IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4778

Automated Essay Scoring and Flexible Learning

Raymond Koon-Ying Li and Kwang-Hoon Oh
School of Multimedia Systems
Monash University
Melbourne, Australia
Tel: +61 3 99052354, +61 3 99052326
Raymond.Li@infotech.monash.edu.au, khoh1@student.monash.edu.au

ABSTRACT

Rapid advancements in networking and computer technology over recent years have changed the ways that education can be delivered. "Flexible learning" is now a reality. Assessment is an important element in the process of learning, irrespective of the delivery methods involved. Test methods, such as multiple-choice tests and matching items, enable simple assessment of a student's learning by a computer. However, there are limitations to these methods. Traditionally, essay testing has been used by educators to assess students' knowledge, especially within higher education. This paper examines an automated essay scoring methods utilizing Latent Semantic Analysis. Computation experience indicates that a LSA system can be used to automatically score long essays in a higher education environment. The findings provide an insight into how LSA works and the problems associated with the application of a LSA model to essay scoring.

INTRODUCTION

Advancements in technology continuously change our lives. With the adoption of the Internet technology, new ways of delivering education, such as Web-based flexible learning or e-learning, have been introduced. Assessment is an important part of learning irrespective of the delivery method. Higher education has traditionally employed the essay test method for assessment purposes. Ideally, flexible learning for higher education students should be provided with an automated essay scoring assessment method.

In this paper, a computer mediated essay scoring method known as Latent Semantic Analysis (LSA) is discussed and computation experience on the use of the method is reported.

ASSESSMENTS IN LEARNING ENVIRONMENTS

Teaching and learning can be improved by well-organized assessment. "Assessment is a process that uses information gathered through measurement to analyse or judge a learner's performance on some relevant work task" (Sarkees-Wircenski & Scott 1995). Evaluation of assessment, however, can be affected by a number of factors and therefore may contain errors. Educators must be aware of where errors are likely to occur. Multiple test methods are often used to eliminate the likelihood of errors. Due to the ease of implementation, objective test methods, such as *multiple-choice tests, true-false tests, matching tests, short-answer tests* and *performance tests* are commonly employed on Web based learning systems to assess student learning. Due to technical difficulties, *essay test* methods are often not used. Each objective test has its own merits but many researchers have concluded that they are not ideal assessment methods for learning and, in particular, for higher educational learning.

Yunker (1999) argues that multiple-choice tests tend "to measure only a very narrow sample of content at a specific point in time and usually require only superficial recognition of information to answer correctly". Oosterhof (1994) suggests that the multiple-choice format is more suitable for testing the lower cognitive skills. According to University of Minnesota (1999), "*True-False* does not provide diagnostic information and is not amenable to questions that cannot be formulated as propositions". *Matching Items* are also poor indicators of student strengths and weaknesses. Matching items are, therefore, not suitable for higher-level students.

Essay Tests

An essay question is best when used to assess a student's ability to communicate ideas in writing (Oosterhof 1994). Essay tests can not only assess student's knowledge but also their skills in writing, content structuring, choosing words, and vocabulary. It can also be used to assess what a student thinks. As essay test format can be adaptable to cater for various types of higher education and are widely accepted as the best higher education testing method. Compared with other test formats, essay tests can provide a better assessment of learners' competence, preventing students from guessing correct answers without any knowledge of the topic. According to Landauer et al. (1999), "…grading and criticizing an essay …….. can be used as a feedback device to help students improve their learning on both content and the skills of thinking and writing"

Curriculum guides often contain complex instructional goals that students are required to achieve. "Often these complex goals can be measured with essay questions, which typically require students to discuss, analyse, compare for similarities and differences, synthesize or evaluate" (Carey 1994). Under the essay test method, students need to determine how they will approach a given problem, plan and organize their responses, and present their ideas.

Objective tests, such as multiple-choice tests, do not require students to handle their responses in the same manner and thus will not allow the students to demonstrate these capabilities.

However, essay testing has several major drawbacks. Page (1996) pointed out that essay scoring by humans is expensive, time-consuming and often unreliable. Fewer questions can be presented in a one hour exam. This often limits the coverage of the material taught. Scoring can be affected by, for example, teachers' standards may shift during scoring, or fatigue may cause lapses in concentration. (Carey 1994)

Automated Essay Scoring

As a result of recent advancements in computer technology, educators have now turned to computers for help. The easiest way to score an essay is by matching words in model answers to words in students' essays. However, students' essays may contain synonyms and homonyms. Even with all the synonyms and homonyms defined and programmed, these words do not carry the same meaning under all situations. Computers must define and select the exact meaning of a word according to the meaning of the sentence or paragraph. For this reason, automated essay scoring is not as simple as scoring multiple-choice questions.

According to Williams (2001) four commonly used conceptual models for automated essay scoring are Project Essay Grade (PEG), Electronic Essay Rater (E_RATER), Text Categorization Techniques (TCT) and the Latent Semantic Analysis (LSA) model.

Williams (2001) has tested these essay scoring methods. In terms of comparison with human markers, E-RATER is the most successful, followed by LSA, TCT, and finally PEG.

The LSA model is a statistical method for capturing relationships amongst words that may have semantic significance. It uses Singular Value Decomposition (SVD). The objective of the method is to identify a hidden (latent) semantic structure within a given document. This enables two documents to be compared and relative fit to be scored amongst a set of document.

For this paper, we focused on LSA.

LSA METHOD AND ITS TECHNICAL DETAILS

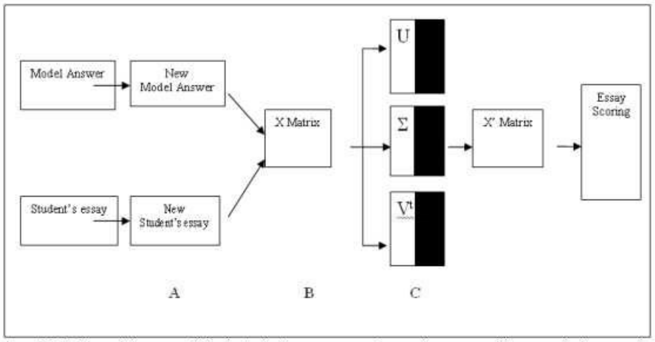
LSA can be trained on materials related to a topic, for example, textbooks, articles or lecturer's note. This result can be represented as information relating to the essay topic. Students' essays are compared against this representation using semantic relatedness. The basic scoring method is how the measurement of a student's essay is similar to the model answer. The degree of difference between the model answer and student's essay is converted into a score.

LSA uses the following method:

Figure 1 provides the overview of the LSA method.

Firstly, all text under consideration was converted into a matrix in which each row stands for a unique word and each column stands for each document (i.e. term by document matrix). Each cell contains the frequency with which the word of its row appears in the document denoted by its column. The cell entries may be subjected to a preliminary transformation in which each cell frequency is weighted by a function that expresses both the word's importance in the particular document and the degree to which the word type carries information in the domain of discourse in general (Landauer et al. 1998).

Next, LSA applies singular value decomposition (SVD) to the matrix. In SVD, the aim is to derive from the "term by document" matrix, a "pseudos-document" which contains a weighted average of the vectors of the words it contain. A document vector in the SVD solution is also a weighted average of the vectors of words it contains, and a word vector weighted average of vectors of the documents in which it appears. Under SVD a term by document matrix is transformed into three matrices. The first matrix is for the original row entities, the second matrix is for the original column entities and the third matrix is for a diagonal matrix containing scaling values. Using a reduced dimension of these three matrices, in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing a close approximation to the original matrix from the reduced dimension component SVD matrices (Landauer et al. 1998). What LSA does is eliminate the obscuring "noise". LSA uses reduced dimensions that erase parts of the matrix in order to estimate latent semantic structure. Two dimensions are normally recommended.



A – Modified model answer (Eliminated all common words, such as prepositions, symbols or marks and the definite articles or the indefinite articles)
B – X matrix based on the frequency of terms by documents
C – Computation using Singular Value Decomposition, then eliminates "noise"

Figure 1. Overview of the LSA method

Finally, the measure of similarity between two documents computed in the reduced dimensional space is usually, but not always, the cosine between document vectors in the SVD solution.

COMPUTATION EXPERIENCE

This session presents the results of applying LSA based software, known as LSI (Latent Semantic Indexing), to automatic scoring of the actual students' response to a question in a 1999 examination paper. The question is drawn from the BUS5150 examination at Monash University. BUS 5150 is a project management subject offered at the Masters level. The students in the program already have Bachelor degrees in various disciplines.

The Question And The Response Sample

The question is "*Although ISO 9000 standards are now widely acceptable, it has recently been subjected to a lot of criticism. Describe the criticism that has been discussed during the lecture*." All (16) student answers (essay documents) from the 1999 batch are included for testing. Length of responses can be classified as long essays. The average length of the response to the question was 176.5 words.

Human Scoring

Responses to all questions were marked by the lecturer and three tutors against a model answer and a marking sheet prepared by the lecturer. All markers have no knowledge of LSA. Table 1 lists the correlations between scoring by individual markers.

Results and Comments

17 documents, which include a model answer, and 16 student responses, are used in the computation. From the 17 documents, there are 143 unique terms that exclude common words. A 143 C 17 matrix, representing the relative frequencies at which each term appears in the documents, was constructed. Singular Value Decomposition technique is then applied to arrive at three matrixes (see Figure 1):

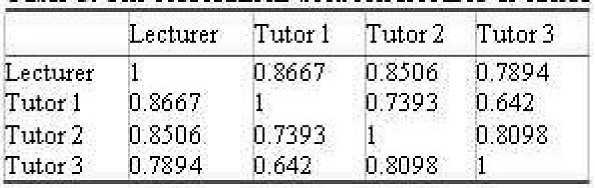
A 143 C 17 term by document matrix (U)
A diagonal 17 X 17 matrix (S)
A 17 X 17 document by document matrix (Vt)

The matrixes are truncated to remove 'noise'. Thus, the dimension of the new matrix will be 2 X 143 matrix (U). The three matrixes are then multiplied together to get the SVD solution matrix.

Table 2 lists the correlation coefficients between the scoring by LSA and individual human markers.

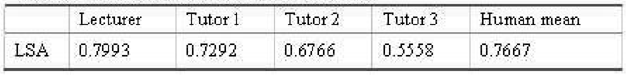
The correlation between LSA scoring and the lecturer's scoring is 0.7993. It means 80% accuracy compared to the lecturer's marking. The correlation coefficient between LSA scoring and the average of three tutors' marking is 0.7263. According to Table 1, even the correlation of human markings is not perfect. The range of human marking is between 64% and 87%. It is interesting to note that the range of LSA against humans is between 56% to 80%. Another point to note here is that the correlation result in Table 2 also agrees with the experience of the markers. The lecturer scores the best, while tutor 3 (who has just

Table 1. The correlations between lecturer & tutors



	Lecturer	Tutor 1	Tutor 2	Tutor 3
Lecturer	1	0.8667	0.8506	0.7894
Tutor 1	0.8667	1	0.7393	0.642
Tutor 2	0.8506	0.7393	1	0.8098
Tutor 3	0.7894	0.642	0.8098	1

Table 2. Correlation between human & LSA



	Lecturer	Tutor 1	Tutor 2	Tutor 3	Human mean
LSA	0.7993	0.7292	0.6766	0.5558	0.7667

started tutoring) scores the worst. Tutors 1 and tutor 2 have four years and two years tutoring experience respectively.

FURTHER ANALYSIS

Reduce Dimensions In The SVD Method

LSA uses reduced dimensions technique in which part of the three matrixes (U,S,Vt) are eliminated in order to estimate the hidden semantic structure. This section presents the analysis carried out to determine the optimal dimensions to be used. Table 3 shows two row extracts from the SVD solution matrix (X¢). The two rows list the derived frequencies of the equivalent terms (Improvement and improvements) against each document. The first document is the model answer.

The correlation between two terms is derived from the two vectors obtained using all figures in each row. Table 4 shows the changes in correlations between terms (equivalent and non-related) and the changes in the dimension. The vector was derived across all documents against different dimensions.

From Table 4, it is evident that as more columns are selected, more 'noise' will be introduced. Thus two dimensional calculations are the best fit in this study.

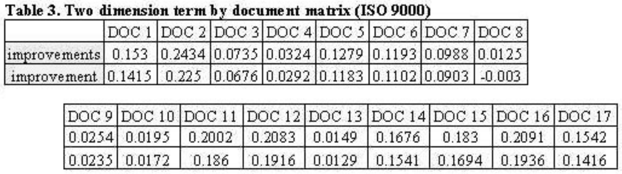
Length Of The Response And Its Score

Table 5 lists the computations result of the LSA on the student responses.

Document 17 (see Table 5) demonstrated that LSA can generate abnormal scoring if the essay is extremely short but with the right terms. This indicates that LSA can establish the hidden relationships among terms within a document derived from the global relationships that exists between terms across all documents. This hidden relationship does not necessarily reflect that the essay contains sentences that are meaningful to human beings. This also highlights that LSA is essay based rather than sentence based. This agrees with the finding of Thompson (1999) and Dennis (2000).

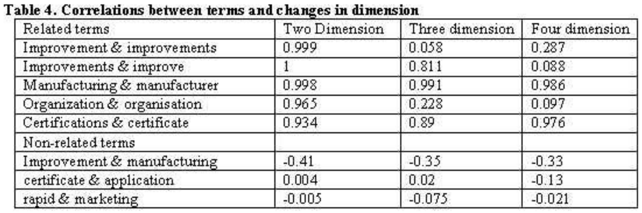
The correlation coefficient of -0.17 between the document word length and the difference between LSA and human scores (the last two columns of Table 5) indicates that the length of essays does not affect the LSA scoring. The correlation coefficient was calculated with Document 17 removed.

Table 3. Two dimension term by document matrix (ISO 9000)



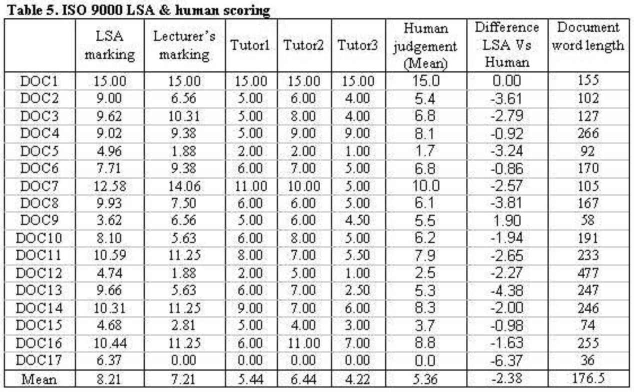
	DOC 1	DOC 2	DOC 3	DOC 4	DOC 5	DOC 6	DOC 7	DOC 8	DOC 9	DOC 10	DOC 11	DOC 12	DOC 13	DOC 14	DOC 15	DOC 16	DOC 17
improvements	0.153	0.2434	0.0735	0.0324	0.1279	0.1193	0.0988	0.0125	0.0254	0.0195	0.2002	0.2083	0.0149	0.1676	0.183	0.2091	0.1542
improvement	0.1415	0.225	0.0676	0.0292	0.1183	0.1102	0.0903	-0.003	0.0235	0.0172	0.186	0.1916	0.0129	0.1541	0.1694	0.1936	0.1416

Table 4. Correlations between terms and changes in dimension



Related terms	Two Dimension	Three dimension	Four dimension
Improvement & improvements	0.999	0.058	0.287
Improvements & improve	1	0.811	0.088
Manufacturing & manufacturer	0.998	0.991	0.986
Organization & organisation	0.965	0.228	0.097
Certifications & certificate	0.934	0.89	0.976
Non-related terms			
Improvement & manufacturing	-0.41	-0.35	-0.33
certificate & application	0.004	0.02	-0.13
rapid & marketing	-0.005	-0.075	-0.021

Table 5. ISO 9000 LSA & human scoring



	LSA marking	Lecturer's marking	Tutor1	Tutor2	Tutor3	Human judgement (Mean)	Difference LSA Vs Human	Document word length
DOC1	15.00	15.00	15.00	15.00	15.00	15.0	0.00	155
DOC2	9.00	6.56	5.00	6.00	4.00	5.4	-3.61	102
DOC3	9.62	10.31	5.00	8.00	4.00	6.8	-2.79	127
DOC4	9.02	9.38	5.00	9.00	9.00	8.1	-0.92	266
DOC5	4.96	1.88	2.00	2.00	1.00	1.7	-3.24	92
DOC6	7.71	9.38	6.00	7.00	5.00	6.8	-0.86	170
DOC7	12.58	14.06	11.00	10.00	5.00	10.0	-2.57	105
DOC8	9.93	7.50	6.00	6.00	5.00	6.1	-3.81	167
DOC9	3.62	6.56	5.00	6.00	4.50	5.5	1.90	58
DOC10	8.10	5.63	6.00	8.00	5.00	6.2	-1.94	191
DOC11	10.59	11.25	8.00	7.00	5.50	7.9	-2.65	233
DOC12	4.74	1.88	2.00	5.00	1.00	2.5	-2.27	477
DOC13	9.66	5.63	6.00	7.00	2.50	5.3	-4.38	247
DOC14	10.31	11.25	9.00	7.00	6.00	8.3	-2.00	246
DOC15	4.68	2.81	5.00	4.00	3.00	3.7	-0.98	74
DOC16	10.44	11.25	6.00	11.00	7.00	8.8	-1.63	255
DOC17	6.37	0.00	0.00	0.00	0.00	0.0	-6.37	36
Mean	8.21	7.21	5.44	6.44	4.22	5.36	-2.38	176.5

It is also evident from Document 12 that the amount of terms used does affect the performance of LSA.

Swapping Of Documents

This section addresses the question "Does the order of essays affect marking?" According to results contained in Table 6, there is no obvious difference to the scoring after swapping the order of the documents in which they are inserted into the LSA setting.

CONCLUSION AND FURTHER RESEARCH

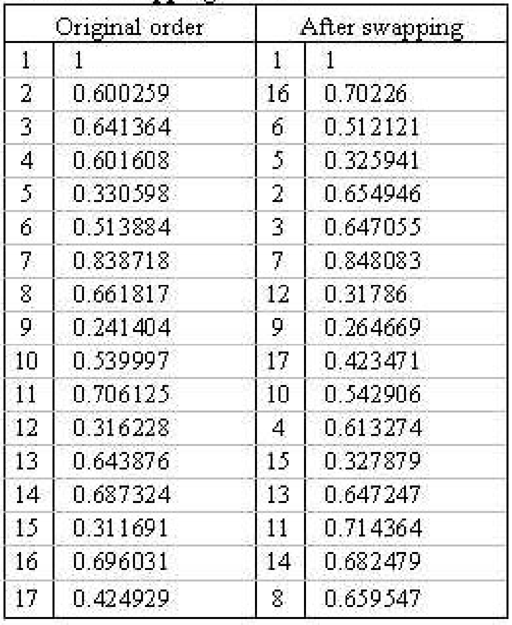
The computation experience combined with a literature search provides the following findings.

- It was found that the range of correlations between LSI scoring and individual human marking is between 58% to 80%, while the variation between human markings range from 64% to 87%. Therefore, LSI's performance can be comfortably compared to human markings. Furthermore, the result of the correlation between LSI scoring and human marking indicates that the variation in the difference between LSI scoring and human markings is related to the experience of the marker themselves.
- It was established that the optimal number of dimensions that we should use to truncate the SVD matrixes (U, S and Vt) is two. This agrees with the recommendation by Landauer (1998).
- It was found that the length of individual student response to a question does not influence the relative scoring. In other words, documents that marked low by humans will be scored low by LSI.
- It can be concluded that the order in which the documents appear on the frequency of terms by documents (X) matrix have no effect on the scoring of the documents.
- LSI can generate abnormal scoring if the essay is extremely short but with the right terms. This also highlights that LSA is essay based rather than sentence based.

The above conclusions are arrived at from testing using a small sample size of student responses to a single question. The above findings may not necessarily be conclusive but do provide an insight into how LSA works and the problems associated with the application of an LSA model to essay scoring. The followings are the recommended future directions for research.

- Further testing using larger samples and more sets of questions should be carried out to validate the above findings and to identify further pointers for future research.
- Capabilities to handle syntax, grammar, vocabularies, surrounding word contexts, as extracted from other models such as E-RATER should be incorporated into the LSA model.
- The LSA model should be modified such that the hidden structure identified has sentence based structure embedded. It should also reflect the relationship between each sentence and paragraph.
- The development of a holistic framework to guide the application of

Table 6. Swapping order.



Original order		After swapping	
1	1	1	1
2	0.600259	16	0.70226
3	0.641364	6	0.512121
4	0.601608	5	0.325941
5	0.330598	2	0.654946
6	0.513884	3	0.647055
7	0.838718	7	0.848083
8	0.661817	12	0.31786
9	0.241404	9	0.264669
10	0.539997	17	0.423471
11	0.706125	10	0.542906
12	0.316228	4	0.613274
13	0.643876	15	0.327879
14	0.687324	13	0.647247
15	0.311691	11	0.714364
16	0.696031	14	0.682479
17	0.424929	8	0.659547

the modified LSA model, including recommendations on types of documents that can be accepted for scoring, a guide on writing the model answer, a pre-filtering mechanism and a post-validation engine.

- Extension of LSA to examine graph and tables embedded within the documents.

REFERENCES

Carey, L. (1994). *Measuring and evaluating school learning*, A Division of Paramount Publishing.

Dennis, S. (2000). Machines can't replace humans yet, Automated Essay Marking : getting the facts straight, Centre for Human Factors and Applied Cognitive Psychology, University of Queensland, retrieved August 1, 2000, from http://www.humanfactors.uq.edu.au/hesarticle.html

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An Introduction to Latent Semantic Analysis, *Discourse Processes*, 25.

Landauer, T. K., Foltz P. W., Laham, D. (1999). Automated Essay Scoring: Applications to Educational Technology, retrieved 1999 from http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html

Oosterhof, A. (1994). *Classroom applications of educational measurement*, New York: Merrill.

Page, E. (1996). Grading Essays by Computer: Why the Controversy? retrieved April 11, 1996 from http://134.68.49.185/pegdemo/ref/WhyContr-96.htm

Sarkees-Wircenski, M., and Scott, J. L. (1995). *Vocational special needs*. Homewood, IL:American Technical Publishers, Inc.

Thompson, C. (1999). New Word Order: The Attack Of The Incredible Grading Machine. Linguafranca, *The Review of Academic Life*. Vol. 9, No. 5 – July/August

University of Minnesota (1999). Writing Ture-False Items, retrieved April 1, 1999 from http://www.ucs.umn.edu/oms/truefalse.htmlx

Williams R. (2001). Automated essay grading: An evaluation of four conceptual models, In A. Herrmann and M. M. Kulski (Eds), Expanding Horizons in Teaching and Learning, *Proceedings of the 10th Annual Teaching Learning Forum*, 7-9 Feb. 2001. Perth: Curtin University of Technology

Yunker, B.D. (1999). Adding authenticity to traditional multiple-choice test formats, *Education*, Fall 1999 v120 i1 p82.