Abstract

Plagiarism is an act of taking the writings of another person and passing them off as one's own. The fraudulence is closely related to forgery and piracy-practices generally in violation of copyright laws. For overcoming the plagiarism, many algorithms have been proposed such as Longest Common Subsequence (LCS), Edit Distance, Document Fingerprinting and Winnowing.

In LCS algorithm, the complexity of matching two documents D1 and D2 with number of tokens m and n respectively is equal to O(m*n). The complexity will increase by the increment of corpus number which will be matched. For decreasing the complexity the decrement of compared tokens number is necessary. This thesis proposed Co-Occurrence Statistical Information and LCS for building plagiarism detector. Co-Occurrence Statistical Information is used for extracting keywords which will be used as a fingerprint of a document. Plagiarism detection will be done by comparing the fingerprint of the related documents using LCS algorithm.

The proposed method requires less complexity compared with LCS without Co-Occurrence Statistical Information. However, the similarity value is relatively similar.

Keyword : Plagiarism, Fingerprint, Co-Occurrence Statistical Information, Longest Common Subsequence.