APPLICATION OF VECTOR SPACE MODEL FOR UNDERGRADUATE THESIS PLAGIARISM DETECTION IN DIAN NUSWANTORO UNIVERSITY

Soraya Arum Rahmasari[1]
[1,2]Bachelor of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University
Jl. Nakula I No. 5-11 Semarang, Indonesia
Telp. (024) 3517261. Fax: (024) 3520165
E-mail: sorayaarum94@gmail.com[1]

Abstract

In the era of globalization, the internet is more advanced. Easy to get data is one of the causes of plagiarism. Same with the college students especially for informatics engineering who is looking for information to do the thesis for a candidate of bachelor, they have an opportunity do the plagiarism by copying another student which has the similar themes or takes some code from the internet. To prevent this problem, there is a need such an application to detect plagiarism. The purpose is implementing plagiarism detection in Dian Nuswantoro University especially for undergraduate thesis is to prevent the student do plagiarism with effectively. This application uses Vector Space Model method combine with Cosine Similarity algorithm. The author used this algorithm because of time consumption used for processing data is more accurate and faster, also this algorithm is suitable for multiple pattern searches. From the problem above, the author will make an application for detecting plagiarism in the field of college student especially for making a thesis. This application will produce a plagiarism report which contains the percentage similarity of document testing. This application can apply the method of Vector Space Model and cosine similarity algorithm to detect plagiarism of the documents. The application can compare between one thesis document and more than one thesis documents.

Keywords: *Plagiarism, Undergraduate Thesis, Vector Space Model, Cosine Similarity*

1. INTRODUCTION

Plagiarism is presenting someone else’s work or ideas as your own, with or without their consent, by incorporating it into the work without full acknowledgment [1]. Plagiarism has happened in academic and non-academic environment. Plagiarism is used to happen in the academic environment. Since students’ awareness in making paper or essay is too average, it triggers a lot of the occurrence of plagiarism. Plagiarism committed by students is often denied by the reason they don’t copy the other people’s work, but they only get inspiration from them.

In era of globalization the internet is more advanced. [2]. Prevention and detection are some ways that can be done in order to reduce plagiarism. Prevention means to hindering the emergence of plagiarism which is concerned with the moral community and the education system. This will provide remarkable long-term effects. The detection means a way to reveal plagiarism.

Some software which designed for detecting plagiarism documents are Turnitin, Plagium, Dupli Checker, iThenticate, Plagiarism Checker, and etc [3]. Plagiarism also accomplished by the students of Information Engineering Bachelors in Dian Nuswantoro University (UDINUS). Because of this, the writer made a plagiarism detection and this system will be applied in information retrieval and the process will use vector space model. Methods of Vector Space Model is one of the simplest methods used to search for documents in common with other documents. This method also has several stages in the search for common document, which looks at the frequency of occurrence of the word in the document preprocessing stage, then calculate the similarity of a document with a document that is compared by calculating the term frequency – inverse document frequency for terming documents and the cosine similarity for similarity of the documents.

2. THEORETICAL FOUNDATION

2.1 Vector Space Model

Vector space model are often used to present a document in a vector space [4]. Vector Space Model is a basic technique in obtaining information that can be used to study the relevance of documents to the search keywords (query) on search engines, document classification and grouping documents. A collection of words and documents are represented in the form of a matrix [5].

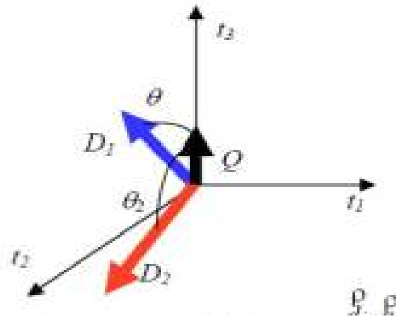


Figure 1. Vector

From figure 1 above, *i* means word in database, D is the document and Q is keyword or query.

2.2 TF-IDF Weighting

The weighing that frequently used in search engines is TF-IDF, the combination of Term Frequency (TF) with Inverse Document Frequency (TDF). Term Frequency – Inverse Document Frequency(TF-IDF) is the weighing that is often used in information retrieval and text mining. TF is a simple weighing which is important whether or not a word is assumed to be proportional to the number of occurrences of the word in the document, while the IDF is the weighing measure how important a word in a document when viewed globally in the whole document. TF x IDF weighing value will be higher if the value of TF great and observed the word is not found in many documents [6].

The formula of TF-IDF are :

a. TF(d,t) = f(d,t)
   f(d,t) is amount of word *t* for each document *d.*
b. IDF(t) = log(n/df(t))
   n is amount of document and df(t) is amount of word *t* in each document.
c. TF-IDF = TF(d,t)*IDF(t)

2.3 Cosine Similarity

Cosine similarity is used to measure the closeness between the two vectors. Cosine similarity is the result of both the vector dot product which is normalized by dividing the Euclidean Distance between two vectors. [5] The formula that can be used as follows:

$$sim(d_x.q) = \frac{d_x.q}{\|d_x\|\|q\|}$$

$$sim(d_x.q) = \frac{\sum_{i=1}^{N} \omega_{i.x}\omega_{i.q}}{\sqrt{\sum_{i=1}^{N} \omega_{i.x}^2}\sqrt{\sum_{i=1}^{N} \omega_{i.q}^2}}$$

Note:

d_x: document x,

q_N : query of document,

$\sum_{i=1}^{N} \omega_{i.x}$: the amount of "*i*" word in the "*x*" document,

$\sum_{i=1}^{N} \omega_{i.q}$: the amount of "*i*" word in the query document.

3. RESEARCH METHODOLOGY

3.1. Data Analysis

The Data were obtained from the thesis of undergraduate Informatics Engineering students, Dian Nuswantoro University Semarang. The data given were a thesis of data from 2010 to 2014. The data that was used consisted of chapter 1 from the thesis.

3.1. Proposed Method

The proposed method of this study is a development method for plagiarism detection in Thesis document. There are 20 documents that will be checked. The format of the document is a *docx. The System will processed the documents and evaluated the similarity of that documents. From Figure 2 there is the step for calculating the plagiarism in several documents. The first is read text document, and the second is preprocessing for separating the word. In preprocessing, the system did tokenization, stopword and stemming. After the preprocessing were done, the documents were calculated in accordance with existing processes in Figure 2, which are calculating term frequency and inverse document frequency, after that calculating cosine similarity, and the last one is doing the test to check the trial document with another document.

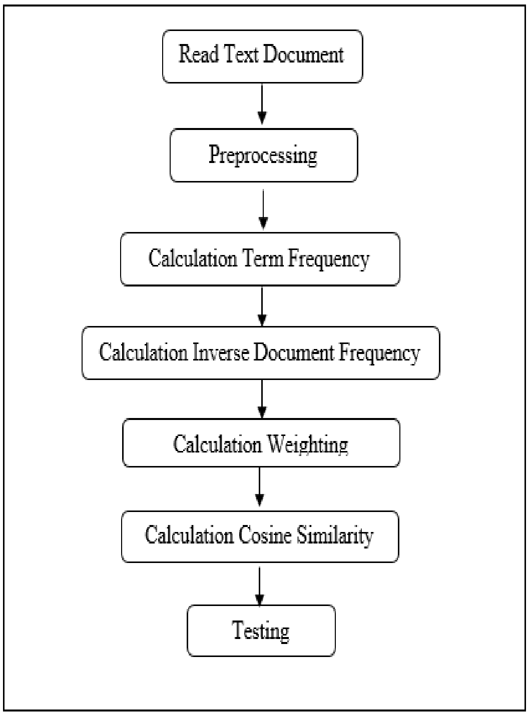

Read Text Document
Preprocessing
Calculation Term Frequency
Calculation Inverse Document Frequency
Calculation Weighting
Calculation Cosine Similarity
Testing

Figure 2 : Proposed Method

1. Preprocessing
   There are some steps that has been done in the prepocessing, the steps are:
   a. Read text document. The program should be able to read the document before doing the process.
   b. Stemming the words. In the stemming, the affixation words should be restored into the basic word.
   c. Stopword is a collection of words that frequently appear in the document. Stopword can be called a conjunctive word. Stopword process is to remove the conjunctive word or unimportant word in the documents.

Table 1:before prepocessing

Keyword	Pengetahuan Logistik
D1	Manajemen dari transaksi logistik
D2	Pengetahuan ini antar individu

D3	Manajemen pengetahuan ini mentransfer pengetahuan logistik

The table 1 is words that have not been preprocessed. It looks more presentable when read, but after done the preprocessing the words change became the basic words as shown as table 2 below.

Table 2: After Preprocessing

Keyword	Tahu logistik
D1	Manajemen transaksi logistik
D2	Tahu individu
D3	Manajemen tahu transfer tahu logistik

2. TF-IDF

In metrics term document there are three step that were used. The first step is calculation of Term Frequency, the second is calculating the Inverse Document Frequency for each term, and the last is calculating the weight of each document.

a. Term Frequency

TF is the frequency of occurrence of a term in the document.

b. Inverse Document Frequency

Document frequency is how much a term appears on the entire document. Inverse Document Frequency is reducing the weight of a term if its emergence is scattered throughout the document collection.

The calculation of TF-IDF was shown in table 3 below.

Table 3:T-IDF

Term	Tf					Idf Log(n/df)
	Keyword	D1	D2	D3	df	
Manajemen		1		1	2	0.176
Transaksi		1			1	0.477
Logistik	1	1		1	1	0.176
Tahu	1		1	1	2	0.176
Individu			1		1	0.477
Transfer				1	1	0.477

c. Weighting

From table 4, the weighting of the term was calculated. Each term has a value of the weight.

Table 4 : Weighting

Term	Weighting			
	Key	D1	D2	D3
Manajemen		0.176		0.176
Transaksi		0.477		
Logistik	0.176	0.176		0.176
Tahu	0.176		0.176	0.176
Individu			0.477	
Transfer				0.477

3. Cosine Similarity

From calculation the weight, the author got the number to calculate the cosine similarity. The number that needed is shown in table 5 below:

Table 5:Cosine Similarity

Term	W_{key}^2	W_{doc1}^2	W_{doc2}^2	W_{doc3}^2	$W_{key}*W_{doc1}$	$W_{key}*W_{doc2}$	$W_{key}*W_{doc3}$
Manajemen		0.031		0.031			
Transaksi		0.228					
Logistik	0.031	0.031		0.031	0.031		0.031
Tahu	0.031		0.031	0.031		0.031	0.031
Individu			0.228				
Transfer				0.228			
Total	0.062	0.29	0.259	0.321	0.031	0.031	0.062

Calculation cosine similarity

a. $\text{Cos Doc}_1 = \dfrac{\sum(Wkey \times Wdoc_1)}{\sqrt{\sum(Wkey^2) \times \sum({Wdoc_1}^2)}}$

$$= \frac{0.031}{\sqrt{0.062 \times 0.29}}$$

$$= \frac{0.031}{0.134} = 0.231$$

b. $\text{Cos Doc}_2 = \dfrac{\sum(Wkey \times Wdoc_2)}{\sqrt{\sum(Wkey^2) \times \sum({Wdoc_2}^2)}}$

$$= \frac{0.031}{\sqrt{0.062 \times 0.259}}$$

$$= \frac{0.031}{0.127} = 0.245$$

c. $\text{Cos Doc}_3 = \dfrac{\sum(Wkey \times Wdoc_1)}{\sqrt{\sum(Wkey^2) \times \sum({Wdoc_1}^2)}}$

$$= \frac{0.062}{\sqrt{0.062 \times 0.321}}$$

$$= \frac{0.062}{0.141} = 0.439$$

From calculation above the rank of similarity could be:

Table 6: Rank

Document	D1	D2	D3
Value	0.231	0.245	0.439
Rank	3	2	1

From the table 6, The third document has bigger percentage than all documents and the percentage is 43,9% so the third document include middle plagiarism. And the first document include minor plagiarism.

4. ANALYIS OF RESULTS AND DISCUSSION

4.1. Application

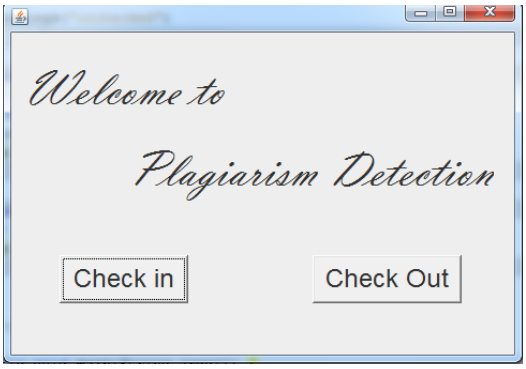


Figure 3, Home Page

From figure 3, this page called home page. There are 2 button in the page, which are check in to entry the check in plagiarism page and check out to quit from the program.

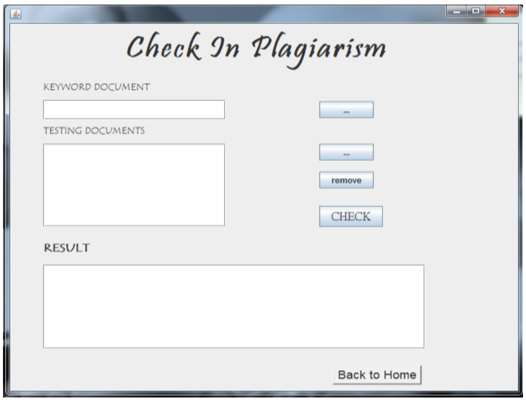
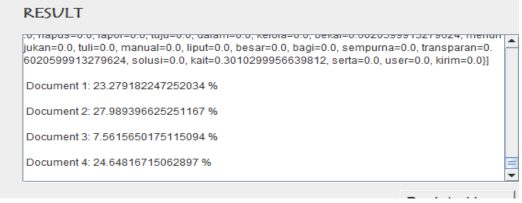


Figure 4, Check in Plagiarism Page

From figure 4, the page called check in plagiarism page. The function of this page is to check the plagiarism between one document to other documents. The user needs to input the keyword document and testing documents, then the result will be shown in text area.

The result of the percentage was shown in figure 5 below



RESULT

```
.0, hapus=0.0, lapor=0.0, tuju=0.0, dalam=0.0, kelola=0.0, bekal=0.6020599913279624, menun
jukan=0.0, tuli=0.0, manual=0.0, liput=0.0, besar=0.0, bagi=0.0, sempurna=0.0, transparan=0.
6020599913279624, solusi=0.0, kait=0.3010299956639812, serta=0.0, user=0.0, kirim=0.0}]

Document 1: 23.279182247252034 %

Document 2: 27.989396625251167 %

Document 3: 7.5615650175115094 %

Document 4: 24.64816715062897 %
```

Figure 5:Result

4.2. Result and Discussion

1. Result

   Result of the research is a creation of the system to find the similarity between one query document with other documents. The system consists of two interface, which are home and check in plagiarism interface. Each interface has a different function. On the program's implementation and interface, user can obtain information such as the percentage of similarity from each documents with a document query comparison using vector space model and cosine similarity.
2. Discussion

   Implementation of the evaluate trials using the accuracy formula which approach retrieve documents and similar as in the Table 7.

Table 7: Accuracy

System	Manual	
	Similar	Not Similar
Similar	0 (TP)	0 (FP)
Not Similar	0 (FN)	19 (TN)

The table shows some of the items needed to measure the performance of the classifier. The items will be used to calculate the accuracy with the formula:

$$A = \frac{(TP+TN)}{(TP+FP+FN+TN)} x100\%$$

$$A = \frac{(0+19)}{(0+0+0+19)} x100\% = 100\%$$

From calculation above the accuracy rate of the system as a percentage is 100%.

5. CONCLUSION AND SUGGESTION

5.1. Conclusion

The conclusions of this research are:

1. This application can apply the method of Vector Space Model and cosine similarity algorithm to detect plagiarism of the documents.
2. The function of this application is to compare between one thesis document and more than one document.
3. The accuracy between manual calculation and the program was 100%.

5.2. Suggestion

Suggestions that can be recommended by the author in completing this thesis are:

1. Completion of the prototype of the application must be made. Currently, users can only compare a document in *docx format.
2. This application to be developed not only on desktop computers but also on the website.

6. REFERENCES

[1]. Oxford, U. o. (2015, May 30). *Plagiarim*. Retrieved from University of Oxford:http://www.ox.ac.uk/students/academic/guidance/skills/plagiarism

[2]. pew. (2015, March 3). *The Mixed Impact of Digital Technologies on Student Research* . Retrieved from pew research: http://www.pewinternet.org

[3]. Team, E. (2015, March 3). *Top 8 Plagiarism Tools for Teacher*. Retrieved from Educational Technology and

Mobile Learning: http://www.educatorstechnology.com

[4] Turney P, Pantel, & Patrick. (2010). From Frequency to Meaning : Vector Space Model of Semantics. *Journal of Artificial Intelegence Reaserch vol 37,* 141-188.

[5] Christopher, D. M. (2008). Introduction to Information Retrieval. *Cambridge University Press*.

[6] K, S. J. (2004). A Statistical Interpretation of Term Specify and Its Application in Retrieval. *Journal of Documentation vol 60*, 493-502.