CHAPTER III
RESEARCH METHODOLOGY

3.1. Research Instrument

3.1.1. Software Specification

There were some software will be used in this study, they are:

- 64-bit Operating System Windows 7
- NetBeans IDE 6.9.1
- Enterprise Architecture

3.1.2. Hardware Specification

There were some hardware will be used in this study, they are: :

- Input devices (keyboard and mouse)
- Installed memory (RAM) 2GB
- Monitor resolution (1024x768)

3.2. Data Sources

There were some data sources used in this study, they are:

0. The primary data that was taken from Undergraduate Program thesis of Informatics Engineering in Dian Nuswantoro University.
1. The secondary data that used were 5 journals, 3 books, and 3 thesis.

3.3. Data Analysis Technique

In this study, the author collected the dataset by using observation. The result of observation was a *Docx document in Bahasa (Indonesian language) from student thesis Informatics Engineering of Dian Nuswantoro University. The chapter one of the thesis was used by the author.

3.4. 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 3, 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.

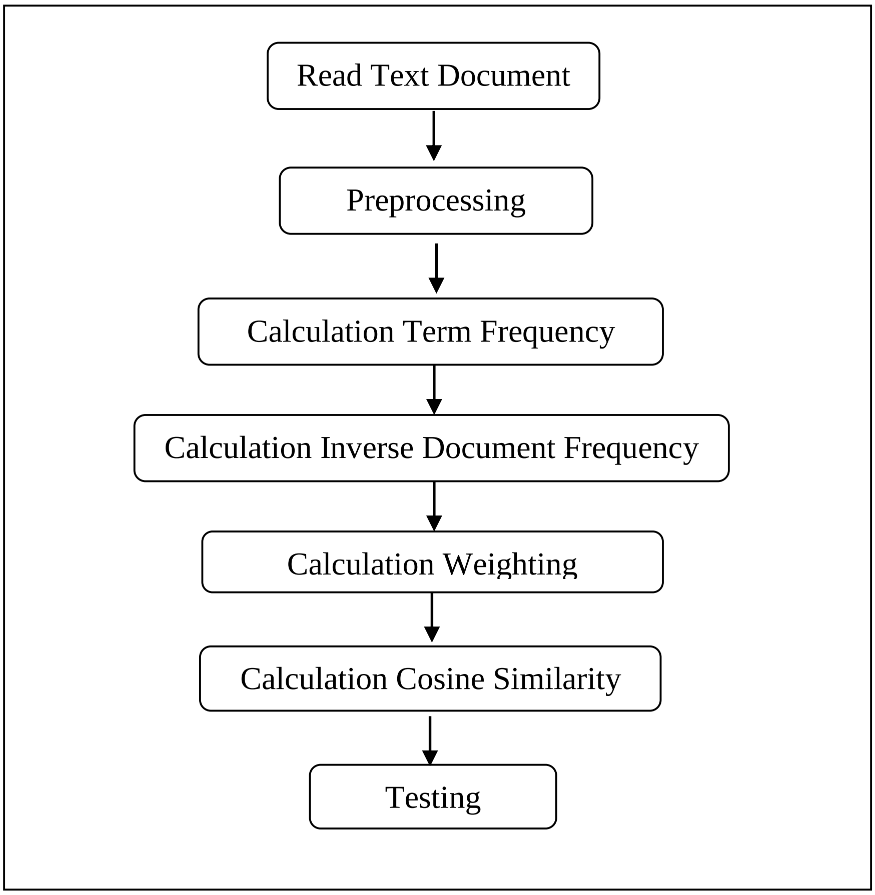
Read Text Document
Preprocessing
Calculation Term Frequency
Calculation Inverse Document Frequency
Calculation Weighting
Calculation Cosine Similarity
Testing

Figure 1 : Step of Study

3.5. Testing Technique

In this phase, the author analyzed the function of this system by doing the test similarity to compare the similarity of one document to another. The algorithm that used are TF-IDF and cosine similarity.

3.5.1. Testing Design

In document testing there are some attributes that used for checking the plagiarism. The tables of design testing are :

Table 1 : Document Inputed

Keyword	Some keyword
D1	Sentence from D1
D2	Sentence from D2
D3	Sentence from D3

Table 2 : TF-IDF Document

	Tf					Idf
Term	Keyword	D1	D2	D3	Df	Log(n/df)

Table 3: Weighting

Term	Weighting			
	Key	D1	D2	D3

Table 4: Cosine

Term	W_{key2}	W_{doc12}	W_{doc22}	W_{doc32}	W_{key}*W_{doc1}	W_{key}*W_{doc2}	W_{key}*W_{doc3}

3.5.2. Manual Calculation

In this part the author uploaded the new key word for checking easily the similarity with the other documents. The data is like below:

Table 5: Input Data

Keyword	Pengetahuan Logistik
D1	Manajemen dari transaksi logistik
D2	Pengetahuan ini antar individu
D3	Manajemen pengetahuan ini mentransfer pengetahuan logistic

From Table 6 the document were inputted. The data that were inputted was preprocessed, in order to form the original word.

Table 6: Preprocessing

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

Table 7: TF-IDF Document

Term	Tf					Idf
	Keyword	D1	D2	D3	Df	Log(n/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

From table 8, there are 1 keyword document and 3 testing document. The D1, D2 and D3 are called *n* or the number of the testing document. In the table 8, TF is the number of words in a document and DF is a total document that contains a word.

Table 8: Calculation 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

From table 9, the weighting of the term was calculated. Each term has a value of the weight. From calculation the weight, the author got the number to calculate the cosine similarity. The number that needed is shown in table 10 below:

Table 9: Calculation 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. Cos $\text{Doc}_1 = \frac{\sum(Wkey \times Wdoc_1)}{\sqrt{\sum(Wkey^2) \times \sum(Wdoc_1{}^2)}}$

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

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

b. Cos Doc2 = $\frac{\sum(Wkey \times Wdoc_2)}{\sqrt{\sum(Wkey^2) \times \sum({Wdoc_2}^2)}}$

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

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

b. Cos Doc3 = $\frac{\sum(Wkey \times Wdoc_1)}{\sqrt{\sum(Wkey^2) \times \sum({Wdoc_1}^2)}}$

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

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

From calculation above the rank of similarity could be:

Table 10: Rank

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

From the table 11, 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.