Measuring the Level of Plagiarism of Thesis using Vector Space Model and Cosine Similarity Methods

IOP Conf. Series: Materials Science and Engineering **662** (2019) 022111 doi:10.1088/1757-899X/662/2/022111

Measuring the Level of Plagiarism of Thesis using Vector Space Model and Cosine Similarity Methods

I Indriyanto[1*], I D Sumitra[2]
[1]Pascasarjana Departemen Sistem Informasi, Universitas Komputer Indonesia, Bandung, indonesia
[2]Information Management Department, Universitas Komputer Indonesia, Bandung, Indonesia

Email : indriyanto@email.unikom.ac.id

Abstract. The purpose of this research is a computer system can assist in detecting and measuring plagiarism of scientific work quickly. There are many techniques that can be used to measure the level of plagiarism of a document. in paper, we will use the Vector Space Model and the research method used namely Cosine Similarity to measure the level of similarity of the thesis. The result of this method was comparing the level of thesis similarity with the dataset using the TF and TF-IDF techniques in graphical form. It can be concluded from the experimental process, using the TF-IDF technique produces a smaller value compared to using the TF technique.

1. Introduction

Plagiarism in scientific papers can only occur due to accidental or intentional. Therefore, it is important to know what the meaning of plagiarism and what is included in plagiarism and the potential for plagiarism. According to Republic of Indonesia Minister of Education Regulation Number 17 of 2010, plagiarism is an intentional or unintentional act of obtaining the value of a scientific work, by quoting part or all of the scientific work of others without including sources [1]. plagiarism as an intellectual use of other people (such as text, ideas, or images) into scientific work that is done intentionally or unintentionally without including the source [2]. The definition of plagiarism in a research report needs to consider factors such as the originality of the material being copied, the position in the report, and references [3].

Text matching software is a tool that is useful for measuring the level of copying text but cannot detect tables or numbers traced plagiarism of ideas, or plagiarism in translations. Vector Space Model (SVM) can be used to measure the level of similarity of a text document, the core of SVM is that each document or query is represented by the words contained therein (indexing). A vector consisting of these words can be defined to describe each part of the document and query, so the document can be determined similarity based on the results of correlation calculations between documents [3, 4]. cosine similarity approach to measure the level of closeness of a sentence [5]. Scientific works that will be published need to be examined beforehand about their scientific value, which should not be plagiarism in them. Checking scientific work documents can be done computerically[3].

The purpose of this research is to create a system that can detect plagiarism of scientific works by combining the Vector Space Model and the method used in this study is the Cosine Similarity method to measure the level of similarity of scientific works with other scientific works.

2. Methods

2.1. Vector Space Model

Vector Space Model is an algebraic model for representing text documents as an identification vector, for example, the word index. VSM is usually used in information filtering, information retrieval, indexing, and relevance ranking [6]. The basic thought of the VSM method is to represent each independent word and each document expressed in a vector so that the complexity of the relationship of words becomes simple and can be calculated. In VSM, each document consists of Term (T1, T2, ..., Tn) and each Term Ti has Wi weight. The term (T1, T2, ..., Tn) is considered as one of the vector elements in the N-dimensional coordinate system [7].

TF-IDF is a weighting scheme that is often used in VSM along with cosine similarity to determine the similarity between two documents. *TF-IDF* considers the frequency of different words in all documents and is able to distinguish documents. In VSM, each vector is composed of terms and weights representing the document. The document similarity can be expressed by the angle or distance between vectors, the smaller the angle or distance means the more similar the two documents. *TF* is Term Frequency and *IDF* is Inverse Document Frequency. The formula is as follows [8]:

$$W_{t,d} = TF_{t,d} * IDF_t \qquad (1)$$

Information:
$W_{t,d}$ = weight of t (*Term*) in one document
$TF_{t,d}$ = frequency of occurrence of t (*Term*) in the document d
IDF_t = Inverse document frequency, where

$$IDF_t = log\left(\frac{N}{n_t}\right) \qquad (2)$$

Information:
N = Amount of all documents
n_t = Number of documents containing *Term* t

The *IDF* reflects the spread of the Term in the entire document so that it can show the difference in Term t in each document. *TF* reflects the spread of the Term in a document. *TF-IDF* can make exceptions for high-frequency words but have little in common, so *TF-IDF* is an effective algorithm for calculating Term weights t.

After weighing each *Term* is done, it takes a calculation to rank to measure the similarity between the query vector and the document vector that will be compared. One method commonly used in similarity calculations is cosine measurements, which determine the angle between document vectors and query vectors and are defined as follows :

$$Similarity\left(\overrightarrow{d_j}, \vec{q}\right) = \frac{\overrightarrow{d_j}.\vec{q}}{|\overrightarrow{d_j}|.|\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij}.w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \sum_{i=1}^{t} w_{iq}^2}} \qquad (3)$$

Where is the weight of the terms, the denominator in this equation is called the normalization factor which serves to eliminate the effect of the length of the document[4] Normalization is needed because where long documents will tend to have greater value because they have a large frequency of occurrence of words.

The ranking process of documents can be considered as a process of selecting (vector) documents that are close to (vector) queries this closeness is indicated by the angle formed. The cosine value that tends to be large indicates that the document is more and more the same as the query. The cosine value equal to 1 indicates that the document matches the query [9].

2.2. Cosine Similarity

In general, the similarity function is a function that accepts two objects and returns a similarity between the two objects in the form of real numbers. Generally, the values generated by the similarity function range in the interval [0 ... 1]. But there are also several similarity functions that produce values that are

outside the interval. To map the results of these functions at intervals [0 ... 1] normalization can be done [10].

Cosine similarity is the calculation of similarities between two vectors and dimensions by looking for cosines from the angle between them and often used to compare documents in text mining [11]. The *Cosine similarity* formula is as follows:

$$Similarity(x,y) = cos(\theta) = \frac{x,y}{\|x\|\|y\|} \qquad (4)$$

Information

x,y : the dot product vector of x and y, calculated by $\sum_{k=1}^{n} x_k y_k$ (5)

$\|x\|$: length of vector x, calculated by $\sum_{k=1}^{n} x_k^2$ (6)

$\|y\|$: length of vector y, calculated by $\sum_{k=1}^{n} y_k^2$ (7)

Pang-NingTan explained that the greater the results of the similarity function, the two objects evaluated are considered to be more similar[12]. If the opposite, then the smaller the result of the similarity function, the two objects are considered to be increasingly different. In a function that produces a value in the range [0 ... 1], the value 1 represents the two objects are exactly the same, while the value 0 represents the two objects is completely different.

3. Results and Discussion

3.1. Plagiarism Detection Technique

The methodology that been used in this research briefly can be seen in Figure 1. The detail explanation from each stage can be seen below:

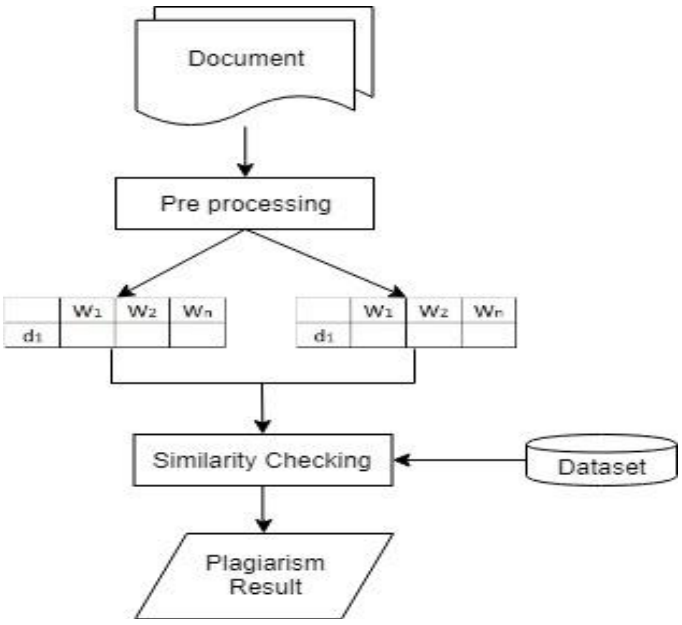

Figure 1. The stage of the Plagiarism Detection Technique.

The dataset is taken from the thesis of informatics students and information systems. Data collected has different file sizes and uses Indonesian. the thesis document consists of several chapters and the greatest possibility of frequent plagiarism is found in Chapter 1 which is the introductory chapter and Chapter 4 of the design used in the study.

Pre-processing is the initial stage in breaking up a thesis document into parts of the chapter that will be processed later. All sentences in the thesis document will be changed into basic words by stemming

the process. The next process is the calculation of *Term Frequency (TF)* is a process to find the number of occurrences of words in a document. And the calculation of the *term frequency inverse document frequency (TF-IDF)* is a joint process between the *term frequency (TF)* and the *inverse document frequency (IDF)* where the *IDF* is a weighting that measures how important a word is in a document when viewed globally throughout the document.

The next step is to calculate the similarity of documents with the cosine similarity technique. The calculation stage of the plagiarism level uses *TF* and *TF-IDF* which are stages to find out how much the level of plagiarism of a document. The final result of the calculation process is the percentage of similarity between documents compared to other documents in the dataset.

3.2. Experiment Result

Trial 1, comparing the level of similarity using *TF* and *TF-IDF* with stemming in the thesis document. Can be seen in Figure 2.

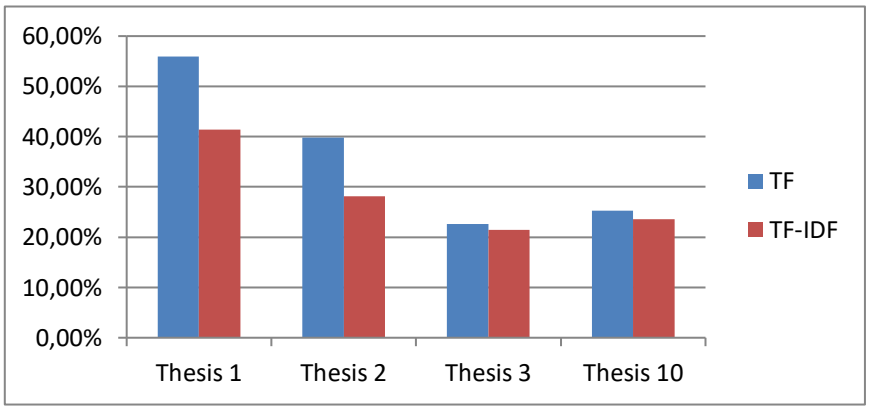

Figure 2. Similarity using *TF* and *TF-IDF* with stemming.

The measurement results using *TF* and *TF-IDF* with stemming to measure the level of similarity of a thesis, *TF-IDF* technique has a smaller value than using the *TF* technique. Trial 2, comparing the level of similarity using *TF* and *TF-IDF* with stemming in the thesis document. Can be seen in Figure 3.

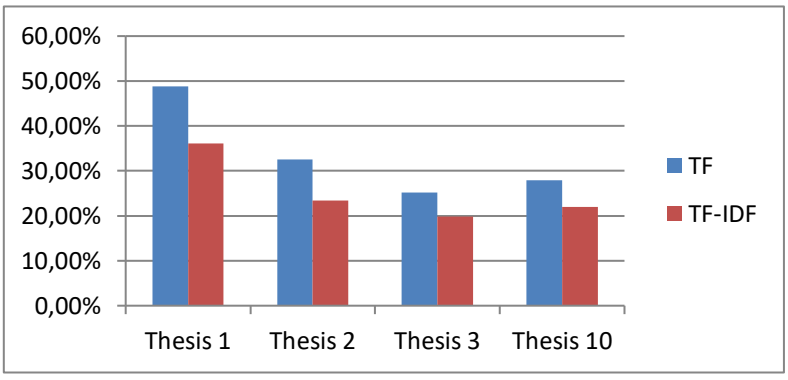

Figure 3. Similarity using *TF* and *TF-IDF* without stemming.

The measurement results using *TF* and *TF-IDF* without stemming to measure the level of similarity of a thesis, *TF-IDF* technique has a smaller value than using the *TF* technique. The trial aims to measure the level of similarity of scientific papers using *TF* or *TF-IDF*. The results of the trial can be seen in Table 1.

Table 1. The level of similarity of documents with *TF* and *TF-IDF*

Dokumen	*TF*		*TF-IDF*	
	Stemming	Whitout Stemming	Stemming	Whitout Stemming
Thesis 1	55,97%	48,82%	41,38%	36,07%
Thesis 2	39,79%	32,49%	28,10%	23,45%
Thesis 3	22,59%	25,20%	21,47%	19,79%
...				
Thesis 10	25,30%	27,89%	23,60%	21,98%

From the results of measuring the similarity of documents using *TF* and *TF-IDF* with stemming and without stemming. shows the percentage similarity using *TF-IDF* has the smallest value compared to using *TF*.

Similar research was conducted by comparing the Jaccard similarity method with the Cosine similarity method on Electronic Text based Assignments. It can be concluded that the cosine similairty approach is suitable for detecting the plagiarism level of text documents rather than the Jaccard similarity method. Although a little longer with the Jaccard similarity method in the process [14].

4. Conclusion

In detecting plagiarism, a thesis in Indonesian text can be done using the Vector Space Model and Cosine Similarity method. TF and TF-IDF techniques are used to calculate the level of similarity of the two documents, where the TF-IDF technique has the smallest value compared to the TF technique.

Acknowledgements

I would like to express my sincere gratitude to Mr. Assoc. Prof. Dr. Ir. Eddy Soeryanto Soegoto, MT as Rector of Universitas Komputer Indonesia (Unikom), and Mr. Dr. Yeffry Handoko Putra, ST., M.T as Head of Master of Information Systems. Special thanks are given to Irfan Dwiguna Sumitra, M.Kom., Ph.D who has given full support and guidance so that this paper can be realized.

References

[1] Muresan, D., & Sinuraya, R. 2018, August. Relation between internet and social media to support sales in business. In IOP Conference Series: Materials Science and Engineering **407**(1) pp. 012062. IOP Publishing.

[2] Helgesson, G., & Eriksson, S. 2015. Plagiarism in research. *Medicine, Health Care and Philosophy*, **18**(1), pp.91-101.

[3] Wager, E. 2014. Defining and responding to plagiarism. *Learned publishing*, **27**(1), pp.33-42.

[4] Triana, A., Saptono, R., & Sulistyo, M. E. 2014. Pemanfaatan Metode Vector Space Model dan Metode Cosine Similarity pada Fitur Deteksi Hama dan Penyakit Tanaman Padi. *ITSMART: Jurnal Teknologi dan Informasi*, **3**(2), pp .90-95.

[5] Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In 2016 4th International Conference on Cyber and IT Service Management (pp. 1-6). IEEE.

[6] Yubo, J., Xing, D., Yi, W., & Hongdan, F. 2011. A Document-Based Information Retrieval Model Vector Space. In 2011 Second International Conference on Networking and Distributed Computing (pp. 65-68). IEEE.

[7] Guo, Q. 2008. The similarity computing of documents based on VSM. In International Conference on Network-Based Information Systems (pp. 142-148). Springer, Berlin, Heidelberg.

[8] Lee, D. L., Chuang, H., & Seamons, K. 1997. Document ranking and the vector-space model. IEEE software, **14**(2), pp. 67-75.

[9] Firdaus, C., Wahyudin, W., & Nugroho, E. P. 2017. Monitoring System with Two Central Facilities Protocol. *Indonesian Journal of Science and Technology*, **2**(1), pp 8-25.

[10] Munir. 2015. the use of multimedia learning resource sharing (MLRS) in developing sharing knowledge at schools. *International Journal of Multimedia and Ubiquitous Engineering*, ***10***(9), pp .61-68.

[11] Zhiqiang, L., Werimin, S., & Zhenhua, Y. 2009. Measuring semantic similarity between words using wikipedia. In *2009 International Conference on Web Information Systems and Mining*. **101**(2), pp. 251-255. IEEE.

[12] Foltz, P. W., Laham, D., & Landauer, T. K. 1999. The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, **1**(2), pp. 939-944.

[13] Soegoto, E. S., & Jayaswara, M. R. 2018. Web and Android Programming Course Information System. In *IOP Conference Series: Materials Science and Engineering* **407**(1),pp. 012063.

[14] Jiffriya, M. A. C., Jahan, M. A., & Ragel, R. G. 2014. Plagiarism detection on electronic text based assignments using vector space model. In *7th International Conference on Information and Automation for Sustainability* (pp. 1-5). IEEE.