Automated Essay Content Analysis Using Concept Indexing

Abigail R. Razon
Prospero C. Naval, Jr.
Dept. of Computer Science
College of Engineering
University of the Philippines
Diliman, Quezon City
cvmig@engg.upd.edu.ph

Belen D. Calingacion
Rowena Cristina L. Guevara
Dept. of Speech Comm. and Theater Arts
College of Arts and Letters
University of the Philippines
Diliman, Quezon City
bcalingacion@gmail.com
gev@eee.upd.edu.ph

ABSTRACT
In this paper, we present a new approach to essay content analysis using the dimensionality reduction algorithm called, Concept Indexing (CI). Experiments are conducted to compare the algorithm’s performance with Latent Semantic Indexing (LSI), which has been known for its empirical success in this area of study. Based on the results, CI has outperformed LSI by -5% in Exact Agreement Accuracy (EAA) measure and by -0.2 in Pearson’s Product-Moment Correlation Coefficient (PMCC) measure. The effects of stemming, removal of stopwords and dimensionality on the algorithms are also presented.

1. INTRODUCTION
Researches on Automated Essay Grading (AEG) have proven its strong potential to be a very powerful educational reinforcement tool, which can be used to supplement teachers in teaching the English language. Writing essays and receiving constructive feedbacks are two of the most important parts in the language learning process. However, for teachers who teach large classes, like in most of our public high schools, grading essay-type examinations or exercises is very time-consuming and labor-intensive, thus, discouraging them from giving such type of activities to students. To address this issue, several AEG systems have already been developed.

Content analysis is one of the major tasks in AEG. In line with this, we present a new algorithm called, Concept Indexing (CI), to accomplish the said task. CI, just like LSI, is a dimensionality reduction algorithm, which captures semantic similarities between pieces of textual information. Its basic implementation involves computation of the p-dimensional representation of a collection of documents or corpus by first clustering the documents into p groups, and then using the centroid vectors of each of the cluster to derive the axes of the reduced p-dimensional semantic space. According to experimental results in [3], CI runs eight (8) to ten (10) times faster than LSI with less memory and storage costs and comparable accuracy results.

The following sections outline the development and implementation of CI to essay content analysis. Section 2 provides the overview existing AEG systems. It also discusses the performance analysis and simulation results of some of these existing systems. Section 3 and 4 describe the steps in the actual implementation of LSI and CI in essay content analysis, respectively. Section 5 presents the results of the experiments conducted on one (1) of four (4) datasets that we have. Lastly, Section 6 gives the conclusion derived from the previous sections and provides suggestions for future work on CI-based essay content analysis.

2. EXISTING AEG SYSTEMS
There are four (4) existing basic conceptual approaches to this technology, namely, Statistical, LSI, NLP, and Bayesian Approaches. These are applied in different prominent AEG systems such as Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), Electronic Essay Rater (E-Rater), and Bayesian Essay Test Scoring System (BetsY), respectively.

2.1 Statistical Approach - PEG
PEG is said to be the very first AEG system. It purely relies on a statistical approach and no NLP technique is utilized. It is based on the concept of proxes which are computer approximations of the intrinsic variables within essays. Such variables include average word length, essay length, counts of each part of speech, number of semicolons or commas, and so on. [16, 15]

The system is made up of two (2) stages, namely, training and scoring. Initially, it is trained on sample essays in which proxy variables are determined for each essay and are entered into the prediction equation. A score is then assigned by computing beta weights or coefficients from the training data [16] – Scoring is done by using multiple regression equation. [18]

Experimental results on PEG only achieved content scores with a correlation value of 0.78. However, it only achieved holistic essay scores with a correlation value of 0.869. These values are about as high as that of the correlation value of five (5) human raters amongst each other. [19]

2.2 LSI Approach - IEA
IEA is a LSI-based approach in AEG systems. In this approach, the text is represented as a 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 row in the matrix stands for a unique word, each column stands for context and each cell represents the frequency of the word, which is considered by a feature – denotes not only the importance of the word in that context but also the degree to which the word type carries information in the domain discourse. The semantics of a word is verified through all the contexts that the word occurs. The number of occurrences of each word in a text determines the semantic space. By reducing this dimension, LSI induces semantic similarities between words. Cosine correlation is used to measure this similarity of the reduced dimensional space constructed from the training data against a test essay. [16, 15]

Experimental results on IEA only achieved content scores with a correlation value of 0.83. On the other hand, the system achieved holistic essay scores with a correlation value of 0.9. [19]

### 2.3 NLP with Statistical Approach - E-RATER with Criterion

E-rater is an embedded AEG system to a web-based real-time application called Criterion. It uses both statistical and NLP techniques to identify specific lexical and syntactic cues in a text to analyze essays. It employs a corpus-based approach to model building, in which actual essay data are used to examine sample essays. [16, 15, 23] The system is made up of five (5) main modules. These are the syntactic, discourse, topical-analysis, model-building and scoring modules. [16, 15, 8] Criterion is an additional component with advisory features that are based on statistical measures and are completely independent from the score generated by E-Rater. It provides additional feedback about qualities of writing related to topic and fluency only. [15]

Experimental results on E-Rater only achieved content scores with a correlation value of 0.69. On the other hand, the system achieved holistic essay scores with a correlation value of 0.89. [19]

### 2.4 Bayesian Approach - BETSY

BetsY is an AEG system that employs the Bayesian Theorem which falls under the Text Categorization Technique Model. [16, 18] There are two (2) Bayesian models widely used in text classification. First, the Multivariate Bernoulli Model in which each essay is viewed as a special case of calibrated features. Second, the Multinomial Model in which each essay is viewed as a sample of calibrated features. [16, 11]

BetsY uses a large set of essay features. These essay features include content related features such as specific words and phrases, frequency of certain content words, form related features including number of words, sentence length, number of verbs, number of commas and others, e.g., the order certain concepts are presented and the occurrence of specific noun-verb pairs. The system learns to classify new essays based on the following steps: train words, evaluate database statistics, eliminate uncommon words, determine stop words, train word pairs, evaluate database statistics, eliminate uncommon word pairs, and perhaps score the training set and trim misclassified training texts. [16]

Experimental results on Bayesian-based AEG system in [12] achieved holistic essay scores with a correlation value of 0.88.

## 2.5 Existing AEG Systems

### Performance Analysis

In evaluating the effectiveness of an AEG system, two (2) metrics are often used. These are Exact Agreement Accuracy (EAA) and Pearson’s Product-Moment Correlation Coefficient (PMCC). EAA is calculated by counting the number of essays in the test set having the same human-graded and computer-graded scores, then, dividing it by the total number of test essays.

\[ EAA = \frac{\text{no. of test essays with } H \text{ score } = \text{AEG score}}{\text{total no. of test essays}} \]

where \( H \) score corresponds to the score given by a human grader and AEG score is the score assigned by the AEG system used.

However, PMCC is calculated using the equation below:

\[ PMCC = \frac{\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} y_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right)^2}} \]

where \( n \) is the number of test essays, while \( x_i \) and \( y_i \) are the two scores assigned by, either two (2) humans or one human and an AEG system, to the same test essay.

Tables 1 to 4 show the simulation results of IEA, E-Rater and a Bayesian-Based System, respectively. For IEA, evaluation study was conducted using 33,205 essays, over seven (7) grade levels, from 6th to 12th. For E-Rater, over 25,000 essays were used for the study, consisting essays from grade levels 6 to 12, and essays from GMAT and TOEFL. The study on Bayesian approach to AEG systems in [12] conducted experiments on Law and Physics essays. Performance evaluation was performed on the Law dataset using 223 essays for training and 50 essays for testing. On the other hand, the Physics dataset was evaluated using 586 training essays and 80 test essays.

### Table 1: OVERALL MEAN INTER-RATER PERFORMANCE METRICS OF IEA ON HOLISTIC* SCORES

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Human-Human</th>
<th>IEA-Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMCC</td>
<td>0.86</td>
<td>0.9</td>
</tr>
<tr>
<td>EAA(%)</td>
<td>61.7</td>
<td>61.1</td>
</tr>
</tbody>
</table>

* - based on all criteria, including content


According to [1], PMCC between two human graders has typical values between 0.6-0.8. From Table 2, however, this is found to be between 0.43-0.59 only. EAA between human graders has a typical value of 60% [2] and based on experiments conducted by Pearson’s Knowledge Technologies Research, it is found to be 61.7% as shown in Table 1.
PMCC values between human graders and E-Rater can be as low as 0.45 as shown in Table 2 while EAA values can be as low as 28% for a Bayesian-based system as shown in Table 4.

### 3. LSI-BASED ESSAY CONTENT ANALYSIS

The LSI-based essay content analysis algorithm implemented in this project is divided into seven (7) major steps. These are as follows:

#### 3.1 Step 1: Preprocessing

After dividing the corpus into training and test sets, the preprocessing stage follows. This is the stage where the terms in the training set are tokenized, lowercased, and sorted alphabetically.

#### 3.2 Step 2: Training and Test Documents Matrix Representation

Next is the construction of the training and test sets frequency matrices $X$ and $Q$, respectively. The rows of these matrices correspond to the terms found in the training or test set, the columns correspond to the documents in that particular set, and the entries on the matrices correspond to the number of times a term is found in each document.

![Figure 1: Step 2 - Training or Test Documents Matrix Representation](image)

#### 3.3 Step 3: Singular Value Decomposition (SVD)

SVD is, then, performed on the training set matrix. SVD is a numerical algorithm in Linear Algebra that decomposes any rectangular or square matrix into three (3) new matrices. For any $m$ by $n$ matrix $X$, SVD is performed by solving the equation:

$$X = UDV^T$$

where $U = XX^T$, $V = X^TX$, and $D$ is a matrix whose diagonals are the singular values of matrix $X$.

#### 3.4 Step 4: Dimension Reduction

Dimension Reduction in LSI is implemented after the decomposition of the original matrix by SVD. This is accomplished by choosing the first $k$ biggest singular values of matrix $D$ and setting the rest of the small values to zero (0), which causes the same number of rows and columns of $U$ and $V^T$ to be disregarded during matrix multiplication. The resulting matrices, $U_k$, $D_k$ and $V_k^T$, are now the new training set vector coordinates in the $k$-dimensional semantic space.
3.5 Step 5: Folding-In
Each column vector of the test set matrix representation is now ready to be projected into the reduced dimensional semantic space. This is accomplished by solving the equation:

\[ q_i = q_i^T U_k D_k^{-1} \]

for all test set column vectors, \( q_i \). This process is called folding-in.

3.6 Step 6: Cosine Similarity Calculation
Since both the training and test set documents are now represented in the reduced dimensional semantic space, cosine similarities between each test document vector, \( q_i \), and all the training document vectors, \( x_j \), can now be calculated. This is done by calculating the dot products between \( q_i \) and \( x_j \), and dividing this by the corresponding document vector lengths.

\[ \text{sim}(q_i, x_j) = \frac{q_i \cdot x_j}{|q_i||x_j|} \]

3.7 Step 7: Score Assignment
The last step in automated essay content analysis is to assign a score to each essay in the test set. After the calculation of cosine similarities, we can now rank these values for each test document. Scores are assigned to each test document based on the score of the training document with the highest cosine similarity to it.

\[ c_j = \frac{m_{j}}{\|m_{j}\|} \]

These normalized concept vectors are placed as columns of the concept matrix, \( C \), otherwise known as the Semantic Space.

4.1.1 Subclustering of Predefined Classes
As mentioned earlier in this paper, in the basic implementation of CI in supervised setting, a concept vector is created for each of the predefined classes or clusters of documents. In which case, the rank of the reduced dimensional space will be the same as the number of these clusters. However, in our datasets, essays belonging to a specific cluster may have independent concepts. As a result, using only one (1) concept vector for each cluster is not enough to represent that cluster in the semantic space. Therefore, subclustering or dividing each cluster into concepts enveloped in it should be implemented. Note that each of the subclusters will still contain essays from the same predefined cluster. [3]

4.2 Step 4: Sparsification (Optional)
Sparsification Strategy can be applied on the \((C_p^T \ast C_p)^{-1}\) matrix. Since this matrix may have many small size or zero entries, these can be set to zero (0) if they are less a certain threshold value. This step is said to be CI’s counterpart of LSI’s truncated SVD. It enhances the overall performance of the system and reduces storage cost.

4.3 Step 5: Concept Decomposition and Folding-In
Concept Decomposition (CD) is, then, performed on both training and test document vectors. Using the Concept Matrix created in Step 3, for any matrix \( A_p \), CD is defined as:

\[ A_p = C_p Z^* \]

where \( Z^* \) is determined by solving the least-squares problem:

\[ Z^* = \arg_{Z} \min \| A - C_p Z \| \]

which has a closed-form solution of:

\[ Z^* = (C_p^T C_p)^{-1} C_p^T A \]

Note that \( Z^* \) is a \( p \times n \) matrix (\( p \)=number of classes, \( n \)=number of documents) whose columns are representations of the documents in the semantic space. Therefore,
to fold-in the training and test sets onto the semantic space, we just need to solve:

\[ X^* = (C_p^T C_p)^{-1} C_p^T X \]

\[ Q^* = (C_p^T C_p)^{-1} C_p^T Q \]

where \( X^* \) and \( Q^* \) are matrix representations of the training and test sets, respectively.

5. SIMULATION DETAILS AND RESULTS

Our experiments concentrated in comparing LSI and CI algorithms using R version 2.9.2 open-source software on MAC OS 10.5.7 and 64-bit Windows 7 Professional operating systems. Since the scoring of the the proposed system would be teacher-dependent, in our experiments, we made sure that the essays were checked by only one teacher. In line with this, we were able to create a dataset, which is consist of two (2) sections of the first (1st) year high school classes of the University of the Philippines Integrated School. This particular dataset is composed of only twenty-eight (28) training documents and sixteen (19) test documents.

In our experiments, we observed the effect of stemming and the removal of stop words. The four (4) cases considered are shown in Table 5.

<table>
<thead>
<tr>
<th>Cases</th>
<th>With Stemming</th>
<th>Stopwords Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Figures 3, 4, 5 and 6 show the EAA and PMCC values from the LSI and CI experiments.

For the LSI experiments, EAA and PMCC results are based on 30% up to 90% LSI semantic space for all the four (4) cases in Table 5. However, for CI experiments, these values are computed using no subclustering (i.e. only one (1) cluster for each predefined class), and two (2) up to seven (7) subclusters for each predefined class for all cases.
As shown on the figures, LSI only achieved a maximum EAA value of -79% and a PMCC value of -0.5 (both in Case 3, Dim=90%). CI, however, achieved an EAA value of -84%, corresponding to its highest PMCC value of -0.7 (both in Case 3, subclusters=7). Based on these, we can conclude that the removal of stopwords (Case3) is very important for both algorithms and that CI can outperform LSI in evaluating essays based on content. Aside from that, as shown on figures 5 and 6, CI’s subclustering of the predefined classes of semantic space has a big impact to its EAA and PMCC values, as evident in the peaking of the said values at subclusters=7.

6. CONCLUSION

In this paper, we have described CI, a new approach in implementing the Content or Semantic Analysis Module of AEG systems. Based on the results of our experiments, we have showed that this algorithm has the potential to replace LSI in evaluating essays according to content. There are also many advantages in using this algorithm such as reduction in storage and memory costs and increase in run-time speed. These advantages are inherent to the algorithm since it requires simpler matrix manipulations and produces sparse matrices, unlike LSI that involves SVD and produces dense matrices.

In the implementation of CI, we have realized the importance of subclustering of the predefined classes of the semantic space to enhance the system’s overall classification performance. Based on the experimental results, we can also say that the removal of stopwords has the biggest impact on the CI algorithm.

For future work, we suggest that simulations on larger datasets should be done to further observe the system’s overall performance. The effects of the topic of the essays in the corpus to the performance measures should also be investigated. We also suggest to implement the subclustering of the predefined classes using different clustering algorithms and to implement different vector space weighting schemes on the datasets, to discover which of these would work best in essay content analysis. It is also highly recommended that the system be tested on other languages, such as Filipino, or on the different dialects of our country, such as Cebuano and Ilonggo. Moreover, it is also interesting to investigate on the effectiveness of the proposed CI-based system in evaluating essays in other high school subjects such as Social Studies and Biology.

7. REFERENCES


