Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment

Alfirna Rizqi Lahitani[1)], Adhistya Erna Permanasari[2)], Noor Akhmad Setiawan[3)]
[(123]Department of Electrical Engineering and Information Technology, Faculty of Engineering
Universitas Gadjah Mada
Jl. Grafika No. 2 Kampus UGM, Yogyakarta 55281, Indonesia
[(1]alfirna.ti14@mail.ugm.ac.id, [(2]adhistya@ugm.ac.id, [(3]noorwewe@ugm.ac.id

***Abstract*- Development of technology in educational field brings the easier ways through the variety of facilitation for learning process, sharing files, giving assignment and assessment. Automated Essay Scoring (AES) is one of the development systems for determining a score automatically from text document source to facilitate the correction and scoring by utilizing applications that run on the computer. AES process is used to help the lecturers to score efficiently and effectively. Besides it can reduce the subjectivity scoring problem. However, implementation of AES depends on many factors and cases, such as language and mechanism of scoring process especially for essay scoring. A number of methods implemented for weighting the terms from document and reaching the solutions for handling comparative level between documents answer and expert's document still defined. In this research, we implemented the weighting of Term Frequency – Inverse Document Frequency (TF-IDF) method and Cosine Similarity with the measuring degree concept of similarity terms in a document. Tests carried out on a number of Indonesian text-based documents that have gone through the stage of pre-processing for data extraction purposes. This process results is in a ranking of the document weight that have closesness match level with expert's document.**

Keywords— Automated Essay Scoring (AES); TF-IDF; Cosine Similarity

I. INTRODUCTION

Evaluation is to determine the extent of competence and capabilities are achieved. In the context of learning process, the evaluation can be conducted by giving the assignment. Usually the assignment given by the lecturer to the students in various forms, for example is the open-asnwer assignment. Open-Answers assignment or usually called essay allows the students to write about ideas, opinions and problem solution. Giving task in essay format provides space to develop communication skills through text description.

Mechanism of assessment is more complicated for essay because considering some components such as knowledge side, opinion side, skills of writing and behavior side. The answers should not a long character description, but right in the core of problem is defined as sentences that have similar meaning, although different construction structure of their words. For contextual content, marking essay assignment requires lecturers to compare the similarity of sentences from student answers.

Technology supports the collection of data, processing of data, until distribution of data in text form. Along with its development, the technology enables a variety of instructional media facilities to the student's assignment using e-learning media that can be done by offline or online. Using e-learning media for assigning task is commonly done in educational purposes, but mechanism of correction and assessment for essay test is done manually in some cases.

Problems appear when the lecturer gives an assessment in text format. It often makes them read it one by one to correct the documents then give a score or value. This will require a special time and also susceptible from subjective assessment.

In this study, the assessment of text data will be analyzed to obtain information retrieval which will measure the level of similarity in order to simplify the correction process based on essay answers. In some studies, the analysis of the information retrieval automation utilized for text-based assessment tasks or commonly known as the Automated Essay Scoring (AES).

To get information retrieval for automated scoring based on text data is in extraction features process, and for this process is also dependent of languages because they have different structure that can be implemented by special technique. Mechanism of essay assessment needs consideration not only about delivered by language, but some contextual component of essay and also process to comparing the answers that have similarity with the expert answers.

A number of methods such as Term Frequency-Inverse Document Frequency (TF-IDF) applied for finding information on social media content [1]. TF-IDF often combined with other techniques to weighting terms on essay online that delivered by English [2]. Some approach to determine similarity level applied such as cosine similarity [3]. Cosine similarity also used for document clustering purposes [4].

This paper presents the implementation of TF-IDF method and cosine similarity approach to measure the

similarity level from Indonesian essay assessment, where the results will be sorted based on documents that have a high similarity level to expert's document. A number of text that have gone through the pre-processing steps will be matched with the query document (q) that contains the keyword terms, further weighting terms using TF-IDF method and cosine similarity to measure the degree of similarity from expert's document to student's document. Thus obtained manner appropriate assessment, more objective according to the accurate portion.

II. OVERVIEW

A. Text Data Mining (TDM)

This research entered the area of data mining analysis. Data mining is a branch of science with a scope broad enough to offer a solution to the anomaly data discovery, sort the set of data, match the data similarity, patterns extraction and data structures [5]. Data mining exploration on large amounts of data to find trends and previously unknown information based on the knowledge and techniques. Utilization data mining applications in higher education is an area of development in the context of education data exploration known as the Educational Data Mining (EDM) [6].

EDM focuses on academic data mining such as Intelligent Tutoring System (ITS), Learning Management System (LMS), Student Behavior Modeling, Predicting Performance and Assessment of student's learning performance [7]. In this study, the text data is the core to explore and analyze, this specific area of data mining also referred as text data mining (TDM). Text mining offers a solution to the problem of data processing, data analysis of structured and unstructured adoption, collaboration techniques of data mining, machine learning, and others [8].

B. Pre-Processing

Pre-processing is the extraction process on a set of words in the document, pre-processing is done with the aim of getting the special features for the purpose of information retrieval [8]. Pre-processing plays a very important role and being the first step in text mining process [9]. The pre-processing steps for text data is performed as follows:

- Case Folding, is the process of changing all the letters in the document to lowercase and remove characters other than letters "a" through "z".
- Tokenizing, is the process of separating each word in the document became the basis of the word, removing prefix, insertion, suffix and duplication.
- Filtering, is the process of removing the words that are considered unimportant commonly called stopwords. Stopwords begins by removing the words that are very common goal to reduce the number of occurrences of words that do not have significant meaning. This process is aimed to get the word to represent the content of the document.
- Indexing, is the process of storing a number of terms in sequence with certain rules. The goal is to accelerate the process of the search term in the document.

Pre-processing steps are depending on language, this happened because every language have a different standard and structure from other language, thus it is necessary for special technique to complete this step.

A number of research that discussed about pre-processing for Indonesian language focused in Tokenizing process or called Stemming for getting token from sentence or document is using Paice/Husk Algorithm with token dictionary as word reference[10]. Tokenizing for Indonesian language is allowing the rule of morphology prefix in Indonesian such as suffix, infix and combination on both for getting appropriate token. Tokenizing using Indonesian morphology approach also called as Nazief –Adriani Algorithm[11].

C. Term Frequency – Inverse Document Frequency

A number of feature extraction results in the pre-processing step known as the term. Weighting term in a document can be determined by the Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF is text based statistical weighting techniques used for the purpose of information retrieval [4]. The studies which implement TF-IDF method to determine the relevance to a query document [12] is TF-IDF weighting for representing the proportion of words in a document. Words with high of TF-IDF weighting shows a strong relationship to the document, including the document query.

D. Cosine Similarity

Weighting term results will be used for the calculation of similarity between expert document to the student documents. The similarity is literally the similarities and are often used in the process of data classification that has similar characteristics. Similarity measure can be exploited to calculate the distance of the similarity of the two things being compared and improve accuracy of information retrieval [13]. Method that can be applied to calculate the similarity distance is the cosine similarity.

Cosine similarity is the method used to measure the degree of similarity, this method is a traditional method that is often used and combined with the TF-IDF. Cosine Similarity is a measure of similarity between two vectors obtained from the cosine angle multiplication value of two vectors being compared [3].

Figure 1 illustrate the comparing similarity level of document using cosine degree concept, where vector coordinate as a documents that compared and cosine degree between vector is similarity degree. Based on cosine principal, cosine 0˚ is 1 and less than 1 to the value of another angle, then the value of the similarity of the two vectors are said to be similar when the value of the cosine similarity is 1.

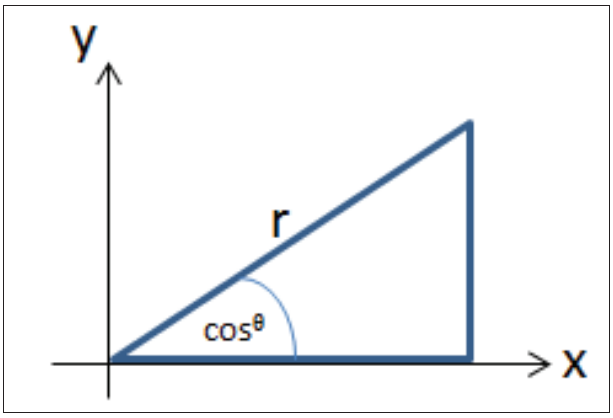

Fig. 1. Cosine degree for similarity concept.

Implementation of cosine similarity method is useful in classifying data on the number of objects that have a certain similarity, as research clustering based on cosine similarity measure [4]. Cosine similarity is used for the calculation of the number of vector cosine angle of the documents and the development of hierarchical clustering algorithm on the document as an effort to increase efficiency and performance.

E. Automated Essay Scoring (AES)

Analysis of text data with TF-IDF method and cosine similarity can be applied for the purpose of Automated Essay Scoring (AES), the system that has been developed to grade the student answers based on marking scheme by applying various algorithms for the implementation of a more accurate scoring.

Some studies with the theme AES them is research by applying Support Vector Regression (SVR) [14] in the CET4 data College English Test for evaluation English writing test. The process is implemented in the system include feature extraction, word length greater than 5, grammar, sentence, off-topic essay and the essay overall similarity score. Based on the process, have the largest value that is closer resemblance to the full-score essays. The test results obtained precision level of 86% which compared to vote manually.

A similar research in the CET4 data but with the different methods, using the K-Nearest Neighbor (KNN) algorithm [15]. Based on these studies, each essay will be represented in the form of Vector Space Model (VSM). The initial step pre-processing of data, then the term vectors are computed using the TF-IDF weight, combined with the calculations Information Gain (IG). Similarity calculation performed by the KNN algorithm. The results of the testing showed that the accuracy obtained more than 76%.

III. METHODOLOGY

In this study, the analysis of text data, terms weighting until the essay document similarity scoring process conducted in stages as follows:

A. Pre-processing Document

A Pre-processing is the first process to be followed by a set of text-based data prior to entry in the processing. The structured or unstructured text data cleaned to reduce noise dimensions of the features of the dataset. The step of pre-processing includes: Case Folding, Tokenizing, Filtering, Indexing.

For Indonesian language, Tokenizing process is following Nazief-Adriani algorithm. The simple steps performed as follows [11]:

- Searching for the word in dictionary, if found then assumed that is root word.
- removing inflection suffixes, for Indonesian language such as “-lah”,”-kah”,”-nya” and etc.
- removing derivation suffixes, for Indonesian language such as “-i”,”-an”,”-kan” and etc.
- removing derivation prefix.

B. TF-IDF Weighting

The calculation of the TF-IDF weighting (w) is done by calculating the term (t) contained in each document (d), the number of terms in each document is indicated by the notation Term Frequency (tf). Total term presence in all the documents collected in the notation (df). Furthermore, to get the value of the Inverse Document Frequency (IDF), the formulation which used as follows:

$$\boldsymbol{idf_i = \log\left(\frac{N}{df_i}\right)} \quad (1)$$

Value (idf) as in (1) is the Inverse Document Frequency is calculated on all the existing term. Notation (N) is the overall number of documents, the notation (df) is total term presence on the entire document. The last step to get the TF-IDF weights on each term is as follows:

$$\boldsymbol{w_{i,j} = tf_{i,j} \times idf_i} \quad (2)$$

Weighting tf-idf (w) as in (2) is the multiplication of term frequency in each document that is denoted by (tf) and the value of the inverse document frequency of a term that is denoted by (idf).

C. Similarity Measure

Cosine similarity is a similarity rate the calculation obtained from the cosine angle multiplication of two vectors being compared, because the cosine 0˚ is 1 and less than 1 to the value of another angle, then the value of the similarity of the two vectors are said to be similar when the value of the cosine similarity is 1 [13]. Cosine similarity calculation performed by the following formula [4]:

$$\boldsymbol{Cos\ \alpha = \frac{A \times B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}} \quad (3)$$

The concept of two-way degree of similarity states that have similarities. As in (3), where A is the weight of each feature of the vectors A and B is the weight of each characteristic in the vector B. Principles cosine similarity, the greater angle formed between two coordinate vector comparison documents, the smaller degree of documents similarity. Conversely the smaller degree of cosine similarity level then degree similarity will be greater.

IV. EXPERIMENT AND RESULT

Figure 2 is the flow of document examination until the final result of the level of similarity ranking test document with an expert document:

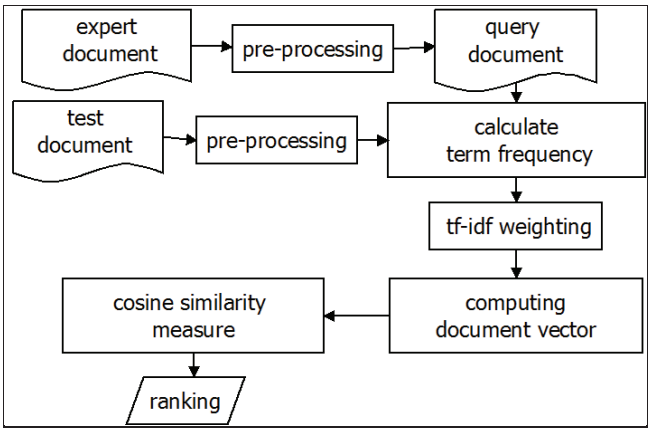

Fig. 2. Flow of document testing.

Based on the flow shown in Figure 2, to test the similarity, prepare a document which is introduced from experts. The Indonesian text-based documents have gone through the stage of pre-processing and contains a number of terms that will be used as a keyword. The document here in after referred to as document query (q). On the other hand, there is a document that will be tested proximity similarity. These documents come from the students' answers text-based in Indonesian language that first going through the stage of pre-processing to get a feature extraction in the form of a term that has been indexed.

Furthermore, along with the document query (q), calculated the number of terms contained in each document (tf) and the total number of times each term contained in the document throughout (df). TF-IDF weighting performed on the entire term in each document. After each document test has been weighted, the next is to calculate vector lengths of each document to then calculate the level of similarity performed on each document the test by comparing the weight of term lock on the document query (q) with a weight terms contained in a document (d).

In this study, the test is done on a set of data derived from e-learning in the form of student response data in the form of an essay to then corrected degree of similarity with an expert document. Here are the stages of testing:

A. Feature Extraction On Pre-Processing Steps

Rate the essay is usually done by comparing the points of the students' answers are contained or a similar approach with expert answers. To get a keyword as a comparison, is provided 5 documents emanating from an expert. The documents are then through the pre-processing stage to get a feature keywords or index terms are here in after referred to as document query (q). The resulting output from the pre-processing 5 expert document is as much as 60 query terms.

On the other hand, 10 student's documents also through pre-processing stage to obtain an index term test document. A total of 166 terms from the query document and test documents have been indexed. Whole number of terms will be denoted by (t).

B. Computing Term Weight Using TF-IDF

After all frequencies term index has been calculated from each document as (tf), the next step is to calculate the total number of times throughout the term contained in the document as (df), there 160 terms are available.

Then calculation from the inverse document frequency as in the formula (1), where N is the number of documents that the testing is as much as 11 documents (including document query) and total of (df) an existing term in the entire document. Inverse will count as much as the existing term, the 160 terms. A number of calculation terms as shown in Figure 3.

term	q	tf										df	idf (d)
		d1	d2	d3	d4	d5	d6	d7	d8	d9	d10		
t1		1	1				1	1	5	3	3	7	0.155
t2	1	2						1	1	6	1	6	0.222
t3	1	1		7	2	3	3	3	4	3	2	10	0
t4	1	1	4	2	3	2		2	2	2	1	10	0
t5	1	1	5	4	5	2	9	5	14	6	8	11	-0.04
t6	1	1	1	1	1	1	1	1	1	1	1	11	-0.04
t7		8	6			2	1			12		5	0.301
t8	1	7	5			2				11		5	0.301
t9		3					1	1	1		1	5	0.301
t10	1	3	2	12			7	2	3		1	8	0.097

Fig. 3. Result of IDF Weighting.

The final result of the weighting is the value of TF-IDF carried out on the entire term in each document. Let the weighting process using equation as in (2). TF-IDF weighting the results shown in Figure 4.

idf (q)	W (tf-idf)									
	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
0	0.15	0.15	0	0	0	0.15	0.15	0.77	0.46	0.46
0.222	0.44	0	0	0	0	0	0.22	0.22	1.33	0.22
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
-0.04	-0	-0.2	-0.2	-0.21	-0.08	-0.4	-0.21	-0.6	-0.2	-0.3
-0.04	-0	-0	-0	-0.04	-0.04	-0	-0.04	-0	-0	-0
0	2.41	1.81	0	0	0.6	0.3	0	0	3.61	0
0.301	2.11	1.51	0	0	0.6	0	0	0	3.31	0
0	0.9	0	0	0	0	0.3	0.3	0.3	0	0.3
0.097	0.29	0.19	1.16	0	0	0.68	0.19	0.29	0	0.1

Fig. 4. Result of TF-IDF Weighting.

C. Similarity Measure Using Cosine Similarity

After each test document has been weighted, the next is to calculate vector lengths of each document. In this study vector length calculation performed on 10 test documents using formula as in (4) and document query (q) using the formula as in (5).

$$\boldsymbol{D_i} = \sqrt{\sum(\boldsymbol{wd_i})^2} \quad (4)$$

$$\boldsymbol{Q} = \sqrt{\sum(\boldsymbol{q})^2} \quad (5)$$

Vector calculation results from each document shown in Figure 5.

Document	Vector Document
|Q|	8.30611
|D1|	6.03276
|D2|	4.88798
|D3|	13.40529
|D4|	4.88031
|D5|	3.02044
|D6|	7.00473
|D7|	3.46447
|D8|	5.62824
|D9|	7.71192
|D10|	3.52832

Fig. 5. Vector of document results.

Cosine similarity do a comparison between the weight of a key term in the document with the query term weights contained in a document. The concept of similarity as shown based on the formula (3) and then for this assessment, A and B are the expert document is a document of the students. Measuring the level of similarity performed on each test document.

$$\boldsymbol{Cosim\ d_i} = \frac{\boldsymbol{q}\times\boldsymbol{d_i}}{|\boldsymbol{Q}|\times|\boldsymbol{D_i}|} = \frac{\sum_{i=1}^{n}\boldsymbol{Wq}\times\boldsymbol{Wd_i}}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{q})^2}\times\sqrt{\sum_{i=1}^{n}(\boldsymbol{d_i})^2}} \quad (6)$$

Based on calculation of equation as in (6), document similarity ranking level of closeness is shown in Figure 6.

Rank	Document	Cosim	Degree
1	d10	0.39026	67°
2	d8	0.22216	77°
3	d4	0.20867	78°
4	d2	0.19146	79°
5	d7	0.17413	80°
6	d9	0.16754	81°
7	d1	0.14895	82°
8	d5	0.12495	83°
9	d6	0.08231	85°
10	d3	0.07784	86°

Fig. 6. Ranking of *Cosine Similarity* documents.

Based on the result that delivered on Figure 6, each document appears based on similarity rank. Whole documents result sorted by document that high similarity level with document expert. Document on the first rank have cosine similarity (cosim) value more high than other document because the cosim value is closeness of 1, where the 1 is value of document expert.

The result that showed on Figure 6 tells that the document with the first rank have low degree value than others. But document with the first rank have cosine similarity (cosim) value more high that others. Based on cosine similarity principle, when the document vector has similarity closeness to 1, so that closeness to be similar.

Based on the result that shown on Figure 6, we try to illustrate the similarity compares level of documents on Figure 7. The degree represented by arrows, document vector represented by lines. Each test documents notes with a number (d1 until d10) and expert document with (q) notation. The angle between expert document vector and test document vector is similarity measure level. For horizontal arrow shows the cosine degree 0˚ and there is a line notation that mean the expert document with cosine degree 0˚.

Based on cosine table [16], cosine degree 0˚ is 1 and the expert document is a reference to be comparing to other test documents. Thus, when there is a test documents that has cosine degree 0˚ and it means that the angle is 1, so that could be concluded similar.

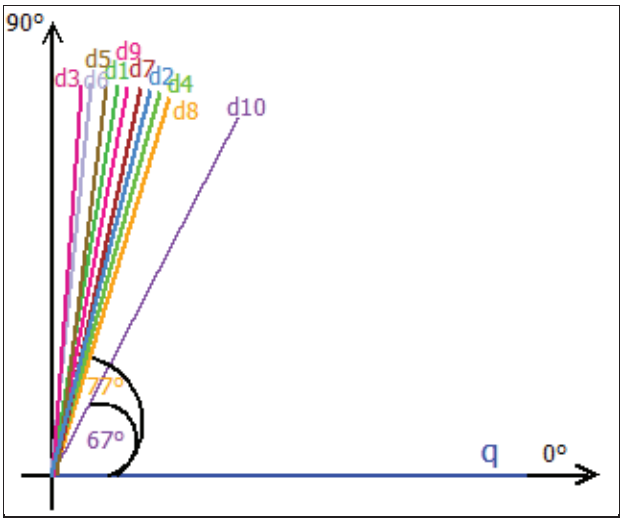

Fig. 7. Illustrations similarity degree of documents.

V. CONCLUSION

This paper describes the use of method for similarity measure using cosine similarity. First, we get extraction of feature from document by pre-processing steps, then calculate the terms weight using TF-IDF method and shows the similarity measure level in rank-based form. By using the similarity principle, the introduction of the character of the relevant text of expert document to test document will provide the assessment results more objective and to accelerate the process of correction in text-based assignment essay category.

For the future woks, conversion from weighting values of document to score values will be defined to improve scoring process from AES system.

REFERENCES

[1] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt, "Learning Semantic Similarity for Very Short Texts," 2015.

[2] Y. Li and Y. Yan, "Automated essay scoring system for CET4," 2010.

[3] W. H. Gomaa, "A Survey of Text Similarity Approaches," vol. 68, no. 13, pp. 13–18, 2013.

[4] K. P. N. V Satya and J. V. R. Murthy, "CLUSTERING BASED ON COSINE SIMILARITY MEASURE," no. 3, pp. 508–512, 2012.

[5] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[6] M. Berland, R. S. Baker, and P. Blikstein, "Educational Data Mining and Learning Analytics: Applications to Constructionist Research," *Technol. Knowl. Learn.*, vol. 19, no. 1–2, pp. 205–220, May 2014.

[7] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 40, 2010.

[8] R. Feldman and J. Sanger, "The Text Mining Handbook," 2006.

[9] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," vol. 5, no. 1, pp. 7–16.

[10] Y. F. A, "Algoritma Stemmer PAICE/HUSK dalam Bahasa Indonesia untuk Pre-processing Text Mining," 2010.

[11] R. V. Imbar, M. Ayub, A. Rehatta, S. Jurusan, S. Informasi, S. Jurusan, and T. Informatika, "Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks," pp. 31–42.

[12] J. Ramos, J. Eden, and R. Edu, "Using TF-IDF to Determine Word Relevance in Document Queries."

[13] W. S. J. Saputra and F. Muttaqin, "PENGENALAN KARAKTER PADA PROSES DIGITALISASI," no. September, pp. 51–56, 2013.

[14] Y. Li, "An effective automated essay scoring system using support vector regression," 2012.

[15] L. Bin, L. Jun, Y. Jian-min, and Z. Qiao-ming, "Automated Essay Scoring Using the KNN Algorithm," pp. 735–738, 2008.

[16] "TABEL SINUS , COSINUS , DAN TANGEN TABEL SINUS , COSINUS , DAN TANGEN," pp. 1–2.