International Journal of Advanced Science and Technology
Vol. 29, No. 6s, (2020), pp. 2130-2136

Plagiarism Detection Using Manber and Winnowing Algorithm

1st Muhammad Faisal
Dept. of Informatics Engineering
UIN Maulana Malik Ibrahim Malang
Malang, Indonesia

2nd Fresy Nugroho
Dept. of Informatics Engineering
*UIN Maulana Malik Ibrahim Mal*ang
Malang, Indonesia

3rd Maulana M. El Sulthan
Dept. of Informatics Engineering
UIN Maulana Malik Ibrahim Malang
Malang, Indonesia

4th Fauziyah Amini
Dept. of Informatics Engineering
UIN Maulana Malik Ibrahim Malang
Malang, Indonesia

5th M. Amin Hariyadi
Dept. of Informatics Engineering
UIN Maulana Malik Ibrahim Malang
Malang, Indonesia

6th Agung Sedayu
Dept. of Architecture Engineering
UIN Maulana Malik Ibrahim Malang
Malang, *Indonesia*

Abstract

Plagiarism is a copy of the essay or opinion of others and does not list the written references, and makes it his own essay or opinion. There are several algorithms that have the ability to detect plagiarism of documents such as Jaro-Winkler algorithm, winnowing, Manber and others. In this study, research conducted on Mamber and Winnowing algorithms in detecting plagiarism. The Manber algorithm is an algorithm that uses K-grams but does not use the formation of a window while the winnowing algorithm is an algorithm that uses the K-grams approach in shaping the fingerprint pool. The app divides the documents into Biword and Triword tokens. These tokens are converted to MD5 value, the tokens have a hash value that has the same length and can be used as a document fingerprint. The Biword and Triword approaches are implanted in the winnowing algorithm, while the Biword is for Manber algorithms. This algorithm can check the phrase of each document, then saved in to an array. At the time of displaying the document will be obtained the same value long, the algorithm is able to display the value of arrays that form a Biword token as a fingerprint.

From the results of the similirity of the 10 test data, the average result for manber algorithm is 90.56%, the Winnowing algorithm is 94% and the Winnowing triword 91.22% algorithm. The average time of generating winnowing triword data is 78.95 seconds and is 5.2% slower than the winnowing biword of 73.75 seconds.

Keywords— *Plagiarism, documents, Manber, winnowing, similirity, generating time*

I. INTRODUCTION

Plagiarism is an act of either intentional or unintentionally gaining or attempting to gain credit or value points on a scientific work, by quoting part or all of the work and/or scientific work of the other party and acknowledged by its scholarly works, without The source appropriately and adequately [1]. Moeliono states that Plagiation takes other people's work and is publicized to be his cause one ignores honour, ignores honesty, tends to cheat and looks down on others [2].

For that it is necessary once designed and built a virtual machine to detect the writings that will be uploaded in the online journal media so as to minimize the presence of plagiarism. The virtual machines that there are mostly foreign-made such as Turnitin, Plagscout, Viper, Plagiarism Checker. To detect any plagiarism lecturers and students must access the network that they provide. Most of the fee Plagiarism detector machines can drain the country's foreign exchange. There are several algorithms that can detect the authenticity of documents. One of the document's authenticity detector algorithms is the winnowing algorithm. According to Diana et al. [3]. Winnowing algorithm detects the

ISSN: 2005-4238 IJAST

similarity of sentences between text files, document finger printing, converts N-grams of text into batches of hash values. Tests test the ability to detect sentences with the same changes as the N-value parameters of N-grams, window W, prime numbers as hashing and plagiarized threshold values. According to Riki [4] The process of similirity of documents uses N-gram parameters, window, synonym recognition, text processing. In this study, authors used winnowing and Manber algorithms to know the resemblance of the document being tested with multiple source documents. Manber algorithms include those used because this algorithm is a text matching algorithm as well and uses fingerprints in the data process, fingerprints are used to detect plagiaration including the smallest similar part in a Documents with the number of words that many Purwitasari, et al. [5]. Manber algorithms can be used to detect the existence of plagiarism, some irrelevant characters will be discarded such as punctuation marks, dots, commas and other marks. Character-based fingerprint techniques are used in this algorithm. By scanning the text on the document. On this research researchers use sentences. The text of the document is collected into two words or called Biword. The Biword is formed to retain the phrase on the document text. The Biword concept gives fewer word tokens than a triword. Biword is formed on each document.

II. LITERATURE REVIEW

Categories of plagiarism have been successfully compiled, among other words plagiation, plagiation changes the word synonym, plagiation of writing style, plagiation of metaphors, plagiation of the idea and plagiarism of Wibowo's own work [6]. The plagiarism detection system that has been done by Meyer [7] consists of similirity test and similirity analysis. Similirity test is intended as a comparison of original documents with comparative documents. Similirity analysis consists of the translation language as structure analysis and the transformation of similirity analysis. The similirity test is done from small documents as well as large documents. [8] conducting research on plagiarism-based citation index by utilizing large-scale corpus data with data has been available and has not been dintegrasikan with all online hosts.

Several studies have been conducted to develop a plagiarism detection system in Indonesia. Salmuasih [9] conducted a research on plagiarism in the text document with the Similarity concept using Rabin Karp algorithm, Kurniawati et al. [10]. Implementing the Jaro-Winkler Distance algorithm to compare the similarities of Indonesian documents. Purwitasari et al [5] examined the existence of the same sentence as an indication of plagiarism with the N-Gram-based hashing algorithm. Liliana et al. [11] Researching about plagiarism source code. The Alfikri et al [12] examines the screening of plagiarism on cross-language. The research will be implemented there are several stages. The initial stage is making a plagiarism detection system sentence with Manber algorithm and subsequent detecting of files. The next stage using the algorithms winnowing and Jaro Winkler distance end stage is the creation of web crawlers aiming to search for documents similar to the existing data on the paper, then with the algorithm. Then the document is compared to the existing one using winnowing and Manber algorithms, to collapse using the algorithm of Jaccard Similirty to search for resemblance of the document.

III. METHODS

The research methodology consists of several stages of research with the diagram block in Figure 1.

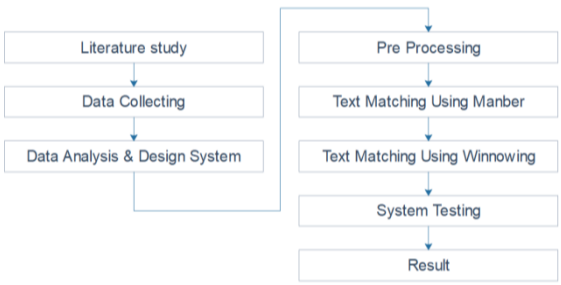

Fig. 1. The research methodology

ISSN: 2005-4238 IJAST

A. Winnowing Algorithm

The Winnowing algorithm is capable of tracing and conducting word similarity testing in an article or in many articles. The technique used is known as rolling hash. The term hash refers to the calculation of the ASCII numerical value of each character Schleimer DKK [13]. The steps that are passed are as follows:

1. Pre-process, i.e. eliminate inappropriate characters, e.g. spaces, punctuation marks
2. Preparation of Gram series based on size K.
3. Calculate the hash value of each gram
4. Mathematical models of hash techniques

$$H_{(c1....ck)} = c_1 * b^{(k-1)} + c_2 * b^{(k-2)} + \cdots + c_{(k-1)} * b^k + c_k \quad (1)$$

with :

c = ASCII

b = calculation based

k = Number of characters

Next calculate the numeric value of the hash of K-grams using these rolling hash mathematical models:

$$H_{(c2....ck+1)} = (H_{(c1....ck)} - c_1 * b^{(k-1)}) * b + c_{(k+1)} \quad (2)$$

5. Divide into different window
6. Select the hash values into document traits.
7. Determine the similirity of the document based on Jaccard coefficient equation

B. Mamber Algorithm

Manber has a solving problem that is almost identical to the Winowing algorithm. Manber's algorithm has fewer steps and the check document time is faster than the Winnowing algorithm, no fingerprint position information is where, the selection of a different fingerprint Kurniawati DKK [10].

Problem solving in Manber algorithm, the fingerprint is taken from each hash-value on the condition:

H mod P = 0

where :

H = Hash-value

P = Divider value.

In the Winowing algorithm hash-value is the minimum in each window. The Manber algorithm is:

1. Forming a group of grams with length N characters.

2. The hash-value calculation of each gram uses a hash function.

Retrieves the hash value to be used as a document fingerprint.

ISSN: 2005-4238 IJAST

C. Jaccard Coefficient

Jaccard Coefficient is a function used to define similarities between text documents. The Hash function is used after the winnowing algorithm process series i.e. calculates the Hash value and selects the smallest fingerprint of the two text documents Schleimer DKK [13]. Just then utilize the function Jaccard Coefficient.

Here is a description of the Jaccard equation Coefficient:

$$\text{Similarity}(\text{di}, \text{dj}) = \frac{|\text{w}(\text{d1}) \cap \text{w}(\text{d2})|}{|\text{w}(\text{d1}) \cap \text{w}(\text{d2})|} \text{x} 100\% \qquad (3)$$

IV. RESULT

System implementation is a manifestation of the analysis and design of the system that has been created. In this implementation, the test is conducted using the abstract document data of the students that have been uploaded to the Internet. To be able to know the level of similarity of the document data, the researcher will randomize the placement of the word position in the abstract. The randomised Abstract is then inserted into the virtual machine to be processed using the winnowing and Mamber algorithms. For the winnowing algorithm the researcher implements the Biword winnowing approach and the winnowing Triword. For Manber algorithm researchers implement a Biword manber approach.

The test phase of the application is a phase that is generated after the creation of completed application programs. The goal is to know if the application that has been made according to the design of the system that was done before with the specific purpose to be gained.

Test the application in the form of a set of sentences with two paragraph abstract English language with the test done several times by changing the composition of paragraphs, changing the writing and placing the words randomly in several sentences.

The method is tested using Manber algorithm, Winnowing Biword and Triword algorithm.

A. Pre-Processing.

Aims to know the processes before processing manber and winnowing algorithms.

The source Data is abstract thesis with two paragraphs in English, as for the source sentence is:

" In this modern era technology is developed rapidly it is decides not only from the hardware but also software sides one of that developments is game game is the sophisticated entertaining shapes for childs adolescents until adults not only purpose for entertaining game also recommended for learning one of lesson which is studied in every stage of education is mathematics many students have difficulties in a counting whether adding decreasing multiply or division here building an arithmetic education game application based on android that is serve functionality in each of exercise level shapes there is also recommendation for studying of every level.

The method of this education game uses naive bayes classifier which is used on the classification of appropriateness degree counting the result shows that application of aritmethic education game using naive bayes classifier is recommended for learning the test on 20 data shows the success amount 85 whereas 15 more are not stable yet. "

The pre-process testing phase is testing by changing capital letters to lowercase, eliminating space, punctuation, punctuation, period marks, commas, semicolon, question marks and others.

The results of pre-process testing are as follows:

" inthismodernera technologyisdevelopedrapidlyitisdecidesnotonlyfromthehardwarebutalso softwaresidesoneofthatdevelopmentsisgamegameisthesophisticatedentertainingshapesforchilds adolescentsuntiladultsnotonlypurposeforentertaininggamealsorecommendedforlearningoneoflessonwhi chisstudiedineverystageofeducationismathematicsmanystudentshavedifficultiesinacountingwhetheraddi ngdecreasingmultiplyordivisionherebuildinganarithmeticeducationgameapplicationbasedonandroidtha

ISSN: 2005-4238 IJAST

tisservefunctionalityineachofexerciselevelshapesthereisalsorecommendation forstudyingofeverylevelthemethodofthiseducationgameusesnaivebayesclassifierwhichisusedontheclassification ofappropriatenessdegreecountingtheresultshowsthatapplicationofaritmethiceducationgameusing naivebayesclassifierisrecommendedforlearningtheteston20datashowsthesuccessamount85whereas15 morearentstableyet "

B. Testing Data and Similirity.

The document text results from the pre-process test are then split tested through multiple stage stages.

The results of Praproses are the tokens of Mamber Biword, winnowing biword and winnowing Triword.

The input value is a prime number 2, Manber algorithm, winnnowing biword algorithm, winnowing triword algorithm. This test aims to find out which method is better with the highest similirity in detecting the presence of plagiarism text documents.

The test of plagiarism data is conducted to know how large the level of plagiarism of a document is traced.

The test Data consists of :

1. AbstractTest1 as source.
2. AbstractTest2 As a data manipulation source with a data modification trial of 10 trial treatments consisting of :
   a. AbstractTest201 : Changing the first paragraph is moved to second paragraph.
   b. AbstractTest202 : Memindah kalimat pertama dan kedua pada paragraph pertama ke bagian akhir pada paragraph pertama.
   c. AbstractTest203 : Move the first sentence in the first paragraph to the end of the first paragraph.
   d. AbstractTest204 : Delete first sentence in first paragraph
   e. AbstractTest205 : Mengurangi kata dalam kalimat pertama pada paragraph pertama
   f. AbstractTest206 : Reduce the word in the first sentence in the first paragraph.
   g. AbstractTest207 : Move the first sentence in the second paragraph to the end of the second paragraph.
   h. AbstractTest208: Deleting the first sentence in the second paragraph
   i. AbstractTest209 : Deleting the first sentence on reducing the word in the first sentence of the second paragraph
   j. AbstractTest210 : Randomize sentences that are in the first and second paragraphs.

The following are the test results of the program application as seen in the table I.

TABLE I. TEST RESULT SIMILIRITY MANBER METHOD BIWORD, WINNOWING BIWORD AND WINNOWING TRIWORD

No	Sentences	*Manber Biword*	*Winnowing Biword*	*Winnowing Triword*
1	abstract Test201	90.91%	100%	98.68%
2	abstract Test202	93.94%	93.63%	98.68%
3	abstract Test203	92.54%	98.04%	91.72%

ISSN: 2005-4238 IJAST

4	**abstract Test204**	**92.19%**	**89.33%**	**86.75%**
5	**abstract Test205**	**95.38%**	**92.31%**	**94.77%**
6	**abstract Test206**	**89.71%**	**100%**	**92.36%**
7	**abstract Test207**	**89.71%**	**97.40%**	**92.36%**
8	**abstract Test208**	**85.94%**	**91.89%**	**84.31%**
9	**abstract Test209**	**92.19%**	**92.16%**	**88.39%**
10	**abstract Test210**	**83.10%**	**86.50%**	**84.15%**
	Average	**90.56%**	**94%**	**91.22%**

TABLE II. OLD TEST RESULTS PROCESS WINNOWING BIWORD AND WINNOWING TRIWORD METHODS

No	**Sentences**	***Winnowing Biword Time***		***Winnowing Triword Time***	
1	abstractTest201	54	ms	64.05	ms
2	abstractTest202	224.76	ms	269.65	ms
3	abstractTest203	57.21	ms	61.37	ms
4	abstractTest204	50.9	ms	48.21	ms
5	abstractTest205	50.8	ms	56.26	ms
6	abstractTest206	57.62	ms	50.42	ms
7	abstractTest207	77.96	ms	56.13	ms
8	abstractTest208	52.19	ms	56.35	ms
9	abstractTest209	53.16	ms	62.78	ms
10	abstractTest210	58.94	ms	64.32	ms
	Average	73.75	ms	78.95	ms

V. CONCLUSSION

Conclusions from the research phase of this research are as follows:

1. Application of plagiarism detection using Manber Biword algorithm, winnowing with Biword and triword approach

ISSN: 2005-4238 IJAST

2. In the process of detection of plagiarism document, the number of the smallest specified value of 2 is used to produce high similirity value between documents.
3. In the process of detection of plagiarism document using winnowing, the value of the window is set at a value of 4.
4. Application of plagiarism detection by performing sentence position change in document.
5. In the test of data 10 documents, the average similarity value of the mamber algorithm amounted to 90.56%, winnowing algorithm 94% and the winnowing algorithm of the Triword 91.22%, so the Biword winnowing algorithm is still better than the algorithm Manber and winnowing Triword

In a 10 document data test, the average data generation time of winnowing Triword amounted to 78.95 seconds and is slower 5.2% than the winnowing Biword by 73.75 seconds

REFERENCES

[1] Kementerian Pendidikan Nasional. (2010). Peraturan Menteri Pendidikan Nasional Nomor 17 tahun 2010 tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi. Jakarta: Kementerian Pendidikan Nasional.

[2] Yahya, I. (2011). Plagiarisme dan [karya] kita. Sarasehan Program Studi Agroteknologi Fakultas Pertanian UPN Veteran (hlm. 1-6). Yogyakarta: UPN Veteran

[3] Diana Purwitasari, Putu Yuwono Kusmawan, Umi Laili Yuhana. *Deteksi Keberadaan Kalimat Sama Sebagai Indikasi Penjiplakan Dengan Algoritma Hashing Berbasis N-gram*. Kursor Vol 6, No.1, 2011.

[4] Riki, Edy, & Maryanto. (2019). Plagiarim Detection Application Uses Winnowing Algorithm With Synonym Recognition For Indonesia Text Documents. Selangor Science & Technology Review (Sester), 3(1), 35-48.

[5] Purwitasari, D., Kusmawan, P. Y., & Yuhana, U. L. (2011). Deteksi keberadaan kalimat sama sebagai indikasi penjiplakan dengan algoritma hashing berbasis N-Gram. Jurnal Ilmiah KURSOR, 6(1), 37-44.

[6] Wibowo, Adik.2012."Mencegah dan Menanggulangi Plagiarisme di Dunia Pendidikan". Departemen Administrasi dan Kebijakan Kesehatan Fakultas Kesehatan Masyarakat Universitas Indonesia.

[7] Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarisme Detection without Reference Collections. In R. Decker, & H. J. Lenz, Advances in Data Analysis and Classification (pp. 359-366). Springer.

[8] Gipp, B. & Meuschke, N. 2011. Citation Pattern Matching Algorithms For Citationbased Plagarism Detection: Greedy Citation Tiling, Citation Chunking And Longest Common Citation Sequence. Proceedings of the 11th ACM Symposium On Document Engineering, pp.249-258.

[9] Salmuasih. (2013). Perancangan Sistem Deteksi Plagiat pada Dokumen Teks dengan Konsep Similarity menggunakan Algoritma Rabin Karp. Yogyakarta: STMIK Amikom.

[10] Kurniawati, A., Puspitodjati, S., & Rahman, S. (2010). Implementasi Algoritma Jaro-Winkler Distance untuk Membandingkan Kesamaan Dokumen Berbahasa Indonesia. Proceeding Seminar Ilmiah Nasional Komputasi dan Sistem Intelejen - KOMMIT. Bali: Universitas Gunadarma

[11] Liliana, Budhi, G. S., Wibisono, A., & Tanojo, R. (2012). Pengecekan plagiarisme pada code dalam bahasa C++. Jurnal Informatika, 11(1), 70-78.

[12] Alfikri, Z. F., & Purwarianti, A. (2012). The construction of Indonesian-English cross language plagiarism detection system using the fingerprinting technique. *Jurnal Ilmu Komputer dan Informasi*, *5*(1), 16-23.

[13] Schleimer, Saul, Daniel S. Wilkerson, dan Alex Aiken. Winnowing: Local Algorithms for Document Fingerprinting. San Diego: In Proceedings of the ACM SIGMOD International Conference On Management Of Data. 2003

ISSN: 2005-4238 IJAST