Rabin Karp And Winnowing Algorithm For Statistics Of Text Document Plagiarism Detection

1st Dedi Leman, 2nd Maulia Rahman, 3rd Frans Ikorasaki, 4th Bob Subhan Riza, 5th Muhammad Barkah Akbbar
[1,2,3,4,5] *Faculty of Engineering and Computer Science*
Universitas Potensi Utama
Medan, Indonesia
[1]lemhan28@yahoo.com, [2]mazrahman18@gmail.com, [3]iko.maknyos@yahoo.co.id,

***Abstract*—The Karp-Rabin algorithm was created by Michael O. Rabin and Richard M. Karp in 1987 who used the hashing function to find patterns in text strings and the Winnowing Algorithm is a method used to detect common subsequence in two or more texts compared. Two texts are known to have similar words/sentences if fingerprint is found in the document, this fingerprint will be used as the basis of comparison between texts, this algorithm will look for fingerprints (similarity in two texts) by changing the n-gram of a text into a numerical value called a hash value, the technique for finding that value is Hashing This algorithm is widely used in determining the close resemblance in a biology sequence. Plagiarism is an action that must be avoided, but there are still many people who know and understand plagiarism. Apart from preventing, detecting plagiarism is an attempt to reduce plagiarism. The problem of plagiarism that is often found among students is plagiarism in text documents. This study aims to build a plagiarism detection system in text documents using the Rabin-Karp algorithm in a computerised manner. This plagiarism detection system is aiding the action of plagiarism with the similarity of sequences of the two documents compared. The system that is compared is a basic process that can be further developed to build better detection applications for plagiarism.**

Keywords—Plagiarism, Rabin Algorithm, Winnowing Algorithm, Letter Similarity.

I. INTRODUCTION

Humans want to ease in everything. These properties will trigger negative actions when motivated by the motivation to cheat and the low ability of people to create and innovate to create original work. In this case, the negative action referred to as Plagiarism.

Plagiarism is a problem of combustion that academics face at all levels vary from the education system. With the advent of digital contests, the challenge of ensuring the academic integrity of work has been strengthened. Plagiarism uses a string matching algorithm, the Rabin Karp algorithm.

The Rabin Karp algorithm is one type of string matching algorithm that can be used for several pattern searches. To reduce processing time and increase accuracy, before calculating the percentage similarity with the dice coefficient similarity method of the string matching value of Rabin

Karp-algorithm, the text document will pass through 3 preprocessing phases is a folding case, filtering and derived using the Nazief-Adriani algorithm. In the application that has been built, text documents abstract the research publications of lecturers at the Research Institute, University of Bengkulu. The recommendations from the publication or journal are by the value of the similarity percentage of abstract publications or journals accessed by visitors. The system development method used to develop this application uses a waterfall model. In the analysis and design phase uses an approaching structure. Finally, to test the application using two methods, namely feasibility tests and comparing the results of manual algorithm calculations and using applications [1].

The winnowing algorithm calculates the hash values of each program, to find the hash value then the hash rolling function is used. Then a window is formed from the hash values. In each window, a minimum hash value is selected. If there is more than one hash with the minimum value, the rightmost hash value is selected. Then all selected hash values are stored as fingerprints of a document. The input of the document fingerprinting process is a text file. Then the output will be a set of hash values called fingerprint. This Fingerprint will be used as the basis for comparing the similarities between the entered texts [2]

The two ways to overcome the problem of Plagiarism, namely by preventing and detecting. Prevention means keeping or obstructing Plagiarism from being done. Businesses like this must be done as early as possible, especially in the education system and community morals. The detecting means making an effort to find out what plagiarism is doing.

Many institutions and teaching staff apply academic sanctions to plagiarists to reduce plagiarism. The problem is how to find out whether a student does plagiarism or not in making a paper. It is necessary to check the results carefully of the student's writing. Therefore, a computerised system of plagiarism detection is needed.

Related Method

A. Design Of Similarity Code Program Determination System In Language C And Pascal Using Rabin-Carp Algorithm

This application can determine the percentage of similarity of code under test. On the application of

"Similarity Checked", there are 5 stages of a process called preprocessing to prepare the program code before checking the similarity of two documents, that are: case folding, filtering, parsing, k-gram, hashing and two-stage process to obtain the percentage of similarity checking code ie: checking the Rabin-Karp algorithm and determination of the value of similarity calculated by dice similarity coefficients. Testing is done by trying out some of the code C and Pascal programs are different, and by using k-gram values are different. Based on the testing that has been done, showed that the application had been designed to determine the percentage of similarity of code C and Pascal programs were tested [3].

B. *Application Of Winnowing Fingerprint And Naive Bayes Methods For Document Grouping*

Documents that are scattered and not well coordinated will make it difficult for information seekers to obtain the desired information, so a system needs to be created that can group documents. This study applies the winnowing method for selecting features such as fingerprint and naive Bayes for grouping. The winnowing algorithm has fulfilled the need for a document similarity detection algorithm, namely, in matching documents not affected by spaces, fonts, punctuation and other characters, or commonly called with whitespace insensitivity. The winnowing algorithm is used to find the fingerprint generated by the system, so that if the word character that appears as a fingerprint between documents is not the same, then the classification process is irrelevant, the percentage of accuracy obtained is 80% [4].

C. *Detection Of Plagiarism Levels Of The Title Of Title Thesis With Winnowing Algorithm*

Two texts are known to have similar words/sentences if fingerprint is found in the document, this fingerprint will be used as the basis of comparison between texts, this algorithm will look for fingerprints (similarity in two texts) by changing the n-gram of a text into a form a number value called a hash value, the technique for finding that value is Hashing. With this system, the Chair of the Study Program or Final Task Coordinator will only enter the title to be submitted by the student, and then the system will automatically check and display the results. These results can be used as a consideration in decision making and can determine whether the title of the thesis is accepted or rejected. [5].

D. *Anti Plagiarism Application With Algorithm Karp-Rabin At Thesis In Gunadarma University*

This thing is possibly unable to be paid attention by the side of campus because of limitation from some interconnected factors for example student amounts Gunadarma University reaching thousands and incommensurate to tester amounts or lecturer the side of campus in charge directs problem thesis. In this paper, an application has been developed to check and look for 5 type percentage similarity from a thesis with another one at certain part or chapters. Percentage got that is 0%, under 15%, between 15-50%, up to 50% and 100%. So it should be expected that the results could be used by thesis advisor and also thesis examiner from the Student at Gunadarma University.[6]

II. METHOD

A. *Karp-Rabin Algorithm*

At repair of method brute force can be classified to follow the sequence comparison of pattern character and character text for every attempt. At the comparison process, there are four categories [6]:

1. From right to left
2. From left to right-
3. In specific order
4. In any order

Based on the four above categories, algorithm Karp-Rabin included in the category from left to right. Algorithm Karp-Rabin applies function of hash, providing simple method to avoid time complexity $\boldsymbol{O}(m2)$. Than checked position every pattern which there is in text, would more efficiently if done only at pattern wanted. Equality checking between two words applies function of hash. The function of must-have hash properties as follows [6]:

1. The ability of efficient computing
2. High discrimination to string
3. Function of hash (y [j+1 .. j+m]) must easy to be computing from hash (y[j .. j+m-1]) hash (y[j+m])

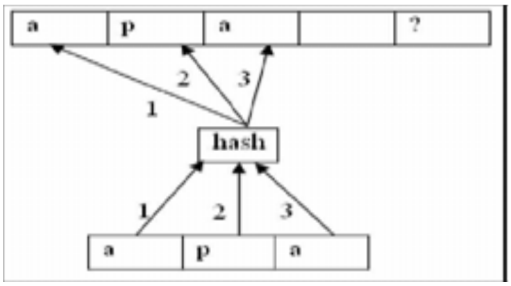

Fig. 1. Illustration of Karp-Rabin Algorithm

Algorithm Karp-Rabin has marked as follows:

1. Applies function of the hash
2. Phase preprocess in time complexity O(m) and constant place.
3. Seeking phase in time complexity O(mn)
4. O(n+m) estimates active time

The function of hash also applied default value index to or key and applied then each time data relating to value or key is taken. In a seeking, hence first time name would be hashing with function of the same hash when saving the data (index) causing yields a value which will be compared to at data is index with the value. Hence in general seeking with ten possibilities (digit 0-9) would be quicker compared to based on 26 possibilities (character a-z)

B. *Winnowing Algorithm*

Winnowing is an algorithm used to process document fingerprinting. Document fingerprinting is a method used to detect the accuracy of copies between documents or only a portion of text. The working principle of this document fingerprinting method is to use the hashing technique. The hashing technique is a function that converts each string to a number. The winnowing algorithm calculates the hash values of each program, to find the hash value then the hash

rolling function is used. Then a window is formed from the hash values. In each window, a minimum hash value is selected. If there is more than one hash with the minimum value, the rightmost hash value is selected. Then all selected hash values are stored as fingerprints of a document. The input of the document fingerprinting process is a text file. Then the output will be a set of hash values called fingerprint. This Fingerprint will be used as a basis for comparing the similarities between the texts that have been entered [7].

C. Implementation Of Karp-Rabin and Winnowing Algorithm

The design of the application that is made is in the form of a system to detect plagiarism of a document. Input in this application is a text document that has an extension .txt. Users will input two documents, namely the original document and the document you want to test. After that, the system will process the two documents and evaluate what the similarity is between the document and how long the process takes. The first time the system does the process is to read the text file entered by the user. From documents entered by the user, the system will check the document so that information will be obtained in the form of words, number of sentences, number of paragraphs and size of the document. After the system gets information from the inputted documents, the system will enter the preprocessing stage. At this stage several processes will be carried out, namely tokenizing, filtering (eliminating unnecessary words) and stemming (cutting words or terms into basic words) The process of filtering is the process of removing words and punctuation that are less important, such as the word ―yang, ― And, that is, spaces, commas and so on. The filtering process used in this system is to use the stopword algorithm where each term will be checked whether the word is in the stopword list. If there is a stopword, the word will be removed so that after the filtering process, a list of unique words will be obtained. After the filtering process processing will be inserted. The process of stemming is a process of cutting off particles such as lah -lah, ,kah, ―- even. Then cut the ownership pronouns like ku- me, ―- you, ―- it. The next step is cutting the affixes such as prefixes and suffixes and confixes (prefixes and endings) in unique words like ―di, ―- even, kankan and so on, so that the basic words will be obtained. Figure 2 is a picture of the document plagiarism checking process carried out by the system.

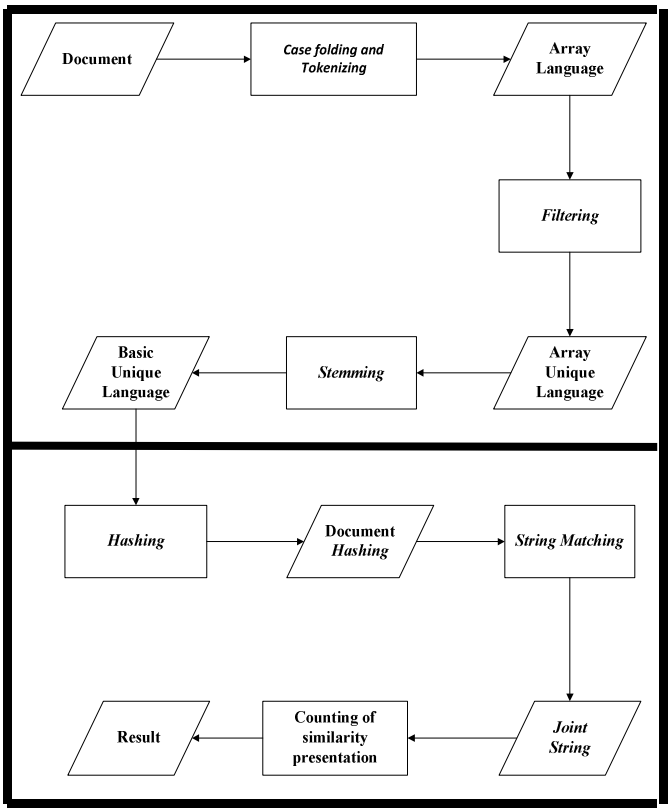

Fig. 2. Flowchart of Karp-Rabin and Winnowing Algorithm

III. RESULT AND DISCUSSION

A case study with Karp-Rabin and Winnowing Algorithm: Winnowing and Karp-Rabin have fulfilled these conditions by removing all irrelevant characters such as punctuation, spaces and other characters so that later only characters in the form of letters or numbers will be processed further. The steps in applying the Winnowing and Karp-Rabin Algorithm are as follows:

Step I Disposal of Irrelevant Characters. Namely the elimination of punctuation, spaces and symbols such as =, #,%, &, (,), -, _, $, @,!, /, ", Such as the example below:

Hello !!! I'm Dedi, how are you beautiful?
Will be changed to
HelloI amhoware you beautiful

Step II Filtering is the stage of taking important words from the results of tokens. Stoplist/stopword is non-descriptive words that can be discarded in the bag-of-words approach. The example of stopwords is ―yang, ―dan, ―di, ―from and so on

HelloIamhow areyou beautiful
Will be changed to
HelloIamhow you beautiful

Step III The stemming stage is the stage of finding the root words of each word filtering results. At this stage, the process of returning various forms of words is carried out into the same representation. The stemming process is used to deal with the problem of the word passive-active and change in word particles.

Step IV Hashing is a way to transform a string into a fixed-length unique value that serves as the string marker, Tokenizing results, filtering and stemming:

TABLE I. RESULTS OF SUBSTRING AND HASHING TEST DOCUMENTS

No	Substring	Hashing
1	hello	
2	iamho	
3	wyoub	
4	eauti	
5	ful	

Test document:

HelloIamhow you beautiful

Authentic document :

HelloIamhow you beautiful

Pattern = "hello"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (107000+9700+1150+117) mod 100
= 117967 mod 100
= 100

Pattern = "iamho"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 101
= 126677 mod 101
= 23

Pattern = "wyoub"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 100
= 126677 mod 100
= 65

Pattern = "eauti"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 100
= 126677 mod 100
= 87

Pattern = "ful"
Hashing = $[115*10^2)+(117*10^1)+ (117*10^0)]$ mod 100
= (10400+1170+107) mod 100
= 126677 mod 100
= 103

TABLE II. RESULTS OF SUBSTRING AND HASHING AUTHENTIC DOCUMENTS

No	Substring	Hashing
1	hello	
2	iamho	
3	wyoub	
4	eauti	
5	ful	

Pattern = "hello"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (107000+9700+1150+117) mod 100
= 117967 mod 100
= 100

Pattern = "iamho"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 101
= 126677 mod 101
= 23

Pattern = "wyoub"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 100
= 126677 mod 100
= 65

Pattern = "eauti"
Hashing = $[(107*10^4)+(97*10^3)+(115*10^2)+(117*10^1)+(117*10^0)]$ mod 100
= (115000+10400+1170+107) mod 100
= 126677 mod 100
= 87

Pattern = "ful"
Hashing = $[115*10^2)+(117*10^1)+ (117*10^0)]$ mod 100
= (10400+1170+107) mod 100
= 126677 mod 100
= 103

This calculation is done on all parsing results so that all substrings have a hash value. The same thing is also done on the training document and then it will be matched to the hash value of each substring on the training document with the test document. Then count the number of substrings found. After that the similarity value will be calculated using the Dice's Similarity Coeeficient formula.

Step V string Matching After forming the hash value, string matching will be done. The function used for string matching is string Matching.

Step VI Calculate Similarity Phase After performing the string matching process, the similarity value is calculated. The percentage calculation of the similarity value is found in the similarity () function. The similarity function above is the implementation of the Dice's Similarity Coefficient. This function is used to determine the similarity value between

the two documents tested. It can be calculated using the formula: S = C / (A + B) * 100%
Information :
A and B = Number of Hash Schemes in documents A and B
C = Number of Hash Schemes in documents A and B

TABLE III. SIMILARITY VALUES

Parsing N=3	Mod 100
Test Documents (A)	270
Authentic Documents (B)	270
Authentic and Authentic Documents (C= 2XA)	540
Formula : S = C/(A+B)*100%	100%

IV. CONCLUSIONS

As a closing for writing this thesis, some things that can be used as conclusions, including:

1. A system has been created that can be used to detect plagiarism against text documents using the Rabin-Karp algorithm.
2. This plagiarism detection application is designed using language
3. The use of stemming influences the accuracy of the similarity value produced. Using stemming produces a value that tends to be less good than without using stemming. But in certain cases such as changing the sentence form of Rabin-Karp's algorithm, which is inserted into stemming, it produces better accuracy of similarity values.
4. This plagiarism detection application is used to calculate plagiarism calculations based on the level of similarity of letters, numbers and symbols per character.

REFERENCES

[1] Oetsch J., Pührer J., Schwengerer M., Tompits H. The System Kato: Detecting cases of plagiarism for answer-set programs. *Theory and Practice of Logic Programming*. 2010; 10(4-6): 759-775.

[2] Purwitasari D., Kusmawan P.Y., Yuhana U.L. Sentence Detection is the same as an Indication of plagiarism with an N-Gram-based Hashing Algorithm. Cursor. 2011; 6 (1).Qamaruzzaman, M.Haris et al. (2016). "Expert system for diagnosing eye diseases in humans using Bayes theorem". STMIK Palangkaraya, Volume: 5, Number 4, ISSN: 2302-5700, Palangkaraya.

[3] Kusumadewi, S., 2003. Artificial Intelligence (Techniques andApplications). Yogyakarta: GrahaIlmu

[4] Anna Kurniawati1, K. A. S., I Wayan Simri Wicaksana3 (2012). "ARCHITECTURE FOR APPLICATION OF DETECTION OF INDONESIAN DOCUMENT FUNCTIONS." 2012 National Information System Conference.

[5] Leman, D., Nurcahyo, GD., Defit, S., 2015, Application Of Plagiarism Detection Statistics Document Text With Rabin Carp Algorithm, Seminar Nasional Informatika

[6] Leman, D., Andesa, Khusaeri., 2015, Implementation Of Vector Space Models To Improve Quality In Library Book Search System, Seminar Nasional Informatika

[7] Haryanto, EV. 2015. Analysis of Wireless Computer Network Security Problems Using Cain, CSRID (Computer Science Research and Its Development Journal) 6 (1), 43-52

[8] Journal of Informatics Engineering, Vol 1 2012, Application of Asthma Disease Diagnosis Using Bayesian Network Web-Based, Yunita Nancy Roselina, Sugeng Purwantoro, and Memen Akbar.

[9] Kadir, Abdul, 2003. Practical Guide Learning Database Using Microsoft Access, Andi, Yogyakarta.

[10] Windah Supartini (2016), Web-Based Expert System With Forward Chaining Method In Early Diagnosis of Tuberculosis Disease in East Java, Vol. 1, No. 3, November 2016: 147-154, ISSN: 2503-2259; E-ISSN: 2503-2267

[11] Adi Radili, 2017. "Application of Winnowing Fingerprint and Naive Bayes Methods for Grouping Documents", CoreIT Journal, Vol. 3, No.2, December 2017, ISSN 2460-738X (Print), ISSN 2599-3321 (Online)

[12] Nur Alamsyah, 2017. "Plagiarism Detection of Thesis Title Similarity with Winnowing Algorithm", "Technologia" Vol. 8, No.4, October - December 2017

[13] Nur Alamsyah, 2017. "Comparison of Winnowing Algorithms with Rabin Karp Algorithm to Detect Plagiarism in Similar Thesis Title Texts", "Technologia" Vol. 8, No.3, July - September 2017

[14] Nur Fadillah Ulfa, 2016. "Making Web-Based Measurement of Similarity Levels Using Applications Using Winnowing Algorithms", Information Journal and Computer Volume 21 No. 3, December 2016