Abstract

Detecting plagiarism in text based documents has been developed in various algorithms. The most popular method is local string matching algorithm, such as Longest Common Subsequence, Smith Waterman, and Edit Distance. It can produce accurate detection. However, the algorithm needed more processing time to compare one with many documents. Thus, a method for reducing the number of irrelevant documents is needed. To achieve the objective, this final project proposed a method to give similarity value between two documents based on term frequency. To decrease the processing time, the whole documents are indexed based on term in each document. The experimentation is carried out on three kinds of system design. Fifty document queries are tested against more than 10000 documents. The system are measured by accuracy and processing time. The combination of winnowing, elimination, and LCS shows best result in both parameters. The accuracy reached 89.78 %, same as the previous system (winnowing and LCS) and decrease processing time until 10-times if compared with previous method.

Keywords: *Plagiarism Detection*, *AVL Tree*, *Term Frequency*, *Information Retrieval*, *Longest Common Subsequence*, *Winnowing*