CHAPTER 1
INTRODUCTION

1.1 Background

Nowadays, the number of scientific journals increase rapidly along with the development of science. A scientific journal is an article that contains knowledge and information on inventions that comply with the rules and published well in electronic media or conventional media. One of the countries that contribute to the number of scientific journals is Indonesia. Quoted from news.okezone.com written by Susi Fatimah, stated that international scientific journals in Indonesia reached 5,125 as of April 6, 2018. This amount is not comparable when compared to the number of scientific journals from various parts of the world. On the web, labs.semanticscholar.org/corpus states that existing scientific journals reach more than 39 million.

The number of journals that grows exponentially creates its own challenges to analyze the topic of scientific journals that are growing rapidly. Difficulties also increase due to the lack of certainty of the topic in a journal. In each scientific journal, there is an opportunity to have more than one topic. Based on all of the challenges, Topic modeling is required to automatically capture the topics of the journal collections.

Topic Modeling can help understand big data and unstructured text. There are a lot of existing methods to do topic modeling. Latent Dirichlet Allocation (LDA) is one of them. Latent Dirichlet allocation (LDA) is a generative probabilistic model that use random distribution words and documents (Blei, et al, 2003).

A lot of topic modeling studies have been carried out using the LDA algorithm. For the example, (Putra and Kusumawardani,2017) analyze social media's topic in Surabaya using Latent Dirichlet Allocation states that the LDA method is able to find a certain pattern in a document and producing several kinds different topics. There is also (Anupriya and Karpagavalli, 2015) who compared two LDA methods to group journals based on abstracts. The two methods are Collapsed Variational Bayes and Gibbs Sampling. The results of the study stated that the performance of the Gibbs sampling method was more effective.

Based on the existing difficulties and results of previous studies, this study uses the Latent Dirichlet Allocation (LDA) algorithm with Gibbs sampling as the sampling's method to cluster scientific journals by title. In this study, the LDA algorithm is centered on the probability distribution of corpus and divides scientific journals into several topics. In one topic, there are similar and continuous scientific journals. Whereas between topics have very far similarity levels. In the end, this system produces a list of document lists for each topic and a list of words that best describes the topic. The corpus of this research was obtained from scientific journal titles through labs.semanticscholar.org/corpus.

1.2 Scope

The scope of the problems discussed in this research are :

1. The data is the title of the scientific journal in English.

2. The results are lists of keywords and lists of documents every topic.

3. Using Latent Dirichlet Allocation Algorithm collapsed Gibbs Sampling.

1.3 Objective

The purpose of this project is to take opportunities based on the topic distribution for every document and word distribution for every topic. The results of these opportunities become dividers for document lists for each topic and keyword list for each topic. In addition, there is testing to determine the number of topics and iterations that are suitable for conducting topic modeling on the existing corpus. All processes carried out are the result of using the LDA algorithm with Gibbs Sampling technique and are influenced by several parameters such as alpha, betta, iteration, and a number of topics.