5thICITB

A Comparison of the Vector Space Model Method and Winnowing Algorithm to Measure the Similarity of Documents

Eva Y Puspaningrum[1], Budi Nugroho[2], Firza Prima A[3]
[1,2,3] *Informatics, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Jl.Rungkut Madya no.01, Surabaya, Indonesia*
Evapuspaningrum.if@upnjatim.ac.id

ABSTRACT

The growth of information and communication technology has increased significantly from year to year. The issue that is developing now is the number of documents that are copied and paste. The amount of text data is constantly increasing in cyberspace so that everyone can easily find the documents they need. Because of these problems, measuring the similarity of the two documents is necessary and is fundamental to detecting plagiarism from many different documents. In this work, we would like to compare the effectiveness of the algorithm used to measure the similarity between two documents. Winnowing and SVM algorithms are widely used to compare documents because the plot is easy to understand and easy to use. The Experiment Result, we can find that the performance of fingerprints and winnowing is better than VSM. Moreover, the winnowing algorithm is more stable than others.

Keywords: Vector Space Model, Winnowing, Similarity of Documents

1. INTRODUCTION

The growth of information and communication technology has increased significantly from year to year. Access to information is very fast and easy. This can have a positive or negative impact. One of negative impact is plagiarism. Plagiarism is the act of plagiarizing or copying other people's work then claiming it as the result of his own work and not including references from the original source or also known as duplicate or copy [1]. Besides that there are also positive impacts one of which is the Search Engine. Search Engine is a system that can be used to find information that is relevant to the needs of its users automatically from a collection of information [2]. This system will receive input in the form of keywords from the information to be searched. With a relatively shorter time will provide the results of several documents or information relevant to the keywords entered by users.

The issue that is developing now is the number of documents that are copied and paste. The amount of text data is constantly increasing in cyberspace so that everyone can easily find the documents they need. There are many documents on the internet that misuse these documents. Because of these problems, measuring the similarity of the two documents is

necessary and is fundamental to detecting plagiarism from many different documents [3]. Comparing similarities between documents has goals such as checking plagiarism, text classification and information retrieval. A lot of research has been done about detecting similarity from documents. There is a lot of research on this field with many different algorithms. The methods can be divided into String-based, Corpus-based and Knowledge-based Similarities [4]. String-based measures determines the similarity by operating on string and character. String-based method is divided into character-based and terms-based approaches. String similarity measures on string sequences and character composition. A string metric is a metric that measures similarity between two text strings for approximate matching or comparison [5]. In term-based apply the cosine similarity measure [6]. The character-based measure uses k-gram which is a sub-string sequence to find fingerprint based [7]. In this paper will calculate the similarity of text with a string-based method. Where in this research, term-based approaches will use the Vector Space Model (VSM) method and character- based approaches will use the winnowing algorithm. The essence of the VSM method is the basis of each document or request represented by the words that are in it (indexing). The vector that consists of these words can be determined to be revised every part of the document and request, then the document can be determined related to the request or not in accordance with the results obtained from them. Documents that have greater relativity with a particular search are considered more related [8]. Winnowing uses the window as the method, i.e. the formation of windows after the hashing process. After the formation of a window containing the hash value, the smallest hash value of each window is selected [9]. Fingerprint is what will be the basis of comparison between text files that have been inputted [10]. A document is said to have plagiarized with other documents if it has a high degree of similarity or exceeds the specified tolerance limit.

2. LITERATURE REVIEW

2.1 Text Pre-Processing

Text pre-processing is the first step in processing text mining. Pre-processing is a process to eliminate parts that are not needed for text processing. Pre-processing method is very important role in text mining techniques [12]. There are several initial stages called pre-processing that need to be done, namely case folding, tokenizing, filtering, stemming [11]. Case folding stage is the stage of changing each letter, where capital letters become lowercase letters. Tokenizing is the stage of truncating an input string based on each constituent word. In this process, the input character cuts into symbols, punctuation marks, or other elements that have meanings called tokens. Filtering is the removal of terms or words that are considered to have no meaning or irrelevance, usually called the stop-word process. Stop-word must be removed from a text because it can make the text heavier and less important for the text mining process. Stemming is used to convert these words into basic words by using certain rules to reduce the index results without having to eliminate the meaning.

2.2 Vector Space Model

Vector Space Model (VSM) is used to present a document in vector space [14]. Vector Space Model is an algebraic model for representing a text document as an identification vector, for example the word index. VSM is usually used in information filtering, information retrieval, indexing, and ranking of relevance [8]. The basis of the VSM method is to represent each word independently and each document expressed in a vector so that the complexity of the relationship of words is simple and can be calculated. In VSM, each document consists of terms (T1, T2, ..., Tn) and each term Ti has a weight of Wi. The terms (T1, T2, ..., Tn) are considered as one of the vector elements in the N-dimensional coordinate system. TF-IDF is a weighting scheme that is often used in VSM together with cosine similarity to determine the similarity between two documents. TF-IDF considers the different frequency of words in all documents and is able to distinguish documents. In VSM, each vector is composed by terms and weights that represent documents. The similarity of documents can be expressed by the distance between vectors, the smaller the distance means the more similar the two documents. The formula is as follows 1:

$$w = tf * idf \tag{1}$$

W is weight of term in one document, tf is frequency of occurrence of term in the document and idf is Inverse document frequency, where formula 2

$$idf = \log(\frac{N}{df}) \tag{2}$$

N is number or number of documents in the collection and df is number of documents containing term.
After weighting each term it will measure the similarity between the query vector and the document vector to be compared. One method commonly used in the calculation of similarities is cosine similarity, which determines the distance between document vector and the query vector. If the cosine value equals 1 indicates that the document matches the query or the cosine value equals 0 that the document does not equal the query [15].

2.3 Winnowing

Winnowing is an algorithm used to process document fingerprint [16]. The winnowing algorithm calculates the hash value on every k-gram. The k-gram method is a method that functions to break a word or sentence into a series of length n characters. Then, windows are formed from windows hash value. Hashing converts a series of characters into a value that becomes a series of values and the resulting value is called a hash value. The minimum hash value is selected in each window [17].
The hashing technique has a formula to calculate the next character hash value, H (c2 ... ck + 1) shown in Formula 3.

$$H(C_2..C_{k+1}) = (H_{(C1..Ck)} - C1 * b^{k-1}) * b + C_{(k+1)} \tag{3}$$

C is ASCII character value, b is base (number of character), k is the number of k-grams.

Window is a series of hash values grouped based on the value of w which shifts one hash to the right until the hash value is used up. Window grouping is adjusted to the window value. For example window value is 4, Hash value is {1 2 3 4 5 6 7} then the window grouping is [1 2 3 4], [2 3 4 5], [3 4 5 6], [4 5 6 7].
From this window, the minimum fingerprint value will be chosen to be the hash value. If there are more than two hashes, the smallest hash value on the far right will be used. After the fingerprint is selected, the similarity is calculated using the Jaccard coefficient. For example, step from winnowing algorithm is shown in Figure 1.

k-gram of lenght 3	informatika → [inf], [nfo], [for], [orm], [rma], [mat], [ati], [tik], [ika]
Hashing	[4683], [4452], [4789], [4855], [4622], [4293], [4913], [4519]
Window of lenght 5	[4683, 4452, 4789, 4855, 4622];
	[4452, 4789, 4855, 4622, 4293];
	[4789, 4855, 4622, 4293, 4913];
	[4855, 4622, 4293, 4913, 4519];
Minimum Fingerprint value	4452, 4293

Figure 1: Example for winnowing algorithm

3. METHOD

The process of processing data in the form of text obtained from documents to look for words that can represent the contents of the document so that it can be analysis the relationship between documents. The process of analysis the text in order to extract useful information for a particular purpose. The initial stage for text processing is called text pre-processing. In this stage there are several stages that must be passed. These stages shown in Figure 2.

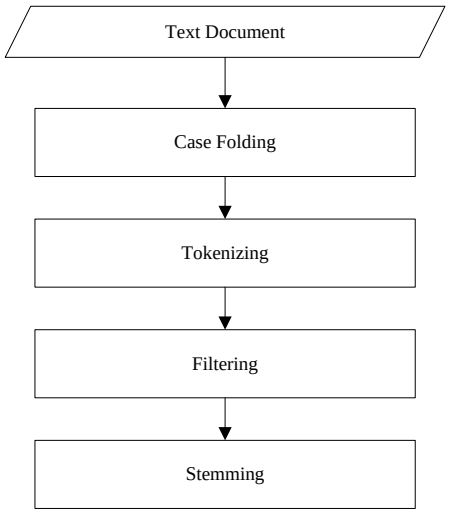

Figure 2: Pre-Processing Text

The data used in this study is in the form of Indonesian language documents. Where in the testing process in this study is to compare 2 documents to calculate the value of similarity.

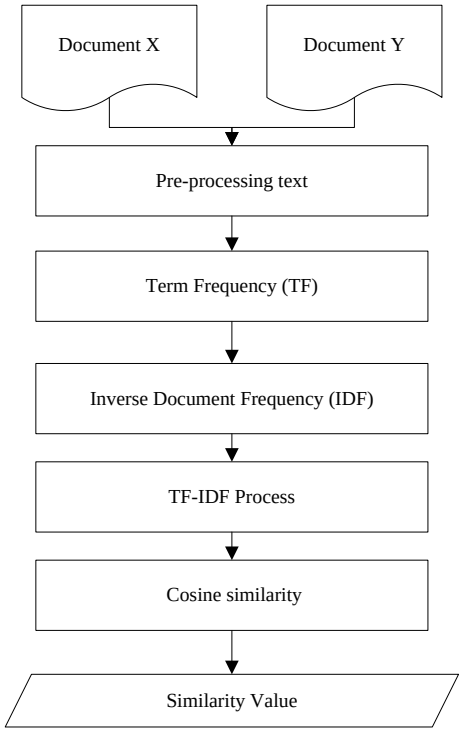

Figure 3: Vector Space Model Algorithm

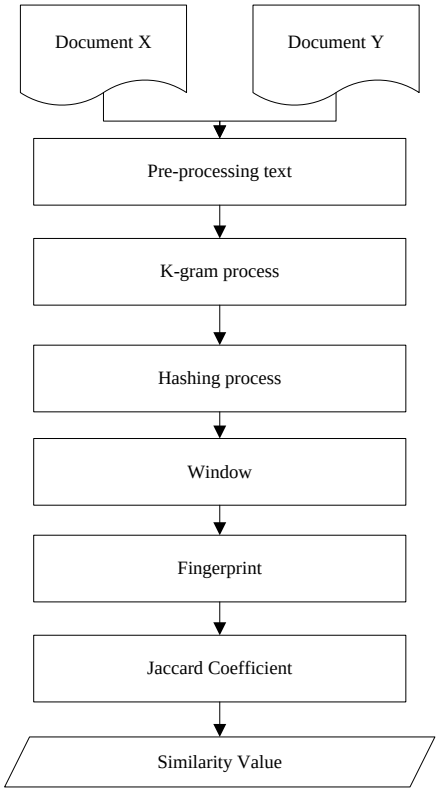

Figure 4: Winnowing Algorithm

In this research comparing two methods in text processing namely SVM and winnowing. The Flowchart of the 2 methods can be seen in Figure 3 and 4. The difference between these two methods is that SVM uses term-based while Winnowing uses character based. Besides that the difference is also in the measurement method similarity. SVM uses cosine similarity while Winnowing uses Jaccard Coefficient.

4. RESULTS AND DISCUSSION

Before comparing documents, we need to pre-processing text. Preprocessing through steps: words are changed to lowercase and separated from strings. Stop words are removed of terms or words irrelevance. Stop words are words that are non-descriptive for the topic of a document such as the, a, and, is. After that, Stemming is used to convert these words into basic words. For example process, processing and processor will be convert to the stem process. In this research, we construct two document X and Y as follows: two documents are consist of 200 words per each document. The experiment was conducted with 5 different test cases, test case 1 used 20 documents to compare with theme A. Test case 2 used documents with theme B, Test case 3 used documents with theme C, Test case 4 used documents with

theme D, Test case 5 uses documents with the theme E. The results of the cosine similarity algorithm are shown in Table 1.

Table 1: Result of SVM Method.

Test cases	The similarity rate (%)
1	31.44
2	32,21
3	34,66
4	35,87
5	21,31
average	31,0125

Winnowing algorithm uses k-gram which the values of k-gram are {2, 3, 4} and window size are chosen as {4, 6, 8, 10}. The pairs of parameter values [k-gram, window] are selected from the combinations of value given the best results.

Table 2: Result of Winnowing Method.

Test cases	Parameter Values [k-gram, window]			
	[2,4]	[2,6]	[3,6]	[4,6]
1	34,31	33,31	32,13	31,76
2	33,18	35,19	35,61	32,68
3	34,87	34,65	34,88	30,26
4	35,01	35,31	35,21	31,54
5	23,12	23,12	23,10	20,54
Average	32,098	32,316	32,186	29,536

From the Table 2 shown that the parameter pair of winnowing algorithm with k-gram = 2 and window = 6 are the best. The winnowing algorithm has a k-gram process and the window can be changed. It was concluded that the lower the size of the k-gram value, the higher the similarity value produced. But that does not mean that a low k-gram value will give an accurate accuracy value. The smaller the k-gram value, the smaller the characters that will match and the more often these characters are found in the text.

5. CONCLUSIONS

Winnowing and SVM algorithms are widely used to compare documents because the plot is easy to understand and easy to use. This research uses this algorithm to measure the similarity between two documents. From the results of the experiment, winnowing algorithm have better performance than SVM. The winnowing algorithm more stable with different parameter pairs, but the weakness is the dependence on k-gram configuration parameters. If winnowing is found by k-gram characters, it requires more time but gives better results than k-gram words. Therefore, choosing the right approach and setting parameters will provide better measurement performance similarities between the two documents.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge that the present research is supported by Departement of Informatics Faculti Computer Science, Universitas Pembangunan Nasional veteran jawa Timur

REFERENCES

[1] Duan, X. Wang, M., Mu, J. 2017. A Plagiarism Detection Algorithm based on Extended Winnowing. *MATEC Web of Conferences 128, 02019* DOI: 10.1051/matecconf/201712802019 EITCE 2017

[2] Rila, M., & Setiawan, H. 2001. Improving Information Retrieval System Performance by Automatic Query Expansion. *Journal of, Departemen Teknik Informatika*, Institut Teknologi Bandung

[3] Tung, K.T. Hung, N.D. and Hanh, L.T.M. 2015. A Comparison of Algorithms used to measure the Similarity between two documents. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 4 June*.

[4] Gomaa, W.H. and Fahmy, A.A. 2013. A Survey of Text Similarity Approaches. *International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013*

[5] Chapman, S. 2006. SimMetrics: a java & c# .net library of similarity metrics, http://sourceforge.net/

[6] Hazen, T.J. 2010. Direct and Latent Modeling Techniques for Computing Spoken Document Similarity. *In Proc. of IEEE Workshop on Spoken Language Technology*, Berkeley, CA, 2010, pp. 12-15.

[7] Kondrak, G. 2005. N-Gram Similarity and Distance. *In proceedings of the 12th international conference on String Processing and Information Retrieval, 2005, pp. 115-126*.

[8] Hongdan, et al. 2011. A Document-Based Information Retrieval Model Vector Space. *IEEE*. 65-68

[9] Elbegbayan, N. 2005. Winnowing: a Document Fingerprinting Algorithm, *TDDC03 Proj*., 2005.

[10] Parapar, J. and Barreiro, A. 2009. Evaluation of text clustering algorithms with n-gram-based document fingerprints. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5478 LNCS, pp. 645–653.*

[11] Robertson, S. 2004. Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation, Vol. 60, 502–520. England.

[12] Vijayarani ,S., Ilamathi , J., Nithya. 2015. Preprocessing Techniques for Text Mining - An Overview. International Journal of Computer Science & Communication Networks, Vol 5(1),7-16 7 ISSN:2249-5789.

[13] Kurniawan, H. (2018, April). Strategy Development Of Human Source Competitiveness Strengthening With Learning Media System Analisis Model. In *Prosidina International conference on Information Technology and Business (ICITB)* (pp. 9-12).

[14] Turney, Peter D. Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37 2010 141-188

[15] Guo, Q. 2008. The Similarity Computing of Document based on VSM. IEEE

[16] Schleimer, S. Wilkerson, D. S. and Aiken, A. 2003. Winnowing: local algorithms for document fingerprinting, in Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 76–85.

[17] Sutoyo, R., Ramadhani, I., Ardiatma, A. D., Bavana, S. C., Warnars, H. L. H. S., Trisetyarso, A., Suparta, W. 2017. Detecting documents plagiarism using winnowing algorithm and k-gram method. 2017 IEEE International Conference on Cybernetics and Computational Intelligence.