AUTOMATIC CORRECTION ESSAY ON DEVELOPMENT E-LEARNING IN SMK PLUS AN-NABA SUKABUMI LATENT SEMANTYC USING ANALYSIS (LSA)

Mashun Sofyan[1]

[1] Information Engineering - University Computer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail : Mashunsofyan@gmail.com[1]

ABSTRACT

Manually correcting essays that spend more time compared to proofread essays automatically. It takes about 1 minute each question to correct essays manually. If the problem will correct 30 then takes about 30 minutes, while the number of students in vocational Plus An-Naba Sukabumi about 200 people.

Based on these data, the need to develop an automated system that helps process essay grading. Development of Automated Essay Scoring System in this final project uses Latent Semantyc Method Analysis (LSA) is to represent the sentence in a semantic matrix that is then performed mathematical calculations to match or map the presence or absence of the word of the group said in a matrix using linear algebra techniques Singular Value Decomposition (SVD).

Based on the results of the research data of alpha and beta can be drawn a conclusion that the development of an essay correction functionality in SMK Plus An-Naba Sukabumi able to carry out the correction process about the type of essay more quickly.

Keywords: LSA, Semantyc Latent Analysis, Singular Value Decomposition (SVD).

1. INTRODUCTION

Manually correcting essays that spend more time compared to proofread essays automatically. It takes about 1 minute each question to correct essays manually. If the problem will correct 30 then takes about 30 minutes, while the number of students in vocational Plus An-Naba Sukabumi about 200 people.

Based on these data, the need to develop an automated system that helps process essay grading. Development of Automated Essay Scoring System in this final project uses Latent Semantyc Method Analysis (LSA) is to represent the sentence in a semantic matrix that is then performed mathematical calculations to match or map the presence or absence of the word of the group said in a matrix using linear algebra techniques Singular Value Decomposition (SVD).

Based on the results of the research data of alpha and beta can be drawn a conclusion that the development of an essay correction functionality in SMK Plus An-Naba Sukabumi able to carry out the correction process about the type of essay more quickly.

Keywords: LSA, Semantyc Latent Analysis, Singular Value Decomposition (SVD).

1.1 E-Learning

The development of increasingly advanced information technologies have been able to help people in all fields. One area that did not escape the utilization of information technology in the process is in the field of education. the development of increasingly advanced learning how to make a paradigm shift and gave birth to new methods in procedural learning. teaching and learning process can now be made anywhere without direct face to face in a room. Their interaction can be done with the help of electronic media such as computers. Learning with the help of electronic media such as this is now known as e-learning.

Basically the concept of e learning can be classified on the basis of the technology it uses into two types, namely computer based training (CBT), which uses a computer program as the main media used by students. And the kind of which is a web-based training (WBT) is a further development of CBT systems using Internet-based technology enhancements. At WBT types of delivery and access subject matter is done through the internet.

1.2 *Learning Management Sistem (LMS)*

In e-learning, there are two main parts that sustain the system, ie Learning Management Systems and e-learning content or subject matter to be learned by the user. Learning Management System (LMS) is a system that helps and serves as a platform for e-learning content.

1.3 Automatic Answer Ratings

Each learning requires an evaluation to determine the extent of absorption of information by students. In general there are two kinds of evaluations that can be performed in the learning

process that is verbally and in writing. Oral exam has the disadvantage of examiners and students must be tested in one room and the same time and require considerable time, especially if the examinees are numerous. While the writing exam is much more efficient because the test can be done at the same time the number of examinees more.

Written test types can be divided into three kinds, namely multiple choice, short stuffing, and essays. Of the three types of tests that can be done in writing the easiest test to be checked automatically is multiple choice. In the multiple-choice type of exam students can simply choose the most correct answers were considered from multiple choice answers. but the weakness of the model multiple choice exam is less failure by examiners determine the extent of knowledge of students tested and the possible presence of the participants answer that gets high marks from the factor of luck by guessing the answers provided. The level of accuracy that can be done by a computer to multiple-choice type of exam can reach 100% because the form is very easy uijan like computer processed simply by comparing the student answer with the answer key in the database.

The next type of written exam is short stuffing. On this kind of field test enough students to fill short answers to questions asked in one or several words. This type of exam is slightly better than multiple-choice because the demands of students to have an answer himself. And other types of written exam is the essay. The essay is an essay that describes the author's opinion on a particular subject is trying judged. Or also can be interpreted as an essay that explores proga in passing but accurate, dense and contains about issues literature, art and culture from the subjective viewpoint of the author. According KBBI essay is paper or essay in prose that describes about a problem from the point of view of the author's straightforward and cursory. The form of essays is a form of testing that accurately able to describe in-depth understanding of students about the topics tested. Furthermore essay exams are also able to stimulate the ability to express opinions in writing.

Exam in the form of essays has shortcomings in terms of resources needed human and time. When an essay exam followed by many participants of the time needed to check the answers will increase anyway. To resolve the problem in the assessment of essay exams, essay grading system automatically using a computer can be a good enough solution.

1.4 Text Preprocessing

Text preprocessing adalah langkah awal sebelum dilakukan sebuah pencocokan string. Proses ini melalui beberapa tahap tindakan yang dilakukan pada sebuah string jawaban dan kunci yang diisi by teachers and students. The stages are carried out in this process are:

1. Case folding / toLowerCase
Case folding is the stage where the string has been inputted by the user who originally is a mix of capital and lowercase letters generalized to be used to lowercase all. This is done because the computer system distinguishes between "A" and "a" so that the "I" and "I" will be assessed differently in the system. Obviously this will have an effect on string matching between the key and the answer of the student.
2. Removal of punctuation
This stage is the removal of punctuation so that remaining string only consist of letters and numbers only. This stage is done because the system would be built only match strings and numbers have not melibarkan mathematical operations.
3. Elimination of common words (stopword)
Stages stopword is the removal of the words are considered public so that the words are not inserted into the string array that will be matched.
4. Tokenizing
Tokenizing is the stage of cutting the string in the text based on the deed of which it is composed and sisimpan in a variable array.
5. Stemming (deletion prefix and suffix)
Stemming is one of the means used to improve the performance of search and matching words with ways to transform the words into a text document to the basic form. Stemming process is done by eliminating all good additive consisting of a prefix, suffix and inserts as well as a combination of the prefix and suffix. There are many algorithms used to perform the process of stemming. Especially for stemming Indonesian frequently used algorithms are algorithms nazrief and andriani, arifin and Setiono, veda algorithms, etc.

1.5 Matriks Singular Value Decomposition

SVD is a so-called linear algebra theorem able to break the block of a matrix A into three new matrix U is an orthogonal matrix, diagonal matrix S, and orthogonal matrix transpose V. SVD theorem is as follows:

$$Amn = Umm \times SMN \times VTnn \quad (1)$$

Where UTU = identity matrix (I), VTV = identity matrix (I). Column matrix U is an orthonormal eigenvector of AAT. While the column matrix V is an orthonormal eigenvector of ATA. and S is a diagonal matrix root of the eigenvalues of matrix U or V in order from the largest.

1.6 Cosine Similarity

Cosine similarity is one similarity calculation method which is the most popular to be

applied to the text document. The main advantages of the method cosigne similarity is not affected on a short length document. Karna is only taken into account the value of the term of each document. Here is the formula for the cosine similarity.

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

Description:
A = matrix A
B = Matirks B
|| A || = Length Matrix A
|| B || = Length Matirks B

1.7 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a mathematical calculation method is fully automated techniques in excavation and inference of the relationship between words and sentences in paragraph [6]. LSA algorithm is one of algorithm development in the field of Information retrieval is able to collect a large number of documents in the database and mengghubungkan relationships between documents by matching a given query. The main function of this LSA is to calculate the similarity (similarity) document by comparing the vector representation of each document. Assessment by the LSA method over the words that exist in writing without regard to the order of words and grammar in the text, so that a sentence is considered is based on key words that exist in the sentence.

To calculate the value of essay using the LSA, the way is to create an essay that became the reference document (query) or key answers and then an essay that will be judged in comparison with the reference essay answers, the more resemblance to the greater value of the answer.

Inspection measures essays with LSA method is the first article represented in the matrix where the row shows the keywords (queries) on the document and each column shows the column documents. Each cell shows the number of occurrences of a keyword which is the matrix of documents in the columns of the matrix. Then the contents of the first cell is transformed where every word frequency weighted with a function that shows the importance of a sentence in a paragraph and also the degree which shows how important the type of words in a sentence.

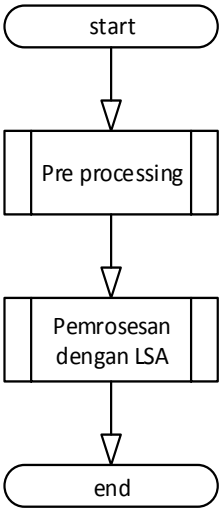
The next step is to do the decomposition of Singular Value Decomposition (SVD) on word-document matrix. In the SVD matrix containing the keyword appearance frequency decomposition into three matrices that if the three matrices are multiplied it will reappear matrix origin. His first matrix describes the entity as a column vector value orthogonal matrix. Both in the form of a matrix diagonal matrix containing a scalar value matrix. Mathematically, the factors that most good is to use the smallest dimension of the matrix first, so that the reconstruction of the best matrix is generated when the value factor is smaller than the number of factors used.

The dimensions of the matrix which has been simplified by removing the matrix diagonal coefficients as many times as desired until the remaining coefficients are chosen as much dimension. The purpose of this simplification is in order to form a matrix that contains the desired correlation value when three matrix reconstructed. Then the assessment will be done by comparing the correlation matrix of query with a correlation matrix of each document by using the cosine α calculation.

2. CONTENTS OF RESEARCH

Development of correction essays included in the training features in a single page along multiple choice questions. Development is done because it found problems in the system that runs that correction manually essay will take a long time as described in the background research that requires an additional feature of automatic correction of essays in the system. Development added in section exercises with automatic correction feature essays which in the previous system already teradapat features exercises but terabatas only multiple choice questions only.

The system to be developed in this study is an essay about automatic correction system with Semantyc Latent Analysis method. In the process before the system performs the calculation method LSA no processes are done first. The development of this essay correction consists of two subprocesses that process preprocessing and LSA process itself. Preprocessing process consists of a folding case, tokenizing, filtering, and stemming. Here is a flowchart of each process.

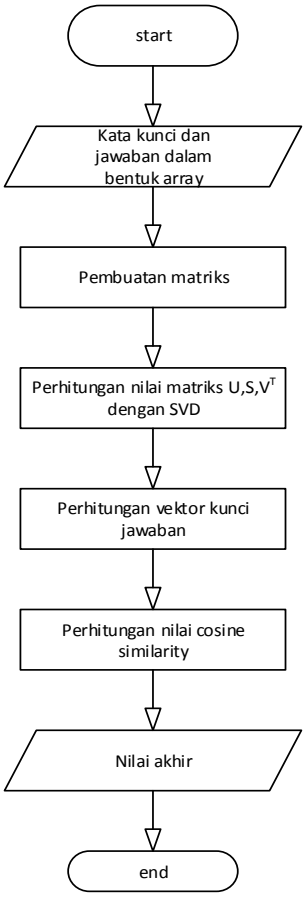


Gambar 1 Flowchart Automated Essay Correction System

To further facilitate the understanding of the process is given an example of matching between the key and answer students' answers. In the following example is one of the replications are held, amounting to 2 essays Indonesian subjects. If a perfect score in this test was 10 then every problem has a weight of 5 in the preparation of these values.

2.1 Analysis methods LSA

Stages are contained in the LSA process consists of four stages in which each stage is no mathematical calculation process. This process is a continuation of the process of preprocessing has been done before to get the value of similarity between the answer and the answer key.



Gambar 2 *LSA* Process Flowchart

2.2 Calculation of the value of the matrix U, S, VT

Having formed a matrix of size mxn then the next step is to find the value of the matrix U, S, VT with SVD method (Singular Value Decomposition). To obtain the following matrices through some fairly complex calculations ranging from doing transpose matrix, searching for eigenvalues and seek vector eigennya. However, in the implementation of the program, the calculation process can use the library of JAMA that has been providing a function to perform matrix calculations instead of MATLAB to obtain the matrix U, S, VT.

From table 3.7 obtained a 9x4 matrix A with dimensions that will do the SVD to find the value of the U, S, VT. The first step to seek the matrix is to find the matrix AT beforehand.

$$A = \begin{vmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{vmatrix}$$

$$\begin{vmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{vmatrix}$$

$$A^T \begin{vmatrix} 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{vmatrix}$$

The next step is to find the value of A * AT and AT * A so we get the following matrix:

$$A^T * A \begin{vmatrix} 7 & 4 & 4 & 4 \\ 4 & 5 & 2 & 2 \\ 4 & 2 & 5 & 3 \\ 4 & 2 & 3 & 5 \end{vmatrix}$$

$$A*A^T \begin{vmatrix} 3 & 1 & 2 & 2 & 2 & 1 & 3 & 1 & 2 \\ 1 & 2 & 1 & 0 & 2 & 1 & 2 & 0 & 1 \\ 2 & 1 & 3 & 2 & 2 & 2 & 3 & 1 & 1 \\ 2 & 0 & 2 & 2 & 1 & 1 & 2 & 1 & 1 \\ 2 & 2 & 2 & 1 & 3 & 2 & 3 & 0 & 2 \\ 1 & 1 & 2 & 1 & 2 & 2 & 2 & 0 & 1 \\ 3 & 2 & 3 & 2 & 3 & 2 & 4 & 1 & 2 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 2 & 1 & 1 & 1 & 2 & 1 & 2 & 0 & 2 \end{vmatrix}$$

Having obtained the multiplication of the matrix A and the transpose of the matrix A above, the next step is to find the eigenvalues by reducing the matrix with lambda multiplied by the identity matrix. The first is to find the eigenvalues of AT * A.

value eigen = $(A^T * A) - \lambda I$

$$\begin{vmatrix} 7 & 4 & 4 & 4 \\ 4 & 5 & 2 & 2 \\ 4 & 2 & 5 & 3 \\ 4 & 2 & 3 & \end{vmatrix} - \begin{vmatrix} \lambda & 0 & 0 & 0 \\ 0 & \Lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \end{vmatrix} \begin{vmatrix} 7 - \lambda & 4 & 4 & 4 \\ 4 & 5 - \lambda & 2 & 2 \\ 4 & 2 & 5 - \lambda & 3 \\ 4 & 2 & 3 & 5 - \lambda \end{vmatrix}$$

From the above calculation continues to find the polynomial equation (many parts) of the matrix above. The trick is to multiply each element is reduced another element of the side. So didapakan the following equation:

$$\lambda^4 - 21\lambda^3 + 300\,\lambda^2 - 858\,\lambda + 1375 = 0$$

Having obtained a polynomial equation as above the next step is to find the root of the equation thus obtained

$\lambda_1 = 15{,}400$ $\lambda_2 = 1{,}235$

$\lambda_3 = 3.366$ $\lambda_4 = 2{,}000$

As for the matrix A * AT to do the same process eigenvalues obtained as follows

$.\lambda_1 = 15{,}400$ $\lambda_2 = 3{,}366$

$\lambda_3 = 2{,}000$ $\lambda_4 = 1{,}235$

$\lambda_5, \lambda_6, \lambda_7, \lambda_8, \lambda_9 = 0{,}000$

Having obtained the eigenvalues of each of the above equation next step is to include the value of λ at the beginning of the equation which is then normalized to obtain the following matrices. For eigen vector of the matrix AT * A will be the value of V which will then be transposed into VT. And the vector of eigenvalues of A * AT will be the value of the matrix U. While the diagonal matrix S is obtained from the roots of eigenvalues that have been sorted from the biggest and not zero.

Matriks U

$$\begin{vmatrix} 0.386 & -0.248 & -0.5 & -0.025 \\ 0.233 & 0.483 & 0 & -0.533 \\ 0.386 & -0.248 & 0.5 & -0.025 \\ 0.269 & -0.489 & 0 & 0.251 \\ 0.395 & 0.413 & 0 & 0.131 \\ 0.279 & 0.171 & 0.5 & 0.408 \\ 0.502 & -0.006 & 0 & -0.302 \\ 0.107 & -0.419 & 0 & -0.433 \\ 0.279 & 0.171 & -0.5 & 0.408 \end{vmatrix}$$

Matriks S

$$\begin{vmatrix} 3.924 & 0 & 0 & 0 \\ & 1.835 & 0 & 0 \\ 0 & 0 & 1.414 & 0 \\ 0 & 0 & 0 & 1.111 \end{vmatrix}$$

Matriks V^T

$$\begin{vmatrix} 0.636 & 0.421 & 0.457 & 0.457 \\ 0.129 & 0.769 & -0.443 & -0.443 \\ 0 & 0 & -0.707 & -0.707 \\ 0.761 & -0.482 & -0.307 & -0.307 \end{vmatrix}$$

$\bar{q}$= | 1 0 1 1 1 1 1 0 1 |

0.386	-0.248
0.233	0.483
0.389	-0.248
0.269	-0.489
0.395	0.413
0.279	0.171
0.502	-0.006
0.107	-0.419
0.279	0.171

0.255	0.00
0.000	0.545

Having obtained the third matrix next step is to do a simplification of the columns of the matrix. Simplification is done on the matrix U by simplifying the number of columns, the matrix S in rows and columns. This simplification measures carried out in accordance with the value of the dimension of the matrix. so we get the matrix U and S with new dimensions as follows.

Matriks U

0.386	-0.248
0.233	0.483
0.389	-0.248
0.269	-0.489
0.395	0.413
0.279	0.171
0.502	-0.006
0.107	-0.419
0.279	0.171

Matriks S

3.924	0
0	1.835

2.3 Perhitungan vektor kunci jawaban

Searches vector matrix Q and each D is done to make the process of calculating the cosine similarity in the next stage. At this stage steps perform matrix multiplication with the set term frequency of each query ditranspose of the matrix U are then multiplied again with the inverse of a matrix S (S-1). The following calculation is an example of the calculation of the key vectors answers (Q).

$$\bar{q} = Q^T . U . S^{-1} \quad (2)$$

$\bar{q} = \bar{q}$ = | 0.617 -0.135 |

So the above matrix calculation results obtained a vector of query answers (Q) is $(q)^{-}$ = | 0617 -0135 |. By performing the same calculations then each answer will result in the value of the vector as follows:

Value vector Q = | 0617-0135 |
Vector values D1 = | 0407 -0813 |
Vector values D2 = | 0444 0470 |
Vector value D3 = | 0445 0470 |

2.4 Cosine Similarity Value Calculation

Cosine similarity value calculation is done in order to get the value or level of similarity between two vectors. In this case the vector which is the vector will be calculated between each key and answer students' answers. This calculation is done using the formula in the equation below.

(3)

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

$$D1 = \frac{(Q_1 . D_{11}) + (Q_2 . D_{12})}{\sqrt{(Q_1)^2 + (D_{11})^2}\ x \sqrt{(Q_2)^2 + (D_{12})^2}}$$

with:

D = the answers students

Q = query answers

A = matrix A

B = Matirks B

|| A || = Length Matrix A

|| B || = Length Matrix B

So that D1 = $\frac{(0{,}617)(0{,}407)+(-0.135)(-0.813)}{\sqrt{(0{,}617)^2+(0{,}407)^2}\ \sqrt{(-0.135)^2+(-0.813)^2}}$ = 0.644

So we get the cosine similarity value of D1 by 0644. For other values are also performed by the same process to obtain respectively cosine similarity value as follows :

$$D2 = \frac{(0,617)(0,444)+(-0.135)(0,470)}{\sqrt{(0,617)^2+(0,444)^2}\ \sqrt{(-0.135)^2+(0,470)^2}} = 0.566$$

$$D3 = \frac{(0,617)(0,445)+(-0.135)(0.470)}{\sqrt{(0,617)^2+(0,445)^2}\ \sqrt{(-0.135)^2+(0.470)^2}} = 0.566$$

Having obtained the similarity value of each answer the next step is multiplication by the weight of each question were tested. In this matter has a weight of 5 out of 10 points that exist in a repeat of the two problems, so that when the answer of the students have a perfect similarity with the given key will have 5 points on that question.

So the value of the recommendations obtained by the teacher of correction functionality in the example problems about this model are as follows:

Value recommendation = cosine similarity value * weighting each question

D1 = 0.644 * 5 = 3.2
D2 = 0.566 * 5 = 2.62
D3 = 0.566 * 5 = 2.62

Resulting Value recommendation of three students who have been working on essays with different answers convinced as follows:
The first student or D1 produces a value on 3:20 rounded to 1 decimal figure to 3.2, second or D2 Students generate value recommendations on 2.62 value rounded to 1 decimal figure to 2.6 and the third student or D3 generate value 2.62 on Value recommendation rounded to 1 decimal numbers into 2.6

3. CLOSING

Based on the description and the results of the analysis conducted, it can be concluded that the existence of automated essay correction functionality teachers are able to carry out the correction process about the type of essay more quickly. Previous takes about 1 minute each question to correct essays manually. If the problem will correct 30 then takes about 30 minutes, while the number of students in vocational Plus An-Naba Sukabumi about 200 people. After development takes about 5 seconds to correct each question. If the problem will correct 30 then takes about 150 seconds or 2.5 minutes, whereas if it would correct the 200 questions then takes about 1000 seconds or 5 minutes

For the development of automated essay correction functionality, suggestions for subsequent research in order to use methods other than Latent Semantyc Analysis (LSA) in performing automatic correction essay questions.

BIBLIOGRAPHY

[1] M. P. B. S. Sahriar Hamza, "Sistem Koreksi Soal Essay Otomatis Dengan Menggunakan Metode Rabin Karp," EECCISS, vol. 7, no. 2, p. 153, 2013.
[2] B. B. D. H. Anak Agung Putri Ratna, "SIMPLE : SISTEM PENILAI OTOMATIS UNTUK MENILAI UJIAN DALAM BAHASA INDONESIA," MAKARA TEKNOLOGI, vol. 11, no. 1, pp. 5-11, 2007.
[3] R. S.Pressman, Software Enginering A Practitioners Approach, New York: McGraw-Hill, 2010.
[4] Munir, "KONTRIBUSI TEKNOLOGI INFORMASI DAN KOMUNIKASI (TIK) DALAM PENDIDIKAN DI ERA GLOBALISASI PENDIDIKAN INDONESIA," Indonesian Comunity on Information and Comunicartion Technology, vol. 2, no. 2, 2009.
[5] D. E. Hartley, Selling e-Learning, ASTD, 2006.
[6] P. W. D. L. Thomas K Launder, An Introduction to Latent Semantic Analysis, 1998.
[7] M. S. Rossa A.S M, Rekayasa Perangkat Lunak terstruktur dan berorientasi objek, Bandung: Informatika, 2013.
[8] A. Nugroho, Analisis dan Perancangan Sistem Informasi dengan Metodologi Berorientasi Objek, Bandung: Informatika, 2005.