

An arrangement of the number of K-grams in the performance of Rabin Karp algorithm in text adjustment

Yuli Astuti¹, Irma Rofni Wulandari²

¹Department of Informatic Management, Faculty of Computer Science, University Amikom Yogyakarta, Yogyakarta, Indonesia

²Department of Information System, Faculty of Computer Science, University Amikom Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Nov 30, 2020

Revised Mar 14, 2022

Accepted Mar 29, 2022

Keywords:

K-gram

Performance

Rabin Karp

Similarity

Text adjustment

ABSTRACT

Rabin Karp algorithm is frequently used to determine the similarity between texts, using the hash function to compare the string identified and the substring in the text. The choice of the k value in the K-gram is often unrestricted. The number of k values used when cutting some terms will take longer if tried one by one. This research will perform a word cutting test on a script using K-gram 0 to 8. The results will cover the effect of the value of each K used on the similarity percentage produced. This research aims to determine the effect of the number of K-grams on the performance of Rabin Karp in text matching. The test underwent 20 sentences and 10 times using the dice coefficient for text similarity testing. The conclusion of this research should not use the K-gram 0 to 2 due to the K-gram basic principle: character deduction. Subsequently, if the character is 0,1,2, it does not have a meaning yet; thus, it gets a high similarity percentage. Based on trials by taking samples of K-gram 0 to 8 from 10 test data sets; the K-gram 3 is the best among K-grams 0 to 8.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Yuli Astuti

Department of Informatic Management, Universiti Amikom Yogyakarta

Jl. Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta, 55283, Indonesia

Email: yuli@amikom.ac.id

1. INTRODUCTION

One of the issues from the encouragement on information and communication technology is plagiarism. The internet and the accessibility of information in one click are often associated with the development of plagiarism [1]. According to Fan in Talib *et al.* [2] the process to extract pattern from a textual data source is a text mining. Retrieving information from the text is the main focus of text mining [3]. To determine the level of similarity between texts and can also be used to compare documents, it needs to be tested with an appropriate algorithm.

Algorithms for text adjustment are very diverse, one of them is Rabin Karp's algorithm which is one of the algorithms used in text mining to match text or strings [4], [5]. This text matching uses the hash function as a comparison between the search string (m) and the substring in the text (n) [6]. K-gram is a method to extract letters from a number of characters from a word and a series of terms with length K where text is continuously read from the beginning until the end of the document [7]. N-Gram, Base and modulo affect the degree of similarity [8]. The K-gram length is a determinant of plagiarism level. Determining the exact K-gram length produces accurate results. Hashing is a means to convert strings to integers [9]. In addition to K-gram, the process of document adjustment can be done using the N-Gram technique and Rabin Karp methods. N-gram is a method to get N piece character of a sentence based on the number of N specified [9], [10].

Many studies utilizing the Rabin Karp algorithm for various cases such as to detect similarities of the participants' answers for essay writing test [11] in addition to taking from the website, the Rabin Karp algorithm can also be used to search for studio locations by generating the category and list of rehearsal studios [12]. The larger the file size is, the longer the time to looking out for similarity. If the file does not undergo an indexing process, the time required is shorter but the similarity of the value decreases, the modulo value affects the processing time, but not the similarity value and smaller K-gram results is better in accuracy of similarity values compared to larger K-gram [13]. The Winnowing algorithm can also be used to detect some sort of plagiarism by searching for fingerprinting documents through converting N-gram sequences from text into a set of hash [14]. Substantial amount of applications apply sequences of N-Gram which weaken its performance [15]. That previously mentioned research has determined which K-gram will be used but the explanation of the reasons for the selection of the k values has not been widely explained. Rabin Karp's algorithm can be used for image or pattern matching, such as fingerprint matching [16]. Rabin Karp also has better performance than other algorithms in the case of semantic-based documents [17] and requires shorter time [18]. In addition to text adjustment, rabin karp images and patterns can also be used to optimize performance on parallel programming algorithms [19], [20].

The value produced by the K-gram is not always an accurate representation of the document [21]. The selection of the k value on K-gram in word cutting is often done freely. The number of k values that can be used when cutting words will take longer time if tried one by one. Studies that discuss the selection of the k value on K-gram are still limited in number, consequently this one will observe the testing of trials of words in the text using K-gram 0 to 8 with the reason for cutting the smallest word from 0 and the longest 8 word cuts, more than 8 fixed deductions can be done but not all text can be done depending on the number of characters of the text. The effect of each k value used on the percentage of similarity generated will be seen as the result of the trial. The results is shown in the form of an evaluation of the performance of the K-gram on the Rabin Karp algorithm. The contribution of this study is to determine the effect of the amount of K-gram on the performance of the Rabin Karp algorithm in text matching.

2. METHOD

This research consists of several steps. The steps of the research are shown in Figure 1. These following steps are the explanation for each process:

- a) Identifying the existing problems from the background, formulation, problem limitation, objectives, benefits, to the methodology used.
- b) Both literature study and literature review are conducted on several references that are relevant to the research topic. The reference referred to in this study is the K-gram on the Rabin Karp algorithm.
- c) Carrying out sentences adjustment by taking two sampling sentences.
- d) Preprocessing is executed in several stages:
 - Tokenization is the process of removing punctuations and changing it to the source text and words that are wanted to be found into words without capital letters.
 - Filtering, that is the deletion of words which often appear such as prepositions, conjunctions, pronouns, as well as affixes.
 - Stemming, the process of converting words into their basic form.
- e) After going through the results of preprocessing, the use of the Rabin Karp algorithm with the initial process of parsing is done subsequently. It is the process of cutting the character letters using the K-gram method.
- f) Hashing, i.e. converts string characters to integers [22]. This process converts text into hash values using ASCII code. As exemplifications: The use of the Rabin-Karp formula with the use of K-gram of 5 with the word clipping is "GUDEG". To get the hash value of the word can be seen from the following calculation:
 - N-gram: 5
 - Basis: 10
 - Modulo: 101

Then determining the ASCII code or ASCII character values from the word GUDEG: G: 103, U: 117, D: 100, E: 101, G: 103. Afterwards, calculating the hash with the (1) [8].

$$H = (\sum_{i=1}^n C(i) * b^{n-i}) \text{ mod } q \quad (1)$$

where:

H : Hash value

C : ASCII character values

- n : n-gram
- b : constant prime number (base)
- q : modulo
- $H = (103 * 10^{(5-1)} + 117 * 10^{(5-2)} + 100 * 10^{(5-3)} + 101 * 10^{(5-4)} + 103 * 10^{(5-5)}) \bmod 101$ H=47
- g) Proceeding to the calculation of similarity using the Dice Coefficient by calculating the value of n-gram [23], [24], with the Calculation as in (2) is done after the adjusting process [25]. Fingerprint hash is unique and non-duplicated hashes.

$$s = \frac{2 * c}{a + b} \times 100\% \tag{2}$$

where:

c: sum between hash a and b fingerprints

a dan b : number of words parsed or fingerprint hashes in text 1 and text 2

s: similarity value

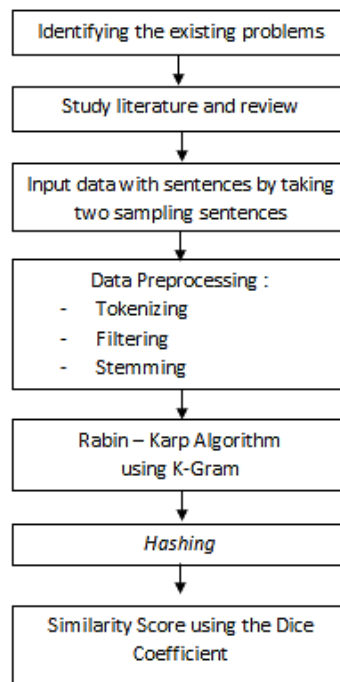


Figure 1. Research flowchart

3. RESULTS AND DISCUSSION

Detecting plagiarism starts from producing a percentage of similarity in the text. Sample text using Indonesian. Table 1 contains an example of the text to be tested. After inputting the two texts, the next step is to pre-process the provided text data. Table 2 is the result of the text preprocessing process which consists of tokenizing, filtering and stemming.

Table 2 describes the results of the pre-processing stage carried out by the system. At the tokenizing stage words are separated based on their order, so that they become tokens. At the filtering stage words that often appear in form of prepositions, conjunctions, pronouns and affixes are removed. Finally stemming is the process of converting the terms into their base form. After doing the preprocessing process, the results will be obtained from the text preprocessing as in Table 3.

Table 1. Text similarity test data 1

Text 1	Text 2
Gudeg adalah makanan khas Yogyakarta	Gudeg merupakan ciri khas makanan dari Yogyakarta
Gudeg is Yogyakarta's signature dish	Gudeg is a signature cuisine from Yogyakarta

Table 2. Text preprocessing

Process	Result
Tokenizing Text 1	gudeg adalah makanan khas Yogyakarta gudeg is yogyakarta signature dish
Filtering Text 1	gudeg makanan khas Yogyakarta gudeg yogyakarta signature dish
Steaming Text 1	gudegmakankhasyogyakarta gudegyogyakartasignaturedish
Tokenizing Text 2	gudeg merupakan ciri khas makanan dari Yogyakarta gudeg is a signature cuisine from yogyakarta
Filtering Text 2	gudeg ciri khas makanan Yogyakarta gudeg signature cuisine yogyakarta
Steaming Text 2	gudegcirikhasmakanyogyakarta gudegsignaturecuisineyogyakarta

Table 3. Text preprocessing results

Text	Clause
Text 1	gudegmakankhasyogyakarta gudegyogyakartasignaturedish
Text 2	gudegcirikhasmakanyogyakarta gudegsignaturecuisineyogyakarta

Table 3 presents the results of pre-processing in the form of basic text. The results of text preprocessing will be used for applying the Rabin-Karp algorithm. This algorithm has stages of K-gram and Hashing, to compare between matching strings. From the sampling data the matching is done using K-gram 4 which is described in Tables 4, 5 and 6. In Table 4, text 1 and text 2 from the pre-processing results are cut into 4 characters using K-gram 4, this process is used to get the character chunks. The result of this truncation is then hashed, which is to convert the string character into an integer. This process converts text into hash values using ASCII code. The results can be seen in Table 5. After the adjusting process then proceed to the calculation of similarity using the dice coefficient. Table 6 is the fingerprint results of the hash text 1 and text 2 and the resulting similarity.

Table 4. Results with K-gram

Text	K-gram Partition
Text 1	{gude}{udeg}{degm}{egma}{gmak}{maka}{akan}{kank}{ankh}{nkha}{khas}{hasy}{asyo}{syog}{yogy}{ogy}{gyak}{yaka}{akar}{kart}{arta}
Text 2	{gude}{udeg}{degc}{egci}{gcir}{ciri}{irik}{rikh}{ikha}{khas}{hasm}{asma}{smak}{maka}{akan}{kany}{anyo}{nyog}{yogy}{ogy}{gyak}{yaka}{akar}{kart}{arta}

Table 5. K-gram hashing result

Text	Hashing
Text 1	152451 169041 146563 148190 151456 158090 143231 155471 143698 160598 156183 151547 144464 169030 175736 161632 152908 174062 143235 155524 144274
Text 2	152451 169041 146553 148088 150341 145833 154811 165720 153943 156183 151535 144318 167428 158090 143231 155485 143859 162375 175736 161632 152908 174062 143235 155524 144274

Table 6. Fingerprint results and similarity level

Process	Result
Fingerprint	152451 169041 158090 143231 156183 175736 161632 152908 174062 143235 155524 144274
Similarity	52.17%

Researchers augmented some additional testing data using several texts with K-gram 0 to 8, the data is in Table 7 (in Appendix). In this paper, the research use text in Indonesian. The data sample uses 10 sets of sentence testing data with different sentence lengths and adds the number of K-grams, from 0-8, described in Table 7 (in Appendix). From the sample, text adjustment applies K-gram 0 to 8. This test uses 20 sentences with 10 iterations. The sentences used have different lengths. Table 7 (in Appendix) presents the similarity values of each K-gram. The result of the similarity percentage shows that there is similarity where the percentage value is getting lower. In testing 2, it stops at K-gram 5 while testing 4 stops at K-gram 7 since the same fingerprint value has not been found anymore, so the similarity value does not exist. 100% similarity is achieved as the results of testing for K-gram 0, as there is no character clipping in words, while in K-gram 1 and 2, the similarity percentage is close to 100% because the character clipping 1 and 2 do not have meaning yet, so that the value of similarity is high. From the 10 tests performed, it can be seen that there

was a very significant distinction in terms of the values of K-gram 2 and 3. K-gram 3 also got a consistent percentage value on each test. From testing with K-grams 4 to 8, it shows that it lowers the value of similarity percentage. Lower percentages will produce lower similarities. The smaller the percentage of similarity, the lower the ability to detect similarities between texts vice versa.

4. CONCLUSION

The similarity of the text can be seen from the results its adjustment. The K-gram 0 to 2 due to the K-gram basic principle: character deduction. Subsequently, if the character is 0,1,2, it does not have a meaning yet; thus, it gets a high similarity percentage. This research resulted in the recommendation of the best K-gram 3 values among K-gram 0 to 8 based upon the trials that have been done. Researchers only took K-gram 0 to 8 test samples with 10 test data sets since the clipping of K-gram 3 already has meaning.

ACKNOWLEDGEMENTS

This research was funded by an internal grant. Thanks to the Faculty of Computer Science, Universitas Amikom Yogyakarta who supported this research.

APPENDIX

Table 7. Test results for several texts and their similarity

Testing	Text 1	Text 2	K-gram	Similarity Results
Testing 1	In Indonesia: Mikroorganisme patogen meginfeksi sel makhluk hidup dapat disebut dengan virus. Virus memiliki kemampuan untuk mereplikasi diri ke dalam sel makhluk hidup. Virus tidak memiliki perlengkapan seluler sehingga tidak dapat bereproduksi sendiri.	In Indonesia: Virus adalah Mikroorganisme patogen meginfeksi sel makhluk hidup yang memiliki kemampuan untuk mereplikasi diri ke dalam sel makhluk hidup karena tidak memiliki perlengkapan seluler sehingga tidak dapat bereproduksi sendiri.	0	100.00%
			1	97.30%
			2	94.37%
			3	90.61%
			4	87.83%
			5	84.82%
			6	81.87%
			7	79.38%
			8	76.92%
			Testing 2	Pathogenic microorganisms that infect living cells are called viruses. Viruses can replicate themselves into living cells. Viruses do not have cellular structure, therefore, they cannot reproduce on their own. In Indonesia: Keahlian untuk membuat karya yang bermutu disebut dengan seni The skill to create quality works is called art.
1	75.86%			
2	37.50%			
3	20.83%			
4	13.04%			
5	4.55%			
6	-			
7	-			
8	-			
Testing 3	In Indonesia: Data mining yaitu sekumpulan data yang diproses sedemikian rupa untuk mendapatkan nilai tambah berupa pengetahuan Data mining is a collection of data that is processed in such a way to obtain added value in the form of knowledge.	In Indonesia: Data mining yaitu sekumpulan data dalam jumlah besar atau kompleks yang dianalisis secara otomatis untuk menemukan pola atau kecenderungan yang penting dan terkadang tidak disadari keberadaannya Data mining is a large or complex collection of data that is automatically analyzed to find important and unknown patterns or trends		
			1	90.91%
			2	45.16%
			3	30.63%
			4	28.07%
			5	28.07%
			6	26.79%
			7	25.45%
			8	24.07%
			Testing 5	In Indonesia: Di dalam tata surya terdapat kumpulan benda langit yaitu sebuah matahari dan benda-benda langit lain yang terikat oleh gaya gravitasi In the solar system, there is a collection of celestial bodies, namely the sun and other celestial bodies that are bound by the force of gravity.
1	94.74%			
2	90.53%			
3	88.50%			
4	85.47%			
5	82.05%			
6	78.63%			
7	75.86%			
8	73.04%			

Table 7. Test results for several texts and their similarity (*continue*)

Testing	Text 1	Text 2	K-gram	Similarity Results
Testing 6	In Indonesia: Mie ayam atau bakmi ayam adalah masakan indonesia Mie Ayam or chicken noodle is Indonesian cuisine	In Indonesia: Mie ayam merupakan salah satu masakan khas indonesia Chicken noodle is one of the Indonesian's signatures cuisine	0	100.00%
			1	86.96%
			2	79.17%
			3	53.85%
			4	43.14%
			5	32.00%
			6	25.00%
			7	17.39%
Testing 7	In Indonesia: Daring merupakan proses pertukaran informasi antar komputer yang telah terhubung melalui internet Online is the process of exchanging information between computers connected to the internet	In Indonesia: Daring merupakan proses pembelajaran atau bertukar informasi melalui hubungan sebuah internet Online is a process of learning or exchanging information through an internet connection	0	100.00%
			1	97.30%
			2	83.54%
			3	72.73%
			4	66.67%
			5	62.22%
			6	56.82%
			7	51.16%
Testing 8	In Indonesia: Biji kopi yang disangrai kemudian dihaluskan sehingga menjadi bubuk kopi dapat dinikmati dengan menyeduhnya The roasted coffee beans are then ground into coffee grounds that can be enjoyed by brewing them.	In Indonesia: Cara menikmati kopi yaitu dengan menyeduh biji kopi yang disangrai kemudian dihaluskan sehingga menjadi bubuk The way to enjoy coffee is to brew coffee beans that have been roasted and then turning them into coffee grounds	0	100.00%
			1	97.14%
			2	87.10%
			3	79.41%
			4	70.59%
			5	64.71%
			6	59.70%
			7	58.46%
Testing 9	In Indonesia: Kementerian Industri dan Teknologi Informasi China mengatakan melalui situs resminya bahwa awal bulan Maret, perusahaan-perusahaan ventilator di China telah mengirimkan sekitar 14.000 ventilator non-invasif dan 2.900 invasif ke Kota Hubei, China China's Ministry of Industry and Information Technology said on its official website that in early March, ventilator companies in China had shipped around 14,000 non-invasive and 2,900 invasive ventilators to Hubei City, China.	In Indonesia: Sebanyak 14.000 ventilator non-invasif dan 2.900 invasif telah di kirimkan ke kota Hubei oleh perusahaan-perusahaan ventilator di china. A total of 14,000 non-invasive and 2,900 invasive ventilators have been shipped to Hubei city, by ventilator companies in China.	0	100.00%
			1	94.44%
			2	72.41%
			3	54.30%
			4	46.91%
			5	41.67%
			6	36.78%
			7	32.95%
Testing 10	In Indonesia: Angka kematian di Italia berjumlah 6,077 kematian dari 63,927 kasus atau setara dengan 9,51 persen, hal ini berbanding terbalik dengan jumlah kasus dan kematian di china. The death rate in Italy is 6,077 deaths of 63,927 cases, equivalent to 9.51 percent, this is inversely proportional to the number of cases and deaths in China.	In Indonesia: Jumlah kasus kematian di cina berbanding terbalik dengan italia yang berjumlah 6.007 kematian dari 63.927 kasus yaitu sebesar 9.51 persen. The number of deaths in China is inversely proportional to Italy which was amounted to 6,007 deaths of 63,927 cases, equivalent to 9.51 percent.	0	100.00%
			1	96.77%
			2	76.67%
			3	62.86%
			4	52.63%
			5	44.16%
			6	34.21%
			7	27.03%
8	22.22%			

REFERENCES

[1] N. Hasan and N. Khan, "Internet and increasing issues of plagiarism," *Shrinkhla Ek Shodhparak Vaicharik Patrika*, vol. 5, no. 12, pp. 125–131, 2018, [Online]. Available: <https://www.researchgate.net/publication/332696789>.

[2] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text mining: techniques, applications and issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp. 414–418, 2016, doi: 10.14569/ijacsa.2016.0711153.

[3] A. Bahrul Khoir, H. Qodim, B. Busro, and A. R. Atmadja, "Implementation of rabin-karp algorithm to determine the similarity of synoptic gospels," in *Journal of Physics: Conference Series*, 2019, vol. 1175, no. 1, pp. 1–7, doi: 10.1088/1742-6596/1175/1/012120.

[4] G. H. Gonnet and R. A. Baeza-Yates, "An analysis of the Karp-Rabin string matching algorithm," *Information Processing Letters*, vol. 34, no. 5, pp. 271–274, 1990, doi: 10.1016/0020-0190(90)90135-K.




[5] B. Leonardo and S. Hansun, "Text documents plagiarism detection using Rabin-Karp and Jaro-Winkler distance algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. 2, pp. 462–471, 2017, doi: 10.11591/ijeecs.v5.i2.pp462-471.

[6] E. Rasywir, Y. Pratama, Hendrawan, and M. Istoningtyas, "Removal of modulo as hashing modification process in essay scoring system using rabin-karp," in *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018*, 2019, vol. 2019-January, pp. 159–164, doi: 10.1109/ICECOS.2018.8605211.




- [7] D. B. Rahmawati, M. L. Irfani, and R. B. R. Purba, "Text mining to detect plagiarism in E-learning system using Rabin Karp algorithm," *Ire 1701953 Iconic Research and Engineering Journals*, vol. 3, no. 8, pp. 183–191, 2020.
- [8] A. P. U. Siahaan, R. Rahim, and D. Siregar, "Parameter adjustment in gaining accuracy of plagiarism detection," *Journal Online Jaringan COT POLIPD (JOJAPS) Parameter*, vol. 10, pp. 22–29, 2017, doi: 10.31219/osf.io/eg74x.
- [9] D. D. Sinaga and S. Hansun, "Indonesian text document similarity detection system using rabin-karp and confix-stripping algorithms," *International Journal of Innovative Computing, Information and Control*, vol. 14, no. 5, pp. 1893–1903, 2018, doi: 10.24507/ijicic.14.05.1893.
- [10] E. Stamatatos, "Intrinsic plagiarism detection using character n-gram profiles," *CEUR Workshop Proceedings*, 2009, vol. 502, pp. 38–46.
- [11] M. Misbah Musthofa and A. Yaqin, "Implementation of Rabin Karp algorithm for essay writing test system on organization XYZ," in *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 2019, pp. 502–507, doi: 10.1109/ICOIACT46704.2019.8938562.
- [12] S. M. Gomez, "MUSICHUB: A web and android based rehearsal studio locator and reservation system in Davao city utilizing geolocation API and Rabin-Karp algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3746–3751, 2020, doi: 10.30534/ijatcse/2020/189932020.
- [13] A. P. U. Siahaan, "Rabin-Karp elaboration in comparing pattern based on hash data," *International Journal of Security and Its Applications*, vol. 12, no. 2, pp. 59–66, 2018, doi: 10.14257/ijisa.2018.12.2.06.
- [14] D. Leman, M. Rahman, F. Ikorasaki, B. S. Riza, and M. B. Akbbar, "Rabin Karp and Winnowing algorithm for statistics of text document plagiarism detection," 2019, doi: 10.1109/CITSM47753.2019.8965422.
- [15] D. Lemire and O. Kaser, "Recursive n-gram hashing is pairwise independent, at best," *Computer Speech and Language*, vol. 24, no. 4, pp. 698–710, 2010, doi: 10.1016/j.csl.2009.12.001.
- [16] S. Kanchana and G. Balakrishnan, "Palm-print pattern matching based on features using Rabin-Karp for person identification," *Scientific World Journal*, vol. 2015, pp. 1–8, 2015, doi: 10.1155/2015/382697.
- [17] C. Supriyanto and A. Syukur, "A comparison of Rabin Karp and semantic-based plagiarism detection," in *International Conferences on Soft Computing, Intelligent System and Information Technology 2012*, 2012, pp. 29–31.
- [18] M. Shabaz and N. Kumari, "Advance-Rabin Karp algorithm for string matching," *International Journal of Current Research*, vol. 9, no. 9, pp. 57572–57574, 2017, [Online]. Available: <https://www.journalcra.com/article/advance-rabin-karp-algorithm-string-matching>.
- [19] O. S. Joshi, B. R. Upadhvay, and M. Supriya, "Parallelized advanced Rabin-Karp algorithm for string matching," *2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2017.8463971.
- [20] L. S. N. Nunes, J. L. Bordim, Y. Ito, and K. Nakano, "A prefix-sum-based Rabin-Karp implementation for multiple pattern matching on GPGPU," in *Proceedings - 2018 6th International Symposium on Computing and Networking, CANDAR 2018*, 2018, pp. 139–145, doi: 10.1109/CANDAR.2018.00026.
- [21] A. P. U. Siahaan, R. Rahim, M. Mesran, and D. Siregar, "K-Gram as a determinant of plagiarism level in Rabin-Karp algorithm," *International Journal of Scientific and Technology Research*, vol. 06, no. 07, pp. 350–353, 2017, doi: 10.31219/osf.io/yxjnp.
- [22] A. Yaqin, A. Dahlan, and R. D. Hermawan, "Implementation of algorithm rabin-karp for thematic determination of thesis," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2019*, 2019, pp. 395–400, doi: 10.1109/ICITISEE48480.2019.9003867.
- [23] W. G. S. Parwita, I. G. A. A. D. Indradewi, and I. N. S. W. Wijaya, "String matching based plagiarism detection for document in Bahasa Indonesia," in *Proceedings of 2019 5th International Conference on New Media Studies, CONMEDIA 2019*, 2019, pp. 54–58, doi: 10.1109/CONMEDIA46929.2019.8981821.
- [24] I. Mardiana, T. Adji, and T. B. Hidayah, "Preface," *Communications in Computer and Information Science*, vol. 516, pp. 155–164, 2015, doi: 10.1007/978-3-662-46742-8.
- [25] R. E. Putri *et al.*, "Examination of document similarity using Rabin-Karp algorithm," *International Journal of Recent Trends in Engineering and Research*, vol. 3, no. 8, pp. 196–201, 2017, doi: 10.23883/ijrter.2017.3404.4sndk.

BIOGRAPHIES OF AUTHORS



Yuli Astuti    earned his Bachelor's degree in Information System from University of Amikom in 2006, Master's degree in Master Informatic Teknik University of Amikom in 2014. Her research interest including information system, prediction, classification and other data mining fields. She can be contacted at email: yuli@amikom.ac.id.



Irma Rofni Wulandari    Graduate from the Informatics Engineering Education undergraduate program at Yogyakarta state university in 2011 and graduate from Gadjah Mada University Information Technology Master's Program in 2016. Her research interest including Information System, Decision Support System and Human Computer Interaction. She can be contacted at email: rofni@amikom.ac.id.