Machine Learning Approaches on External Plagiarism Detection

Imam Much Ibnu Subroto
Department of Informatics Engineering
Faculty of Industrial Engineering
Semarang, Indonesia
imam@unissula.ac.id

Ali Selamat
Department of Computer Science
Faculty of Computing
Johor, Malaysia
aselamat@utm.my

Badieah Assegaf
Department of Informatics Engineering
Faculty of Industrial Engineering
Semarang, Indonesia
imam@unissula.ac.id

Abstract

External plagiarism detection is a technique that refers to the comparison between suspicious document and different sources. External plagiarism models are generally preceded by candidate document retrieval and further analysis and then performed to determine the plagiarism occurring. Currently most of the external plagiarism detection is using similarity measurement approaches that are expressed by a pair of sentences or phrase considered similar. Similarity techniques approach is more easily understood using a formula which compares term or token between the two documents. In contrast to the approach of machine learning techniques which refer to the pattern matching and cannot directly comparing token or term between two documents. This paper proposes some machine learning techniques such as k-nearest neighbors (KNN), support vector machine (SVM) and artificial neural network (ANN) for external plagiarism detection and comparing the result with Cosine similarity measurement approach. This paper presented density based that normalized by frequency as the pattern. The result showed that all machine learning approach used in this experiment has better performance in term of accuracy, precision and recall.

***Keywords**: plagiarism detection, machine learning*

1. Introduction

Progression the Internet in the academic world quite encouraging. Many scientific articles are now using electronic documents are published electronically as well. This case there are two sides less well or not well in the world of plagiarism. The good side is that not what documents are easy to find making it easier to copy text and ideas by anyone. If done without citation, then can be categorized as a crime is theft of intellectual property. The good side is the easier it is to find the perpetrators of plagiarism that can prevent the author to do so. They will fear the sanctions have been many cases of plagiarism that ever happened. Plagiarism detection technique has been widely applied research, of course, to get a better performance so that it can make actors of a plagiarism deterrent.

From several plagiarism detection techniques that exist today, almost all of them use the similarity approach. Even, there are researchers who classify plagiarism techniques are all based on similarity [1]. The author feels there are other approaches that are not discussed in a while has adequate performance potential, the machine learning approach. This paper will review the various plagiarism detection techniques with all its advantages and disadvantages. Broadly, the author divides plagiarism approach into two major parts, namely similarity approaches and machine learning approaches. Classification will be discussed in more detail in other chapters in this paper.

2. Comparison of Similarity and Machine Learning Approaches

Both approach's similarity and machine learning approaches have the same goal, which is to get the best performance. Performance can be measured by measuring accuracy, precision and recall. Accuracy indicates the percentage of correct detection results are revealed. Precision denominates what percentage of all instances that a detection method reports as suspicious is plagiarism. Recall denominates what percentage of all plagiarized instances in the collection a detection method report.

The first comparison is modeling. Similarity approach is basically a measurement of distance. The closer a couple then is said that the two objects are similar. If the distance is 0 means that the two pairs are identical. Typically, the results based on similarity measurement using the values were normalized by the inverse calculation, meaning the value equal one is the zero distance. In contrast, if the value is close to zero means the distance approaches infinity. In the representation, this approach typically uses token n-grams, graphs, line sequence. Obviously, the formula is different for each technique and representation. The accuracy of the technique is highly dependent on the design formulas are made.

In contrast to the similarity approach, the accuracy of machine learning approaches not only depend on the formula used, but rather depends on the strength of expert data used during the training process. Formula on this approach is general in nature, even the real formula cannot be explained clearly. The better the data validation by experts then will get better accuracy; the more data that is used in then training process generally will increase accuracy.

The result of calculations on the similarity approach is a continuous value, which is generally a value between 0 and 1. Thus, the possibility of the output value is infinite. To determine the classification of plagiarism, we need a minimum limit (threshold) of the value of similarity can be categorized as plagiarism class. Threshold value can be adjusted according to our wishes. Threshold value should have a combined performance of accuracy, recall and precision are the best. Strategies are needed to determine the optimal threshold in similarity approach [2].

The output machine learning is a value that represents a particular class. Briefly, plagiarism class or not. No further process is required to determine the optimal threshold in the machine learning approach.

Another weakness that similarity approach is that not all similar sentences that can be considered to plagiarism. For example, compare the following two sentences. The first sentence reads "The plagiarism is a crime", while the second sentence reads "The plagiarism is not a crime". With the similarity approach, of course, the two sentences are similar, suppose 85%. But It may not be considered plagiarism by reason of having different meanings. In machine learning approach, the problem can be solved by inserting the negative words as one or the features in the training data.

Table 1. Comparison of Plagiarism Detection Approach

	Plagiarism Detection Approach	
	Similarity Measurement	**Machine Learning**
Modeling	Based on mathematical formula, Formula is easy to understand	Base on training/learning process, Common Modeling and unknown real formula
Result	Unlimited value (range 0 to 1)	Discrett value [yes; no]
Decision	Must determines a threshold value i.e. >0.8, >0.9,>0.95 etc.	Explicitly stated i.e. two categories: plagiarism|not plagiarism
Techniques	Cosine, Jaccard, Hamming distance, fingerprint, sequence alignment, graph similarity, ontology etc.	k-nearest neighbours, artificial neural network, support vectors machine
Performance	The performance depend on the formula accusation itself	The performance depend on the model design and training data (experts person and amount of data)

Table 1 is a summary of comparative approaches in plagiarism detection methods, which between similarity approaches and machine learning approaches.

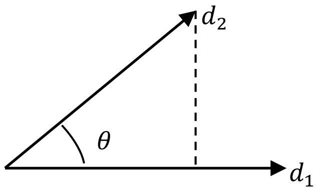
Classification of Plagiarism Detection

By looking at a fairly significant comparison between the similarity approach and machine learning approaches then this paper proposes a classification of plagiarism detection based on method approach as figure 1. In outline, the classification is divided into two, namely similarity approaches and machine learning approaches. Similarity approaches split some methods a fairly different, especially in a way to represent data further adapted by means of measuring similarity.

3. Similarity Approach

Similarity-based plagiarism detection is a method that uses similarity measurement to calculate the degree of plagiarism. Usually the range is between 0 and 1 or between 0% and 100%. In this approach usually an ordinary data represented in the form of vector, graph, or as a sequence. Based on the approach of how to represent relationship between two objects of this paper divides similarity approach into 3 kinds, namely the distance vector, graph similarity, and bioinformatics.

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. The more similar of two documents then the greater possibility of plagiarism. This is the principle that be used to determine the degree of plagiarism of a documents by using the similarity measurement.



Angle between documents

The measure reflects the degree of closeness or separation of two documents and should correspond to the characteristics that are believed to distinguish the plagiarism category. If the value of θ equal to zero mean two documents is exactly the same. The greater value of θ means that the two documents were less similar.

This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Given two documents $\overrightarrow{t_a}$ and $\overrightarrow{t_b}$, their cosine similarity is refer to formula

$$\mathrm{SIM_C}\left(\overrightarrow{t_a},\overrightarrow{t_b}\right) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\overrightarrow{t_a}|\times|\overrightarrow{t_b}|} \qquad (1)$$

Cosine similarity is the famous techniques. There are several other techniques that can be used in the detection of plagiarism such as Jaccard coefficient, Euclidean distance, Hamming distance, etc..

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

Sequence alignments are also used for non-biological sequences, such as those present in document plagiarism detection. Text document actually is the sequence of words. Some researcher such as [3-6] use sequence alignment to detect source code plagiarism and some other for free-text plagiarism detection such as Horton et.al [7].

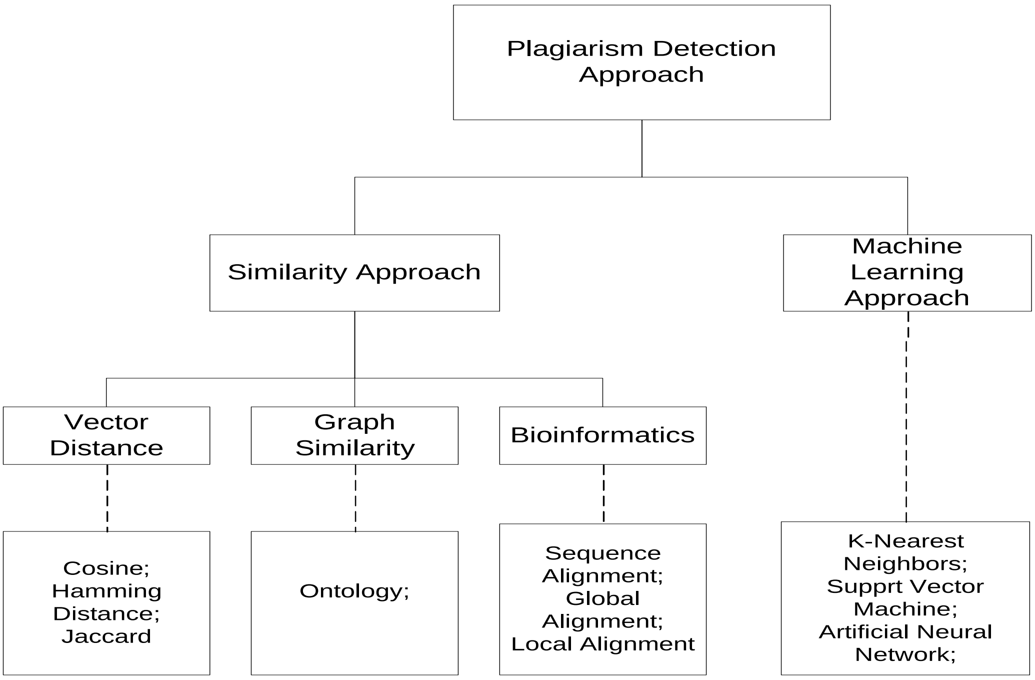

Plagiarism Detection Approach
Similarity Approach
Machine Learning Approach
Vector Distance
Graph Similarity
Bioinformatics
Cosine; Hamming Distance; Jaccard
Ontology;
Sequence Alignment; Global Alignment; Local Alignment
K-Nearest Neighbors; Supprt Vector Machine; Artificial Neural Network;

Fig. 1. Classification of Plagiarism Detection Approuch

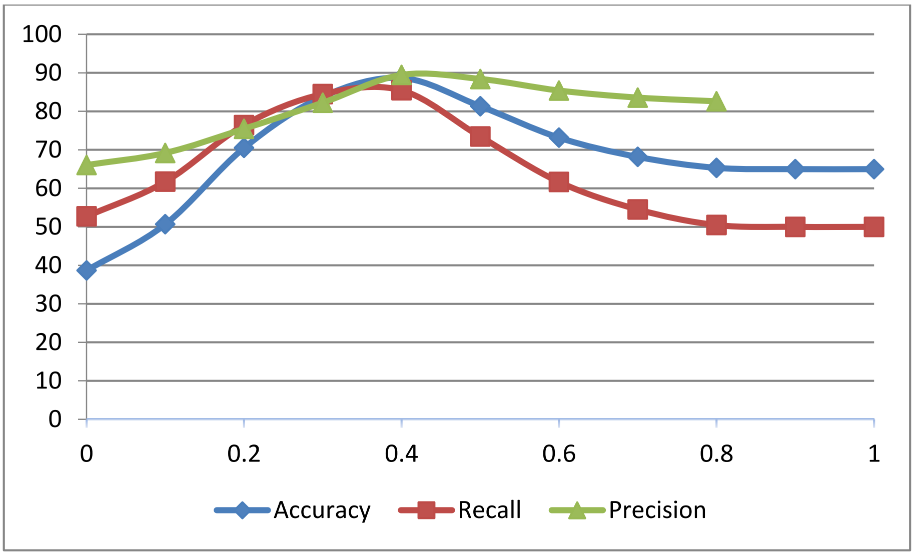


Figure 2. Cosine Similarity Performance based on Threshold

Figure 2 tells how to determine the optimal threshold on the similarity approach. The dataset is taken in a way from web crawler, the amount of data compared to some of the search engines with queries in the form of sentences. The measurement results obtained by comparing the similarity between the query phrase and some phrases from search engine results. In the cosine similarity, the data is represented as a vector comparison between two pieces of text. Each vector is calculated based on the term frequency (tf) or inverse term frequency (idf). The calculation of the term through the pre-processing called tekonization and stemmer. Overall the process generate output similarity values ranging between 0 and 1 experiment was continued by calculating the performance of accuracy, recall and precision with some variation of the threshold. The result can be seen in Table 1. For ease of understanding, the illustration in Figure 2.14 can be explained clearly. The best threshold point is determined by a combination of accuracy, recall and precision of the best as well. Starting from the smallest value, the larger the threshold value then the performance will increase. After the peak value of the performance, then the performance will decrease along with the magnitude of the threshold value. The decrease is due to the many errors that occurred in the classification of cases of plagiarism that is not considered plagiarism. This is because the higher the specified criteria. In the Figure shows that the best conditions occurred at about 0.40 threshold. With accuracy, recall, and precision each worth 88.76%, 85.46% and 89.45%.

4. Machine Learning Approach

It will appear on the results of the above questions: Is the similarity value of 0.40 can be regarded as a sufficient indicator of plagiarism? It looks like the numbers are too low to be similar. There are at least two things are the reason that the 0.4 threshold be quite good. The first reason, the experimental results in Table 2.3 is an indisputable fact. The second reason is related to the source data comes from the search engines, which is a secondary data source. Like most search engines, in this case Google, the search results are usually longer than the query entered. Usually at the front and at the rear there is an additional piece of wording the phrase. The example has been shown in figure 2.15.

In contrast to the similarity measurement approach, the result machine learning is a value that represents a particular class. Briefly, plagiarism class or not (honest). No further process is required to adjust the optimal threshold on the machine learning approach. Good or not result depends on the validity of the data collected as well as the architectural design of machine learning are used.

Another weakness on similarity approach is that not all similar sentences that can be considered to plagiarism. For example, compare the following two sentences. The first sentence reads ”The plagiarism is a crime”, while the second sentence reads

”The plagiarism is not a crime”. With the similarity approach, of course, the two sentences are similar, suppose 85%. But It may not be considered plagiarism by reason of having different meanings. In machine learning approach, the problem can be solved by including the negative words as one or the features in the training data.

Table 2: Prelimeniary Result of Basic Machine Learning and Similarity

Method	Accuracy	Precision	Recall
KNN	92.33	91.50	91.68
SVM	93.00	92.86	91.99
ANN	91.67	91.73	89.56
Similarity (Cosine)	88.67	85.46	89.46

Referring to the many techniques used in the two approaches, through the table 2 we show the initial results in this chapter which is one more reason to dig deeper plagiarism detection study by machine learning approaches. Cosine similarity is one of the most famous techniques in addition to similarity Jaccard and HMM. Almost every experiment uses this technique as a benchmark. Machine learning is used as a comparison in this case is KNN, SVM and ANN. These techniques have been proven both in performance to be used in a variety of fields. From the 600 datasets we tested, the results showed that in general machine learning has a better performance than the similarity measurement. All results from machine learning approaches have performance above 91%, while the performance of the approach is approximately 88% similarity. Preliminary results, it makes the author more interested to learn more and do some enhancement techniques to obtain improved performance.

In the machine learning approach, the performance is strongly influenced by at least 3 things. The first is how to represent the data in a numeric format, ie starting from the data capture and pre-processing. If it is not appropriate in the design of data representation then the performance will never be good. The second is the validity of the data that is retrieved. Data plagiarism must be validated by experts. Validation error can disturbing the learning process machine, then the performance will not be optimal. An expert on plagiarism are human beings, so it is likely wrong in doing validation also. That is why the amount of data to be enough to reduce statistical errors. The third is the architecture design. In machine learning like SVM or ANN, there are several variants. For example, there is variation in the SVM kernel: linear, sigmoid, polynomial, etc.. In the existing design of a multi-layer ANN, the hidden layer, and also there is some variation fitness function. Each case is not necessarily suitable for all kinds of variants. It will require experimentation and a more detailed analysis. All three thing is exactly what will be discussed in this thesis, especially in the case of SVM and ANN improvement.

5. Conclusion

From the study and the preliminary result has been done, it can be concluded that both approaches have good performance. Surely a machine learning approach has better performance. By analyzing the advantages and disadvantages of existing similarity measurement approaches or machine learning approach, the machine learning approach is attractive for further study. Although the accuracy of machine learning SVM and ANN is high at around 92%, but it still looks the challenges to improve performance by way of modification methods. Hybrid method is a choice that we propose to improve the performance of SVM and ANN engine.

References

[1] N. Meuschke and B. Gipp, "State of the Art in Detecting Academic Plagiarism," *International Journal for Educational Integrity,* vol. 9, p. 12, 2013.

[2] B. Stein*, et al.*, "Strategies for retrieving plagiarized documents," presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007.

[3] N. Meuschke and B. Gipp, "State of the Art in Detecting Academic Plagiarism," International Journal for Educational Integrity, vol. 9, p. 12, 2013.

[4] B. Stein, et al., "Strategies for retrieving plagiarized documents," presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007.

[5] J. Jeong-Hoon, et al., "Evolution Analysis of Homogenous Source Code and its Application to Plagiarism Detection," in Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007, 2007, pp. 813-818.

[6] R. Brixtel, et al., "Language-Independent Clone Detection Applied to Plagiarism Detection," in Source Code Analysis and Manipulation (SCAM), 2010 10th IEEE Working Conference on, 2010, pp. 77-86.

[7] J.-S. Lim, et al., "Plagiarism detection among source codes using adaptive local alignment of keywords," presented at the Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, Seoul, Korea, 2011.

[8] J. Jeong-Hoon, et al., "A Plagiarism Detection Technique for Java Program Using Bytecode Analysis," in Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on, 2008, pp. 1092-1098.

[9] R. Horton, et al., "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections," Digital Studies / Le champ numérique, vol. 10, 2010.