Detection of Text Similarity for Indication Plagiarism Using Winnowing Algorithm Based K-gram and Jaccard Coefficient

Eva Y Puspaningrum¹, Budi Nugroho², Ariyono Setiawan³, Nuraini Hariyanti⁴
¹²,⁴ Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur
³ Aviation Polytechnic of Surabaya

Email: evapuspaningrum.if@upnjatim.ac.id

Abstract. One of the digital data is a document. Documents can be easily copied and deleted. Anyone can retype or copy parts of the document. In this paper will detect text similarity. The more similarity of words there is the more indicated the document is plagiarism. Winnowing algorithm performs the calculation of hash values of each k-gram. This method improves the search time with more accuracy in the detection process. All data selected hash values will be fingerprints of a document. Fingerprint will be used as a basis for comparing similarities between text data. The fingerprint value of the winnowing process for each document will be matched by using the Jaccard Coefficient to measure the similarity of the text. In this paper results show that the adjustment of the k-gram and window values can affect the final result of the similarity percentage value. The smaller the k-gram value, the greater the percentage value.

Keywords. Text Similarity, Winnowing, K-Gram, Jaccard Coefficient

1. Introduction

One component that is in digital data is a text document. Documents in digital form make it easy to save, search, but very easy in terms of plagiarism. Documents can be easily copied and deleted. Anyone can retype or copy parts of the document. There are many problems in the detection of plagiarism, because plagiarism does not always contain all the clear documents or paragraphs. There are several ways for plagiarism, for example by copying smaller pieces of content, which can then be changed to make it very difficult to detect [1]. Document is easy to manipulate and distributed. That is also easy to search and obtain information could lead to terrible things such as: plagiarism [2].

Plagiarism detection identifies documents that may be traced from source documents. The system will analyze the level of similarity between texts and if the level of similarity between the two documents is high, then the method of reporting the two documents is suspicious of plagiarism [3]. Currently, there are several free and commercial applications to detect plagiarism such as: Turnitin.com, Plagiserve.com, Word Check Keyword [4] and Plagiarism Checker.

Steven has conducted research on detection plagiarism using the Rabin-Karp algorithm [5]. Rabin-Karp is a search algorithm that searches for substring patterns in text using the hashing method and k-gram. K-gram is a series of terms with length k. The method is applied to form words or characters. This method continues to read from the source text to the end of the document[6]. The algorithm works well, but will produce a relatively long computational time if the text contains many unequal hashing comparisons [7].
Broder et al. [8] proposed a method that generates new fingerprints by hashing the order of longer fingerprints and determining the similarity of documents. There are two categories of algorithms in plagiarism detection. That is an algorithm that analyzes the structure of text grammar and an algorithm that analyzes text fingerprints [9]. Using grammar-based method to detect verbatim copying can get better results than using it to detect the copied text or rewriting [10]. In this paper use fingerprint-based, because data used text of the document. There are many methods used to detect document text plagiarism. One of them is algorithm that text fingerprints. This approach is to detect plagiarism by dividing documents into a set of 'fingerprints'. A set of fingerprints contains pieces of text that might overlap one another. The fingerprint is then used as a request to search the database, to estimate the level of similarity with other documents [11]. The Winnowing algorithm in this paper is a way to detect the existence of the same sentences or often referred to as common subsequence problems. Winnowing algorithm uses the document fingerprint approach. Documents are summarized by a small group of character sequences called fingerprints which can be efficiently used to find copies of parts of a document in large documents [12]. The plagiarism detection process uses which Winnowing algorithm the output is a set of hash values obtained through the k-gram method. K-gram is a series of k-items sequentially from a data, usually in the form of text. The value of k can vary depending on needs, ranging from one to as long as the existing text. The k-gram method is a method that functions to break a word or sentence into a series of length n characters. The Winnowing algorithm uses the K-Gram d method to improve the accuracy of plagiarism checkers. The results are very satisfying especially for plagiarized articles which have experienced a lot of vocabulary changes. In addition, the Winnowing algorithm also shows good performance [13]. The fingerprint value of the winnowing process for each document will be matched by using the Jaccard Coefficient to measure the similarity of the text. Jaccard coefficient to measure the level of general similarity based on term-matching which is good on plagiarism with no-rewrite and low obfuscation, and produces high precision performance [14].

2. Methodology

Document texts contain usually lots of noise and uninformative [15]. There are several steps that are carried out in this process. Pre-processing include case folding, tokenizing, filtering, and stemming. Case folding stage is the stage of changing each letter, where capital letters into lowercase letters. Tokenizing stage is the stage of cutting the input string based on each of its constituent words. In this process, the input character cuts into symbols, punctuation marks, or other elements that have meanings called tokens. Filtering is the removal of terms or words that are judged to have no meaning or irrelevant, usually called the stop-word process. Stemming process transform the words contained in a document into the root words. The data used is abstract from Indonesian language journals. Input for this system is abstracts in Indonesian journals. Data testing is abstract to be compared with data sets in the database. Comparative data is data that has a similar theme. This is done so as to find out the percentage of similarity of the text data. Figure 1 shows a general description of this system. There are several processes for Winnowing algorithm yield a percentage of similarity. The series of processes must be carried out in order. At each stage produces a value as a result of the process, which will then be processed and reused in the next process. Figure 2 explains the stages of the winnowing algorithm to detect text similarities. Stages begin with text that is first processed to produce basic words. Next is the k-gram process for breaking a character into a series of grams with a number of n-gram values. The hash value is obtained from the hashing process, which is changing a series of grams into a value. The hash value functions as a window forming a number of values w.
A series of hash values in a window is used to find the smallest hash value in each window called the fingerprint process. After that the process continues with the calculation of the percentage of similarity based on the fingerprint value that has been obtained. The percentage of similarity is calculated using the Jaccard similarity coefficient method.

Step 1: Define data. The data used as testing material is in the form of text. The text which used to obtain a percentage value, can be seen in Table 1.

| Table 1. Data |
|---------------|
| Testing Data | Comparison data |
| Teknik Informatika | Informatika |
| Informatika | Teknik Mesin |

Step 2: K-gram. After pre-processing, continued with the character separation process, the k-gram process, can be seen in Table 2. For example the k-gram value used is 3.
Step 3: Hashing. This process of forming the hash value obtained from each gram is calculated using the ascii table. From Table 2 the results obtained from the hashing calculation process using the equations (1) and (2) can be seen in Table 3.

Step 4: Windowing. The hash value is divided into groups according to the number of window values specified. The number of windows used is 5 can be seen in Table 4.

Step 5: Fingerprint. The hash value is selected in each window to find the smallest value, if there is the same value then select the value to the right. The process in step 5 can be seen in Table 4 and the results of the smallest hash value of each window can be seen in Table 5. In this process later it will be used in the fingerprint step to find the same value between the two texts compared to obtain a similarity value.
Step 6: calculation of text similarity. Calculation to obtain the percentage value of text similarity using equation (3) is the Jaccard Coefficient formula.

**Test Data and Comparative Data:**

a. Similarity = \( \frac{2}{5} \times 100\% = 40\% \)

b. Similarity = \( \frac{2}{6} \times 100\% = 33\% \)

The results obtained are comparative data that have similarities with test data by 40%.

### 3. Result and Discussion

Testing is conducted by comparing abstracts with several abstracts that have the same theme. For example a scientific work with the theme "Naive Bayes" then the comparison is also a similar journal theme. The trial was conducted with several test scenarios, namely k-gram and 2,3,4,5,8,9,10 window 2,3,4,5,8,9,10. The average percentage value can be seen in table 6.

| k-Gram | Average percentage similarity (%) |
|--------|----------------------------------|
| 2      | 40.023                           |
| 3      | 23.898                           |
| 4      | 14.307                           |
| 5      | 14.085                           |
| 8      | 11.901                           |
| 9      | 11.517                           |
| 10     | 11.229                           |

The test results in Table 6 show the average similarity value of 7 different test scenarios for k-gram and window. It can be concluded that the tested data have low plagiarism level values below 50%. From the above experiments it can be observed that this value is a general detection of frequently used words appears in the abstract with a similar theme. Even though text filtering has been done this can still happen because preprocessing is not easy because there are several syntax writing formats that are still passed from the filter so that the detected object becomes irrelevant. But to detect plagiarism textually the winnowing algorithm is quite effective because it can detect plagiarism like copy paste and relocate words. The adjustment of k-gram and window on the winnowing algorithm has a big influence in finding percentage values.

Winnowing algorithm has a k-gram and window process can be changed It can be concluded that the higher the k-gram value, the lower the similarity value produced and the lower the k-gram value, the higher the similarity value.

But that does not mean that a low k-gram value will provide an accurate accuracy value. The smaller the k-gram value, the smaller the characters will be matched and the more often these characters will be found in the text.

### 4. Conclusion

Winnowing algorithm can be able to detect plagiarism in a fairly fast time. Textually, this algorithm is very effective in dealing with copy paste and relocation plagiarism. The weakness of this algorithm is that it cannot provide guarantees and evidence of plagiarism found. The choice of k-gram value will affect the similarity value. A small k-gram value will give a greater similarity value the smaller the substring, the smaller the character will be affect the hash value and can give a fingerprint value
identical so matching is irrelevant. The use of this algorithm to detect plagiarism requires further checking of the file to be processed, because it requires a text file without writing errors.

References

[1] Khalid Shams, 2010 Plagiarism Detection using Semantic Analysis. Dhaka: BRAC University.
[2] Y.-T. Liu, H.-R. Zhang, T.-W. Chen, and W.-G. Teng, “Extending Web search for online plagiarism detection,” in Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on, 2007, pp. 164–169.
[3] L Sindhu, Sumam Mary Idicula, 2016 Fingerprinting based detection system for identifying plagiarism in Malayalam text documents International Conference on Computing and Network Communications (CoCoNet) IEEE Published Februari 2016
[4] C. K. Kent and N. Salim, “Web based cross language plagiarism detection,” in Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on, 2010, pp. 199–204.
[5] S. Steven, “Perancangan Program Aplikasi Pendeteksian Plagiarisme Dokumen Berbasis Teks Menggunakan Algoritma Rabin-Karp.” BINSUS, 2009.
[6] Sutoyo, R., Ramadhani, I., Ardiatma, A. D., Banana, S. C., Warnars, H. L. H. S., Trisetyarso, A., … Suparta, W. (2017). Detecting documents plagiarism using winnowing algorithm and k-gram method. 2017 IEEE International Conference on Cybernetics and Computational Intelligence
[7] Z. A. Khan and R. K. Pateriya, “Multiple pattern string matching methodologies: a comparative analysis,” Int. J. Sci. Res. Publ., vol. 2, no. 7, pp. 1–7, 2012.
[8] Mohamed Elkhidir, Mohannad M, Ibrahim, Tarig A Khalid, Shawgi Ibrahim, Mohamed Awadalla, 2016 Plagiarism detection using free-text fingerprint analysis World Symposium on Computer Networks and Information Security 2015 (WSCNIS) IEEE Published January 2016
[9] D Zou, W Long, Z Ling , 2010 A cluster-based plagiarism detection method Notebook Papers of CLEF 2010 Lab Report for PAN at CLEF 2010
[10] Oetsch J, Pührer J, Schwengerer M, and Tompits H, 2010 The System Kato: Detecting Cases of Plagiarism for Answer-Set Programs Theory and Practice of Logic Programming, 10(4-6), pp 759-775, 2010.
[11] Agung Toto Wibowo , Kadek W Sudarmadi , Ari M. Barmawi , 2013 Comparison Between Fingerprint and Winnowing Algorithm to Detect Plagiarism Fraud on Bahasa Indonesia Documents International Conference of Information and Communication Technology (ICoICT)
[12] Shuai W Ang, Haoliang Qi, Leilei Kong, Cuixia Du2, 2013 Combination Of Vsm And Jaccard Coefficient For External Plagiarism Detection Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013
[13] B Stein, S Eissen, 2006 Fingerprint-based Similarity Search and its Applications. Advances in Information Retrieval 28th European Conference on IR Research (pp. 565–569). London: Bauhus University Weimar
[14] Emma Haddia , Xiaohui Liu , Yong Shib, 2013 The Role of Text Pre-processing in Sentiment Analysis International Conference on Information Technology and Quantitative Management, Published by Elsevier, Procedia Computer Science 17 (2013 ) 26 – 32
[15] C Ramasubramanian, R Ramya, 2013 Effective PreProcessing Activities in Text Mining using Improved Porter’s Steming Algorithm International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013, ISSN (Online) : 2278-1021.