IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

The optimalization of cosine similarity method in detecting similarity degree of final project by the college students

To cite this article: R A Purba *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **830** 032003

View the article online for updates and enhancements.

This content was downloaded from IP address 124.123.177.82 on 15/08/2020 at 09:34

IOP Conf. Series: Materials Science and Engineering **830** (2020) 032003 doi:10.1088/1757-899X/830/3/032003

The optimalization of cosine similarity method in detecting similarity degree of final project by the college students

R A Purba*, S Suparno and M Giatman

Department of Engineering, Technology and Vocational Education, Padang State University, Padang, Indonesia

*ramen_purba@yahoo.com

Abstract. Technology development makes everything unlimited. Everyone easily to gets information. There are negative and positive impacts. The one negative impact is plagiarism the work of others. Of course this brings a bad impact. This study aims to examine the similarity of Student Final Reports at the Politeknik Unggul LP3M Medan. The method used is the Cosine Similarity Method. This method was chosen because it works based on mathematical calculations. How it works is by comparing the final project done by students with the final project that has been there before. With the Cosine Similitary Method, percentage of similarity will be obtained. If the similarity is high, the final project is said to be tracing.

1. Introduction

Scientific work is a person result thoughts obtained through a research process. Scientific work is not produced easily, but through stages in accordance with established rules [1]. One of the said scientific works is the student's final project. The final project is undertaken by students as one of the requirements for completing their education in college. The final project is carried out through the stages of data collection, data analysis, and continued with the guiding process [2]. The final project must also be maintained before the board of examiners. Politeknik Unggul LP3M is one of the tertiary institutions which requires students to prepare a final project as a condition for completing their education. Conditions that occur in Politeknik Unggul LP3M, many student's plagiarisms their final project from students who have graduated before. So that many similarities are found between one final project with another Final Project. This condition certainly cannot be tolerated because it violates the ethics of publication of scientific papers. The solution must be sought because it is closely related to the academic culture and leadership in Politeknik Unggul LP3M [3]. When there is a similarity finding, of course the one blamed by the leadership for not carrying out monitoring and evaluation of the student's final project [4].

To detecting authenticity scientific work, it will usually be examined by seeing the similarity of a scientific work with scientific work that has already been published. The degree of similarity obtained is usually subjective according to the person who examined it [5]. So the level of accuracy obtained is not necessarily right to decide whether a final project is a plagiarism or not. To detect the similarity of scientific works requires a long time and high accuracy if done manually. Sometimes the student just replaces the word with another word that has the same meaning and meaning between the word. To check this requires time, accuracy and understanding of the language used [6-9].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

To find out whether a scientific work is plagiarism from someone else work will require time and accuracy in analyzing it. Therefore, information technology based tools must be used to do this. The rapid development of information technology makes it possible to do a combination of information technology devices with rapid methods that allow information technology devices to be combined with mathematical concepts. One method that can be used is the Cosine Similitary method [10,11]. This method works based on mathematical calculations. The cosine similarity method is a method for calculating the similarity between two objects expressed in two vectors by using the keyword of a document as a measure. Cosine Similarity method is a method used to calculate the level of similarity between two objects. This method is calculated based on vector space similarity measure. Calculate the similarity between two objects expressed in two vectors by using the keyword of a document as a measure [12].

To increase the speed and accuracy in this research, a combination of Cosine Similarity Method and a website-based computer programming language will be carried out to support the effort to do a similarity test on a scientific work. Work steps based on the Cosine Similarity Method will be entered into the programming language. With a combination of the Cosine Similarity Method with a website-based programming language, the similarity testing process can run quickly, accurately, efficiently, and effectively. The results will also be very high accuracy [13,14].

2. Method

The research method carried out in accordance with the steps of completion by using the method of cosine similarity. The performance steps of the Cosine Similitary algorithm are as follows [15,16]:

- Determined each query, namely the query from the answer (D), the query from the answer key (Q) and the combination of both (Queries).
- All three queries are eliminated stoplist or symbols that do not affect the assessment, such as periods, commas, exclamation points.
- The three queries are eliminated stopwords or common words that are commonly used in a query, as "*dan*", "*jika*", "*di*", "*namun*", "*tetapi*", and other.
- The term frequency query answer and the answer key query to the queries are calculated. So the term calculation in the answer query and the answer key query refers to the terms contained in the queries.
- Calculated the value of document frequency (n) or the number of files (N) that has a term for each term in the queries count inverse document value frequency use formula: log (N/n).
- Multiplied by the term frequency value with the inverse document frequency value for each term in Q or D.
- The scalar multiplication results for each answer query are calculated against the answer key query. The result of multiplication of each answer by query is summed (according to the numerator in the formula above).
- The results of vector multiplication are calculated for each answer key query and answer query.
- Similarity values (different vector values between D and Q) are calculated using the formula:

$$simlsa(d,q) = \sum_{k-1}^{i}(weight_{ik}weight_{qk}) \Big/ \sqrt{\sum_{k-1}^{i}(weight_{ik}{}^{2}weight_{qk}{}^{2})} \quad (1)$$

The steps of the Cosine Similarity Method algorithm can be represented in the flowchart as follows [17]:

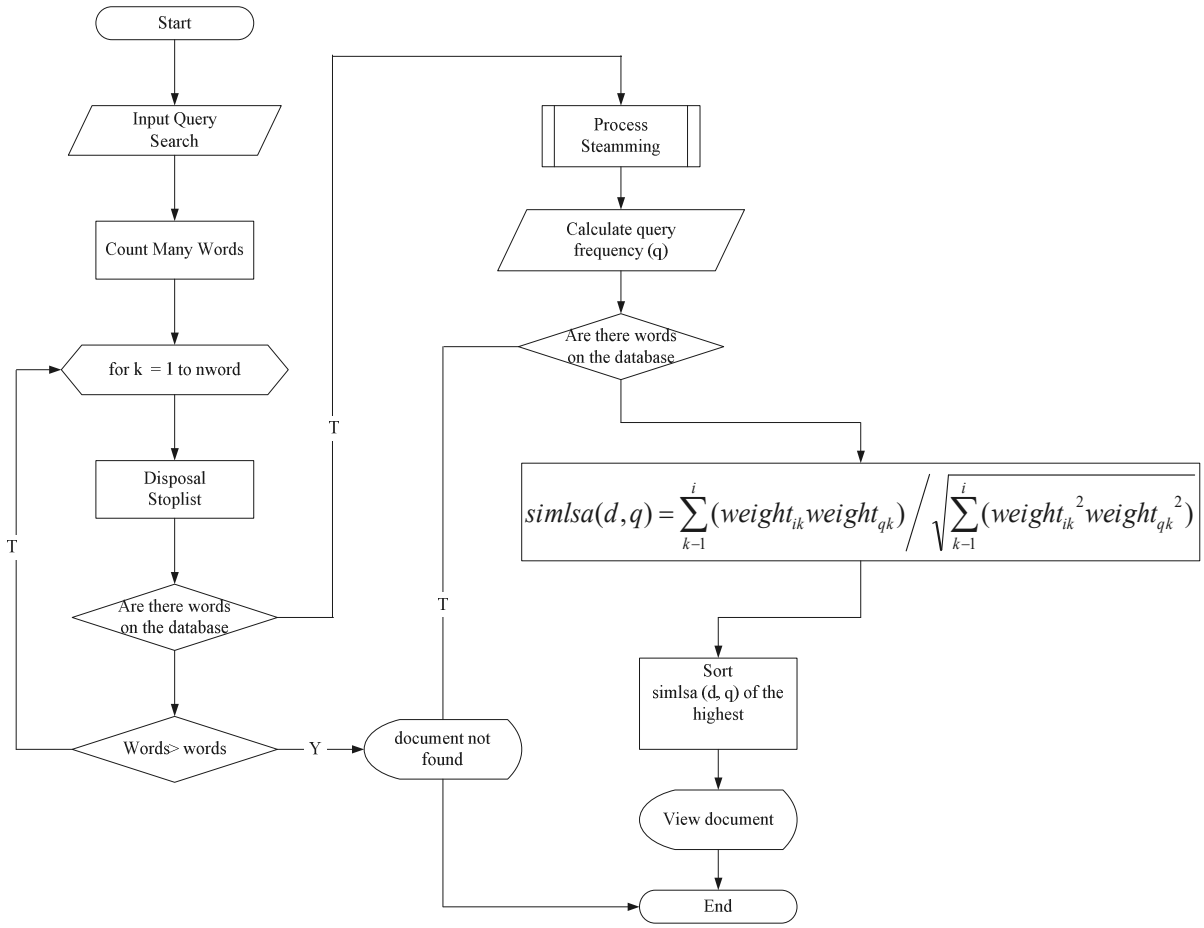

Figure 1. Flowchart cosine similarity algorithm method.

3. Results and discussion

3.1. Results

3.1.1. Define the query sample. Determine query sample, namely the query from the answer (D), the query from the answer key (Q) and the combination of both (Queries):

Table 1. Query sample**.**

Document	***Term that represents a document***
Q	raw information data that has not been processed yet cannot be used as a reference take a decision
D1	a raw data set
D2	raw data that has not been processed
D3	data used as a reference for decision making information

3.1.2. The value of inverse document frequency. Perform the calculation of inverse document frequency with the formula: log (N/n) +1:

Table 2. Calculating inverse value.

No	Term	tf				Df	idf
		Q	D1	D2	D3		Log(n/df)+1
1	reference	1	0	0	1	2	1.30103
2	not yet	2	0	2	2	3	1.12493874
3	can	1	0	1	1	3	1.12493874
4	data	1	2	1	1	4	1
5	made	1	0	0	1	2	1.30103
6	processed	1	0	1	1	3	1.12493874
7	fact	0	0	1	0	1	1.60205999
8	information	0	0	1	1	2	1.30103
9	decision	1	0	0	1	2	1.30103
10	note	1	0	0	0	1	1.60205999
11	take	1	0	0	0	1	1.60205999
12	raw	1	1	1	1	4	1
13	show	0	0	1	0	1	1.60205999
14	making	0	0	0	1	1	1.60205999
15	group	0	1	0	0	1	1.60205999
16	thing	1	0	0	0	1	1.60205999

3.1.3. Frequency term and inverse document frequency. Determine frequency term value with inverse document frequency value:

Table 3. Calculating term frequency and inverse document frequency.

No.	Term	Q	D1	D2	D3
1	Reference	1.30103	0	0	1.30103
2	not yet	2.24987748	0	2.24987748	2.24987748
3	Can	1.12493874	0	1.12493874	1.12493874
4	Data	1	2	1	1
5	Made	1.30103	0	0	1.30103
6	Processed	1.12493874	0	1.12493874	1.12493874
7	Fact	0	0	1.60205999	0
8	Information	0	0	1.30103	1.30103
9	Decision	1.30103	0	0	1.30103
10	Note	1.60205999	0	0	0
11	Take	1.60205999	0	0	0
12	Raw	1	1	1	1
13	Show	0	0	1.60205999	0
14	Making	0	0	0	1.60205999
15	Group	0	1.60205999	0	0
16	Thing	1.60205999	0	0	0

3.1.4. Vector difrent value between D to Q. Calculation similarity value (Vector difrent value between D to Q):

Table 4. The calculation results.

No.	Text	Similarity Value	Value (Similarity Value * weight)
1	D1 : raw data set	0.23058480	23.06 %
2	D2 : raw data that has not been processed	0.50054143	50.05 %
3	D3 : data used as a reference for decision making information	0.71291907	71.29 %

3.2. Discussion

The presence of a final assignment similarity checking system that combines website based programming with the Cosine Similarity Method of the final project examination process of students in the Politeknik Unggul LP3M becomes faster and more effective. Students also feel more challenged and serious in working on their final project. No need to manually check. The system that was built can help Politeknik Unggul LP3M in improving the quality and quality of the final project of its graduate students. The results of testing with a number of student's final project documents show that the system can be used well with a percentage of testing of 85.6%. 30 final project documents where tested, 27 Final Project documents were processed well and 3 data were processed well. To further improve system performance and accuracy, more vocabulary must be entered. The similarity checking process will also be more effective and efficient.

4. Conclusions

Based on the testing of 30 final project documents that have been carried out can be concluded:

- The accuracy of the similarity detection process is greatly influenced by the number of words tested, so that the test results can be maximized, the Final Project being tested is required to use Indonesian in accordance with official spelling.
- The accuracy of the classification of final project documents and the vocabulary of the number of words to be used as testing material are influenced by the complete lack of the number of Indonesian words that have been entered into the system.
- Trials conducted on 30 different student final project documents obtained the highest similarity value of 41%.
- Based on the test results obtained if the Cosine Similarity Method adheres to the context of normalization of vector length by comparing parallel to each other from the 2 documents compared.

References

[1] Maarif A A 2015 Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah *Teknik Informatika Universitas Dian Nuswantoro, Semarang*
[2] Ariantini D A R, Lumenta A S and Jacobus A 2016 Pengukuran Kemiripan Dokumen Teks Bahasa Indonesia Menggunakan Metode Cosine Similarity *Jurnal Teknik Informatika* **9**(1)
[3] Hoekstra A and Newton P 2017 Departmental leadership for learning in vocational and professional education *Empirical Research in Vocational Education and Training* **9**(1) 12
[4] Hauge T E and Norenes S O 2015 Collaborative leadership development with ICT: Experiences from three exemplary schools *International Journal of Leadership in Education* **18**(3) 340-364
[5] Oghbaie M and Zanjireh M M 2018 Pairwise document similarity measure based on present term set *Journal of Big Data* **5**(1) 52
[6] Sohangir S and Wang D 2017 Improved sqrt-cosine similarity measurement *Journal of Big Data* **4**(1) 25
[7] Darmawan R and Wahono R S 2015 Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization *Journal of Intelligent Systems* **1**(2) 109-114
[8] Nandhini K and Balasundaram S R 2014 Extracting easy to understand summary using differential evolution algorithm *Swarm and Evolutionary Computation* 1–9
[9] Satya K P N V and Murthy J V R 2012 Clustering Based On Cosine Similarity Measure *International Journal of Engineering Science & Advanced Technology* **2**(3) 508–512
[10] Wang Y, Li H, Wang H and Zhu K Q 2012 Concept-based web search *In International Conference on Conceptual Modeling* 449-462
[11] Agirre E, Cuadros M, Rigau G and Soroa A 2010 Exploring Knowledge Bases for Similarity *In*

LREC
[12] Zaware S N, Gautam A, Nashte S and Khanuja P An Effectual Approach For Calculating Cosine Similarity *International Journal of Advance Engineering and Research Development* 2348-4470 13-18
[13] Bhavsar V C, Boley H and Yang L 2004 A Weighted-Tree Similarity Algorithm for Multi-Agent Systems in E-Business Environments *Computational Intelligence* **20**(4) 584-602
[14] Corley C and Mihalcea R 2005 Measuring the semantic similarity of texts *In Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment* 13-18
[15] Imbar V, Adelia R, Ayub M and Rehatta A 2014 Implementasi Cosine Similarity dan Algoritma Smith Waterman untuk Mendeteksi Kemiripan Teks *Jurnal Informatika* **10**(1)
[16] Sugiyamta 2015 Sistem Deteksi Kemiripan Dokumen dengan Algoritma Cosine Similarity dan Single Pass Clustering *Jurnal Informatika* **7**(2)
[17] Rizki, Dhidik and Supraptono E 2017 *Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen* (Semarang: Jurusan Teknik Elektro. Fakultas Teknik. Universitas Negeri Semarang Kampus Sekaran, Gunung pati)