Journal of Theoretical and Applied Information Technology
20th September 2015. Vol.79. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IDENTIFYING TEXT DOCUMENT PATTERN FOR TWO TERMS APPEARANCES VIA LATENT SEMANTIC ANALYSIS (LSA) METHOD AND TERM DISTANCE BETWEEN TWO DOCUMENTS

[1]SOEHARDJOEPRI, [2]NUR IRIAWAN, [3]BRODJOL SUTIJO SU, [4]IRHAMAH

[1]PhD student in Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Assoc. Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[4]Assoc. Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: [1]joepri@matematika.its.ac.id, [2]nur_i@statistika.its.ac.id, [3]brojol_su@statistika.its.ac.id, [4]irhamah@ statistika.its.ac.id

ABSTRACT

Pattern of writing a document is often said to be strongly influenced by the mother tongue, but do not guarantee produce writing that is always the similar pattern. If the trend similarity patterns caused by intentional copying of documents, then it is necessary to be created a detection tool to identify the terms pattern in those documents. This phenomena initiate this paper to acquaint the further investigation on text document pattern recognition for terms appearances by employing latent semantic analysis (LSA) method couple with terms distance between two documents. This study also describes determination of text documents similarity, which in turn can be used for early plagiarism detection.

Keywords: *Text Document, Pattern, Latent Semantic Analysis, Term Distance, Plagiarism*.

1. INTRODUCTION

Nowadays, technology especially information technology has concentrated on their impressive development and expansion. The development in this mean fairly meet people's needs and lead to rising new branch of science, namely information retrieval. This development can have positive and negative impact. Technology of information retrieval could be as one of the positive impact, which makes one easily find, see, and learn documents.

Communication supposed greatly influenced by someone mother tongue, that is assumed also greatly affect the pattern of writing a document. How can one recognize the pattern of a text document? Patterns identification of text documents have been conducted by Soehardjoepri, et al. [1] that searches the first order terms in each text document obtained by the emergence of the terms. This paper interested in continuing this identification by developing the identification of the pattern of the first and second order terms in each text documents by recognizing the emergence of these two first order terms. The work of the pattern detection would be demonstrated in this paper by looking at the emergence of the first two terms of each text document using latent semantic analysis (LSA) method couples with the distance between the two documents. This study also describes the determination of similarity of text documents, which in turn can be used to detect early plagiarism.

2. LATENT SEMANTIC ANALYSIS (LSA) METHOD

LSA is a well-known method for detecting document similarity. This method does the detection by extracting and representing the contextual-usage meaning of words by statistical computations applied to large corpus of text [2]. In LSA method, on the other hand, semantic text can

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be determined by calculating semantic similarity which involves various sciences like language, syntax of computer programming, mathematical logic, and related domains of sciences. The semantic similarity is based on the terminological similarity which works in three major steps, namely text parsing, term frequency - inverse document frequency (TF-IDF) weighting and singular value decomposition (SVD) [3].

2.1 Text Parsing

Text parsing is a process which is carried out to make a more meaningful sentence by breaking a sentence into words or phrases. Parsing text is divided into three consecutive processes, namely tokenizing, filtering and stemming.

- **Tokenizing**

Tokenizing is the process of identifying the smallest units (tokens) of a sentence structure. The purpose of this tokenizing is to get the terms that will be indexed. The classification is done for text tokens separated by a space or enter into a document. As for some of the cases handled by tokenizing namely: 1. Handling Special Character, for making patterns using regular expresion. 2. Phrase, in addition to special characters, tokenizing also can handle multiple patterns of phrases such as names, places and adjectives. 3. Whitespace, this character is ignored by tokenizing and regarded as a token separator.

- **Filtering**

Filtering is a process in which tokens are obtained from tokenizing process will be selected from the tokens that are considered unimportant (stoplist). Stoplist is a word that often appear and can be ignored in the process of filtering

- **Stemming**

Stemming is a process that aims to take root word of the word increase or a single word of word formation. It reduces the number of different terms in the collection.

After parsing the text of sentences in the original document, it would be collected into two collections; words and separators between words like dot (.), comma (,), space and other separators.

A	:	matrix would be decomposed
U	:	orthogonal left singular matrix
S	:	diagonal orthogonal matrix
V	:	orthogonal right singular matrix
m	:	number of rows
n	:	number of columns

The collection of words would be tabled as terms dictionary remarking the order of appearance of the term in a
sentence.

2.2 Term Frequency-Inverse Document Frequency (TF-IDF) Weighting

TF-IDF is a calculation of weights that are often used in text mining. TF-IDF is a way to give weight to the relationship of a word (term) of the document. This method combines the two concepts for the calculation of weight namely frequency of occurrence of a word ($\boldsymbol{tf}$) and the inverse document

$\boldsymbol{W_{ij}}$	:	weight of said term to - ***j*** and documents to - ***i***
$\boldsymbol{tf_{ij}}$	:	number of occurrences of the word / term to - ***j*** in the document to - ***i***
$\boldsymbol{N}$	:	the number of all documents in the database
$\boldsymbol{n}$	:	the number of documents that contain the term to - ***j***

frequency ($\boldsymbol{idf}$) containing the word, as shown in equation (1).

$$\boldsymbol{W_{ij}} = \boldsymbol{tf_{ij}} \text{ x } \boldsymbol{idf}$$
$$= \boldsymbol{tf_{ij}} \text{ x } (\boldsymbol{log}\ (\boldsymbol{N/n}) + \mathbf{1}) \quad (1)$$

wherein:

2.3 Singular Value Decomposition (SVD)

Text parsing to such document produces a collection of indexed terms counted as frequency that can be written in a matrix, called matrix **A**. The next stage, this matrix will be decomposed using the Singular Value Decomposition (SVD) into three matrices, namely orthogonal matrices **U** and **V,** and one diagonal orthogonal matrix **S**. The orthogonal matrix **U** forms a left singular matrix, obtained by multiplying the matrix **A** with its transpose. Whereas, the orthogonal matrix **V** forms a right singular matrix, obtained by multiplying the transpose matrix **A** with matrix **A**. These two orthogonal matrices **U** and **V** then are calculated their eigenvectors. While the diagonal orthogonal matrix **S** is formed by taking the roots of eigenvalues of matrix **A** multiplied by the transpose of matrix **A**, or vice versa. These three new construction matrices are multiplied as a re-construction of matrix **A** [2], as shown in equation (2).

$$\mathbf{A}_{mn} = \mathbf{U}_{mn} \text{ x } \mathbf{S}_{nn} \text{ x } \mathbf{V}_{nn}^{T} \quad (2)$$

wherein:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. DISTANCE BETWEEN TWO TERMS

Couple of term and frequency of occurrence will be used to form a point which can be plotted in three-dimensional Cartesian coordinates (the axis namely: *X*, *Y* and *Z*). The 1st and the 2nd terms of each sentence in each document will be put in *X* and *Y* axis, while their frequency of occurrences will be put in *Z* axis. The closer the two point, represent the more similar of two sentences from two documents. This closeness can be represented by a distance of these two points, which is calculated as the lengthwise of line that connects three dimensionally end to end of points. To do this, the euclidean distance as in equation (3) can be employed [4].

$$|PQ| = \sqrt{\left(X_P - X_Q\right)^2 + \left(Y_P - Y_Q\right)^2 + \left(Z_P - Z_Q\right)^2}, \quad (3)$$

where $|PQ|$ is the distance between point *P* in (X_P, Y_P, Z_P) and point *Q* in (X_Q, Y_Q, Z_Q).

4. THE PROPOSED PATTERN IDENTIFICATION ALGORITHM

In previous studies, Ozgur & Gungor [5] proposed a novel text classification approach based on two main concepts, lexical dependency and trimming. They extend the standard bag-of-words method by incorporating the feature vector pattern dependence. They analyzed the effect of pruning (filtering features with low frequency) for both feature words and dependency features. Parameter tuning is done by eight different levels of pruning to determine the optimal level. The experiment was repeated on three datasets with different characteristics. They observed a significant increase in the level of success as well as a reduction in the dimension of feature vectors. They argue that, in contrast to the works in the literature, the higher the level of pruning should be used in text classification. By analyzing the results from the perspective of the dataset, they also indicate that the dataset at the same level of formality has the same leading dependency and demonstrate that the behavior was closely with various levels of pruning.

In another study, some researchers worked on linguistic pattern. Some linguistic features, e.g. part-of-speech information, complex nominal, proper nouns, and word senses, were considered in addition to the words, but no significant improvement was observed [6]. Later, by referring to the negative effect of a specific dependency (subject-object-verb), Moschitti [7] mentioned that linguistic processing does not improve the bow approach in general. A related study extracted dependencies by capturing frequently occurring keyword combinations within short segments using a rule-based algorithm, has been done by Ghanem et al. [8]. This algorithm yielded successful results, but this experiment was done only on a specific and not widely used dataset. Another study increased the success rates of the classifier by accompanying the bow approach with a combination of noun modifier dependencies and word senses [9].

Text document identification pattern looks has not previously been studied. This paper demonstrates the proposed pattern identification by using LSA couple with distance between pair terms. Pattern identification is mainly done using the emergence of two consecutive terms, namely the 1st and 2nd terms of each sentence in each document. There are four sequential algorithms to identify the pattern, called Dictionary of Terms, Pair Term, Measure Distance and Similarity Reference Document. Each of the four proposed algorithms is as follows:

Algorithm 1: Dictionary of Terms

LSA above produces the terms are sorted by the number of occurrence of the term in each document. This algorithm will create a dictionary of terms obtained from the LSA. The Construction of this dictionary is very important in identifying patterns of text documents. There are five steps to construct the dictionary of terms in this algorithms, which can be described as follows:

Step 1. Take *K* documents which are being tested.
Step 2. Identify and named each document orderly, namely Doc-1 to Doc-*K*.
Step 3. Extract all *K* documents using LSA to obtain terms contained in each sentences on every document.
Step 4. Arranged these terms according to sentences and document orders come from, then coding each term by using T_i with i = 1, 2, 3, …, n
Step 5. Put each terms and their coding in the Term Dictionary table.

Algorithm 2: Term Pair

The consecutive appearance of terms in such sentence are set as a term pair. These pair are mapped to the dictionary to construct the table of term pair. There are three steps to construct these table of term pair, which can be described as follows:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 1. Set 1st and 2nd term of every sentence of each document as a term pair.
Step 2. Tabulate consecutive term partner based on the appearance of sentences in each document in the table term partner.

Algorithm 3: Measuring Distance

Each of term pair obtained in Algorithm 2 above and their frequency appearance would be described in the 3-dimensional Cartesian coordinates (*X*, *Y* and *Z*). The 1st term of each sentence in every document are plotted at *X*-axis and 2nd term in *Y*-axis. Assume that the appearance of terms in each sentence of each document is incremented by 1 unit of distance in *X*-axis and *Y*-axis. Frequency of term pair appearance is plotted at *Z*-axis which is calculated from zero. Two documents which are compared the pattern are plotted together in the same plot. The points of each document term pair of these two documents are connected and calculated their distance. The five steps in the following algorithm would show how to calculate this distance. These steps can be described as follows:

Step 1. The shape of each pair of terms in each document and the frequency of occurrence for forming a point represented in the 3-dimensional Cartesian coordinate system (*X*, *Y*, and *Z* axis).
Step 2. Locate the 1st term of each sentence in every document at *X*-axis.
Step 3. Locate the 2nd term of each sentence in every document at *Y*-axis.
Step 4. Locate frequency of term pair of each sentence in every document at *Z*-axis.
Step 5. Calculate the distance of terms (points) that is formed by two sentences from two documents to describe the similarity of the two sentences from two documents.

Algorithm 4: The Reference Pattern Similarity of Documents

All three algorithms were compiled together to build the algorithm of reference pattern similarity of documents. This algorithm would help to identify and make decision about the pattern similarity of two documents. There are two steps to construct the reference pattern similarity algorithm of documents. These steps can be described as follows:

Step 1. Tabulate all distances from each point of the two documents.
Step 2. Determine the maximum distance (d_j, j = 1, 2,..., n), write as (D = maximum distance of (d_j)). If D is not more than 3, then the two documents have the same pattern, implying that the two documents have the same pattern.

5. IMPLEMENTATION

This section discusses the implementation of the proposed four consecutive algorithms above are applied to the Indonesian text documents. Table 1 lists six text documents as data test which are designed with a special structure and content.

Applying Algorithm 1, each document in Table 1 is parsed by employing LSA and their resulted terms are listed in Table 2 and Table 3. Term pair which are containing 1st and 2nd terms of each sentence in each document can be found by implementing Algorithm 2. Table 4A, 4B, 4C, 4D, 4E and 4F show term pair of tested six documents. Pattern comparison as the next step is then done by implementing Algorithm 3 which calculate the distance between the coordinate points for each pair of terms in each document.

The most possible pattern comparison is firstly done between Doc-1 and Doc-2, because Doc-2 is previously designed the same as Doc-1 with some changes in the order of sentences in a few paragraphs, but they have the same meaning. Secondly, comparison have been done between Doc-1 and Doc-5. This comparison could also represents directly the pattern similarities between Doc-2 and Doc-5. Thirdly, comparison between Doc-1 and Doc-6 have to be performed. Finally, Doc-3 and Doc-4 are checked to see their pattern similarity. All comparison results are tabulated in Table 5.

Finally, the last step to decide the document similarity between two documents is by applying Algorithm 4 which measures the distance between two text documents. The result can be inferred that two pair documents between doc-1 and doc-2, doc-3 and doc-4, and doc-1 and doc-5 have the same pattern. It is primarily shown by the distances among all points of these pair documents are less than 3. In contrast, distances of all points in pair of documents doc-1 and doc-6 are greater than 3. It means that doc-1 and doc-6 have different pattern, even though some sentences are same.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. CONCLUSIONS

All four algorithms have succeed to demonstrate the pattern similarity identification. Latent Semantic Analysis (LSA) method couple with term distance calculation building these four algorithms can be used to identify the pair text documents pattern similarity.

7. ACKNOWLEDGEMENT

The authors are grateful to BPPDN-DIKTI which has provided scholarships for our doctoral program and the Department of Statistics Institut Teknologi Sepuluh Nopember (ITS) which provide the facility.

APPENDIX

Table 1. Six text documents tested for identifying the pattern

No	Documents	The structure and content of documents
1	Doc-1	the original document contains words and sentences into paragraphs.
2	Doc-2	the same document as Doc-1 with some changes in the order of sentences in some paragraphs but have the same meaning.
3	Doc-3	the same document as Doc-1 with some changes in the order and structure of words in some paragraphs.
4	Doc-4	the same document as Doc-3 with some changes on some words with their synonym.
5	Doc-5	merged of Doc-1 and Doc-4.
6	Doc-6	a new document which is different with Doc-1, Doc-2, Doc-3, Doc-4, and Doc-5.

Tabel 2. List of terms in each text document after parsing by LSA

No	Document	Terms (in Indonesian)
1	Doc-1 & Doc-2	ruang, bunyi, gema, cegah, gedung, serap, keras, langit, dinding, bahan, akustik.
2	Doc-3 & Doc-4	musik, budaya, lampung, festival, adakan, hingga, tradisional, daerah, akustik.
3	Doc-5	ruang, bunyi, gema, cegah, gedung, serap, keras, langit, dinding, musik, budaya, lampung, festival, adakan, hingga, tradisional, daerah, bahan, akustik
4	Doc-6	negosiasi, dasar, ikut, kantor, rapat, orang, lebih, manusia, teknik, kalangan, bangun

Table 3. Term Dictionary (in Indonesian)

No term	Term	No term	Term	No term	Term	No term	Term
T1	Ruang	T10	Musik	T19	Akustik	T28	Teknik
T2	Bunyi	T11	Budaya	T20	Negosiasi	T29	Kalangan
T3	Gema	T12	Lampung	T21	Dasar	T30	Bangun
T4	Cegah	T13	Festival	T22	Ikut		
T5	Gedung	T14	Adakan	T23	Kantor		
T6	Serap	T15	Hingga	T24	Rapat		
T7	Keras	T16	Tradisional	T25	Orang		
T8	Langit	T17	Daerah	T26	Lebih		
T9	Dinding	T18	Bahan	T27	Manusia		

Table 4A. Term Pair (Doc-1)

Doc-1	Term	
Sentence	1st	2nd
1	T19	T2
2	T9	T7
3	T1	T4
4	T1	T5

Table 4B. Term Pair (Doc-2)

Doc-2	Term	
Sentence	1st	2nd
1	T1	T4
2	T1	T5
3	T19	T2
4	T9	T7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4C. Term Pair (Doc-3)

Doc-3	Term	
Sentence	1st	2nd
1	T17	T12
2	T10	T15
3	T10	T10
4	T10	T11
5	T13	T14
6	T13	T13

Table 4D. Term Pair (Doc-4)

Doc-4	Term	
Sentence	1st	2nd
1	T10	T17
2	T12	T10
3	T10	T19
4	T10	T11
5	T11	T10
6	T12	T14

Table 4E. Term Pair (Doc-5)

Doc-5	Term	
Sentence	1st	2nd
1	T10	T17
2	T12	T10
3	T10	T19
4	T10	T11
5	T11	T10
6	T12	T14
7	T19	T2
8	T9	T7
9	T1	T4
10	T1	T5

Table 4F. Term Pair (Doc-6)

Doc-6	Term	
Sentence	1st	2nd
1	T30	T26
2	T23	T20
3	T22	T22
4	T20	T27
5	T25	T29
6	T20	T23
7	T21	T28

Table 5A. The distance (d) from Doc-1 and Doc-2

Doc-1			Doc-2			d
X	Y	Z	X	Y	Z	
1	1	0	3	3	0	2,83
1	1	1	3	3	1	2,83
2	2	1	4	4	1	2,83
2	2	2	4	4	2	2,83
3	3	2	1	1	2	2,83
3	3	3	1	1	3	2,83
3	4	3	2	2	3	2,24
3	4	4	2	2	4	2,24

Doc-2			Doc-1			d
X	Y	Z	X	Y	Z	
1	1	0	2	3	0	2,24
1	1	1	2	3	1	2,24
1	2	1	3	4	1	2,83
1	2	2	3	4	2	2,83
2	3	2	1	1	2	2,24
2	3	3	1	1	3	2,24
3	4	3	1	2	3	2,83
3	4	4	1	2	4	2,83

Table 5B. The distance (d) from document-3 and document-4

Doc-3			Doc-4			d
X	Y	Z	X	Y	Z	
1	1	0	2	7	0	6,08
1	1	1	2	7	1	6,08
2	2	1	4	3	1	2,24
2	2	2	4	3	2	2,24
2	3	2	2	8	2	5,00
2	3	3	2	8	3	5,00
2	4	3	2	4	3	-
2	4	4	2	4	4	-
3	5	4	5	3	4	2,83
3	5	5	5	3	5	2,83
4	5	5	5	4	5	1,41
4	5	6	5	4	6	1,41

Doc-4			Doc-3			d
X	Y	Z	X	Y	Z	
1	1	0	4	6	0	5,83
1	1	1	4	6	1	5,83
2	2	1	1	7	1	5,10
2	2	2	1	7	2	5,10
1	3	2	1	2	2	1,00
1	3	3	1	2	3	1,00
1	4	3	1	4	3	-
1	4	4	1	4	4	-
3	2	4	5	5	4	3,61
3	2	5	5	5	5	3,61
2	5	5	5	8	5	4,24
2	5	6	5	8	6	4,24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Tabel 5C. The distance (d) from document-1 and document-5

Doc-5			Doc-1			d	Doc-1			Doc-5			d
X	Y	Z	X	Y	Z		X	Y	Z	X	Y	Z	
1	1	0	4	6	0	5,83	1	1	0	4	5	0	5,00
1	1	1	4	6	1	5,83	1	1	1	4	5	1	5,00
2	2	1	5	7	1	5,83	2	2	1	5	6	1	5,00
2	2	2	5	7	2	5,83	2	2	2	5	6	2	5,00
1	3	2	6	8	2	7,07	3	3	2	4	7	2	4,12
1	3	3	6	8	3	7,07	3	3	3	4	7	3	4,12
1	4	3	6	9	3	7,07	3	4	3	4	8	3	4,12
1	4	4	6	9	4	7,07	3	4	4	4	8	4	4,12
3	2	4	6	9	4	7,62	3	4	4	6	6	4	3,61
3	2	5	6	9	5	7,62	3	4	5	6	6	5	3,61
2	5	5	6	9	5	5,66	3	4	5	5	9	5	5,39
2	5	6	6	9	6	5,66	3	4	6	5	9	6	5,39
4	6	6	6	9	6	3,61	3	4	6	1	1	6	3,61
4	6	7	6	9	7	3,61	3	4	7	1	1	7	3,61
5	7	7	6	9	7	2,24	3	4	7	2	2	7	2,24
5	7	8	6	9	8	2,24	3	4	8	2	2	8	2,24
6	8	8	6	9	8	1,00	3	4	8	3	3	8	1,00
6	8	9	6	9	9	1,00	3	4	9	3	3	9	1,00
6	9	9	6	9	9	-	3	4	9	3	4	9	-
6	9	10	6	9	10	-	3	4	10	3	4	10	-

Table 5 D. The distance (d) from document-1 and document-6

Doc-6			Doc-1			d	Doc-1			Doc-6			d
X	Y	Z	X	Y	Z		X	Y	Z	X	Y	Z	
1	1	0	7	8	0	9,22	1	1	0	4	5	0	5,00
1	1	1	7	8	1	9,22	1	1	1	4	5	1	5,00
2	2	1	8	9	1	9,22	2	2	1	5	6	1	5,00
2	2	2	8	9	2	9,22	2	2	2	5	6	2	5,00
3	3	2	9	10	2	9,22	3	3	2	6	7	2	5,00
3	3	3	9	10	3	9,22	3	3	3	6	7	3	5,00
4	4	3	9	11	3	8,60	3	4	3	7	8	3	5,66
4	4	4	9	11	4	8,60	3	4	4	7	8	4	5,66
5	5	4	9	11	4	7,21	3	4	4	8	9	4	7,07
5	5	5	9	11	4	7,28	3	4	4	8	9	5	7,14
4	6	5	9	11	4	7,14	3	4	4	7	10	5	7,28
4	6	6	9	11	4	7,35	3	4	4	7	10	5	7,28
6	7	6	9	11	4	5,39	3	4	4	9	11	6	9,43
6	7	7	9	11	4	5,83	3	4	4	9	11	7	9,70

REFERENCES:

[1] Soehardjoepri, Iriawan, N., Sutijo, S.U.B., and Irhamah, "On the Text Documents Pattern Recognition Using Latent Semantic Analysis and Kolmogorov-Smirnov Test", *South East Asian Conference on Mathematics and Its Applications*, Department of Mathematics, FMIPA-ITS, Surabaya-Indonesia, 2013.

[2] Landauer, T.K., Foltz, P.W., and Laham, D., "Introduction to Latent Semantic Analysis", *Discourse Processes*, 25 (1998), pp. 259 - 284.

[3] Ardiansyah, A., "Application Development Using Plagiarism Detector Method of Latent Semantic Analysis (LSA)", *Plagiarism Case Study of Scientific Speak Indonesian*, Universitas Pendidikan Indonesia, Bandung, 2011.

[4] Dattorro, "Convex Optimization & Euclidean Distance Geometry", Mεβoo Publishing USA, 2005, v2014.04.08.

[5] Ozgur, L., and Gungor, T., "Text classification with the support of pruned dependency patterns", Elsevier B.V, (2010), 1598–1607, doi:10.1016/j.patrec.2010.05.005.

[6] Moschitti, A., and Basili, R., "Complex linguistic features for text classification", *In: Proc. European Conf. on Information Retrieval* (ECIR 2004), (2004), pp. 181–196.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[7] Moschitti, A., "Kernel methods, syntax and semantics for relational text categorization", *In: Proc. ACM 17th Conf. on Information and Knowledge Management (CIKM)*, Napa Valley, California, 2008.

[8] Ghanem, M., Guo, Y., Lodhi, H., and Zhang, Y., "Automatic scientific text classification using local patterns", ACM SIGKDD Explore News. 4 (2), (2002), pp. 95–96.

[9] Nastase, V., Shirabad, J.S., and Caropreso, M.F., "Using dependency relations for text classification", *In: Proc. Nineteenth Canadian Conf. on Artificial Intelligence*, Quebec, Canada, 2006.

[10] Konig, A.C., and Brill, E., "Reducing the human overhead in text categorization", *In: Proc. 12th ACM SIGKDD Internet Conf. Knowledge Discovery and Data Mining*, Philadelphia, USA, (2006), pp. 598–603.

[11] Chakravart, I.M., Laha, R.G., and Roy J., "Handbook of methods of applied statistics", Volume 1, Wiley, 1967.

[12] Dumais, S.T., "Latent Semantic Analysis", Annual Review of Information Science and Technology 38: 188 (2005), doi: 10.1002/aris.1440380105.

[13] Peacock, J.A., "Two-dimensional goodness-of-fit testing in astronomy", Monthly Notices Royal Astronomy Society 202, (1983), pp. 615 - 627.

[14] Hansen, P. C., "The truncated SVD as a Amethod for regularization", BIT 27 (1987), pp. 534 - 553.

[15] Law, A. M., and Kelton, W. D., "Simulation Modeling and Analysis, McGraw-Hill International Series", Singapore, 2000.

[16] Beeferman, D., Berger, A., and Lafferty, J., "Statistical models for text segmentation, Machine Learning", *Special Issue on Natural Language Learning* 34 (1999), (1-3): pp. 177-210.

[17] Deerwester, S., Dumais, S.T., Landauer, T., Furnas, G. and Harshman, R, "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science* 41 (1990), (6): 391- 407.

[18] Kasim, S., "Making Application Method to Detect Plagiarism with Latent Semantic Analysis", *Final, Department of Informatics*, University of Surabaya, 2012.

[19] Lopes, R.H.C., Reid, I., and Hobson, P.R., "The two-dimensional Kolmogorov-Smirnov test", *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*, Amsterdam, the Netherlands, April 23–27, 2007.

[20] Thomas L., Peter W. F., and Darrell L., "Introduction to Latent Semantic Analysis", Discourse Processes 25 (1998), 259-284, doi: 10.1080/01638539809545028.

[21] Kolmogorov, A., "On Tables of Random Numbers", *Theoretical Computer Science*, (1998), 387–395, doi: 10.1016/S0304-3975(98)00075-9.

[22] Dag, J.S., Dag, B.T., and Nils, G.K., "A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality", *Monthly Weather Review*, 135, (2007), pp. 1151–1157, doi: http://dx.doi.org/10.1175/MWR3326.1.

[23] Emmert-Streib F, "Statistic Complexity: Combining Kolmogorov Complexity with an Ensemble Approach", PLoS ONE 5(8): e12256, (2010), doi: 10.1371/ journal.pone.0012256.

[24] Song-Hee K. and Ward W., "the Power of Alternative Kolmogorov-Smirnov Tests Based on Transformations of the Data". *ACM Trans. Model. Comput. Simul*. V, N, Article A, (2013), 1-20, doi: http://dx.doi.org/10.1145/0000000.0000000.

[25] Talib, "Making Program Application for Text Document Similarity Detection with the Smith-Waterman algorithm", Gunadarma University, Depok-Indonesia, 2010.

[26] Yoga, "Detection Application Development Plagiarism in Text Documents Using Rabin-Karp algorithm", Vol.1, No. 4, KARMAPATI, 2012.

[27] Hendri W., "Text Document Content Similarity Detection Algorithm Using Levenshtein Distance", State Islamic University of Maulana Malik Ibrahim, Malang, 2012.