.M. and A.S.M.L. Hogue, 2012. Automated essay scoring using generalized latent semantic analysis. J. Comput., 7: 616-626.

Jensen, R. and Q. Shen, 2008. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. Wiley, New Jersey, USA., ISBN:978-0-470-22975-0, Pages: 259.

Jolliffe, I.T., 2002. Principal Component Analysis. 2nd Edn., Springer, New York, USA., Pages: 478.

Kakkonen, T. and E. Sutinen, 2004. Automatic assessment of the content of essays based on course materials. Proceedings of the 2nd International Conference on Information Technology: Research and Education (ITRE'04), June 28-July 1, 2004, IEEE, London, England, UK., pp: 126-130.

Kakkonen, T., N. Myller, J. Timonen and E. Sutinen, 2005. Automatic essay grading with probabilistic latent semantic analysis. Proceedings of the 2nd Workshop on Building Educational Applications using NLP, June 29, 2005, Association for Computational Linguistics, Stroudsburg, Pennsylvania, pp: 29-36.

Kaplan, R.M., S. Wolff, J. Burstein, C. Li and D. Rock *et al*., 1998. Scoring essays automatically using surface features. Master Thesis, Technical Report 94-21P, Educational Testing Service, New Jersey, USA.

Landauer, T.K., D. Laham, B. Rehder and M.E. Schreiner, 1997. How Well can Passage Meaning be Derived Without using Word Order? A Comparison of Latent Semantic Analysis and Humans. In: Proceedings of the 19th Annual Meeting of the Cognitive Science Society, Shafto, M.G. and P. Langley (Eds.). Cognitive Science Society, Stanford, USA., pp: 412-417.

Larkey, L.S., 1998. Automatic essay grading using text categorization techniques. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28, 1998, ACM, Melbourne, Australia, ISBN:1-58113-015-5, pp: 90-95.

Lehmann, E.L. and G. Casella, 1998. Theory of Point Estimation. 2nd Edn., Springer, New York, USA., Pages: 574.

Mason, O. and I. Grove-Stephenson, 2002. Automated Free Text Marking with Paperless School. In: Proceedings of the 6th International Conference on Computer Assisted Assessment, Danson, M. (Ed.). Loughborough University, Loughborough, England, UK., pp: 85-95.

Mitchell, T., T. Russel, P. Broomhead and N. Aldridge, 2002. Towards Robust Computerized Marking of Free-Text Responses. In: Proceedings of the 6th International Conference on Computer Assisted Assessment, Danson, M. (Ed.). Loughboroug University, Loughborouh, England, UK., pp: 150-200.

Oduntan, O.E. and I.A. Adeyanju, 2017. A comparative analysis of modified principal component analysis and generalized latent semantic analysis approach to automated marking of theory-based exams. IOSR. J. Mob. Comput. Appl., 4: 31-41.

Rencher, A.C. and W.F. Christensen, 2002. Methods of Multivariate Analysis. 2nd Edn., John Wiley and Sons, Inc., New York.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.

Valenti, S., F. Neri and A. Cucchiarelli, 2003. An overview of current research on automated essay grading. J. Inform. Technol. Educ., 2: 319-330.

Whittington, D.H. and H. Hunt, 1999. Approaches to the Computerized Assessment of Free Text Responses. In: Proceedings of the Sixth International Computer Assisted Assessment Conference, Danson, M. (Ed.). Loughborough University, UK., Loughborough, England, UK., pp: 1-13.