# SIMILARITY DETECTION OF STUDENT ASSIGNMENTS USING ROCCHIO METHOD

1st Dewi Soyusiawaty
Informatics Engineering
Ahmad Dahlan University
Yogyakarta, Indonesia
dewi.soyusiawaty@tif.uad.ac.id

2nd Anna Hendri Soleliza Jones
Informatics Engineering
Ahmad Dahlan University
Yogyakarta, Indonesia
annahendri@tif.uad.ac.id

3rd Panggah Widiandana
Informatics Engineering
Ahmad Dahlan University
Yogyakarta, Indonesia
widiandanapanggah95@gmail.com[3]

***Abstract*—** Student assignments is one component that influences the final score of a course given by the lecturer. The assignment of a course is charged at 20% of the final assessment of a course. Each lecturer has an average of 40 students in each class, while the lecturer can teach more than one class of courses. The large number of students will make the number of assignments too large because the ratio of the number of lecturers and students is very different. This is a burden on the lecturer because it has the responsibility to check each of student work assignments. The level of similarity of course assignments is one of the ways lecturers can assess a subject's assignment. This study aims to help lecturers detect the similarity of student tasks by using a task similarity detection system. This study uses Rocchio method to detect the similarity of words. The type of document tested is formatted in .pdf. The language used in the course is Indonesian. The preprocessing stages in this research consists of Case Folding, Tokenizing, Filtering, Sorting. The TF.IDF uses as the word weighting and the similarity measurement with the Rocchio method so that it gets the percentage value of the similarity between documents. The development stage started from planning, analysis, design, implementation and testing. The testing stage including testing with dummy data and usability data. Based on the research that has been done, the Rocchio method can classify documents that are similar to the results of the accuracy of 74,6% and detection errors from the similarity of queries with documents of 2.54 or 25.4%.

***Keywords— similarity, course assignment, TF. IDF, Rocchio***

## I. INTRODUCTION

According to Indonesian Dictionary, Plagiarism is the activity of taking the essays (opinions) of others and broadcast as an essay (opinion) by themselves. Research conducted by McCabe [7] was put forward that 70% of students admitted doing plagiarism in which half felt guilty of cheating on written assignments, 40% of students claimed to use cut-paste when completing their assignments, the results of this study proved that the seeds of plagiarism had existed since they were in school.

The assignment course is one component that influences the final grade of a course. Based on the results of a questionnaire on the Informatics Engineering lecturers that conducted by the researcher, 67.885% of the lecturers ask their assignment to be in the format of (.doc) and (.pdf) files. Then 75% of lecturers had difficulty in checking assignments, 92,857% of lecturers who had given punishment to the student but the plagiarism still occurs, therefore a system that could assist lecturers to detect similarity of course assignment was needed.

## II. METHOD

### A. Information Retrieval System

Information retrieval (IR) is finding material (usually documents) of an unstructured nature that satisfies an information need from within large collections [6].

In the Information Retrieval System Architecture, there are two tasks that handled by the system, namely database preprocessing and method applied to calculate the relevance or similarity between documents in the database that have been preprocessed by querying users. The flow of Information Retrieval can be seen in Figure 2.1 [1].

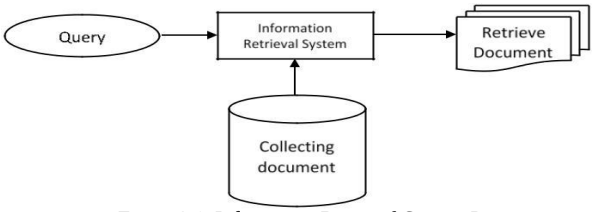


Figure 2.1. Information Retrieval System Process

In computer architecture, preprocessing is needed to detect the similarity. Preprocessing flow can be seen in figure 2.2

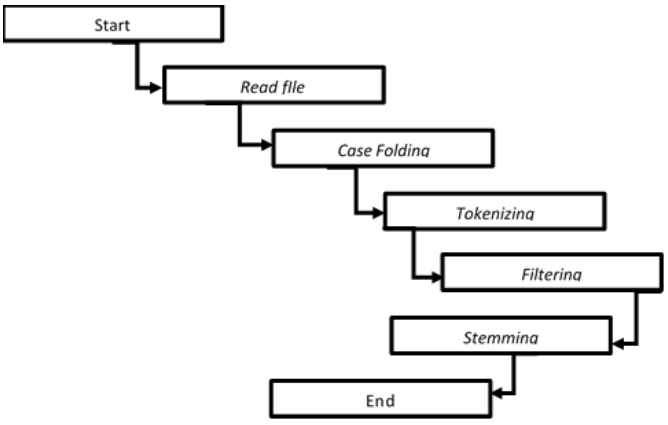


Figure 2.2.: Flowchart of preprocessing

**978-1-5386-6940-2/18/$31.00 ©2018 IEEE**

### B. *Tf.Idf Weighting*

The weighting of words in a document is done after preprocessing using Term Frequency (TF). Term Frequency (TF) is the weight of a word t in a document and Inverse document frequency (IDF) is made to reduce the effect of the word whose frequency is too high in the document. The Idf equation can be seen in formula 2.1.

$$Idf_t = 1 + log\frac{N}{N_t} \ldots\ldots\ldots\ldots\ldots (2.1)$$

Information:
$Idf_t$ = inverse document frequency
N = the total number of documents
$N_t$ = number of documents containing terms

### C. *Rocchio Method*

Rocchio Relevance Feedback method is a query reformulation strategy used for help novice users for information retrieval systems. Users are presented with the results of the search for relevant documents [9]. Rocchio Relevance Feedback equation can be seen in formula (2.2).

$$\mathbf{R_n = N + \beta \left(\left(\frac{D_P}{N_P}\right) - \left(\frac{D_n}{N_n}\right)\right) \ldots\ldots\ldots\ldots\ldots (2.2)}$$

Information:
R: term similarity.
N: Number of terms for each document.
β: Term weight value.
*Dp*: Term of relevant documents
*Np*: Total number of relevant documents
*Dn*: The term of the document is irrelevant
*Nn*: The total number of documents is not relevant.

### D. *Min-Max Normalization*

Min-max normalization is a transformation process in which numerical attributes will be scaled into a smaller size, such as between -1 to 1 or 0 to 1. Min-max normalization equation can be seen in equation (2.3)

$$\mathbf{d' = \frac{d - min(p)}{max(p) - min(p)} \ldots\ldots\ldots\ldots\ldots (2.3)}$$

Information:
d '= scale value
d = initial value
min (p) = minimum value
max (p) = maximum value

### E. *MAD (Mean Absolute Deviation)*

MAD is an average absolute error over a certain period regardless of whether the forecasting result is greater or smaller than the reality. Equations can be seen in equation (2.4).

$$\mathbf{MAD = \sum \left|\frac{A_t - F_t}{n}\right| \ldots\ldots\ldots\ldots\ldots (2.4)}$$

Information:
At = actual request
Ft = demand forecasting
n = number of forecasting involved

### F. *Library*

Library is a collection of programs or functions to facilitate the creation of a program.
1. Python literary library
2. Python library scikit-learn

This study uses scikit learn libraries to calculate word weight using the Tf.Idf method.

### G. *Testing*

This study uses dummy data test for accuracy. Dummy data is used to compare document 1 with document 2, predicted and actual value, predictive value is obtained from the system, while Actual value is obtained from the value of checking student assignments manually.

### H. *Similarity Flow System*

The process of detecting the similarity of student assignments can be seen in figure 3.1

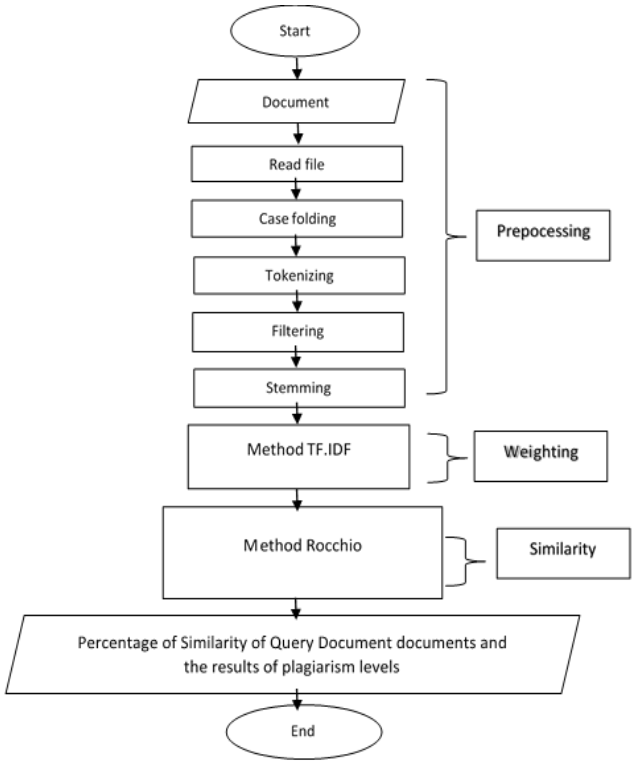


Figure 3.1.: Flowchart of similarity detection process

## III. RESULT

### A. *Rocchio Method Calculation*

After weighting is carried out in each term, the calculation of the document similarity can be done using the rocchio method. From the data that has been obtained, the calculation is done using equation (2.4), the calculation can be seen as follows:

$$R_n = N + \beta \left( \left( \frac{D_P}{N_P} \right) - \left( \frac{D_n}{N_n} \right) \right) \ldots\ldots\ldots\ldots\ldots (2.2)$$

From the equation (2.2), the calculation can be seen as following:

R : Relevant Document 1
N : 7
β : 10,60
$D_p$ : 7
$N_p$ : 27
$D_n$ : 0
Nn : 45

$$R_1 = 7 + 10{,}60 \left( \left( \frac{7}{27} \right) - \left( \frac{0}{45} \right) \right) = 9{,}75$$

Then the calculation result of R1 is 9.75. This calculation is applied to other documents so as to obtain the following results:

$$R_2 = 7 + 13{,}74 \left( \left( \frac{3}{27} \right) - \left( \frac{5}{45} \right) \right) = 7{,}00$$

$$R_3 = 7 + 8{,}30 \left( \left( \frac{4}{27} \right) - \left( \frac{2}{45} \right) \right) = 7{,}86$$

$$R_4 = 7 + 14{,}44 \left( \left( \frac{1}{27} \right) - \left( \frac{7}{45} \right) \right) = 5{,}29$$

$$R_5 = 7 + 16{,}12 \left( \left( \frac{4}{27} \right) - \left( \frac{6}{45} \right) \right) = 7{,}24$$

$$R_6 = 7 + 12{,}52 \left( \left( \frac{2}{27} \right) - \left( \frac{5}{45} \right) \right) = 6{,}54$$

$$R_7 = 7 + 6{,}87 \left( \left( \frac{3}{27} \right) - \left( \frac{2}{45} \right) \right) = 7{,}46$$

$$R_8 = 7 + 7{,}17 \left( \left( \frac{2}{27} \right) - \left( \frac{3}{45} \right) \right) = 7{,}05$$

$$R_9 = 7 + 16{,}87 \left( \left( \frac{1}{27} \right) - \left( \frac{9}{45} \right) \right) = 4{,}25$$

$$R_{10} = 7 + 7{,}35 \left( \left( \frac{1}{27} \right) - \left( \frac{6}{45} \right) \right) = 6{,}29$$

After the results of the rocchio are obtained, the next step is to normalize min max to determine the level of similarity of the query with the document. Min max normalization is calculated based on equation 2.3, with the calculation example:

$$D1 = \frac{9{,}75 - 4{,}25}{9{,}75 - 4{,}25} = 1 \text{ x } 100\% \qquad = 100\%$$

the same way is used for each document, so the results can be seen as in Table I.

TABLE I. CLASSIFICATION

| Doc | Rocchio Score | Percentage Value | Classification |
|---|---|---|---|
| D1 | 9,75 | 100 % | High |
| D2 | 7,00 | 49,52 % | Medium |
| D3 | 7,86 | 65,78 % | Medium |
| D4 | 5,29 | 19,25 % | Low |
| D5 | 7,24 | 53,84 % | Medium |
| D6 | 6,54 | 41,34 % | Medium |
| D7 | 7,46 | 58,20 % | Medium |
| D8 | 7,05 | 50,50 % | Medium |
| D9 | 4,25 | 0,00 % | Low |
| D10 | 6,29 | 36,27 % | Medium |

In Table I, can be seen the level of similarity of the query to the document.

### *B. System Testing*

System testing is done by using dummy data, which is comparing the results of manual handling with system calculations. The result of data dummy testing using the rocchio method data from Table I, can be seen in Table II.

TABLE II. DATA DUMMY

| Doc | Prediction Score | Actual Score | Difference |
|---|---|---|---|
| D1 | 11,22 | 9,75 | -1,47 |
| D2 | 7,14 | 7,00 | -0,14 |
| D3 | 8,29 | 7,86 | -0,42 |
| D4 | 4,38 | 5,29 | 0,91 |
| D5 | 7,57 | 7,24 | -0,33 |
| D6 | 4,42 | 6,54 | 2,11 |
| D7 | 7,67 | 7,46 | -0,21 |
| D8 | 7,12 | 7,05 | -0,07 |
| D9 | 2,94 | 4,25 | 1,31 |
| D10 | 5,44 | 6,29 | 0,85 |

Dummy data test is used to determine the level of system error.

Information:

At = 9,75+7,00+7,68+5,29+7,24+6,54+7,46+7,05+4,25+6,29 = 66,19

Ft = 11,22+7,14+8,29+4,38+7,57+4,42+7,67+7,12+2,94+5,44 = 68,73

n = 10.

The calculation can be seen:

$$MAD = \left| \frac{66{,}19 - 68{,}73}{10} \right| = 2{,}54$$

Based on the calculation of MAD, it is found that the average results of system calculation errors with manual calculations in detecting the similarity of queries with documents are 2.54 or 25,4 %.

## IV. Conclusion

This research has produced an application for detecting the similarity of student assigments using rocchio method. Similarity testing with the rocchio method will be carried out after the preprocessing text process, that is, tokenization, filtering and stop removal is done on the query data indexing. The results of this document are similar from query. Through testing the system is able to detect the similarity of student tasks with the results of the accuracy value of 74.6% and the value of detecting the error similarity of queries with documents of 25.4%.

## References


[1] Barber, A.S., Barraclough, E.D. and Gray, W.A. 'On-line information retrieval as a scientist's tool', Information Storage and Retrieval, 9, 429-44- (1973)

[2] Ganguly, D., Jones, G.J.F., Ramírez-de-la-Cruz, A. et al. (2018). "**Retrieving and classifying instances of source code plagiarism**" Inf Retrieval J (2018) 21: 1. https://doi.org/10.1007/s10791-017-9313-y

[3] Jameel, S., Lam, W. & Bing, L. (2015). "**Supervised topic models with word order structure for document classification and retrieval learning**" Inf Retrieval J (2015) 18: 283. https://doi.org/10.1007/s10791-015-9254-2

[4] Jayadianti, Herlina. Solving Problem Of Semantic Terminology In Digital Library. **International Journal Of Advances In Intelligent Informatics**, [S.L.], V. 3, N. 1, P. 20-26, Mar. 2017. Issn 2548-3161. Available At:Http://Ijain.Org/Index.Php/Ijain/Article/View/70%7cto_Array%3a0>. Date Accessed: 27 Aug. 2018. Doi:Https://Doi.Org/10.26555/Ijain.V3i1.70.

[5] Kotlerman, L., Dagan, I. & Kurland, O. (2018). "Clustering Small-Sized Collection of Short Text" Inf Retrieval J (2018) 21: 273. https://doi.org/10.1007/s10791-017-9324-8.

[6] Manning, Christopher D., Raghavan, Prabhakar., Schutze, Hinrich. "Introduction to information retrieval", 2008. Cambridge University Press 32 Avenue of the Americas, New York, NY 10013-2473, USA. 2008.

[7] McCabe, Donald L., Linda Klebe Trevino, and Kenneth D. Butterfield. "Cheating in Academic Institutions: A Decade of Research." Ethics & Behavior 11.3 (2001): 219-232.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

[9] Prasetya, Didik Dwi; Prasetya Wibawa, Aji; Hirashima, Tsukasa. The performance of text similarity algorithms. **International Journal of Advances in Intelligent Informatics**, [S.l.], v. 4, n. 1, p. 63-69, mar. 2018. ISSN 2548-3161. Available at: <http://ijain.org/index.php/IJAIN/article/view/152%7Cto_array%3A0>. Date accessed: 27 aug. 2018. doi:https://doi.org/10.26555/ijain.v4i1.152.

[10] Selberg, E. W. 2011. Information Retrieval Advances Using Relevances Feedback. Thesis. Department of Computer Science and Engineering University of Washington.

[11] Soyusiawaty, Dewi., Aribowo, Eko., 2016. "Designing and Implementing Parsing for Ambiguous Sentences in Indonesian Language", Journal of Theoretical and Applied Information Technology 29th February 2016. Vol.84.No.3.http://www.jatit.org/volumes/Vol84No3/5Vol84No3.pdf.

[12] Yunianta, Arda Et Al. Semantic Data Mapping Technology To Solve Semantic Data Problem On Heterogeneity Aspect. **International Journal Of Advances In Intelligent Informatics**, [S.L.], V. 3, N. 3, P. 161-172, Dec. 2017. Issn 2548-3161. Available At:Http://Ijain.Org/Index.Php/Ijain/Article/View/131%7cto_Array%3a0>. Date Accessed: 27 Aug. 2018. Doi:Https://Doi.Org/10.26555/Ijain.V3i3.131.