Citation Analysis on Scientific Articles Using Cosine Similarity

Ulfa Mardatillah
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
ulfamardatillah@gmail.com

Wildan Budiawan Zulfikar
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
wildan.b@uinsgd.ac.id

Aldy Rialdy Atmadja
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
aldy@if.uinsgd.ac.id

Ichsan Taufik
Department of Informatics
UIN Sunan Gunung Djati
Bandung, Indonesia
ichsan@uinsgd.ac.id

Wisnu Uriawan
Department Informatics and
Mathematics
INSA, Lyon, France
wisnu_u@uinsgd.ac.id

Abstract— ***Citations have a vital role in scientific articles. Every sentence that is listed must be obtained from the reference. There is a measure that is similarity. The problem is, the likeness of a sentence in a scientific article to the source it refers to can have different meanings. The high level of similarity can undoubtedly increase the similarity value if it is checked using the plagiarism (similarity checker) tool, where the higher the value is more similar to other articles on the contrary with low similarity remaining. However, a low level of similarity can also mean that a statement in a scientific paper has referred to a completely irrelevant source. This research aims to analyze the level of similarity between scientific articles and other articles or sources it refers to. Checking the input text starts with text preprocessing, weighting the TF-IDF, and determining the similarity level using Cosine Similarity. Based on the test results, it is found that the results in a scientific article document have a high level of similarity with the article it refers to, and the other 30% have a low level of similarity, or in other words, there is no relationship.***

Keywords—scientific articles, journals, TF-IDF, cosine, text mining, citation, incorrect citation

I. Introduction

The development of information technology is increasing and has a positive impact. One of the positive impacts of information technology development is the ease of exchanging information. This convenience is often misused by someone or several people in completing work. The misused usually occurs in Scientific Documents. Scientific documents are documents of reasoning and research results using a variety of objects and methods. A good document has explicit references and sources as a reference, for example, a scientific paper or journal that has explicit references and citations and does not misuse the original citation of the document [1]–[4].

Scientific journal manuscripts are the author's scientific research manuscripts and consist of direct and indirect quotations. Direct quotations are quotations that are written exactly like the source, both in language and spelling [5], [6]. An indirect quotation is a quotation that is not the same as the original, and the citer only takes the main idea of the quoted source to be restated with the sentence compiled by the quote. Quotations are considered important as valid references and are included in the writing of scientific papers [2], [3], [7], [8].

Mistakes that harm scientific journal scriptwriters are called plagiarism. Based on previous research that uses a cosine similarity method to check for plagiarism only, this research complements and refines it by weighing each word in scientific documents and testing the text-similarity value. So, documents that have a high similarity value are abused quotes and a high similarity value. The low similarity is a document that has been cited with little abuse or towards correctness.

The website-based bibliographic reference citation checking system for scientific journals will be used in checking scientific documents. Then, this study uses the TF-IDF weighting method, which functions to determine the weighting value of the validity of scientific journal manuscripts, and the Cosine Similarity Algorithm, which looks for word equations according to scientific journal manuscripts.

The basic principle of the method used in this research is Text Preprocessing using Case Folding, Tokenizing, Filtering, Stemming. Furthermore, the value of word weights in scientific documents, namely TF-IDF, is carried out after the stemming. Also, stopword removal stages are calculated on the value or weight of a word (term) in the document. After obtaining the word weight value, then testing the level of similarity using Cosine similarity, which compares the similarity between documents. In this case, what is being compared is a query with a training document. In calculating cosine similarity, the first thing to do is do a scalar multiplication between the query and the document, add it up, multiply the length of the document and the square of the query length, and then calculate the square root. Furthermore, the scalar multiplication result is divided by the multiplication result of the length of the document and the query. In related works, this method is widely used and shows a good performance in several cases [9]–[15]. Also, it possible to add any improvement to increase its accuracy and performance [16]–[19].

978-1-6654-4402-6/21/$31.00 ©2021 IEEE

II. Methodology

A. Text Processing

Text mining or by other names such as intelligent text analysis, text data mining, or knowledge discovery in a text can simply be interpreted as the process of finding patterns that were not previously seen in certain text documents or sources.

The initial step of text mining is preprocessing. Some of them are case folding or changing text to lowercase and without punctuation or special characters, filtering, namely by removing conjunctions, conjunctions, tokenizing the separation of a document or sentence into each word or term, and Stemming or separating affixes in words into basic word [15], [20].

TABLE I. CASE FOLDING

Document	Result
D1	ada beberapa penelitian terdahulu yang memanfaatkan mikrokontroller arduino uno dan sensor seperti untuk pengontrolan suhu dan ruang
D2	perancangan dan implementasi pengontrol suhu ruangan berbasis mikrokontoller arduino uno dengan sensor suhu lm3s layak digunakan dan diaplikasikan sebagai sistem pengontrol suhu ruangan

The next step is tokenizing. This process converts sentences into words with spaces as the separating feature as described in Table 2.

TABLE II. TOKENIZING

Document 1 (D1)	Document 2 (D2)
ada	perancangan
beberapa	dan
penelitian	implementasi
terdahulu	pengontrol
yang	suhu
memanfaatkan	ruangan
mikrokontroller	berbasis
arduino	mikrokontroller
uno	arduino
dan	uno
sensor	dengan
seperti	sensor
untuk	suhu
pengontrolan	lm3s
suhu	layak
ruang	digunakan
	dan
	diaplikasikan
	sebagai
	sistem

Filtering is removing unnecessary words in the input text. So, the filtered text will be easily processed at a later stage. This study uses 758 words stored in an external dataset. After that is stemming. This step changes the filtered word into a root word and removes the affixes. The results of the filtering and stemming processes can be seen in table 3.

TABLE III. D1 AND D2 AFTER FILTERING AND STEMMING

Document 1 (D1)	Document 2 (D2)
teliti	rancang
manfaat	implementasi
mikrokontroller	control
arduino	suhu
uno	ruang
sensor	basis
control	mikrokontroller
Suhu	arduino
Ruang	uno
	sensor
	suhu
	lm3s
	guna
	Aplikasi
	Sistem

B. TF-IDF and Cosine Similarity

TF is obtained from the number of words that appear in a text, so you can find out how many times the word appears. And after this TF calculation will be continued with the calculation of DF. Table 4 is the process of calculating the weight value to determine the number of words that appear in the citation text and the source of the quote so that it can be calculated by calculating the cosine similarity [21], [22].

TABLE IV. TERM FREQUENCY

No	Word	Term Frequency (TF)	
		Document 1 (D1)	Document 2 (D2)
1	Teliti	1	0
2	Manfaat	1	0
3	mikrokontroller	1	1
4	Arduino	1	1
5	uno	1	1
6	sensor	1	1
7	control	1	2
8	suhu	1	3
9	ruangan	1	2
10	rancang	0	1
11	implementasi	0	1
12	basis	0	1
13	lm3s	0	1
14	guna	0	1
15	aplikasi	0	1
16	sistem	0	1

Then calculate the value of DF (document frequency), which is the number of words that appear in the two documents tested as described in table 5.

TABLE V. DOCUMENT FREQUENCY

No	Word	Document Frequency
1	teliti	1
2	manfaat	1
3	mikrokontroller	2
4	arduino	2
5	uno	2
6	sensor	2
7	control	3
8	suhu	4
9	ruangan	3
10	rancang	1
11	implementasi	1
12	basis	1
13	lm3s	1
14	guna	1
15	aplikasi	1
16	sistem	1

In table 6 are the results of the IDF (*inverse document frequency*) which is the value of the word analysis using (1) which is the value of the times with the TF (*term frequency*) in order to get the weight value.

$$1+Log(N/df+1) \tag{1}$$

TABLE VI. IDF RESULT

Word	DF	IDF 1	IDF 2
Teliti	1	1	0
Manfaat	1	1	0
Mikrokontroller	2	0.5	0.5
Arduino	2	0.5	0.5
Uno	2	0.5	0.5
Sensor	2	0.5	0.5
Control	3	0.5	0.823908741
Suhu	4	0.5	1.301
Ruangan	3	0.5	0.823908741
Rancang	1	0	1
Implementasi	1	0	1
Basis	1	0	1
lm3s	1	0	1
Guna	1	0	1
Aplikasi	1	0	1
Sistem	1	0	1

After calculating the IDF (inverse document frequency), the next step is to analyze the IDF, the TF value, and the IDF value that have been obtained from the above calculations, then the TFIDF analysis is a calculation to get the weighted value of the quoted text and the source of the citation tested. The following table 7 is a TFIDF analysis table.

TABLE VII. TF-IDF RESULT

No	Word	TF		IDF		TF IDF	
		D1	D2	D1	D2	D1	D2
1	Teliti	1	1	1	1	1	0
2	Manfaat	1	1	1	1	1	0
3	mikrokontroller	1	0.5	0.5	0.5	0.5	0.5
4	Arduino	1	0.5	0.5	0.5	0.5	0.5
5	Uno	1	0.5	0.5	0.5	0.5	0.5
6	Sensor	1	0.5	0.5	0.5	0.5	0.5
7	Control	1	0.5	0.5	0.5	0.5	0.82
8	Suhu	1	0.5	0.5	0.5	0.5	1.30
9	Ruangan	1	0.5	0.5	0.5	0.5	0.82
10	Rancang	0	0	0	0	0	1
11	Implementasi	0	0	0	0	0	1
12	Basis	0	0	0	0	0	1
13	lm3s	0	0	0	0	0	1
14	Guna	0	0	0	0	0	1
15	Aplikasi	0	0	0	0	0	1
16	Sistem	0	0	0	0	0	1

The cosine similarity calculation is described in table 8 where document 1 is the citation text and document 2 is the source text of the citation. The citation text is multiplied by the citation source text, the citation text and the citation source are squared so that the total is obtained to find the text-similarity value.

TABLE VIII. CALCULATION OF COSINE SIMILARITY

Word	TF IDF				
	D2	D2	D1.D2	D1^2	D2^2
Teliti	1	0	0	1	0
Manfaat	1	0	0	1	0
mikrokontroller	0.5	0.5	0.25	0.25	0.25
Arduino	0.5	0.5	0.25	0.25	0.25
Uno	0.5	0.5	0.25	0.25	0.25
Sensor	0.5	0.5	0.25	0.25	0.25
Control	0.5	0.82	0.41	0.25	0.67
Suhu	0.5	1.30	0.65	0.25	1.69
Ruang	0.5	0.82	0.41	0.25	0.67
Rancang	0	1	0	0	1
implementasi	0	1	0	0	1
Basis	0	1	0	0	1
lm3s	0	1	0	0	1
Guna	0	1	0	0	1
Aplikasi	0	1	0	0	1
Sistem	0	1	0	0	1
			2.47	3.75	11.05

$$Crossproduct = 6.13 \qquad (2)$$

$$A.B = 2.4744\ ,\ |A|.\ |B| = 6.13 \qquad (3)$$

$$Cos\ (A,B)= 0.4036 \qquad (4)$$

Based on the results of the manual algorithm calculation above (4), the similarity value in the text is 0.4036. In this study, the level of similarity will be divided into 3 categories, namely high, medium, and low. The low level is in the range of 0-0.3, the medium level is in the range of 0.3-0.6, and the high level is in the range of 0.6-1.0. The result of (4) is 0.4036 so that it is known that the text has a medium level.

III. RESULT AND DISCUSSION

The system test scenario defines the amount of data used and what processes are carried out in the system to determine the error in a citation of this scientific journal. The process carried out in testing this system is data collection, text preprocessing, weighting, and text similarity Cosine Similarity. The data used in testing this system is an online journal available on the internet by looking for references to the bibliography listed in the document to be tested and inputting the citation text. 10 of the 50 test samples can be seen in the table below.

TABLE IX. TEST RESULT

No.	Citation	Result	Similarity Level
1	T001	0.43352308	Medium
2	T002	0.12427994	Low
3	T003	0.07988254	Low
4	T004	0.16726948	Low
5	T005	0.76533020	High
6	T006	0.08811075	Low
7	T007	0.02087507	Low
8	T008	0.05383300	Low
9	T009	0.16374133	Low
10	T010	0.24542823	Low

In the table above, there is a threshold, that is, if the similarity of the text is 0-0.3, it means that the similarity level is low. If the similarity of the text is 0.3-0.6 then the similarity is medium. If the similarity text results in a value of 0.6-1.0, the similarity level is high as described in the section on methodology.

Based on the 50 samples the data tested has a different similarity value for each text input. The thing that affects the similarity value is the number of word weights tested. In the first data sample, the number of text 1 has 9 words and the input text 2 has 15 words, with the number of word weights that are not too far in the range. In the second input text test, between text 1 and text 2, the number of word weights is not proportional, so it is dominated by text 2. And the similarity value is tested based on the weight of the words that appear in the two input texts. Documents with medium and high levels concerning the suitability of the quotation and the source of the citation are 70%, and the low level is 30%, unequal documents are documents that do not have a match between the quotation and the source of the citation. The results of data validation checking scientific journal documents are that 70% of the citations in the article are true and 30% of the citations are incorrect.

IV. Conclusion

Checking scientific journal citation in testing sample data, it is found that the text is not similar. The citation text that does not have a similarity value, the text close to similar is the text that has a value close to similar, and similar is the text that has a similarity value. TF-IDF weighting in testing one citation text with one citation source text that comes from the bibliography, then the weighting greatly affects the value of the similarity of the text. The similarity of Cosine Similarity in testing one citation text with one citation source text that comes from the bibliography. Then the text-similarity value is obtained by the threshold value and knowing the text is included in several categories, namely not similar, close to similar, and similar with similar accuracy and Approaching similar, which is 70% and not 30% similar because of similarity and approaching similar. There is an agreement between the quote and its source and the word weight value that affects it.

References

[1] M. Golosovsky, *Citation Analysis and Dynamics of Citation Networks*. Cham: Springer International Publishing, 2019.

[2] H. F. Moed, "Citation analysis in research evaluation," in *Proceedings of ISSI 2005: 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005, vol. 2, pp. 437–441.

[3] A. Shabani and R. Saadat, "Examining the citations received by DOAJ's journals from ISI Web of Science's articles (2003-2008)," in *Proceedings - 3rd International Conference on Information Sciences and Interaction Sciences, ICIS 2010*, 2010, pp. 109–113.

[4] X. Su, A. Prasad, M. Y. Kan, and K. Sugiyama, "Neural multi-task learning for citation function and provenance," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2019, vol. 2019-June, pp. 394–395.

[5] M. Roman, A. Shahid, S. Khan, A. Koubaa, and L. Yu, "Citation Intent Classification Using Word Embedding," *IEEE Access*, vol. 9, pp. 9982–9995, 2021.

[6] Y. Yang and Y. Zheng, "The effect of open access journals on citation impact: A citation analysis of open access journals using Google Scholar," in *4th International Conference on Cooperation and Promotion of Information Resources in Science and Technology, COINFO 2009*, 2009, pp. 278–280.

[7] H. H. Bi, J. Wang, and D. K. J. Lin, "Comprehensive citation index for research networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1274–1278, 2011.

[8] P. A. H. Kazi and M. S. Patwardhan, "Context based citation summary of research articles: A step towards qualitative citation index," in *IEEE International Conference on Computer Communication and Control, IC4 2015*, 2016.

[9] P. P. Gokul, B. K. Akhil, and K. K. M. Shiva, "Sentence similarity detection in Malayalam language using cosine similarity," in *RTEICT 2017 - 2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, 2017, vol. 2018-January, pp. 221–225.

[10] W. Darmalaksana, C. Slamet, W. Budiawan Zulfikar, F. Fadillah, D. Sa'adillah Maylawati, and H. Ali, "Latent semantic analysis and cosine similarity for hadith search engine," vol. 18, no. 1, pp. 217–227, 2020.

[11] S. Roopak and T. Thomas, "A novel phishing page detection mechanism using HTML source code comparison and cosine similarity," in *Proceedings - 2014 4th International Conference on Advances in Computing and Communications, ICACC 2014*, 2014, pp. 167–170.

[12] S. Pattnaik and A. K. Nayak, "Summarization of odia text document using cosine similarity and clustering," in *Proceedings - 2019 International Conference on Applied Machine Learning, ICAML 2019*, 2019, pp. 143–146.

[13] S. Zhu, J. Wu, and G. Xia, "TOP-K cosine similarity interesting pairs search," in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, 2010, vol. 3, pp. 1479–1483.

[14] D. Soyusiawaty and Y. Zakaria, "Book data content similarity detector with cosine similarity (case study on digilib.uad.ac.id)," in *Proceeding of 2018 12th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2018*, 2018.

[15] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8206 LNCS, pp. 611–618.

[16] K. C. Kirana, S. Wibawanto, N. Hidayah, and G. P. Cahyono, "The improved artificial neural network based on cosine similarity in facial emotion recognition," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2019, pp. 45–48.

[17] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," in *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019*, 2019.

[18] C. Beşiktaş and H. A. Mantar, "Real-time traffic classification based on cosine similarity using sub-application vectors," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7189 LNCS, pp. 89–92.

[19] M. Kryszkiewicz, "Determining Cosine Similarity Neighborhoods by Means of the Euclidean Distance," *Intell. Syst. Ref. Libr.*, vol. 43, pp. 323–345, 2013.

[20] M. Biba and M. Mane, "Sentiment analysis through machine learning: An experimental evaluation for Aslbanian," *Adv. Intell. Syst. Comput.*, vol. 235, pp. 195–203, 2014.

[21] M. Irfan, Jumadi, W. B. Zulfikar, and Erik, "Implementation of Fuzzy C-Means algorithm and TF-IDF on English journal summary," in *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017, pp. 1–5.

[22] M. M. Saad, N. Jamil, and R. Hamzah, "Evaluation of Support Vector Machine and Decision Tree for Emotion Recognition of Malay Folklores," *Bull. Electr. Eng. Informatics*, vol. 7, no. 3, pp. 479–486, Sep. 2018.