Journal of Advanced Research in Cloud Computing, Virtualization and Web Applications
Volume 1, Issue 2 - 2018, Pg. No. 8-10
Peer Reviewed Journal
Research Article

Detection of Plagiarism by KMP and Boyer Moore Algorithm

Rajat Tiwari[1], Nishant Sharma[2]

[1]B.Tech Scholar, CSE Dept. Nothern Institute of Engineering Technical Campus, Alwar, Rajasthan, INDIA.
[2]Asst. Prof., CSE Dept. Nothern Institute of Engineering Technical Campus, Alwar, Rajasthan, INDIA.

Abstract

Now a days plagiarism i.e. copying from others work and using it without any permission from the right owner has become so common that so many people do that and publish documents on their own name which is a crime. So as many software tools exist to find out plagiarism and detect that in any documents if something is copied or not. We have also seen this as a major problem in academics where students of UG ,PG and Phd courses copying some part of original document. In this paper my composition of algorithm divides the submitted article in small pieces and scans it to compare with connected databases and internet.

Keywords: Plagiarism, KMP Algorithm, Boyer Moore Algorithm

Introduction

This has always been a very crucial task for people to compare two documents that whether they are copied fully or just some data of one document is copied to another document. Plagiarism is a very deplorable threat to our information system and education system as stealing someone's work is a crime. Plagiarism can be of many types like;

To totally copy someone's work.

To take some data from different documents and mix them afterwards.

Some changed some words but the basic idea is same as of original document.

Plagiarism can also be seen in today's world that most of the students and people due to laziness and may be shortage of time do copying and show it to others, the same thing applies in case of students they never drew particularly attention to practical knowledge and all that research work instead they mostly rely on copy paste criteria.

To find out similarity between two sequences in documents is plagiarism, but it is very difficult to understand what is the main content that is copied from where thus if we use the KMP algorithm along with boyer moore algorithm then we can divide the document in pieces and scan it precisely.

Pre Work on Related Topics

This plagiarism problem is not new for our system so many people have done research on this particular topic and how to overcome these problems and so many algorithms has also been developed to reconsider the situation of copying data from one to others.

A hard step against plagiarism has been taken by technical university, Madrid, Spain, who developed a structure metric tool PK2 for this problem. The only thing that PK2 was unable to identify was when the data is copied in a very small fragments of so many documents. This tool may give false result sometimes when they are used in a group of shuffled statements, thus it was not a very big success.

The PK2 tool is based on structure metric system developed by the students of computer science which were said to make a minor project using c language, assembly language and microprogramming.

It is a metric based tool so when considering the java language for comparing then it only most used keywords and library functions are considered.

Corresponding Author: Rajat Tiwari, CSE Dept. Nothern Institute of Engineering Technical Campus, Alwar, Rajasthan, India.
E-mail Id: rajattiwarirudra007@gmail.com
Orcid Id: https://orcid.org/0000-0002-5851-0586
How to cite this article: Tiwari R, Sharma N. Detection of Plagiarism by KMP and Boyer Moore Algorithm. *J Adv Res Cloud Comp Virtu Web Appl* 2018; 1(2) 8-10.

Another great discovery in this field was CCPE (Cheater Cheater Pumpkin Eater) discovered by James A McCart and Jay Jarman in 2008. It was developed ton determine whether the microsoft databases are duplicated.

System Algorithm

- *Knuth-morris-pratt algorithm*

This algorithm commonly known as KMP algorithm searches for an occurence of a word in a particular string or main string by doing an observation that when a mismatch occurs, the word itself contain enough information to determine whether the next match could begin thus bypassing re-examination of previously mismatched words, The most strainght forward algorithm is that which looks for a character match at successive values in a string as the index value reach the end of the string then if there is no match then search is said to fail. At each location the algorithm firstly searches the character from a word if match found then the successive word is searched ,but this process take so much time on the other hand KMP algorithm has a better worst case performance than normal search algorithm.

The difference is that KMP uses previous matched

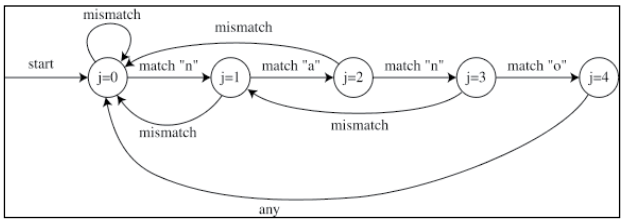

Figure 1.The figure shows the procedure for matching characters by KMP algorithm by taking an example of "nano" word

information which normal search does not.

KMP algorithm complexity is O(n+k). thus we can also use hash functions for faster comparison, thus comparison can be done from left to right in a document or journal. Here 'n' is the length of word and 'k' is the length of pattern.

We are using dependency matrix for comparison of same size matrix, it can be assumed that if the person has changed text position and variables but total variables in a function would remain same and thus such a code can be detected by this algorithm ,only limitation of this is that the same size of matrix can be used or in other words the same word limit should be there.

System Overview

Block diagram of figure 2 illustrates the basic functioning of the algorithm

System Architecture and Working

system working in plagiarism detection.

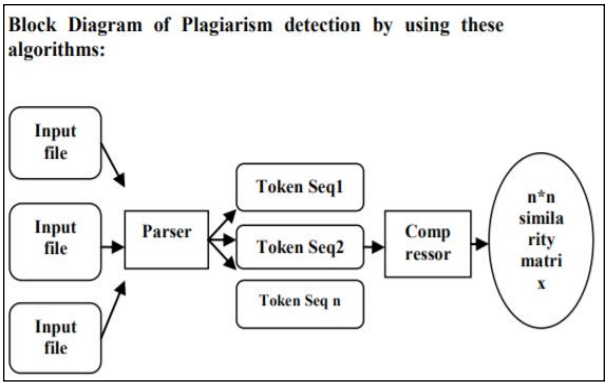

Figure 2.System Architecture

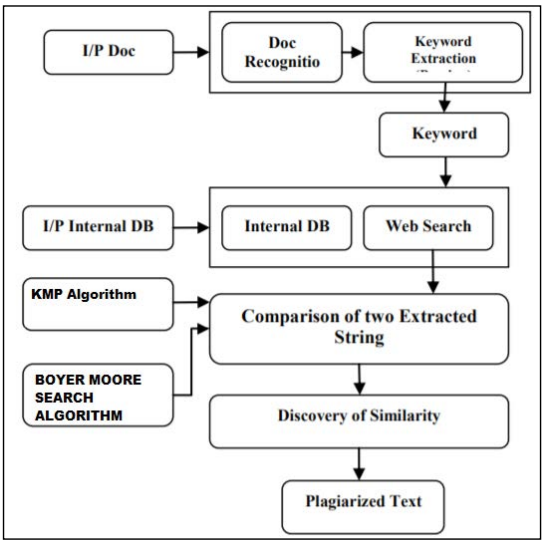

Figure 3.System Architecture of Plagiarism Detection

- Input document

Input document is given for recognition the document can be any journal or paper.

- Keyword extraction

The keywords from that document has been extracted for comparing it with the internet and other databases.

- Algorithms

Here we have used the KMP algorithm along with boyer moore algorithm to find out any suspicious matter in the document.

- Discovery of similarity

After going through the algorithms any similarity can be judged or discovered.

We have computer Plagiarism detection (PD) accuracy by using Precision value and Recall value as a follows;

$$\text{PRECISION VALUE}(P) = \frac{\text{NUMBER OF CORRECT DOC RECOVERED FOR PD}}{\text{NUMBER OF TOTAL DOC RECOVERED FOR PD}}$$

$$\text{RECALL VALUE}(R) = \frac{\text{NUMBER OF CORRECT DOC RECOVERED FOR PD}}{\text{NUMBER OF TOTAL RELEVANT DOC FOR PD}}$$

S.No.	Approach	Precision Value %	Recall Value %
1.	MLSOM	61%	60%
2.	LSI	61%	65%
3.	Proposed System	71% and above	71% and above

Conclusion

This paper proposes new plagiarism technique using KMP algorithm with boyer moore algorithm. My proposed system gives value of approximately 80% accuracy with precision.

References

1. https://en.wikipedia.org/wiki/Boyer%E2%80%93Moore_ string_search_algorithm.
2. https://www.google.co.in/search?q=all+about+K-MP+algor. ithm&oq=all+about+KMP+algorithm+&aqs=-chrome..69i57.17627j0j7&sourceid=chrome&ie=UTF-8.
3. https://www.google.co.in/search?q=complexi-ty+analysis +ofkmp+algorithm&oq=complexity+of+K-MP&aqs=chrome.1.69i57j0l2.11697j0j7&sourceid=-chrome&ie=UTF-8.
4. https://www.ics.uci.edu/~eppstein/161/960227.html.
5. https://www.geeksforgeeks.org/searching-for-pat-terns-set-2-kmp-algorithm/.
6. https://www.researchgate.net/publica-tion/220975322_Knuth-Morris-Pratt_Algorithm_An_ Analysis.

Date of Submission: 2018-11-13
Date of Acceptance: 2018-11-23