CHAPTER 1
INTRODUCTION

1.1 Background

At this time, information technology (IT) rapidly developed and becoming part of human life. IT is considered a subset of information and communications technology (ICT). After modern ICT being presented, nowadays a lot of people have access to the internet and get many information from it. Because of that people now consider to read online news than printed news like newspaper. It make news number keep increasing every time. And all the news comes with many topic. It will be hard for people to find specific news and require a lot of time and effort.

To solve the problem, needed a program to automate the news clustering. K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem[1]. It is one of the algorithm for clustering which has long existed. K-means algorithm is the algorithm that clustering data with iterative process.

To make the K-means algorithm work with the news article, used text preprocessing and term weighting. The result is a program that can clustering news article based on user news articles. With the program, user can easily get related news with the topic they want.

1.2 Scope

The scope covered by this project are :

1. How to do the news article calculation with K-Means?
2. How to make TF-IDF calculation more accurate?

1 https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

3. How to calculate the distance between online news data with user news data?

1.3 Objective

The objective of this project is to make a program that can automatically clustering online news. The benefit of this project are reader can read news based on the topic and make clustering job faster and easier.