

ON THE IDENTIFICATION OF THE STRUCTURAL PATTERN OF TERMS OCCURRENCE IN A DOCUMENT USING BAYESIAN NETWORK

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

¹PhD student in Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

³Assoc. Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

⁴Assoc. Prof., Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

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

ABSTRACT

The pattern of text documents is strongly influenced by the advent of the first term in composing term structure of each sentence. When two documents have the same pattern, then the second and the following terms tend to be same. This paper would create a special tool for detecting the similarity of structural pattern of two text documents. Latent Semantic Analysis(LSA) couples with Bayesian Network (BN) are employed as the main engine to build the algorithms. The work of these approaches is demonstrated to detect the similarities of the appearance of the term in the sentence in any text documents.

Keywords: *Text Pattern Document, Term, Latent Semantic Analysis, Bayesian Network.*

1. INTRODUCTION

Current information technology has been concentrating very impressively on the development and expansion of text analysis. Development meets the needs of society and led to increase new branch of science including information retrieval. This development can have positive and negative effects. Information search technology could be one positive effect, which makes people easily find, view, and learn about a document.

Communication is primarily influenced by the mother tongue, which is assumed to affect the pattern of written document. How can one recognize a pattern of text documents? Soehardjoepri, et al. [1] have succeeded to develop text pattern identification in order to find the two first order terms in any text document. The distance of these two first order terms between any of two documents are calculated to measure their similarities.

This paper demonstrates another approach in detecting the similarity of two documents. The first three terms from LSA are constructed as a BN structure and applying the likelihood principle to measure the similarities.

2. Latent Semantic Analysis(LSA)

LSA has a significant contribution in detecting the document similarity. The capability to extract and represent the contextual meaning to the word-usage statistics of such document can be applied to a large corpus of text [2].

LSA involves two main stages, namely Parsing Text and Singular Value Decomposition (SVD). Parsing text, which breaks the sentence into terms by ignoring period (.), comma(,), space and other separators, will be employed in this study. The sequence of terms as a result of the LSA process to such master documents could be used to create a dictionary term [1]. The algorithm for constructing the dictionary is as follows:

Algorithm 1: Term order appearance

This algorithm would list the sequence of terms taken from the terms dictionary according to terms appearance in a sentence. The steps are as follows:

- Take the terms from the sequence and put them into the related sentences.
- Arrange the terms orderly follow the term appearance in each sentence.

- Characterized the term order appearance by the number, i.e.1, 2, 3, ...,n.

The results of this algorithm is a table arrangement term based on the appearance of the term in the sentence in each document.

3. Bayesian Network(BN)

Term order occurrence in term dictionary as a result of Algorithm 1 is defined as a BN structure. In this structure, every node represents the emergence of terms. Network of nodes represents the order of the emergence of term in a sentence. All terms in the first appearance can be calculated their probability among those first terms. All terms in the second appearance can be calculated their probability among those second terms. The probability of the following terms appearance could be calculated as the same way. Term order in such BN then its probability which represents the computational sequence of appearance of the first term until the last term in each sentence can be calculated.

Consider a BN containing with n nodes, namely node X_1 to node X_n , taken from the term dictionary, then the joint occurrence could be represented as $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ or $P(x_1, x_2, \dots, x_n)$. The order of terms in BN denotes the conditional sequence of terms in the sentence. The conditional probability concept, therefore, should be employed for calculating the probability of BN structure. The chain rule of probability theory allows us to factorize the joint probabilities to be

$$\begin{aligned}
 &P(x_1, x_2, \dots, x_n) \\
 &= P(x_1) \cdot P(x_2|x_1) \dots \dots P(x_n|x_1, \dots, x_{n-1}) \\
 &= \prod_i P(x_i|x_1, \dots, x_{i-1}).(1)
 \end{aligned}$$

The order of terms in a BN could follow the Markov processes. Applying the Markov property, therefore, the structure of BN in equation (1) implies that the conditional probability of a particular node is depending only on its parent nodes. The equation (1) could be written as $P(x_1, x_2, \dots, x_n) = \prod_i P(x_i | \text{Parents}(X_i))$,(2)

where $\text{Parents}(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$ [3].

Consider when n = 3 nodes, node X_1, X_2 , and X_3 . There could be three possibilities BN structure as shown in the graph on Figure 1. The joint distribution according to this graph will follow the model given in Chow and Liu [5] as an example of a (first-order) dependence tree or as a singly-connected DAG which is directly representing the

Markov properties. The joint distribution of each graph in Figure 1 can be written as $(x_1, x_2, x_3) = P(x_2|x_1) \cdot P(x_3|x_1) \cdot P(x_1)$ for Figure 1.(a), $P(x_1, x_2, x_3) = P(x_3|x_1, x_2) \cdot P(x_1) \cdot P(x_2)$ for Figure 1.(b), and $P(x_1, x_2, x_3) = P(x_3|x_2) \cdot P(x_2|x_1) \cdot P(x_1)$ for Figure 1.(c).

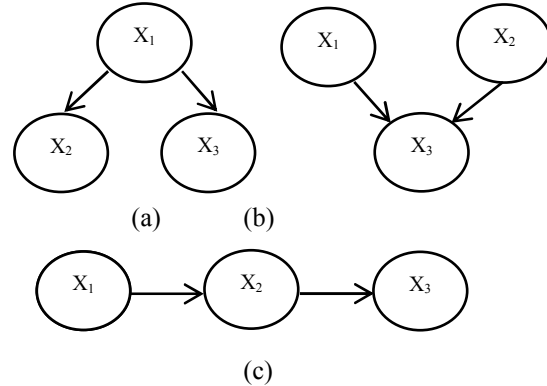
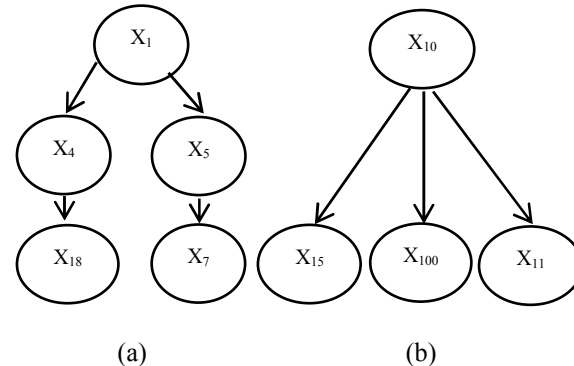


Figure 1. Directed Acyclic Graph (DAG) For Three Nodes

Suppose three sentences are used to express the emergence of a term taken from a dictionary term shown in Figure 2. X_i in Figure 1 will be changed by T_i according to the existence of term in each sentence on the dictionary. The joint distribution of each graph in Figure 2 can be written as $P(T_1, T_4, T_5, T_{18}, T_7) = [P(T_7|T_5) \cdot P(T_5|T_1)] \cdot [P(T_{18}|T_4) \cdot P(T_4|T_1)] \cdot P(T_1)$ for Figure 2.(a), $P(T_{10}, T_{15}, T_{10}, T_{11}) = P(T_{15}|T_{10}) \cdot P(T_{10}|T_{10}) \cdot P(T_{11}|T_{10}) \cdot P(T_{10})$ for Figure 2.(b), and $(T_{10}, T_{15}, T_{10}, T_{11}, T_{12}, T_{19}, T_{11}) = (T_{12}|T_{15}) \cdot P(T_{15}|T_{10}) \cdot P(T_{19}|T_{10}) \cdot P(T_{10}|T_{10}) \cdot P(T_{11}|T_{11}) \cdot P(T_{11}|T_{10}) \cdot P(T_{10})$ for Figure 2.(c).



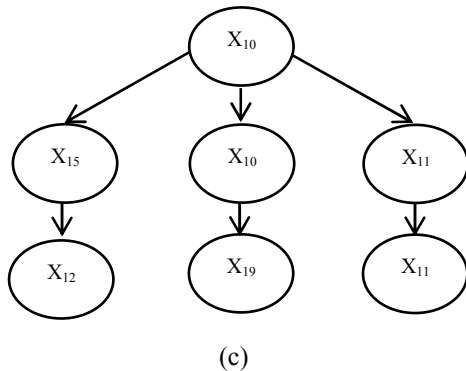


Figure 2. DAG for three sentences

Each sentence contains some terms. The first three terms in Indonesian are generally contains the main principal of a sentence (e.g. subject, predicate, and object). Thus, in this study, these first three terms of each sentence are taken to be processed as a pattern of sentences. Algorithm 2 shows the step in constructing a dictionary of the term emergence pattern of documents.

Algorithm 2: Term emergence patterns

This algorithm would build a dictionary term emergence patterns of a master document. The steps are as follows:

- Take all appropriate term in each sentence from the term order emergence using Algorithm 1.
- Arrange those all appropriate terms follow the structure of every sentence in each document as the term emergence pattern.
- Take the first three terms from the term emergence pattern in every sentences of each document.
- Construct the joint distribution of the first three terms emergence pattern in every sentence of each document.

The emergence of term in a sentence is always dependent on the appearance of the previous term. The series of emergence of term would construct a network of term. This network could be called as Bayesian Network due to the emergence of term series shows one come prior the other. Therefore, the series of term would represent the series of prior and posterior sequence which ensemble the Bayesian Network.

Algorithm 3 shows the steps to calculate the likelihood of the first three emergences as the joint distribution of the terms laid out in the sentence as a Bayesian Network.

Algorithm3: Likelihood of the First Three Term Emergences

This algorithm would explain how to calculate the joint probability distribution of the first three emergences of terms in every sentence as its likelihood. The steps are as follows:

- Take the first three emergences of terms from every sentence based on the term emergence pattern as representation of a sentence using Algorithm 2.
- Collect and calculate the individual term
 - Collect and calculate the frequency of the first emergences of terms from all sentences as a group then calculate the probability of each term amongst the first terms emergence in this group.
 - Collect the second and the third emergences of terms from all sentences and calculate their probabilities in and amongst their groups as done for the first term.
- Calculate the likelihood of each sentence by multiplying all probability of the first three emergences of terms in the sentence

The document contained some sentences, therefore, can be calculated its likelihood based on their sentence probability calculated by Algorithm 3. In addition, the principle of the sequence of sentences in such document can also be assumed as a Bayesian Network structure, but those sentences can be assumed to be independent. The probability of such pattern of document can be represented by probability of those sequences of sentences as representing in a likelihood of document calculated by multiplication of all likelihood of each sentences. When all of documents can be seen their likelihood, then they could be compared their pattern by calculating the likelihood ratio of each pair documents. The algorithm to calculate the likelihood ratio between two documents can be seen in Algorithm 4.

Algorithm4: Likelihood Ratio of Document

This algorithm would explain how to calculate the likelihood ratio between two documents. The steps are as follows:

- Calculate the likelihood of each sentence represented by the first three terms in each document using Algorithm 3.
- Calculate the likelihood of each document by multiplying the likelihood of each sentence in the related document.
- Calculate the likelihood ratio of two documents by dividing likelihood of the preference document with the indifferent document.



- Decide the comparison between these two documents by using their likelihood ratio based on Bayes Factor presented in Table 1 (Kass and Raftery, 1995).

Table 1: Likelihood Ratio

L_R	Conclusion
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
>150	Very strong

where L_R = Likelihood ratio of Doc-I and Doc-j ($j \neq i$), for $i = 1, \dots, k$ and k is the number of documents.

Algorithm 4 would be firstly applied to test the pattern similarity of five tested documents as has been used in Soehardjoepri, et.al. [1] and secondly to six documents, after one differently additional tested documents is added. The process is discussed in the following section.

4. Numerical Implementation

Five documents which has been used in Soehardjoepri, et.al. [1], are used in this section to show the work of this algorithm. Doc-1 and Doc-2 are previously designed to have an almost similar structure. Firstly, of the five documents will be peeled all their terms by using LSA as stated in Algorithm 1. The result of this algorithm is the emergence of terms in each sentence of each document, which are shown in Table 2.

Implementation of Algorithm 2 couple with the rule based of the first three terms in Indonesian sentences to those five documents having term sequence listed in Table 2, give the structure of pattern emergence terms as shown in the second column of Table 3 to Table 7. By applying the form of Bayesian Network, the probability of the structure of each sentence is given in the third column of those tables.

In order to find the probability of each sentence structure, Algorithm 3 could be applied. Calculation of the probability of each term in the structures has to be prepared by counting the frequency of each term occurrence which appears in every sentence of whole five documents. The frequency of each term occurrence in every sentence of five documents can be seen in Table 8. Based on this frequency, the probability of each term occurrence can be found by calculating the related term in each appearance, as stated in the second step of Algorithm 3. By employing Table 8 and inputting it to the last step of Algorithm 3, the probability of the first three term occurrence for

each sentence in five documents is shown in Table 9.

Comparing between two documents, e.g. Doc-1 and Doc-2 which are previously designed to have similar structure, can be done by calculating their likelihood ratios. The likelihood of each document can be calculated by employing Table 9 and inputting it to Algorithm 4. Supposing that the Doc-1 is more preferable than Doc-2, the likelihood ratio can be found by dividing likelihood of Doc-1 with likelihood of Doc-2. Where based on Table 9, the likelihood of Doc-1 and Doc-2 are 0.1×10^{-10} respectively. Therefore, the likelihood ratio is 1 and do not reject the null hypothesis (e.g. Doc-1 and Doc-2 have similar structure).

Comparison between other pair ways documents can be done in the same way, and their likelihood ratios can be seen in Table 10. Table 1 can be used to make the interpretation of these likelihood ratios. It can be seen that only Doc-1 and Doc-2 shows the closest similarity pattern, due to their perfect likelihood ratio ($L_R = 1$). Other likelihood ratios show that the null hypothesis has to be rejected and those documents have significantly different pattern [6].

Table2: The Emergence Of terms in each Sentence of Each Document (Doc)

Docu ment	Senten ce	Occurrence order									
		1	2	3	4	5	6	7	8	9	10
Doc-1	1	T ₁₉	T ₂	T ₁							
	2	T ₉	T ₇	T ₁	T ₂	T ₁	T ₃				
	3	T ₁	T ₄	T ₁₈	T ₉	T ₈	T ₆	T ₂			
	4	T ₁	T ₅	T ₇	T ₂	T ₁₈	T ₆	T ₂	T ₅	T ₄	T ₃
Doc-2	1	T ₁	T ₄	T ₁₈	T ₉	T ₈	T ₆	T ₂			
	2	T ₁	T ₅	T ₇	T ₂	T ₁₈	T ₆	T ₂	T ₅	T ₄	T ₃
	3	T ₁₉	T ₂	T ₁							
	4	T ₉	T ₇	T ₁	T ₂	T ₁	T ₃				
Doc-3	1	T ₁₇	T ₁₂	T ₁₀	T ₁₆	T ₁₅	T ₁₁	T ₁₀			
	2	T ₁₀	T ₁₅	T ₁₂							
	3	T ₁₀	T ₁₀	T ₁₉							
	4	T ₁₀	T ₁₁	T ₁₁							
	5	T ₁₃	T ₁₄	T ₁₁	T ₁₀	T ₁₆					
	6	T ₁₃	T ₁₃	T ₁₄	T ₁₂	T ₁₂					
Doc-4	1	T ₁₀	T ₁₇	T ₁₂	T ₁₇	T ₁₆	T ₁₅	T ₁₁	T ₁₀		
	2	T ₁₂	T ₁₀	T ₁₅							
	3	T ₁₀	T ₁₉	T ₁₀							
	4	T ₁₀	T ₁₁	T ₁₁							
	5	T ₁₁	T ₁₀	T ₁₆	T ₁₄	T ₁₃					
	6	T ₁₂	T ₁₄	T ₁₃	T ₁₂	T ₁₃					
Doc-5	1	T ₁₀	T ₁₇	T ₁₂	T ₁₇	T ₁₆	T ₁₅	T ₁₁	T ₁₀		
	2	T ₁₂	T ₁₀	T ₁₅							
	3	T ₁₀	T ₁₉	T ₁₀							
	4	T ₁₀	T ₁₁	T ₁₁							
	5	T ₁₁	T ₁₀	T ₁₆	T ₁₄	T ₁₃					
	6	T ₁₂	T ₁₄	T ₁₃	T ₁₂	T ₁₃					
	7	T ₁₉	T ₁₂	T ₁							
	8	T ₉	T ₇	T ₁	T ₂	T ₁	T ₃				
	9	T ₁	T ₄	T ₁₈	T ₉	T ₈	T ₈	T ₆	T ₂		
	10	T ₁	T ₅	T ₇	T ₂	T ₁₈					

Table 3: Structure Of The Term Pattern Emergence And Its Probability For Doc-1

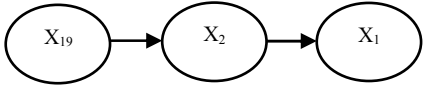
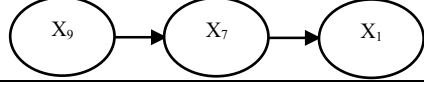
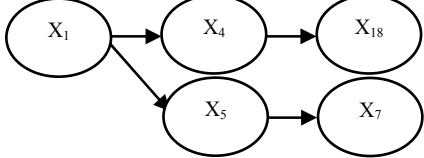
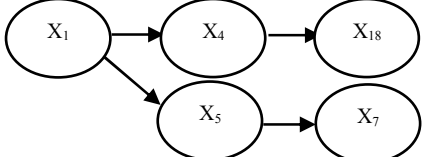
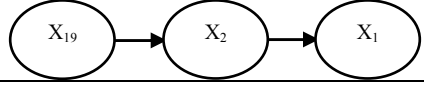
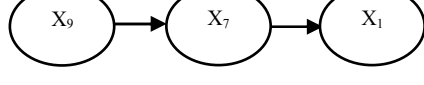
No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_{19}, T_2, T_1) = P(T_1 T_2).P(T_2 T_{19}).P(T_{19})$
2.		$P(T_9, T_7, T_1) = P(T_1 T_7).P(T_7 T_9).P(T_9)$
3.		$P(T_1, T_4, T_{18}, T_5, T_7) = P(T_{18} T_4).P(T_4 T_1) \\ P(T_7 T_5).P(T_5 T_1). P(T_1)$

Table 4: Structure Of The Term Pattern Emergence And Its Probability For Doc-2

No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_1, T_4, T_{18}, T_5, T_7) = P(T_{18} T_4).P(T_4 T_1) \\ P(T_7 T_5).P(T_5 T_1). P(T_1)$
2.		$P(T_{19}, T_2, T_1) = P(T_1 T_2).P(T_2 T_{19}).P(T_{19})$
3.		$P(T_9, T_7, T_1) = P(T_1 T_7).P(T_7 T_9).P(T_9)$

To show the work of this proposed method, we take another document (called Doc-6) which is designed as almost similar to documents five but it is designed as totally different with the four previous documents. After LSA has succeeded to process all of six documents and applying Algorithm 1 followed by cutting the first three terms of them, then the term emergence patterns of Doc-6 is shown in Table 11.

The frequency of each occurrence of the term in every sentence of the six documents can be seen in Table 12, revising the information in Table 8. Based on the frequency in Table 12 and applying Algorithm 3, the likelihood of each sentence in every document for all of six documents would be

changed from Table 9 to Table 13. The pairwise likelihood ratio amongst the six documents, therefore, can be found by applying Algorithm 4 and the result can be seen in Table 14.

Based on Table 14, the testing to similarity pattern of Doc-6 to those previous five documents will not change the previous decision amongst the previous five documents. This table also shows the prove that Doc-6 is significantly different with all of documents except between Doc-5 and Doc-6. These last two documents have almost similar pattern, due to their likelihood ratio that can be stated as 'Not worth more than a bare mention' as in Table 1.

Table 5: Structure Of The Term Pattern Emergence And Its Probability For Doc-3

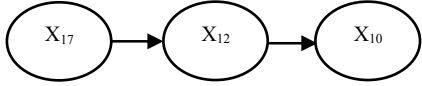
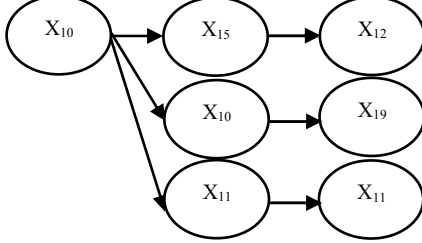
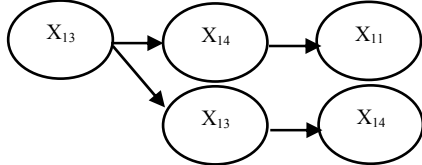
No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_{17}, T_{12}, T_{10}) = P(T_{10} T_{12}) \cdot P(T_{12} T_{17}) \cdot P(T_{17})$
2.		$P(T_{10}, T_{15}, T_{12}, T_{10}, T_{19}, T_{11}, T_{11}) =$ $P(T_{12} T_{15}) \cdot P(T_{15} T_{10}) \cdot P(T_{19} T_{10}) \cdot P(T_{10} T_{10}) \cdot$ $P(T_{11} T_{11}) \cdot P(T_{11} T_{10}) \cdot P(T_{10})$
3.		$P(T_{13}, T_{14}, T_{11}, T_{13}, T_{14}) = P(T_{11} T_{14}) \cdot P(T_{14} T_{13}) \cdot$ $P(T_{14} T_{13}) \cdot P(T_{13} T_{13}) \cdot P(T_{13})$

Table 6: Structure Of The Term Pattern Emergence And Its Probability For Doc-4

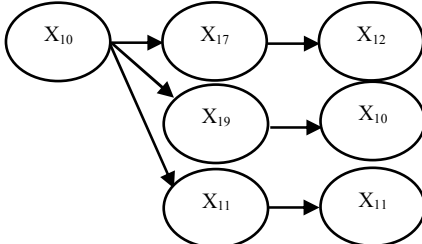
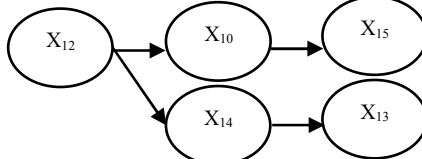

No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_{10}, T_{17}, T_{12}, T_{19}, T_{10}, T_{11}, T_{11}) =$ $P(T_{12} T_{17}) \cdot P(T_{17} T_{10}) \cdot P(T_{10} T_{19}) \cdot P(T_{19} T_{10}) \cdot$ $P(T_{11} T_{11}) \cdot P(T_{11} T_{10}) \cdot P(T_{10})$
2.		$P(T_{12}, T_{10}, T_{15}, T_{14}, T_{13}) = P(T_{15} T_{10}) \cdot P(T_{10} T_{12}) \cdot$ $P(T_{13} T_{14}) \cdot P(T_{14} T_{12}) \cdot P(T_{12})$
3.		$P(T_{11}, T_{10}, T_{16}) =$ $P(T_{16} T_{10}) \cdot P(T_{10} T_{11}) \cdot P(T_{11})$

Table 7: Structure of the term pattern emergence and its probability for Doc-5

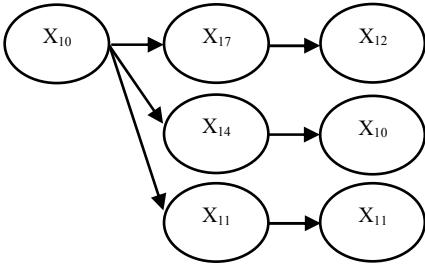
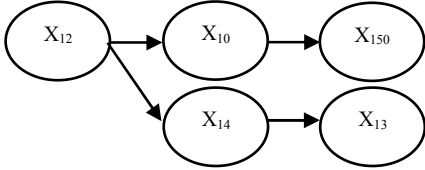
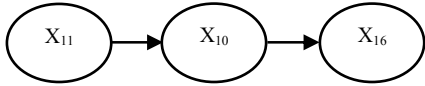
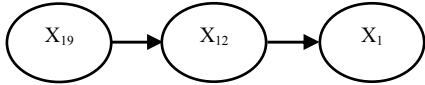
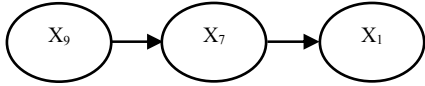
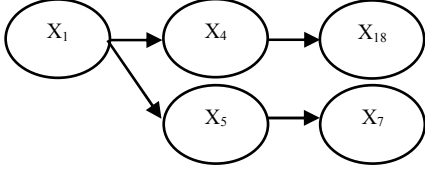
No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_{10}, T_{17}, T_{12}, T_{14}, T_{10}, T_{11}, T_{11}) = P(T_{12} T_{17}) \cdot P(T_{17} T_{10}) \cdot P(T_{10} T_{14}) \cdot P(T_{14} T_{10}) \cdot P(T_{11} T_{11}) \cdot P(T_{11} T_{10}) \cdot P(T_{10})$
2.		$P(T_{12}, T_{10}, T_{15}, T_{14}, T_{13}) = P(T_{15} T_{10}) \cdot P(T_{10} T_{12}) \cdot P(T_{13} T_{14}) \cdot P(T_{14} T_{12}) \cdot P(T_{12})$
3.		$P(T_{11}, T_{10}, T_{16}) = P(T_{16} T_{10}) \cdot P(T_{10} T_{11}) \cdot P(T_{11})$
4.		$P(T_{19}, T_{12}, T_1) = P(T_1 T_{12}) \cdot P(T_{12} T_{19}) \cdot P(T_{19})$
5.		$P(T_9, T_7, T_9) = P(T_1 T_7) \cdot P(T_7 T_9) \cdot P(T_9)$
6.		$P(T_1, T_4, T_{18}, T_5, T_7) = P(T_{18} T_4) \cdot P(T_4 T_1) \cdot P(T_7 T_5) \cdot P(T_5 T_1) \cdot P(T_1)$

Table 8: The frequency of term occurrence in five documents

Occurrence Sequence					
1 st		2 nd		3 rd	
Ter m	Freq	Term	Freq	Term	Freq
T ₁₀	9	T ₁₀	5	T ₁	6
T ₁	6	T ₇	3	T ₁₁	4
T ₁₂	4	T ₄	3	T ₁₈	3
T ₉	3	T ₅	3	T ₇	3
T ₁₉	3	T ₁₁	3	T ₁₀	3
T ₁₁	2	T ₁₄	3	T ₁₂	3
T ₁₃	2	T ₂	2	T ₁₅	2
T ₁₇	1	T ₁₂	2	T ₁₆	2
		T ₁₇	2	T ₁₃	2
		T ₁₉	2	T ₁₉	1
		T ₁₅	1	T ₁₄	1
		T ₁₃	1		



Table 9: Likelihood Of Each Sentence In Each Document

		Denominator					
Numerator		L _R	Doc-1	Doc-2	Doc-3	Doc-4	Doc-5
Doc-1				1	6.01E+08	6.41E+06	4.02E+16
Doc-2					6.01E+08	6.41E+06	4.02E+16
Doc-3						9.37E+01	6.69E+07
Doc-4							6.27E+09

Table 10: Likelihood ratios for pair ways of 5 documents

Document	Sentence	Likelihood
Doc-1	1	0.00133
	2	0.00200
	3	0.00200
	4	0.00200
Doc-2	1	0.00200
	2	0.00200
	3	0.00133
	4	0.00200
Doc-3	1	0.00004
	2	0.00100
	3	0.00167
	4	0.00400
	5	0.00089
	6	0.00007

Document	Sentence	Likelihood
Doc-4	1	0.00200
	2	0.00200
	3	0.00400
	4	0.00148
	5	0.00089
	6	0.00074
Doc-5	1	0.00200
	2	0.00300
	3	0.00400
	4	0.00148
	5	0.00089
	6	0.00074
	7	0.00133
	8	0.00200
	9	0.00200
	10	0.00200

Table 11: Structure of the term pattern emergence and its probability for Doc-6

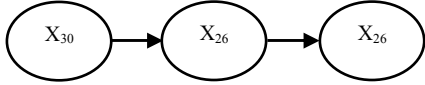
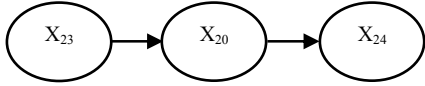
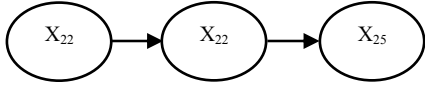
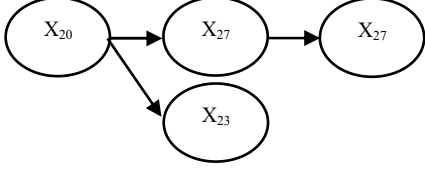
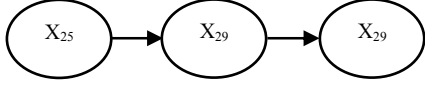
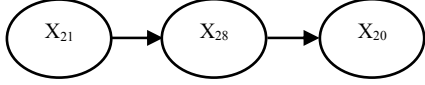
No.	Structure of the term pattern emergence	Joint probability of the structure
1.		$P(T_{30}, T_{26}, T_{26}) = P(T_{26} T_{26}) \cdot P(T_{26} T_{30}) \cdot P(T_{30})$
2.		$P(T_{23}, T_{20}, T_{24}) = P(T_{24} T_{20}) \cdot P(T_{20} T_{23}) \cdot P(T_{23})$
3.		$P(T_{22}, T_{22}, T_{25}) = P(T_{25} T_{22}) \cdot P(T_{22} T_{22}) \cdot P(T_{22})$
4.		$P(T_{20}, T_{27}, T_{27}, T_{23}) = P(T_{27} T_{27}) \cdot P(T_{27} T_{20}) \cdot P(T_{23} T_{20}) \cdot P(T_{20})$
5.		$P(T_{25}, T_{29}, T_{29}) = P(T_{29} T_{29}) \cdot P(T_{29} T_{25}) \cdot P(T_{25})$
6.		$P(T_{21}, T_{28}, T_{20}) = P(T_{20} T_{28}) \cdot P(T_{28} T_{21}) \cdot P(T_{21})$

Table 12: The frequency of term occurrence in six documents

Occurrence Sequence					
1 st		2 nd		3 rd	
Term	Freq	Term	Freq	Term	Freq
T ₁₀	9	T ₁₀	5	T ₁	6
T ₁	6	T ₇	3	T ₁₁	4
T ₁₂	4	T ₄	3	T ₁₈	3
T ₉	3	T ₅	3	T ₇	3
T ₁₉	3	T ₁₁	3	T ₁₀	3
T ₁₁	2	T ₁₄	3	T ₁₂	3
T ₁₃	2	T ₂	2	T ₁₅	2
T ₁₇	1	T ₁₂	2	T ₁₆	2
T ₂₀	2	T ₁₇	2	T ₁₃	2
T ₃₀	1	T ₁₉	2	T ₁₉	1
T ₂₅	1	T ₁₅	1	T ₁₄	1
T ₂₃	1	T ₁₃	1	T ₂₆	1
T ₂₂	1	T ₂₆	1	T ₂₄	1
T ₂₁	1	T ₂₀	1	T ₂₅	1
		T ₂₂	1	T ₂₇	1
		T ₂₇	1	T ₂₉	1
		T ₂₉	1	T ₂₀	1
		T ₂₃	1		
		T ₂₈	1		

Table 13: Likelihood Of Each Sentence In Each Document

Docu ment	Sentence	Likelihood
Doc-1	1	0.00073
	2	0.00110
	3	0.00110
	4	0.00110
Doc-2	1	0.00073
	2	0.00110
	3	0.00110
	4	0.00110
Doc-3	1	0.00002
	2	0.00055
	3	0.00091
	4	0.00219
	5	0.00049
	6	0.00004
Doc-4	1	0.00041
	2	0.00110
	3	0.00110
	4	0.00219
	5	0.00081
	6	0.00049

Docu ment	Sentence	Likelihood
Doc-5	1	0.00110
	2	0.00164
	3	0.00219
	4	0.00081
	5	0.00049
	6	0.00110
	7	0.00110
	8	0.00041
	9	0.00073
	10	0.00110
Doc-6	1	0.00002
	2	0.00002
	3	0.00002
	4	0.00004
	5	0.00146
	6	0.00002
	7	0.00002

Table 14: Likelihood Ratios for Pair Ways Of Six Documents

		Denominator						
		L _R	Doc-1	Doc-2	Doc-3	Doc-4	Doc-5	Doc-6
Numerator	Doc-1		1	2.19E+09	2.28E+06	1.58E+18	4.71E+18	
	Doc-2			1	2.19E+09	2.28E+06	1.58E+18	4.71E+18
	Doc-3				1	9.60E+02	7.23E+08	2.16E+09
	Doc-4					1	6.94E+11	2.07E+12
	Doc-5						1	2.98E+00
	Doc-6							1

6. CONCLUSIONS

Finally it can be concluded that the Latent Semantic Analysis, Bayesian Network and Likelihood Ratio which are structured in four algorithms in this paper strongly able to identify pattern similarities term emergence of two documents.

7. ACKNOWLEDGEMENT

The authors thank BPPDN-College which has provided scholarships for doctoral programs and the Department of Statistics Institute of Technology (ITS) which provide facilities, giving a boost to publish in international journals and motivate doctoral program is immediately resolved. Thanks also to the anonymous reviewer who has reviewed this paper.

REFERENCES:

[1] Soehardjoepri, Iriawan, N., Ulama, B.S.S., 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), 259 - 284.

[3] Kevin, B.K. and Ann, E.N., "Bayesian Artificial Intelligence Second Edition", Computer Science and Data Analysis Series, CRC Press, New York, 2011

[4] S.K.M. Wong and C.J. Butz, "A Bayesian Approach to User Profiling in Information Retrieval", Saskatchewan, Canada, S4S 0A2 Ottawa, Ontario, Canada, K1N 6N5, 2000.

[5] C.K. Chow and C.N. Liu, "Approximating discrete probability distributions with dependence trees", *IEEE Transactions on Information Theory*, IT-14, 3, 462-467, 1968.

[6] Kass R.E. and Raftery, A.E., "Bayes Factors", *Journal of the American Statistical Association*, Vol.90, No. 430 (Jun., 1995), 773-795.

[7] Ben-Gal I., "Bayesian Networks", in Ruggeri F., Faltin F. & Kenett R., *Encyclopedia of Statistics in Quality & Reliability*, Wiley & Sons, 2007.

[8] Soehardjoepri, Iriawan, N., Ulama, B.S.S., and Irhamah, "Identifying Text Document Pattern For Two Terms Appearances VIA Latent Semantic Analysis (LSA) Method And Term Distance Between Two Documents", *Journal of Theoretical and Applied Information Technology*, 20th September 2015. Vol.79. No.2, 322 – 329, ISSN: 1992-8645, E-ISSN: 1817-3195, JATIT & LLS, Pakistan, 2015.

[9] Pawan G., Laxmidhar B. and T.M. McGinnity, "Application of Bayesian Framework in Natural Language Understanding", Vol.25, ISSUE 2008, 251 – 269, IETE TECHNICAL REVIEW, doi: 10.4103/0256-4602.44656.

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

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

[12] Dumais, S.T., "Latent Semantic Analysis", *Annual Review of Information Science and*



- Technology 38: 188 (2005), doi: 10.1002/aris.1440380105.
- [13] 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, (6): 391- 407, 1990.
- [14] Thomas L., Peter W. F., and Darrell L., "Introduction to Latent Semantic Analysis", *Discourse Processes* 25, 259-284, 1998, doi: 10.1080/01638539809545028.
- [15] Law, A. M., and Kelton, W. D., "Simulation Modeling and Analysis", McGraw-Hill International Series, Singapore, 2000.
- [16] Thomas L., Peter W. F., and Darrell L., "Introduction to Latent Semantic Analysis", *Discourse Processes* 25, 259-284, 1998, doi: 10.1080/01638539809545028.
- [17] Kasim, S., "Making Application Method to Detect Plagiarism with Latent Semantic Analysis", *Final, Department of Informatics*, University of Surabaya, 2012.