Journal of Informatics and Computer Science No. 2 Vol.2 Oktober 2016
Universitas Ubudiyah Indonesia
ISSN : 2442-4714

PLAGIARISME DETECTION SYSTEM ON THIEF FACULTY OF COMPUTER SCIENCE UNIVERSITY U'BUDIYAH INDONESIA BASED TEXT MINING

Desita Ria Yusian TB, Fathiah
Department Of Informatics Engineering, Faculty of Computer Science,
Universitas Ubudiyah Indonesia,
Jalan Alu Naga, Desa Tibang, Banda Aceh, Indonesia
Email : desita@uui.ac.id, fathiah@uui.ac.id

ABSTRACT
Plagiarism is the act of intentionally or unintentionally in obtaining or trying to obtain credit or value to scientific work., citing part or all of the work / or scientific work of others who are recognized as scientific works, co without stating the source appropriately and adequately. Problem of plagiarism has become a phenomenom that is being discussed in academic forums. Based on a variety of information about the practice of plagiarism is increasingly disturbing, this study conducted at the Ubudiyah University Indonesia (UUI). This system will implemented the detection system plagiarism that will be creating using a document apllied method of vector space model to analyze similarities in document displayed on the web browser. Input on this system is a document in the form of text format whith the extention test performed on real data that is usig data documents with vector space model algorithm value is above 85% to 100% for high level document. As for a document with a low degree of similarity values below 50%.
Keyword : Plagiarism, Vector Space Model, Similarity

INTRODUCTION
According to the regulation of the Minister of National Education RI. 17 of 2010 on prevention and prevention of plagiarism in universities. Chapter 1 Article 1 Paragraph 1 (Jonatan, 2012), plagiarism is a deliberate or intentional act of obtaining or attempting to obtain credit or value for a scientific work, by quoting part or all of the work and / or scientific work of another party acknowledged as a work scientific, without declaring the source appropriately and adequately.

The rapid development of information systems has become an important element in the development of the mindset of the community, especially students and researchers for the time to the future. One of the elements of the support element is to help the process of making the paper more easily and quickly. Not only does it bring about a positive impact on technological advancement, it also brings an almost unavoidable negative impact of plagiarism.

At college nowadays it has been demanded to have competitive advantage by utilizing all resources owned. In addition to facilities, infrastructure and human resources, information systems are one of the resources that can be used to increase competitive advantage. Information systems can be used to obtain, process and disseminate information to

support daily operational activities as well as support strategic decision-making activities one of them by having a system detection in correcting thesis topics to avoid plagiarism.

The issue of plagiarism has become a hotly discussed phenomenon in academic forums. Based on various info about the practice of plagiarism which is increasingly troubling the researchers, this research is considered quite important to be done nowadays since Ubudiyah University (UUI) Campus is one of the campuses that dedicates itself to creating professional scientists and researchers for Aceh in the future will come. This system will be implemented on one faculty owned by UUI namely computer science faculty with four study programs that is D-III.

Management Informatics, DIII computerized accounting, S-1 Informatics Engineering and S-1 Information Systems. Given the vision and mission of UUI is something strategic in the year 2015-2025 to become a world class cyber university where as a first step students will be facilitated by various facilities based on information technology one of them by creating a facility of detecting plagiarism to make the best graduates who have high scientific level in producing a research result.

During this UUI campus is still using the facility of data storage thesis in a format that uses large memory, so the longer it will be difficult in collecting, retrieving and reading data. The thesis repository system is not available in centralized containers nor has it been a difficulty in detecting plagiarism from a new thesis with that previously researched. Until now the detection process still relies on the teaching staff through the guidance stage of each college student. However, this is not a matter of reaching effective review because the process is still done manually.

STUDY LITERATURE

The detection of plagiarism is also done in text documents using the rabin-karp algorithm with synonym recognition. The algorithm that implemented this hash function proved to be powerful enough to detect plagiarism through word equality. And through the variant of the modified Rabin-Karp algorithm, the system not only compares the remainder of modul but also compares the results to the modul itself. This way it can avoid spurious hits (error in matching). So that can be obtained a percentage of similarities and better processing time. And to anticipate the word that is replaced by the synonym used synonym recognition approach. (Sandy Dewanto, et al: 2013) Detection system at this time has begun to be done and growing because of the number of fraud and dishonesty in making a writing of the research results. The creation of an application program for the detection of similarity of text documents with the Smith-Waterman algorithm is one way to detect plagiarism from a thesis by comparing between two documents suspected as a plagiarism. (Farid Thalib, et al: 2014)

In further research, the jaro-winkler distance algorithm is applied to the plagiarism detection system for Indonesian text documents by comparing the similarity between Indonesian text documents, so that it can be determined a document is plagiarism or not. The result is that the detection of plagiarism against the corpus document through the stemming and

query stages through stages without stemming has a better detection value of 30.58%. (Ahmad kornain, et al: 2014)

Text Mining

Text mining can be defined as the discovery of new information and not previously known by the computer, by automatically extracting information from different unstructured text source sources. The key to this process is to combine the information that has been successfully extracted from various sources (Tan, 1999). The main purpose of text mining is to support the process of knowledge discovery in large document collections.

Text mining is a multi-disciplinary field of science, involving information retrieval (IR), text analysis, information extraction (IE), clustering, categorization, visualization, database technology, natural language processing (NLP), machine learning, and data mining. It can also be said that text mining is one form of application of artificial intelligence (artificial intelligence / AI). Text mining is also used in some spam email filters as a way of determining message characteristics that may be advertisements or other unwanted material.

Plagiarism detection systems can be developed to:

1. Text data such as essays, articles, journals, research and so on.
2. A more structured text document

Such as programming languages. Some types of plagiarism are:

1. Word-for-word plagiarism is copying every word directly without any modification.
2. Plagiarism of authorship is to acknowledge the work of others as a result of their own work by putting their own names in the name of the real author.
3. Plagiarism of ideas is to recognize the thoughts or ideas of others. Plagiarism of sources, if an author uses quotes from other authors without specifying the source. (Parvatti, 2005)

Document Extraction

Text that will be text mining process, generally have some characteristics such as having high dimension, there is noise in data, and there is not good text structure. The way used in studying a text data, is to first determine the features that represent each word for each feature in the document. Before determining the representative features, a general preprocessing stage is required in the text mining of the document, ie case-folding, tokenizing, filtering, stemming, tagging and analyzing.

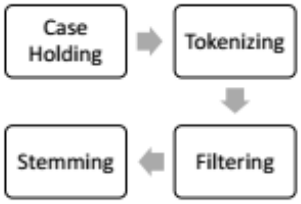

Figure 2.1 Pre-processing

Vector Space Model (VSM)

Vector space model (VSM) is one form of algebraic modeling that is used to describe text and documents into vector form. This method is commonly used to determine the value of document resemblance to a query word. The value of the similarity of a string or word will be greater if the string more often appear in a document. VSM is commonly used in information filtering, indexing, and relevancy ranking.

As the name implies VSM uses vectors as a simile of documents and queries. Document is vectored **d** (di1,di2,…dit) and *query* be vectored **q** (wq1,wq2,…wqt). Each vector corresponds to a term that can be defined differently depending on the usage of the VSM. The term definitions are generally words, phrases, or keywords. If the term is defined as a word then the dimension of the vector space is the number of words in the entire text. Calculating the value of the term and its relation to each vector is also called term weighting. The advantages of VSM method are as follows:

a. A simple model with a linear algebraic base that is easy to calculate
b. The weights of terms are not in binary form
c. Can determine the relevance rating between queries and documents

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a method of weighting the relevance of a term with a document. In the TF-IDF method the calculated element is the term occurrence frequency of a document (TF) and also the inverse document frequency containing the word. In the case of TF-IDF calculation on a single document then there is also a sentence as a benchmark calculation.

The IDF of a word indicates the importance of a word to the entire document. The smaller the IDF value, the word is increasingly not considered important for the document as a whole. While the TF of a word signifies how many words are spoken in one sentence. In other words a word will be of greater relevance to a sentence if the word appears in many sentences in question and there is no or little appearance on other sentences in the same document. The following equation 2.1 is the equation for the calculation of TF-IDF.

Similarity Coefficient

In the final result VSM method which is the value of similarity or relevance of a document expressed in a value that is similarity coefficient. The formula of calculating the value of similarity or known as the similarity coefficient (SC) is to use the formula as follows:

$$sim(d_j, q) = \frac{d_j . q}{\left|\left|d_j\right|\right| \left|\left|q\right|\right|}$$

METHOD

The experimental research will be experimental based on the type of data obtained by collecting the research material in the form of a thesis document document that has been studied previously from the storage place (database) faculty of computer science at UUI campus. The data tested on this plagiarism detection system is a text document with a .pdf extension. The stages of what will be done to build the system to detect plagiarism can be seen in the figure below 3.1:

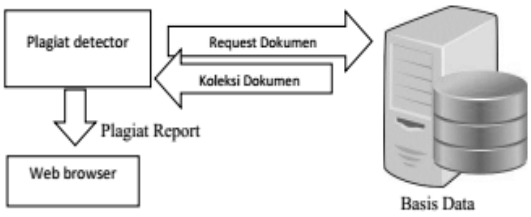

Figure 3.1 Plagiarism Detection System Scheme

The design of a plagiarism detection system to be created using a document applies the Vector Space Model method to analyze the similarity levels in documents displayed on a web browser.

Input on this system is a document in the form of text format with .doc extension. Users will input the document to be tested into the system and the system to request to the database server to send a collection of training documents that will be processed through several stages of tokenisasi, save the training documents from the filtering, stoplist and stemming for documents to be tested, calculate the Term Frequency (TF), calculate the Inverse Term Frequency (ITF), find the similarity level and display the percentage of plagiarism in the form of a report (report) to the web browser.

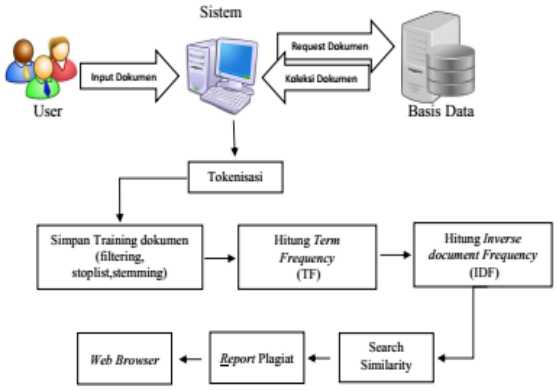

Figure 3.2 Scheme of plagiarism detection system stage

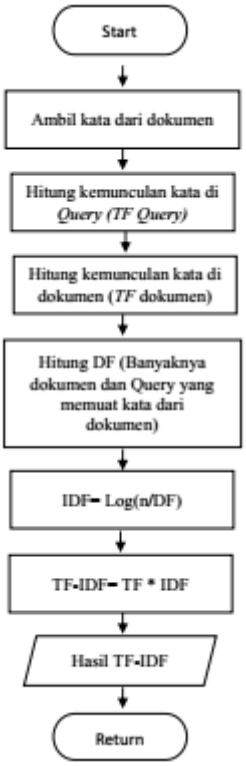

Figure 3.3 Flowchart process Calculation of Term Frequency - Inverse Document Frequency (TF-IDF)

ANALYSIS AND DISCUSSION

The results of the analysis of the data obtained before the system performed show some weaknesses that can be concluded that the required system that can overcome the deficiencies of existing systems, namely systems that have the ability:

a. A system capable of detecting plagiarism against the final task / thesis, ranging from verbatim plagiarism, if possible to semantical plagiarism.
b. The system has facilities for storing data assignments end / thesis to make it easier in accessing the data.
c. The system implements a data representation method that can reduce memory usage for massive data storage.
d. The system performs the task of plagiarism without the need for specific human skills (observation, caution or reasoning), but performs detection operations automatically using computer processing resources.

4.1 System Interfaces

Plagiarism detection system implemented in faculty of computer science UUI is a display is the initial appearance of the system using the

Vector Space Model Method, in the main menu view there are 6 tab menu, the main menu, training document, corpus training, plagiarism detection, about us, help. The main menu view can be seen in Figure 4.1.

Figure 4.1 The main view page of the plagiarism detection system

The training document menu is used as the page used to include the document as a reference to the test document. Display the training document menu can be seen in Figure 4.2. The training document menu is used to perform the preprocessing process of the document to be used as training data.

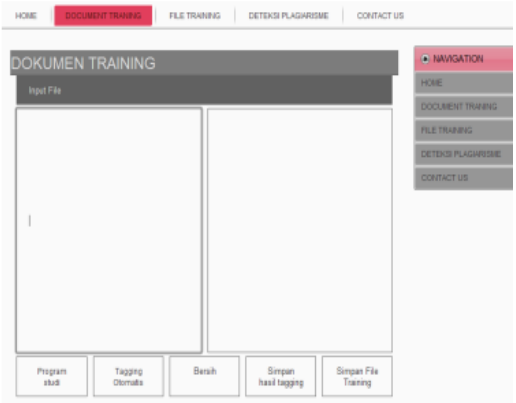

Figure 4.2 Display the Input menu of the document training

The corpus training menu is used to store tokenisasi, stopword, stemming or non stemming results that will be used for training documents. The training file menu view can be seen in Figure 4.3. Training files are used as a place to store documents that will be used as training data.

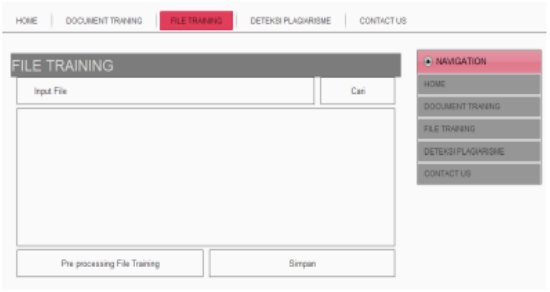

Figure 4.3 Display the File Training menu

The plagiarism detection menu is used to perform plagiarism detection of how much percentage similarity a document is tested with training documents. The menu view of plagiarism detection can be seen in Figure 4.4. The output of this menu is a percentage of the value of the document tested against the training document.

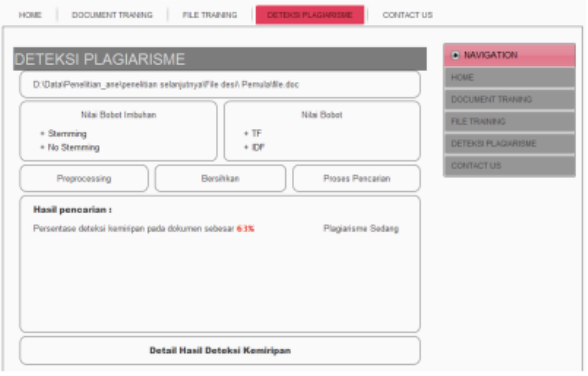

Figure 4.4 Display Detection Menu Plagiarism

Experiment 1

Experiment 1 aims to find out the results of a better similarity calculation between similarity results performed calculations using TF or TF-IDF. The results of trial 1 can be seen in Figure 4.5. Based on the results of test 1 can be concluded that by using TF weighting the result will be better than using TF-IDF weighting.

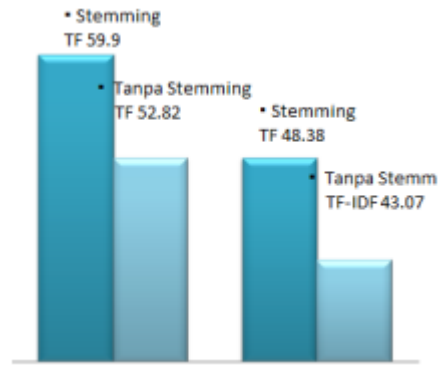

Figure 4.5 Result of the experiment 1

Experiment 2

Experiment 2 aims to find out the results of a better similarity calculation between the similarity results performed calculations using training documents stemming or without stemming. The test results of 2 training documents can be seen in Figure 4.6 and the training document without stemming can be seen in Figure 4.7.

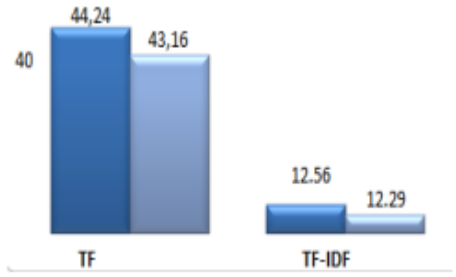

Figure 4.6 Test results with 2 documents in Training Stemming

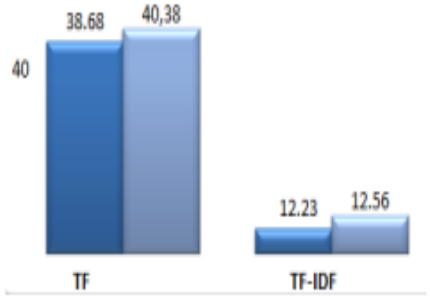

Figure 4.7 Test results with 2 documents in Training without Stemming

CONCLUSION

Based on the test that has been done to detect the presence of elements of plagiarism in the computer science faculty then got the following conclusions:

a. Testing using real data is document data berplagiat with menngunakan Vector Space Model algorithm produce high similarity value that is above 85% until 100% for documents that are of similar magnitude. As for documents with a low level of similarity or not containing plagiarism then produce a value similarity below 50%.
b. Based on test 1 conducted in Chapter 4, the results of the Vector Space Model application in system testing to detect plagiarism levels can use TF or TF-IDF, IDF, where the results using TF-IDF are much better than TF.
c. Based on test results 2 results plagiarism by using TF- IDF would be better if the query stemming and training files are not in stemming or otherwise.

BIBLIOGRAPHY

[1] Ridhatillah, A dkk **(**2003) Dealing with plagiarism in the information. System reasearch Community: A Look at Factors That Drive Plagiarism and Ways to Address Them, MIS Quarterly; Vol. 27, No. 4, p. 511-532/December 2003.

[2]Farid Thalib, Ratih K. (2014). Pembuatan Program Aplikasi untuk Pendeteksian Kemiripan Dokumen Teks dengan Algoritma Smith-Waterman,

[3]Sandy Dewanto, dkk. (2013) Deteksi Plagiarisme Dokumen Teks Menggunakan Algoritma Rabin-Karp Dengan *Synonym Recognition*, Program Studi Ilmu Komputer, Program Teknologi Informatika dan Ilmu Komputer, Universitas Brawijaya Malang

[4] Ahmad kornain, dkk (2014) Penerapan
Algoritma Jaro-Winkler *Distance* Untuk Sistem Pendeteksi Plagiarisme Pada Dokumen Teks Berbahasa Indonesia, Program Studi Teknik Informatika, STMIK GI MDP