Comparison Between Fingerprint and Winnowing Algorithm to Detect Plagiarism Fraud on Bahasa Indonesia Documents

Agung Toto Wibowo
Faculty of Informatics
Telkom Institute of Technology
Bandung, Indonesia
atw@ittelkom.ac.id

Kadek W. Sudarmadi
Faculty of Informatics
Telkom Institute of Technology
Bandung, Indonesia
wispukako@gmail.com

Ari M. Barmawi
Faculty of Informatics
Telkom Institute of Technology
Bandung, Indonesia
mbarmawi@melsa.net.id

Abstract— **Plagiarism detection has been widely discussed in recent years. Various approaches have been proposed such as the text-similarity calculation, structural-approaches, and the fingerprint. In fingerprint-approaches, small parts of document are taken to be matched with other documents. In this paper, fingerprint and Winnowing algorithm is proposed. Those algorithms are used for detecting plagiarism of scientific articles in Bahasa Indonesia. Plagiarism classification is determined from those two documents by a Dice Coefficient at a certain threshold value. The results showed that the best performance of fingerprint algorithm was 92.8% while Winnowing algorithm's best performance was 91.8%. Level-of-relevance to the topic analysis result showed that Winnowing algorithm has got stronger term-correlation of 37.1% compared to the 33.6% fingerprint algorithm.**

Keywords— Fingerprint, Winnowing, Plagiarism Detection, Dice Coefficient

I. INTRODUCTION

Recently, plagiarism cases —especially in scientific writing— were on the news. This can bring down public's credibility towards quality of research in an institution where individuals involved. Due to fast developments of information technology, any kind of information can be found easily through the internet. These information are widely abused by various parties which lead to plagiarism.

Development of algorithm to detect plagiarism fraud has actually been proposed by other researchers. Some of the conventional approach to the detect of plagiarism indications including character-based, structural-based, cluster-based, syntax-based, cross-language based, and semantic-based [8].

Furthermore, there are some character based algorithm [8] such as:

a. Fingerprint. Using parts of document to be checked to plagiarism. The checked parts will be proceed by certain mechanism (character-matching) to determine the fraud. There are two algorithms based on fingerprint which are fingerprint algorithm and winnowing algorithm [1, 9, 10].

b. String similarity. Using this approach overlapped document will be found using string-matching and sentence-matching algorithm. This approach was developed and applied to COPS (copy protection system) [2]. In addition, detection system is also developed using Longest Common Subsequence (LCS) [3], Levenshtein Distance and also Smith-Waterman [13].

c. Structural-based methods. While the two previous approaches above focused on exploiting lexical features of a document, structure-matching method focused on structural things such as header, section, paragraph, and reference [8]. Tree-structure feature with POS tagging is also used in this study. Other approaches such as cluster-based, syntax-based, cross language-based and semantic-based also used by other researcher [8].

This paper is focusing on the use of fingerprint (fingerprint and winnowing algorithm) in document-matching. Moreover, at the end of the section, characterization of fingerprint and winnowing algorithm on plagiarism detection in Bahasa Indonesia documents is also stated. Characterization here means linkage-level of selected term as a fingerprint to the document's topic.

This research used 112 documents. Fraud documents are built by using paraphrase process with a 20%-80% rate and conversing active-passive voice.

The plagiarism process started from document pre-processing stage. Pre-processing steps are applied to [7]: case folding, parsing, filtering, stemming[11], and tokenizing. Document is then proceeded to obtain sequence N-Gram token based. This N-Gram token based is used on fingerprint and winnowing algorithm as an input to get certain keyword of a document.

Determination of plagiarism is done by dice coefficient[6]. At the end of the process, detection accuracy is counted. Even more, resulted keyword will be analyzed to specify linkage-level of selected term as topic's keyword.

II. BASIC ALGORITHM

This study focused on the implementation of fingerprint and winnowing algorithm. Both algorithms are used to check a

978-1-4673-4992-5/13/$31.00 ©2013 IEEE

fingerprint-technique plagiarism. The basis of the algorithms are described as following.

A. *Fingerprint Algorithm*

Fingerprint search [9] in a document can be done by using a hash function in N-Gram of a document. N-Gram is a substring of k length by merging symbols/tokens in a document. For instance, if there is a string s1 = “very intriguing”, then N-Gram of s1 can be obtained by eliminating irrelevant symbols/features. By eliminating space as an irrelevant feature, the process will get s1’ = “veryintriguing”.

From this s1’, if k = 3 is chosen, there will be some N-Gram = {‘ver’, ‘ery’, ‘ryi’, … ‘ing’}. Sub-string resulted from those N-Gram will then be inputted into hash function. Particular hash value is selected which meet a hash(N-Gram) condition modulo p = 0. Further explanation can be seen in the illustration on figure 1 [9]:

A do run run run, a do run run

(a) Some Text.

adorunrunrunadorunrun

(b) The text with irrelevant features removed.

adoru dorun orunr runru unrun nrunr runru unrun

nruna runad unado nador adoru dorun orunr runru

unrun

(c) The sequence of 5-grams derived from the text.

77 72 42 17 98 50 17 98 8 88 67 39 77 72 42 17 98

(d) A hypothetical sequence of hashes of the 5-grams.

72 8 88 72

(e) The sequence of hashes selected using 0 mod 4.

Figure 1. Ilustration Fingerprint Algorithm [9]

There are some advantages and disadvantages in the fingerprint-searching process written on Figure I; Advantage from that fingerprint algorithm is document-matching and document-storage can only be done in fingerprint feature only. This resulted partial copy of document can be detected.

One of the disadvantages from fingerprint algorithm is when hash(N-Gram) mod p = 0 can find neither fingerprint. When this happens, document-matching process was as if a vain since there are no chosen fingerprint. Therefore, winnowing algorithm is created—to complete.

B. *Winnowing Algorithm*

Winnowing algorithm utilized fingerprint concept with the addition of window concept. From this window, hash value chosen will be the minimum fingerprint value. When there are more than 2 values, smallest hash value on the rightmost will be chosen. Steps a-d on winnowing algorithm are similar to the previous illustration on Figure 1. Different steps of winnowing algorithm are [9]:

(77, 74, 42, **17**) (74, 42, 17, 98) (42, 17, 98, 50)
(17, 98, 50, 17) (98, 50, **17**, 98) (50, 17, 98, **8**)
(17, 98, 8, 88) (98, 8, 88, 67) (8, 88, 67, 39)
(88, 67, **39**, 77) (67, 39, 77, 74) (39, 77, 74, 42)
(77, 74, 42, **17**) (74, 42, 17, 98)

(e) Window of hashes of length 4.

17 17 8 39 17

(f) Fingerprints selected by winnowing.

[17,3] [17,6] [8,8] [39,11] [17,15]

(g) Fingerprint paired with 0-base position information.

Figure 2. Illustration Winnowing Algorithm [9]

Chosen fingerprint (both in fingerprint and winnowing algorithm) is then calculated by local similarity. If we have 2 documents A and B, this local similarity will firstly look for a specific fingerprint from both documents. A’s fingerprints can be denoted by h(A) while B’s h(B). Values of |h(A)| and |h(B)| are the cardinalities from document’s fingerprint which later will be used to count the local similarity [1].

$$\varphi local(A,B) = \frac{|h(A)\cap h(B)|}{|h(A)\cup h(B)|} \qquad (1)$$

III. DESAIN

This study conducted a modification in hash function. The process used MD5 function on pre-proceed terms/sequences. Local similarity will not used in this study. Dice coefficient is used to determine plagiarism.

A. *Preprocessing Document*

Before a document is matched, it has to go through preprocessing steps. Preprocessing steps used in this study are: Case Folding, Parsing, Filtering, Stemming, Tokenizing

Case folding: All letters have to be case-insensitive. As for this study, all letters are changed to lowercase [7]. Parsing process is needed so that textual contents of each paragraph are obtained. This process resulted a term in the article [7].

Filtering also called stop word removal. Important terms from previous process are decided. Unimportant terms are discarded so that stop list can be made[7].

Steming. This step reduced words into stem. For example, ‘invented’, ‘inventor’, ‘invention’ will be reduced and changed into their stem, invent. Stemming algorithm used in this study is Porter stemmer for Bahasa Indonesia [11]. Tokenizing step changed words into IDs. These IDs are called tokens and later are used as symbol in matching process [7].

B. *Dice Coefficient*

Matching process is done on the resulted fingerprint. This study did not use local similarity to calculate similarity level. Dice coefficient is used instead.

Calculation of document similarity between document one and two is counted with such Dice coefficient [6] :

$$Dice(A,B) = \frac{2 \times |ngram(A) \cap ngram(B)|}{|ngram(A)| + |ngram(B)|} \quad (2)$$

C. Fingerprint Block Diagram

In this fingerprint block diagram explain the N-Gram token based and MD5 position in the whole system. Fingerprint of a document is selected from a divisible-by-a-certain prime-number hash value. This fingerprint is kept and later matched with another document's fingerprint. The bigger fingerprint-similarity level is, the stronger plagiarism fraud is. Figure 3. Fingerprint Algorithm Block Diagram explains the flow process.

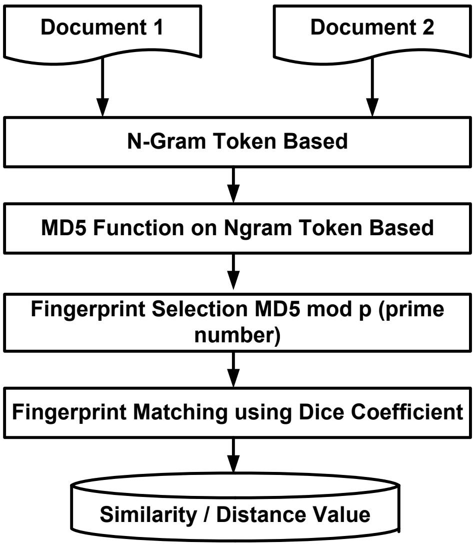

Document 1
Document 2
N-Gram Token Based
MD5 Function on Ngram Token Based
Fingerprint Selection MD5 mod p (prime number)
Fingerprint Matching using Dice Coefficient
Similarity / Distance Value

Figure 3. Fingerprint Algorithm Block Diagram

D. Winnowing Block Diagram

Winnowing block diagram is almost similar to the fingerprint block diagram. A complete block diagram for winnowing algorithm can be shown at Figure 4 Winnowing Algorithm block diagram.

Winnowing algorithm will used a certain size of window. Each window resulted term as fingerprint which later stored for use in document matching process. The bigger fingerprint-similarity level is, the stronger plagiarism fraud is.

IV. Dataset And Testing

Trials of both algorithms are done on the derivative dataset. Quantitative comparison between fingerprint and winnowing algorithm are done using accuracy measurement. Classification of term-correlation level to the topic in an article is also done in this study.

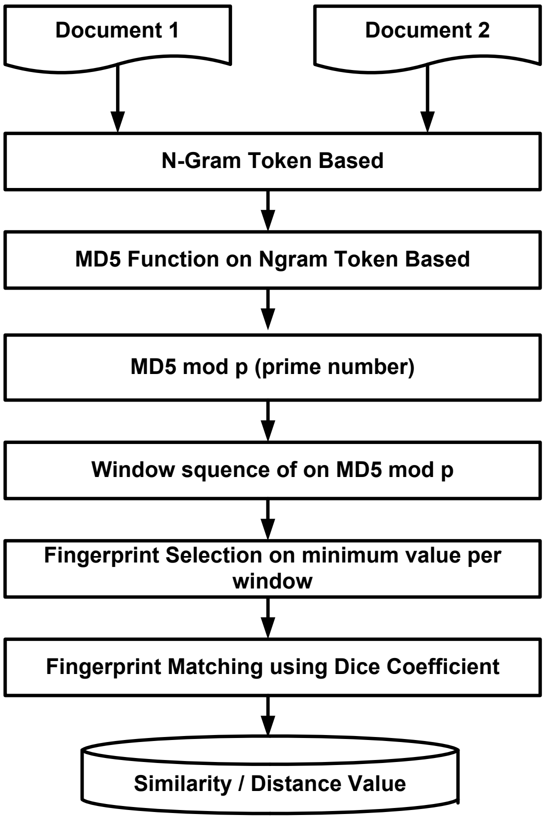

Document 1
Document 2
N-Gram Token Based
MD5 Function on Ngram Token Based
MD5 mod p (prime number)
Window squence of on MD5 mod p
Fingerprint Selection on minimum value per window
Fingerprint Matching using Dice Coefficient
Similarity / Distance Value

Figure 4. Winnowing Algorithm Block Diagram

A. Dataset

This study used 112 titles of faculty of informatics' final projects. Fraud documents are built gradually by paraphrasing 32 titles with the following mechanism:

a. Original datasets as part of fundamental theory in the final project is randomly chosen. To fulfill a good plagiarism quality, each document is only taken as many as 2 pages for the plagiarism.

b. The copy-paste document of plagiarism is added 20% of sentences from original document (a). This sentence-adding process can be done by several forms: (1) In the beginning, (2) In the end, (3) In the middle.

c. Paraphrasing the document (20-80%) is constructed on the copy-paste document (b). This paraphrasing process is built gradually from the lowest 20% and increased by 10% on an ongoing basis until 80%.

d. Active-passive voice change is conducted on the overall content of document (a).

A total of 80 documents are still treated as original documents with no plagiarism done. Resume of the dataset test is shown on the Table I Testing Dataset:

TABLE I. TESTING DATASET

No	File Category	Quantity
1	Non-Plagiarism Document	80
2	Original Document	32
3	Sentence Adding by 20%	32
4	20% of paraphrase	32
5	30% of paraphrase	32
6	40% of paraphrase	32
7	50% of paraphrase	32
8	60% of paraphrase	32
9	70% of paraphrase	32
10	80% of paraphrase	32
11	Changing the Active-passive voice	32
Total Document		**400**

B. *Fingerprint Algorithm Test Analysis*

In fingerprint algorithm, trial parameter is conducted on N-Gram and threshold value. Number of tested N-Gram were [2, 3, 4, 5] while on threshold value were [0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9].

TABLE II. ACCURACY IN FINGERPRINT ALGORITHM

Threshold	N-Gram			
	2	3	4	5
0.05	81.0%	89.3%	90.5%	68.9%
0.10	87.5%	92.8%	90.5%	70.3%
0.15	90.8%	89.3%	90.5%	68.2%
0.20	91.0%	88.0%	86.8%	66.8%
0.25	89.8%	89.8%	82.5%	61.8%
0.30	90.0%	82.3%	79.8%	58.4%
0.35	87.8%	77.8%	75.3%	52.9%
0.40	85.8%	73.8%	69.5%	49.2%
0.45	83.3%	70.0%	64.3%	45.3%
0.50	74.0%	65.5%	58.3%	40.8%
0.55	70.8%	59.3%	53.3%	35.3%
0.60	63.8%	54.3%	46.8%	30.5%
0.65	58.0%	48.5%	41.5%	23.9%
0.70	52.3%	43.3%	38.3%	19.7%
0.75	45.8%	36.0%	33.0%	14.7%
0.80	37.5%	29.3%	28.5%	9.2%
0.85	31.3%	25.8%	27.8%	5.5%
0.90	24.3%	24.0%	24.3%	3.4%

Trial's performance is calculated by using accuracy value. This is applied in the system and obtained by adding the plagiarism detected document with the non-plagiarism document. The result is then divided by total existent document. The resume from this fingerprint algorithm trial is shown in Table II. Accuracy in Fingerprint Algorithm.

From the experiment result in Table II, fingerprint algorithm gave best performance at 92.8% of accuracy. This value was obtained from 0.1 threshold value and N-Gram score of 3. The graph of data from Table II is shown in Appendix A.

From the trial result, even though highest accuracy was obtained from N-Gram score of 3, N-Gram score of 2 (bigram) showed a more stable performance. This bigram give accuracy value more than 80% in threshold range 0.05-0.45. The value of n-grams are used in the subsequent analysis is 2.

C. *Winnowing Algorithm Test Analysis*

In winnowing algorithm, trial parameter will be performed on the N-Gram, window size and threshold values. The numbers of N-Gram tested were [2, 3, 4, 5], number of the window size was [4, 6, 8, 10, 12] while the threshold value were [0.05, 0.1, 0:15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9].

TABLE III. ACCURACY MEASUREMENT ON WINNOWING ALGORITHM (N-GRAM = 2)

Threshold	Window Size				
	4	6	8	10	12
0.05	83.3%	82.3%	81.3%	79.5%	62.8%
0.10	88.0%	87.3%	87.3%	84.5%	68.1%
0.15	91.5%	91.8%	91.5%	88.0%	71.8%
0.20	91.3%	91.0%	91.0%	88.3%	72.0%
0.25	91.8%	91.8%	91.0%	88.3%	72.0%
0.30	91.3%	90.8%	89.8%	88.0%	71.8%
0.35	88.8%	87.5%	86.0%	85.3%	68.9%
0.40	84.0%	82.5%	81.8%	79.8%	63.1%
0.45	80.0%	78.3%	78.5%	74.5%	57.5%
0.50	75.5%	72.5%	72.8%	68.5%	51.2%
0.55	71.5%	67.8%	65.8%	64.3%	46.7%
0.60	65.3%	62.3%	61.8%	57.3%	39.3%
0.65	57.5%	55.3%	55.3%	50.8%	32.5%
0.70	50.0%	49.0%	48.3%	46.3%	27.7%
0.75	42.3%	41.5%	40.5%	38.0%	19.0%
0.80	37.3%	35.5%	34.8%	31.3%	11.9%
0.85	38.8%	26.5%	27.0%	26.8%	7.1%
0.90	21.5%	20.5%	21.0%	23.0%	3.2%

In this study, the accuracy's average produced by threshold combination was analyzed. Accuracy's averages were grouped depends on N-Gram value which can be seen in Appendix B.

From the trials, it is shown that N-Gram score of 2 (bigram) had a better performance than any other N-Gram values. As for the window-size accuracy, score of 4 gave the best performance among all. Therefore, the study focused on N-Gram = 2 and window-size = 4. Table III showed accuracy result of N-Gram = 2.

From Table III, it can be stated that stable condition of winnowing algorithm had the same value as fingerprint algorithm, which is N-Gram = 2 with threshold range 0.05-0.45. This condition leads to more than 80% accuracy level. Figure 5 indicates comparison result from both algorithms on stable condition.

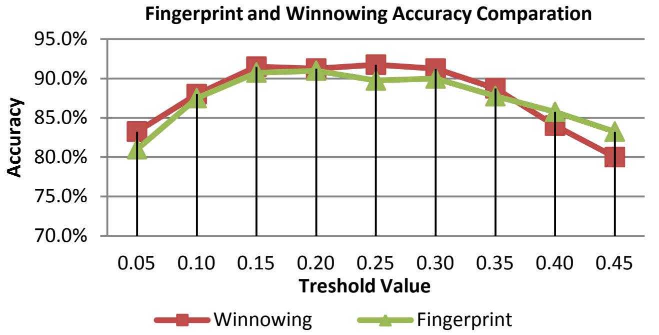


Figure 5. Fingerprint (N-Gram = 2) and Winnowing (N-Gram =2, Window Size = 4) Accuracy Comparation

Figure 5 show the accuracy of winnowing algorithm is better than fingerprint algorithm.

D. Term (Words) Analysis

Analysis of term (words) is done by regarding the chosen term in each algorithm. Words are sorted by frequency of occurrence from the largest to the smallest. 30% of the frequent terms are selected to be analyzed. The selected term will then be determined whether it is a topic-specific word or just commonly-used word.

TABLE IV. TERM ANALYSIS COMPARISON FOR FINGERPRINT AND WINNOWING ALGORITHM

Fingerprint Algorithm	Specific	General
Distinct Term	31.7%	68.3%
Term Frequency	33.6%	66.4%
Winnowing Algorithm	Specific	General
Distinct Term	34.7%	65.3%
Term Frequency	37.1%	62.9%

From Table IV, term-relevancy level of winnowing algorithm resulted a higher score than frequency or distinct term. By increasing the level of specific term in plagiarism detection, we hope that we can have better result in public's credibility towards quality of a study/ research.

V. CONCLUSION

Based on the discussion, it can be concluded that : although maximum accuracy value of fingerprint algorithm is higher (92.8%) than winnowing algorithms' (91.8%), winnowing algorithm has better and more stable performance. Winnowing algorithm also has a better performance in keyword-specifying as indicated by a higher correlation-level of fingerprint term.

ACKNOWLEDGMENT

This study is part of *Hibah Bersaing* research entitled "Kajian Information Retrival Guna Membangun Sistem Deteksi Indikasi Plagiat" founded by Direktorat Jendral Pendidikan Tinggi Kementerian Pendidikan dan Kebudayaan Republik Indonesia.

REFERENCES

[1] Benno Stein, 2007, "Fingerprint-based Similarity Search and its Applications", Universität Weimar 2007

[2] Brin S., Davis J., H. Garcia-Molina, Copy detection mechanisms for digital documents, in: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, ACM, San Jose, CA, United States, 1995, pp. 398–409.

[3] Chow K., Salim N., Features based text similarity detection, Journal of Computing 2 (1) (2010) 53–57.

[4] Dıez Silvia G., Foussy F., Shimboz M., Saerens M., 2010, "Normalized Sum-over-Paths Edit Distances", International Conference on Pattern Recognition

[5] Freire M., Cebrian M., 2007, "Design of the AC Academic Plagiarism Detection System", Universidad Aut´onoma de Madrid

[6] Kondrak Grzegorz, "N-gram similarity and distance", University of Alberta

[7] Nugroho Eko, Ridok Achmad, Rahayudi Bayu, 2011, "System Design Of Plagiarism Text Document Detection Using Rabin-Karp Algorithm", Brawijaya University

[8] Osman A. H., et al., 2012, "An improved plagiarism detection scheme based on semantic role labeling", Appl. Soft Computing Journal

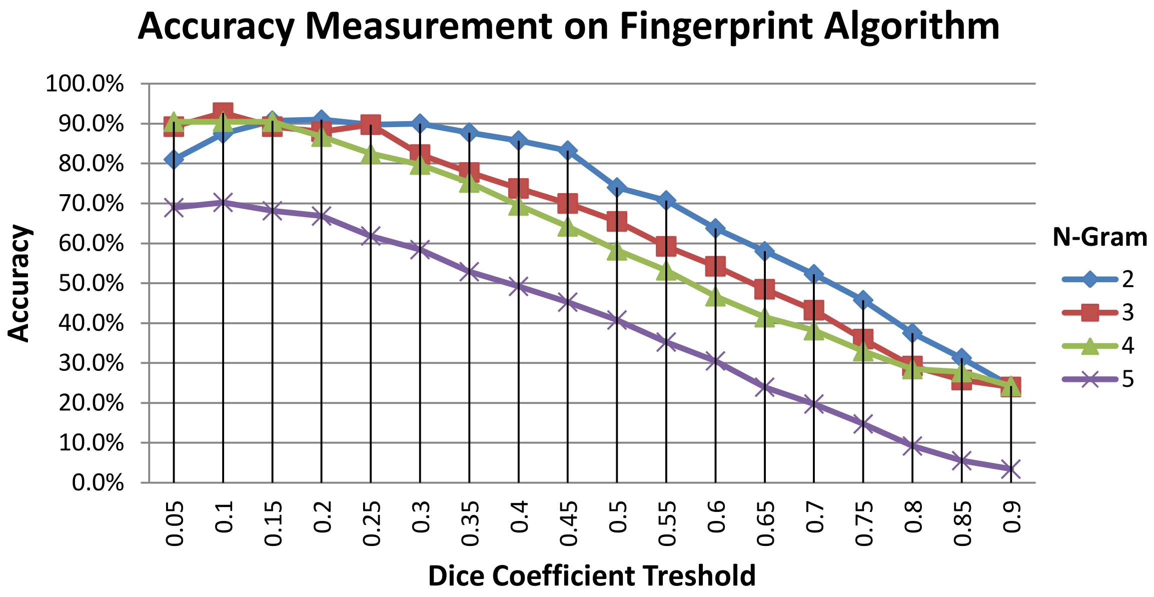
[9] Schleimer Saul, Wilkerson Daniel S., Aiken Alex, 2003, "Winnowing: Local Algorithms for Document Fingerprinting", Special Interest Group on Management Of Data (SIGMOD)

[10] Stamatatos Efstathios, 2009, "Intrinsic Plagiarism Detection Using Character n-gram Profiles", University of the Aegean

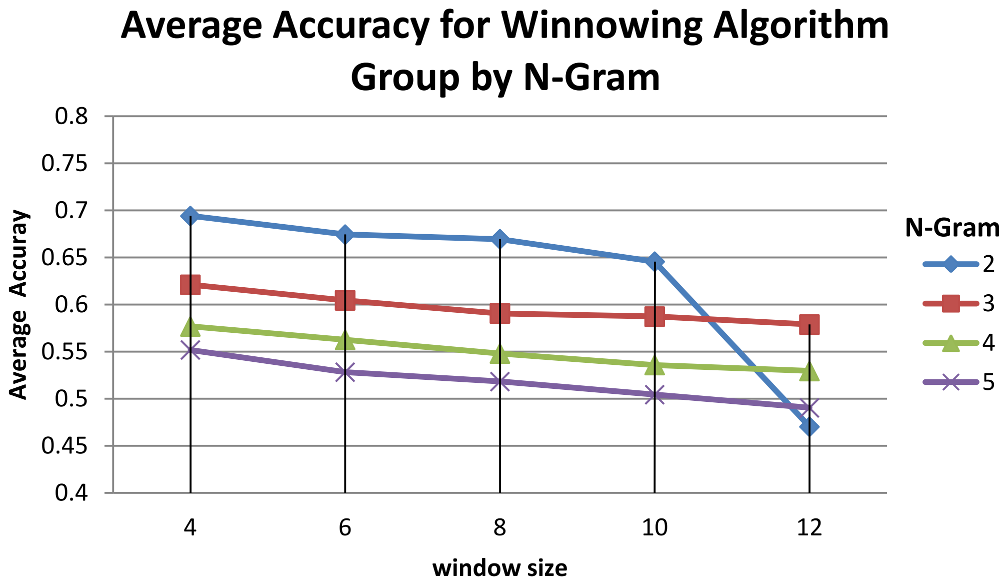
[11] Tala F.Z., 2003, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia", University of. Amsterdam

[12] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, Introduction to Algorithms, Second Edition, ISBN-10: 0-262-03293-7 September 2001

[13] Zhan Su, B.Y. Ahn, K.Y. Eom, M.K. Kang, J.P. Kim, M.K. Kim, Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm, International Conference on Innovative Computing Information and Control, 2008



Appendix A. Accuracy Measurement on Fingerprint Algorithm



Appendix B. Average Accuracy for Winnowing Algorithm Group by N-Gram