Evaluation of the accuracy of winnowing, rabin karp and knuth morris pratt algorithms in plagiarism detection applications

by Ida Widaningrum, Dyah Mustikasari, Rizal Arifin, H A Pratiwi

Submission date: 19-Sep-2023 11:05AM (UTC+0700)
Submission ID: 2170319024
File name: 12._Evaluation_of_the_accuracy_of_winnowing,_rabin_karp.pdf (495.52K)
Word count: 2813
Character count: 15079

Journal of Physics: Conference Series

Evaluation of the accuracy of winnowing, rabin karp and knuth morris pratt algorithms in plagiarism detection applications

Evaluation of the accuracy of winnowing, rabin karp and knuth morris pratt algorithms in plagiarism detection applications

I Widaningrum[1*], D Mustikasari[1], R Arifin[2], And H A Pratiwi[1]

[1] Department of Informatics Engineering, Universitas Muhammadiyah Ponorogo, Ponorogo, Indonesia
[2] Department of Mechanical Engineering, Universitas Muhammadiyah Ponorogo, Ponorogo, Indonesia

*Email: iwidaningrum.as@gmail.com

Abstract. The unethical behaviour of acts of plagiarism has been a disgrace in the educational realm. Using the internet, people can easily find articles or documents that are relevant to their current work, and simply duplicate the sentences or paragraphs without paraphrasing or giving correct citations. Such action falls into the area of plagiarism. In order to minimize the problem, especially in the educational field, it is necessary to develop plagiarism detection applications. The algorithm employed in the application plays an important role in obtaining accurate plagiarism detection results. To the best of our knowledge, three algorithms are commonly used in plagiarism detection applications, namely Winnowing, Rabin Karp and Knuth Morris Pratt, which are all employed in our application. To specify the accuracy of each algorithm, the percentages of the plagiarism detection results are compared to the results from examination by a human expert. From our results, we found that the order of the accuracy from highest to lowest corresponded to the Winnowing algorithm, Rabin Karp algorithm and Knuth Morris Pratt algorithm, with value differences of 1.19%, 53.91% and 83.91% respectively.

1. Introduction

Plagiarism involves presenting and claim the work of another person as one's own [1]. In the technological era, there is widespread plagiarism, especially in the world of education, and the practice has greatly increased due to easy internet access. Therefore, it is necessary to devise the most optimal detection tools. The optimization of detection is affected by the algorithm [2]–[4], which plays an important role in obtaining accurate results. This paper will discuss three algorithms that are often used in plagiarism detection, namely Rabin Karp, Winnowing and Knutt Morris Pratt (KMP) [5]. In the research, we compare the efficacy of the three algorithms in detecting plagiarism in texts in two languages, Indonesian and English.

The investigations to detect the degree of similarity were conducted using the Winnowing algorithm [6]–[8], Winnowing and Rabin Karp [9], and KMP algorithm [2], [10]–[12]. The other previous research detect documents [13]–[16] using the Rabin Karp algorithm to evaluate the similarity of hash and k-gram values [17], [18]; the Rabin-Karp algorithm was employed to calculate the percentage of document similarities; and a combination of the Rabin-Karp and Levenshtein distance algorithms was used to assess the levels of similarity [19], [20]. Plagiarism detection in Indonesian text can be performed by Winnowing, preceded by the bypass pre-processing stage, Rabin-

Karp and the Confix-Stripping algorithm [21]. In this investigation, plagiarism detection in the documents will be analysed using three algorithms, namely the Winnowing algorithm, the Rabin Karp algorithm and the Knuth Morris Pratt algorithm, followed by data processing to obtain the results of the performance of each of these.

2. Method

We compared three algorithms, namely Winnowing, Rabin Karp and the KMP algorithms, in detecting plagiarism. The comparison process used in the research is described in Figure 1.

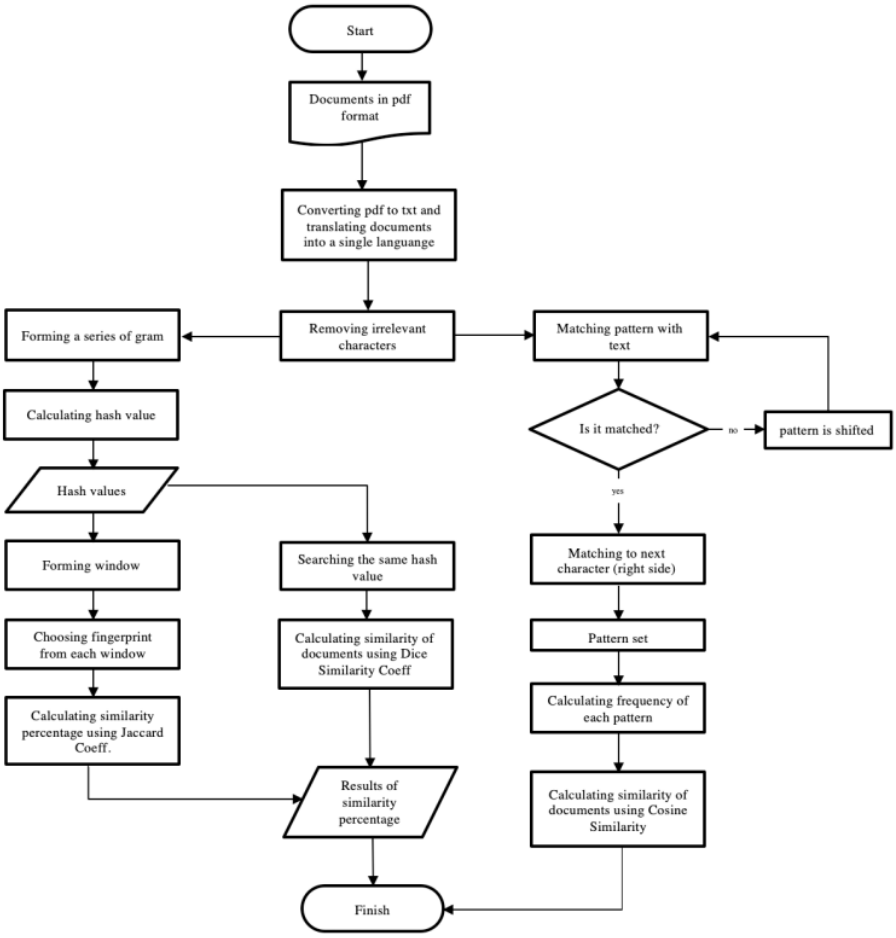

Figure 1. Process of Winnowing, Rabin Karp and KMP algorithms

Winnowing [22]–[24] is implemented by the following steps. First, unnecessary characters in the text, such as symbols, punctuation and spaces are deleted, capital letters changed to lower case. The second step is the formation of grams; for example, the value of grams with a size of 7. The third step is the Rolling Hash process to produce hash value from each gram formed. The hash values are then divided according to the window; for example, the specified window value is 4. The window selection is almost the same as the formation of grams. The next step is to choose the smallest hash value of

each window to use as a fingerprint. The following sentences are used to explain the calculation of the hash value.

Sentence 1: "Jurusan Teknik Informatika adalah salah satu jurusan yang ada di Fakultas Teknik Universitas Muhammadiyah Ponorogo"

Sentence 2: "Jurusan Teknik Informatika adalah salah satu jurusan terfavorit di FT Unmuh Ponorogo"

From the calculation, we obtained 38 fingerprints for the first sentence, and 28 for the second. The percentage of similarity was then calculated using the Jaccard similarity coefficient equation (1). The total of equal fingerprints is 17, and the number of different fingerprints is 32; therefore, the total of all fingerprints is 49.

$$Similarity = \frac{Total\ of\ equal\ fingerprint}{Total\ of\ all\ fingerprint}\ x\ 100\% \quad (1)$$

$$= \frac{17}{49}\ x\ 100\% = \mathbf{34.7}\%$$

In the Rabin Karp algorithm [25]–[31], the first to third steps are the same as in the Winnowing algorithm. However, after obtaining the hash value of the Rolling Hash process [32], it is determined the equal hash of two documents. In this case, we used the same two sentences as before. From the calculation, 42 hashes were obtained. The percentage of similarity was calculated using the Dice similarity coefficient equation. The number of equal hashes is 42, the number of hashes in sentence 1 is 99, and the number of hashes in sentence 2 is 67. We then calculated the similarity using the Dice similarity coefficient equation (2).

$$Similarity = \frac{2\ x\ sum\ of\ equal\ hash}{total\ of\ hash\ from\ two\ document}\ \text{x}\ 100\% \quad (2)$$

$$= \frac{2\ x\ 42}{99+67}\ \text{x}\ 100\%$$

$$= \frac{84}{166}\ \text{x}\ 100\%$$

$$= 50.6\%$$

With the same two sentences, the steps applied in the Knuth Morris Pratt algorithm are as follows. The first step is the same as the previous two algorithms. The next step is to eliminate the stop word. The third is forming a collection of words or patterns. The result of the patterns in sentences 1 and 2 are shown in Table 1.

Table 1. Formation of words patterns collection in sentence 1 and 2

jurusan	jurusan
teknik	teknik
informatika	informatika
jurusan	jurusan
di	terfavorit
fakultas	di
universitas	ft
muhammadiyah	unmuh
ponorogo	ponorogo

The following steps are the process of weighting the frequency of words. Once the word collections are formed, they are combined into one group and the frequency of the document is calculated. The results of the frequency of sentences 1 and 2 is shown in Table 2.

Table 2. Determination of the frequency of words or patterns

Words	Frequency in sentence 1	Frequency in sentence 2
teknik	2	1
informatika	1	1
jurusan	2	2
terfavorit	0	1
di	1	0
fakultas	1	0
ft	0	1
unmuh	0	1
universitas	1	0
muhammadiyah	1	0
ponorogo	1	1

The final step is calculation of the percentage of similarity with the Cosine Similarity equation (3).

$$\text{CosSim} = \frac{(2x1)+(1x1)+\cdots+(1x0)+(1x1)}{\sqrt{2^2+1^2+\cdots+1^2+1^2}\ x\ \sqrt{1^2+1^2+\cdots+0^2+1^2}} \text{x100\%} \quad (3)$$

$$= \frac{2+1+\cdots+0+1}{\sqrt{4+1+\cdots+1+1}\ x\ \sqrt{1+1+\cdots+0+1}} \text{ x100\%}$$

$$= \frac{8}{\sqrt{14}\ x\ \sqrt{10}} \text{x100\%} = \frac{8}{11.83} \text{ x100\%}$$

$$= \mathbf{57\%}$$

3. Result and Discussion

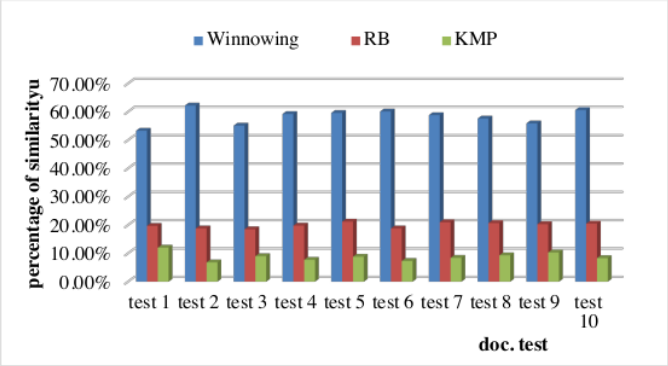

Figure 2. Percentage of similarity

The documents used in the research were written in Bahasa and English on grouping. Ten test documents were used and 20 comparison documents. The three algorithms were applied to the documents in .pdf format with the following conditions:

- For the Winnowing algorithm, gram is 5, window is 7 and base is 3, with use of the similarity equation of the Jaccard coefficient.

- For the Rabin Karp algorithm, gram is 7 and base is 3, with use of the equation of the coefficient of similarity of dice.
- For Knuth Morris Pratt (KMP), the equation used was the cosine similarity.

Figure 2 shows the results of the three algorithms for the detection of plagiarism in the first test document of 10 comparison documents. From the figure, it can be concluded that the use of the Winnowing algorithm produced is approximately 40% to 75%. It used 10 test documents, compared to 20 comparison documents with grouping issues. The value of gram is 5, window 7 and base 3.

Regarding the use of the Rabin Karp algorithm, percentage of similarity between the documents was produced, on average between 15% and 25%. Finally, regarding the use of the Knuth Morris Pratt algorithm, there was a low percentage of similarity between the documents, or only a small degree of similarity between the documents, because the percentage produced was less than 15%.

Further, we performed tests to determine the level of accuracy of the measurement results of the three algorithms used. These were performed by comparing document 1, which had 17,605 characters, and document 2, obtained from the process of removing several sentences from document 1. The number of characters in document 2 was consequently 16,393. According to manual calculations, the percentage of similarity between the two documents was $\frac{16,393}{17,605}$ x 100% = 93.11%.

Table 3. Result of the similarity percentage from tests using the Winnowing algorithm, Rabin Karp algorithm and Knuth Morris Pratt algorithm

Algorithm	Similarity (%)	Reference (%)	Difference (%)
Winnowing	94.30		1.19
Rabin Karp	39.20	93.11	53.91
Knuth Morris Pratt	9.20		83.91

Table 3 shows the result of the similarity percentage of the three algorithms. From the table, it can be seen that they produce similarity percentages of 94.3%, 39.2% and 9.2% respectively. When they are compared to the results of the manual calculations, the order of accuracy from highest to lowest was obtained from the Winnowing algorithm, Rabin Karp algorithm and Knuth Morris Pratt algorithm, with value differences of 1.19%, 53.91% and 83.91% respectively.

4. Conclusion

The Winnowing algorithm can detect high levels of plagiarism in documents, the Rabin Karp algorithm at a medium level, and the Knuth Morris Pratt algorithm can detect plagiarism at a low level in documents that have the same subject, with gram arrangements = 5, window = 7 and base = 3. The analysis confirms that the optimal algorithm of the three is the Winnowing algorithm, which produced 94.3% of similarity. It also exhibits the highest accuracy, with only a 1.19% difference from the examination by the human expert, at 93.11%.

Acknowledgment

This work is part of the PTUPT research grant work funded by the Directorate General for Research and Development (DRPM), Ministry of Research, Technology of Higher Education of the Republic of Indonesia in fiscal year 2019, under contract No. 025 / SP2H / LT / MULTI / L7 / 2019.

References

[1] D. Steveson, H. Agung, dan F. Mulia, “Rabin Karp Plagiarisme Detection Applications For Tasks and Problems in School Using Rabin Karp Algorithm,” no. 1, hal. 12–17, 2018.
[2] M. Syarif, “Implementasi Algoritma String Matching Dalam Pencarian Surat Dan Ayat Dalam Al-Quran Berbasis Web,” *iIndonesian J. Netw. Secur.*, vol. 6, no. 2, hal. 70–76, 2017.
[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, dan C. Stein, *Introduction to Algorithms*, vol. 42, no. 9. 1991.

[4] M. GOU, "Algorithms for String matching," in *Conference Proceedings*, 2014, hal. 1–8.
[5] T. Lecroq, "Handbook of Exact String-Matching Algorithms Christian Charras," no. May, 2014.
[6] A. Yudhana, "Implementation of Winnowing Algorithm with Dictionary English-Indonesia Technique to Detect Plagiarism," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, hal. 183–189, 2018.
[7] N. Alamsyah, "Deteksi Plagiarisme Tingkat Kemiripan Judul Skripsi dengan Algoritma Winnowing," *Technologia*, vol. 8, no. 4, hal. 205–213, 2017.
[8] R. K. Wibowo dan K. Hastuti, "Penerapan Algoritma Winnowing Untuk Mendeteksi Kemiripan Teks pada Tugas Akhir Mahasiswa," *Techno.COM*, vol. 15, no. 4, hal. 303–311, 2016.
[9] N. Alamsyah, "Perbandingan Algoritma Winnowing Dengan Algoritma Rabin Karp Untuk Mendeteksi Plagiarisme Pada Kemiripan Teks Judul Skripsi," *Technologia*, vol. 8, no. 3, hal. 124–134, 2017.
[10] W. Astuti, "Analisis String Matching Pada Judul Skripsi Dengan Algoritma Knuth-Morris Pratt (Kmp)," *Ilk. J. Ilm.*, vol. 9, no. 2, hal. 167–172, 2018.
[11] R. Alamanda, C. Suhery, dan Y. Brianorman, "Aplikasi Pendeteksi Plagiat Terhadap Karya Tulis Berbasis Web Menggunakan Natural Language Processing Dan Algoritma KMP," *J. Coding, Sist. Komput. Untan*, vol. 04, no. 1, 2016.
[12] D. E. Knuth, J. H. Morris, dan V. R. Pratt, "Fast pattern matching in strings," *Am. Paleontol.*, vol. 6, no. 2, hal. 323–350, 1977.
[13] M. Fakhrurrozi, "Implementasi algoritma rabin-karp pada aplikasi pendeteksi plagiat untuk dokumen," Malang, 2016.
[14] Salmuasih dan A. Sunyoto, "Implementasi Algoritma Rabin Karp untuk Pendeteksian Plagiat Dokumen Teks Menggunakan Konsep Similarity," *Semin. Nas. Apl. Teknol. Inf. 2013*, hal. 23–28, 2013.
[15] S. Suhada dan S. Bahri, "Implementasi Algoritma Rabin Karp Dan Stemming Najief Andriani Untuk Deteksi Plagiarisme Dokumen," *Swabumi*, vol. 5, no. 1, hal. 84–89, 2017.
[16] T. Mardiana, T. B. Adji, dan I. Hidayah, "The Comparation of distance-based similarity measure to detection of plagiarism in Indonesian text," *Commun. Comput. Inf. Sci.*, vol. 516, hal. 155–164, 2015.
[17] A. Putera Utama Siahaan dan Sugianto, "Analisis k-gram, basis dan modulo rabin-karp sebagai penentu akurasi persentase kemiripan dokumen," in *SENASPRO 2017 | Seminar Nasional dan Gelar Produk*, 2017, hal. 198–206.
[18] Y. Arifin, S. M. Isa, L. A. Wulandhari, dan E. Abdurachman, "Plagiarism Detection for Indonesian Language using Winnowing with Parallel Processing Plagiarism," in *2nd International Conference on Computing and Applied Informatics 2017*, 2017, hal. 0–6.
[19] A. H. Purba dan Z. Situmorang, "Analisis Perbandingan Algoritma Rabin-Karp Dan Levenshtein Distance Dalam Menghitung Kemiripan Teks," *J. Tek. Inform. Unika St. Thomas*, vol. 02, hal. 24–32, 2017.
[20] R. E. Putri, A. Putera, dan U. Siahaan, "Examination of Document Similarity Using Rabin-Karp Algorithm," *Int. J. Recent Trends Eng. Res.*, vol. 3, no. 8, hal. 196–201, 2017.
[21] D. D. Sinaga dan S. Hansun, "Indonesian text document similarity detection system using rabin-karp and confix-stripping algorithms," *Int. J. Innov. Comput. Inf. Control*, vol. 14, no. 5, hal. 1893–1903, 2018.
[22] N. Elbegbayan, "Winnowing , a Document Fingerprinting Algorithm," *TDDC03 Proj.*, 2005.
[23] S. Schleimer, D. S. Wilkerson, dan A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," *Proc. 2003 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '03*, hal. 76–85, 2003.
[24] J. Parapar dan Á. Barreiro, "Winnowing-based text clustering," *Proceeding 17th ACM Conf. Inf. Knowl. Min. - CIKM '08*, hal. 1353, 2008.
[25] M. R. Pratama, E. B. Cahyono, dan G. I. Marthasari, "Aplikasi Pendeteksi Duplikasi Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Winnowing Dengan Metode K-Gram Dan Synonym Recognition," hal. 6, 2012.
[26] A. P. U. Siahaan, Mesran, R. Rahim, dan D. Siregar, "K-Gram As A Determinant Of Plagiarism

Level In Rabin-Karp Algorithm," *Int. J. Sci. Technol. Res.*, vol. 6, no. 07, hal. 350–353, 2017.
[27] E. Nugroho, "Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp," *J. Strateg. Stud.*, vol. 34, no. 2, hal. 281–293, 2011.
[28] A. Prastyanti, "Sistem Deteksi Kemiripan Kata Pada Dua Dokumen Menggunakan Algoritma Rabin-Karp," Universitas Diponegoro Semarang, 2014.
[29] A. B. Mutiara dan S. Agustina, "Anti Plagiarism Application with Algorithm Karp-Rabin at Thesis in Gunadarma University," *arXiv Prepr. arXiv0811.4349*, hal. 9, 2008.
[30] G. H. Gonnet dan R. A. Baeza-yates, "An Analysis Of The Karp-Rabin String Matching Algorithm," *Inf. Process. Lett.*, vol. 34, hal. 271–274, 1990.
[31] R. M. Karp dan M. O. Rabin, "Efficient randomized pattern-matching algorithms," *IBM J. RES. Dev.*, vol. 31, no. 2, hal. 249–260, 1987.
[32] A. Sabor Bostan, "Winnowing Algorithm for Program Code," no. July, 2017.

Evaluation of the accuracy of winnowing, rabin karp and knuth morris pratt algorithms in plagiarism detection applications

ORIGINALITY REPORT

5%	2%	8%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dian Rachmawati, Amalia Amalia, Muhammad Rinaldi. "Analysis of Maximal Shift Algorithm and Rabin-Karp Algorithm in Graphic Design Dictionary", 2021 5th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), 2021 Publication	3%
2	centaur.reading.ac.uk Internet Source	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		