CHAPTER II
THEORITICAL BACKGROUND

2.1. Related Work

This chapter will discuss about the related work from the other papers, there are 3 papers which are taken by the author:

2.1.1. *Sistem Deteksi Plagiarisme Dokumen Bahasa Indonesia Menggunakan Metode Vector Space Model* [4].

In this era of globalization, the plagiarism has already happened in academics, especially for students who create a thesis, because of the availability of the facilities to copy and paste from one document to others. With this reason, the writer want to propose the document plagiarism detection system in Indonesian language using Vector Space Model Method. The documents, which are the similarity percentage level is tested, are the journal document of Informatics and Information Systems program, in which the detection process plagiarism through preprocessing stage, such as the tokenization process, the wiping out of stopword, and stemming, the next step is scoring calculation and cosine similarity. In the development of this system, the author uses the Java programming language. The methodology of system development uses iterative model approaching of incremental development. The main purpose of this system is to determine the level of similarity or plagiarism of

a journal. This application is expected to be able to detect and provide similarity percentage documents from the process of student plagiarism action in completing the final project.

2.1.2. *Mengukur Tingkat Kesamaan Paragraf Menggunakan Vector Space Model untuk Mendeteksi Plagiarisme* [5]

The writing or pieces of writing taken from the writing of someone else, intentionally or unintentionally, whether it is not properly referenced, can be categorized as plagiarism. Therefore, it is needed an application that could determine the similarity of documents or parts of documents quickly. If the similarity between documents or parts of documents is in the high level can be estimated, there have occurred the act of plagiarism in the documents or parts of the documents.

Comparing a document with all the documents in the database are also needed much time even though done with the help of computers. If there are several *n* of documents in the database, the complexity to calculate the degree of similarity between the documents in pairs are $O(n^2)$. When *n* is large enough then the calculation process takes a long time. For that reason, the determination of the amount of *k* from the relevant documents must be done first before the transcription in the document being compared one by one. The steps of this work are to:

1. Describe how swish-e arranges the index of documents in PDF format and displays the relevant documents based on the query that is inputted by the user.
2. Build a vector space model and using the cosine angle measurements to determine the similarity of paragraph in the document.
3. Make an application prototype to detect the similarity between documents. The process of comparing documents, query is needed. Query is the keyword of a document. In one document, we can find single query or more. The results of the study using a single query word indicate that the couple of paragraphs in a group of high similarity is more than the couple of paragraphs with low and moderate similarity. The average number of paragraphs in a pair of high similarity group is 21.6 pairs, followed by a group of low similarity with 15.0 and 8.2 for each similarity groups. The two words query's results are also similar to the results of a word query. Total pairs of paragraphs with high similarity are more than the low and moderate similarity. The pair paragraphs' average number with high similarity is 1461.6 pairs, followed by a group of low similarity, 581.0, then the moderate similarity groups with 6.0. The study results of three words query and all kind of queries show that the proposed algorithm detects similarities with both of the couple paragraph

successfully. The study results show that the pairs of paragraphs with high similarity can be detected well.

2.1.3. External and Intrinsic Plagiarism Detection Using Vector Space Models [6]

Plagiarism states precisely the meaning of intellectual stealing, has became a serious keys not only in academic institutions for ages. There are some different kinds of plagiarism which are occurred, from the easiest copy paste to paraphrased and translated plagiarism without rewrite the sources or the authors.

External plagiarism detection is similar to textual information retrieval (IR) (Baeza- Yates and Ribeiro-Neto, 1999). This kind of plagiarism detection gives a unit of query terms in an IR system to revert a ranked set of documents from a corpus that best synonym the query terms. Giving the answer of such queries in the most common structure is an index reverse. An external plagiarism detection system using an inverted index indexes passages of the reference corpus' documents. For each passage in a suspicious document a query is send to the system and the returned ranked list of the reference passages is analyzed. Such a system was presented in (Hoad and Zobel, 2003) for finding duplicate or near duplicate documents. For external plagiarism there are three stages that consist of the system. The first is vectorization of the passages of each document in reference corpus and partitioning of the reference

corpus vector space. Second, vectorization of the passage of a suspicious document and finding the closest ones from each passages using corpus vector space. The last one is, post processing of detected plagiarized passages, combining subsequent plagiarized passages into a single block.

On the other hand, the recent method of plagiarism detection which is considered in the scientific community is Intrinsic plagiarism detection. It was first introduced in (Meyer zu Eissen and Stein, 2006) and defined as detecting plagiarized passages in a suspicious document without a reference collection or any other external knowledge. A suspicious document is first decomposed into passages.

The system is divided into three stages. Vectorization of each sentence in the suspicious document becomes the basic stages. Next is determination of outlier sentences according to the document's mean vector. The final stage is post processing of the detected outlier sentences.

According to Mario, Markus and Roman in External and Intrinsic Plagiarism Detection Using Vector Space Model Journal, they had done the experiment for each external and intrinsic plagiarism detection system in order to determine which system has higher discriminative power. As the result is the research was focused in every single word in a suspicious document in external

plagiarism detection; on the other hand, in intrinsic plagiarism detection research focused in punctuation, part of speech tags, pronouns and closed class words from the documents. Both of them have their own specific function to detect the plagiarism itself. For that reason, it is better to use the combined vector space to evaluate on the complete development corpus with varying parameters, showing significant improvements in all measurements compared to the separates feature spaces.

2.1.4. State of the Art

The summary of the papers above will be explained in the table 1 below.

Table 1 : Summary of Related Work

No	Title	Author	Year	Problem	Method	Result
1	Sistem Deteksi Plagiarisme Dokumen Bahasa Indonesia Menggunakan Metode Vector Space Model	Tudesman, Enny Oktalina, Tinaliah, Yoannita	2014	Plagiarism	Vector Space Model	Plagiarism Detection System Indonesian Documents that used the Vector Space Model where the user can understand the flow of the application process and can determine the level of similarity or plagiarism document journal

2	Vector Space Model untuk mendeteksi plagiarisme	Taufiq M. Isa, Taufik Fuadi Abidin	2013	Plagiarism	Vector Space Model	The study results of three words query and all kind of queries show that the proposed algorithm detects similarities with both of the couple paragraphs successfully. The study results show that the pairs of paragraphs with high similarity can be detected well.
3	External and Intrinsic Plagiarism Detection Useing Vector Space Models	Mario Zechner, Markus Muhr, Roman Kern, Micheael Granitzer	2009	Plagiarism	Vector Space Model	The research was focused in every single word in a suspicious document in external plagiarism detection; on the other hand, in intrinsic plagiarism detection research focused in puntuation, part of speech tags, pronouns and closed class words from the documents. Both of them have their own specific function to detect the plagiarism itself. For that reason, it is better to use the combined vector space to evaluate on the complete development corpus with varying parameters, showing significant improvements in all measures compared to the separates feature spaces.

2.2. Theoretical Foundation

2.2.1. Text Format

Text format is divided into two kinds of text, which are [7] :

a. Plain text is an unformatted text, that text does not contain styles such as fonts, font size, bold, italic and etc. Notepad (*.txt) is the most popular example of plain text.
b. Formatted text. The example of the text is microsoft word (*.doc).

2.2.2. Corpus

The corpus is a collection of some of the texts as a source language and literature research requirements of the text block is used as an object of study of language and literature. McEnery and Wilson (2001) said that “the corpus is the contents of each text. In principle, any collection of more than one text can be called the corpus”.

2.2.3. Plagiarism

According to the Oxford Dictionary [8], Plagiarism is the act of taking essay or opinion of others and make it as if a bouquet or opinions as his own property. Plagiarism is a form of crime that can be charged under criminal law because it could be considered as theft of copyrighted property of others. Plagiarism action undertaken in the academic world can be entangled with the

sanction of dismissal from academic institutions, revocation of a degree, and so on.

In case of various kinds of plagiarism definition, many people make "Classification" of plagiarism types with a different base. According Sudigdo Sastroasmoro[9], the propotion classification or the percentage of word in sentences, the plagiarism of paragraph is divided into several classifications, that's are :

a. Minor Plagiarism: < 30%
b. Medium Plagiarism: 30% - 70%
c. Heavy Plagiarism: > 70%

2.2.4. Information Retrieval

Information Retrieval (IR) is the process of finding material (usually documents) of structured (usually text) that meets the information needs of the large collections (usually stored on computers) [10]. Information Retrieval means a search process of unstructured data from several major collections, which later it would find the results of the information which is required, or a search on the computer's internal data storage media contained on the Internet. Information Retrieval System (IRS) is a system used to reinvent (retrieve) documents that are relevant to the needs of users from a collection of information based on keywords or query from the user. In addition to finding relevant documents to the query, the IRS also conducted the rank to the search results. A document that

has a higher ranking than the other documents will be considered more relevant to the query.

Information retrieval is different with data retrieval. The aim of data retrieval is to determine document which has suitable keyword with the query that given by user in the several documents. It has not been able to solve the user's problem when need an information. While the aim of information retrieval to find all documents which has relevant keyword with query

2.2.5. 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[11].

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)

2.2.6. Vector Space

Vector space model are often used to present a document in a vector space [12]. 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 [10].

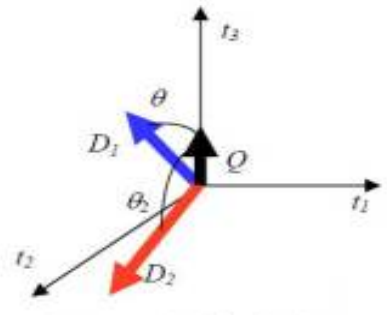

Figure 1 : Vector

Note :

I = word in database

D = Document

Q = Keyword / Query

The advantages using vector space model are [12]:

a. Term Weighted not binary
b. The model is simple because it is a linear algebra
c. The documents rank is more relevant
d. Allow partial matches

2.2.7. 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. [10] 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

2.3. Framework of Study

In conducting the research study, this framework of study is make this research gradual and consistent. The problem is many students in university who did make plagiarism in the thesis and there are no tools to detect one document to other documents. 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. The framework of study will be shown in figure 2 below:

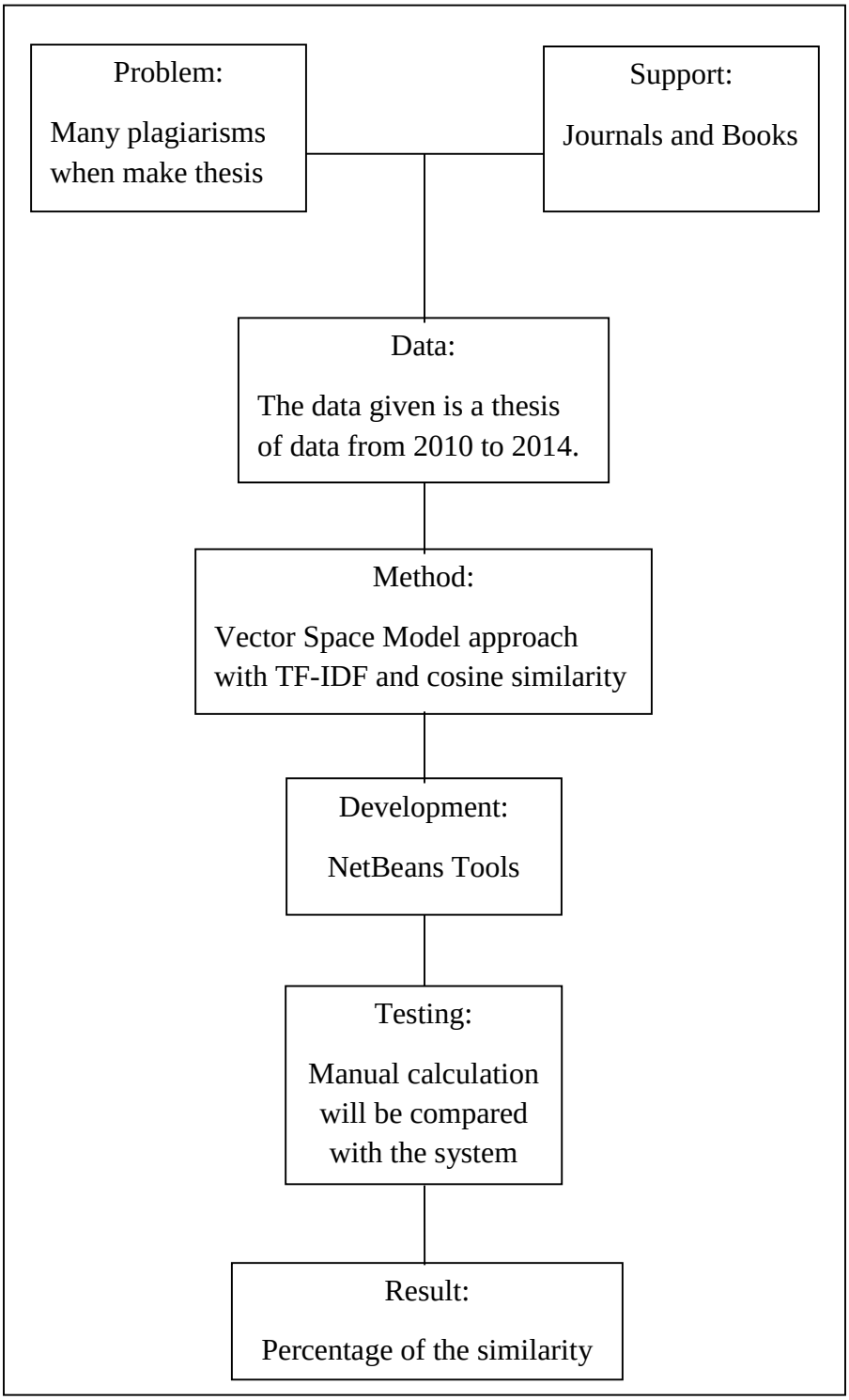

Figure 2 : Framework of Study