A SEMANTEC ANALYSIS ON PLAGIARISM DETECTION: METHODS AND TOOLS

[1]Ashita Mahajan, [2]Arshiya, [3]Neeraj Sharma, [4]Jashanpreet Singh
[1]Research Scholar, [2] Research Scholar, [3]Assistant Professor, [4]Assistant Professor
[1]Information Technology,
[1]Chandigarh Engineering College, Landran, India.

Abstract: **Plagiarism is an illegitimate stealing or publication of another artist, writer or performer in any field be it science, art, music, video, text, etc. As it is a assurgent hitch this paper includes an outlook of all the methods for detecting plagiarism. Each plagiarism detection method adopts several different techniques.**

Index Terms **- plagiarism, method, tools, original, detection and technology.**

I. INTRODUCTION

Due to large scale technological amelioration , free resources and information related to every motif is easily accessible. This elevation in technology increases the extortion of original credentials as it becomes very easy. When an original document is plagiarized , for it to become laborious to detect , some of the original semantics or contentions are alternated .

Thus plagiarism detection methods require distinct techniques that work on peculiar and divergent propositions . It is very vital to provide information but in a restricted way to ensure that the primeval document cannot be plagiarised . The different plagiarism detection techniques can focus on different aspects of the document which may include the basic semantics , topic of the document , paragraph wise comparisons , texts comparisons , text parsing , keyword comparison , metrics , parse tree comparisons , etc.

These distinct techniques target divergent countenances of a document to be sure not to leave a single plagiarized form behind . According to Culwin and Lancaster plagiarism detection includes four stages –

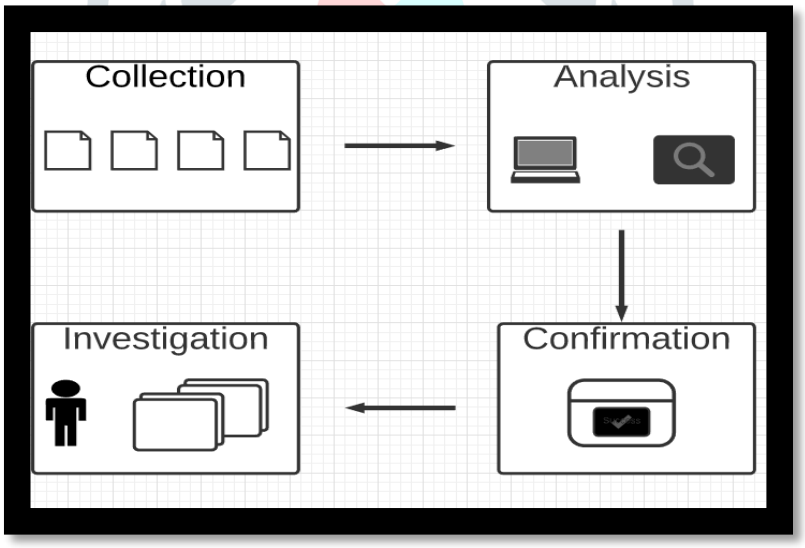

Figure 1

- Collection :

In this the new document to be checked for plagiarism is obtained or collected .

- Analysis:

This includes the comparison of the new document to be checked and the original document . These two documents are compared on the basis of distinct aspects depending on different plagiarism detection methods.

- Confirmation:

In this phase , all the semantics and aspects are altogether compared to confirm plagiarism .

- Investigation :

In this phase the writer of the plagiarized credential is investigated for the document .

II. PLAGIRISM DETECTION METHODS

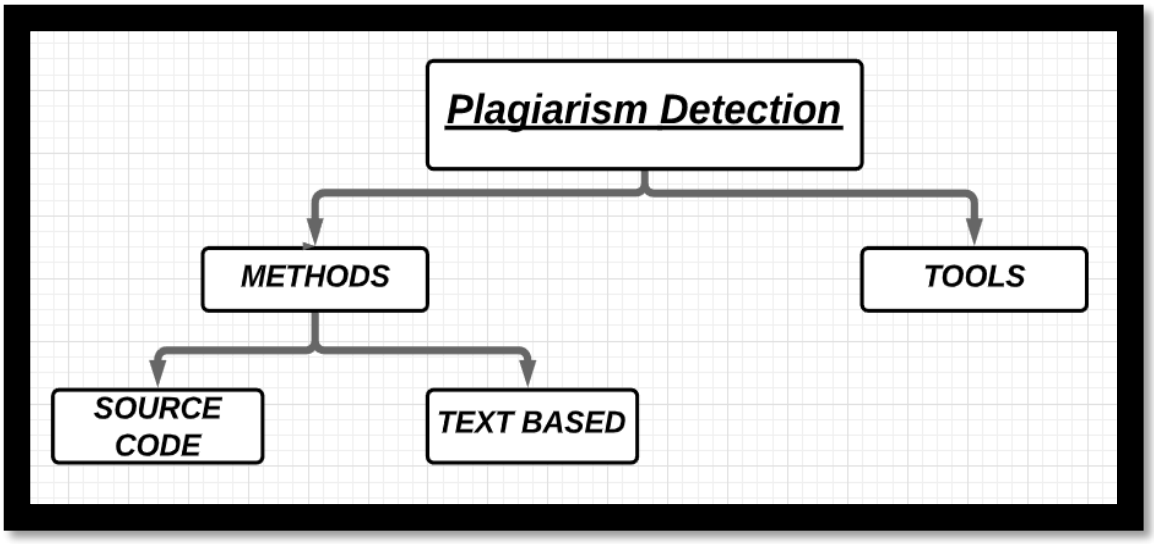

Figure 2

TEXT BASED DETECTION TECHNIQUES :

1. SUBSTRING MATCHING:

THIS METHOD IS USED TO DETECT PLAGIARISM BY COMPARING THE STRINGS OF THE NEW DOCUMENT TO BE CHECKED WITH THE ORIGINAL CREDENTIAL . THIS USES THE STRING AS AN OMEN . IT USES GRAPH AS A BASE AND THE SUBSTRINGS HERE ARE REPRESENTED AS SUFFIX TREES . IT IS WIDELY USED IN PLAGIARISM DETECTION OF SOFTWARE CODES . JARO – WINKLER ALGORITHM IS ONE OF THE EFFICIENT ALGORITHMS USED FOR SUBSTRING MATCHING .

JARO – WINKLER ALGORITHM MAKES USE OF THREE MAIN PARAMETERS FOR COMPARING SUBSTRINGS OF THE TWO CREDENTIALS .

THE LENGTH OF STRINGS TO BE COMPARED.

NUMBER OF SIMILAR STRINGS IN BOTH THE DOCUMENTS .

FINDING THE NUMBER OF INTERSECTIONS .

2. KEYWORD SIMILARITY:

THIS DOCUMENT FIRST CHECKS SIMILARITY BETWEEN DOCUMENTS . IN THE SIMILAR DOCUMENTS FIRST A THRESHOLD POINT FOR THE KEYWORDS IS SET . THEN THE KEYWORDS OF THE ORIGINAL DOCUMENT ARE COMPARED WITH THE KEYWORDS OF THE STATED DOCUMENT . IF THE COUNT OF SIMILAR KEYWORDS IS LESS THAN THE THRESHOLD THEN THE WHOLE STATED DOCUMENT IS FURTHER SEGREGATED INTO SUBSECTIONS AND THEN RECURSIVELY CHECKED FOR SIMILAR KEYWORDS .

3. EXACT FINGERPRINT MATCHING:

THIS METHOD MAKES USE OF HASH TABLE . FIRST THE STATED DOCUMENT TO BE CHECKED FOR PLAGIARISM IS SPLIT UP INTO PROGRESSIONS CALLED CHUNKS. THEN THE PATTERN OF THE STATED DOCUMENT IS CHECKED DIGITALLY . THIS MOULD IS THEN INSERTED IN HASH TABLE . THE TOTAL DILAPIDATIONS IN THE HASH TABLE SHOW THE SEQUENCING OF THE TWO DOCUMENTS . THIS METHOD IS NOT COST EFFICIENT .

4. TEXT PARSING:

THIS METHOD IS USED TO DETECT PLAGIARISM BY CHECKING THE STRUCTURE OF A SENTENCE IN THE STATED DOCUMENT . IT IS MAJORLY CONCERNED WITH COMPARING THE SEMANTICS OF SENTENCES OF THE STATED DOCUMENT WITH THE ORIGINAL DOCUMENT . TO MAKE THE PROCESS MORE EFFICIENT AND RELIABLE , THE SENTENCE TO BE CHECKED IS SUBSTITUTED IN THE FORM OF A TREE . THE TREE IS SUBDIVIDED INTO NODES THAT COMPARE DIFFERENT SEMANTICS OF THE SENTENCE . THIS TREE IS KNOWN AS A PARSE TREE . INITIALLY ONLY TRACES OF THE PARSE TREE ARE TAKEN WHICH INCLUDE BRANCHING , ASSIGNMENTS , ETC. JUST LIKE A TREE , EACH NODE OF TREE IS SUBDIVIDED INTO CHILDREN NODES FOR RECURSIVE COMPARISONS . THIS WHOLE PROCEDURE IS FOLLOWED UNTIL WE REACH THE LEAF NODES AFTER WHICH NO SEGREGATION IS POSSIBLE . THIS TREE CONFIRMS PLAGIARISM .

SOURCE CODE DETECTION TECHNIQUES:

1.LEXICAL SIMILARITIES :

THIS METHOD OF PLAGIARISM DETECTION , ADOPTS SEVERAL DIFFERENT TECHNIQUES FOR DETECTING PLAGIARISM ON A WORD LEVEL . LEXICAL SIMILARITIES IS AN IMPROVED VERSION OF STRING MATCHING TECHNIQUE . IT CAN WORK ON DIFFERENT ASPECTS LIKE SENTENCES , TOKENS , KEYS , SYNONYMS , CASE OF WORDS ,IDENTIFIERS , RESERVED WORDS , ETC.

2.PARSE TREE SIMILARITIES:

AS THIS IS A METHOD OF PLAGIARISM DETECTION FOR SOURCE CODES , WHENEVER A PROGRAM IS COMPILED , THE COMPILER CONVERTS THE PROGRAM INTO A PARSE TREE . WHEN THE PROGRAM IS BEING CONVERTED INTO THE PARSE TREE , EACH STEP OF THE PROGRAM IS SUBDIVIDED INTO PARSE TREE FORMATION . THIS GIVES A BETTER REPRESENTATION OF THE PATTERN AND THE SEMANTICS OF A PARTICULAR PROGRAM . IN THIS WAY A PROGRAM CAN BE EASILY COMPARED WITH THE PARSE TREE OF THE ORIGINAL PROGRAM. EVEN IF SOME PORTIONS OF THE PROGRAM ARE CHANGED , THE PARSE TREE SHOWS OTHER SIMILARITIES THAT MAKES THIS METHOD MORE RELIABLE AND EFFICIENT FOR DETECTING PLAGIARISM .

3.PROGRAM DEPENDANCE GRAPHS (PDG):

AS THE NAME SUGGESTS PDG IS A GRAPHICAL REPRESENTATION OF THE PROCEDURES TAKING PLACE IN A PARTICULAR PROGRAM . A PDG COMPRISES OF TWO TYPES OF LINES :

- DASHED LINES WHICH ARE RESPONSIBLE FOR REPRESENTING THE CONTROL DEPENDENCIES .
- SOLID LINES WHICH ARE RESPONSIBLE FOR SHOWING THE DATA DEPENDENCIES IN A PROGRAM .

IT TELLS THE MANNER IN WHICH ONE STATEMENT IS DEPENDANT ON ANOTHER STATEMENT FOR ITS EXECUTION AND ITS RESOURCES . A PDG IS A DIRECTED GRAPH THAT SHOWS THE FLOW OF INFORMATION . IT CONSISTS OF ALL THE BASIC FUNCTIONS OF A PROGRAM LIKE FUNCTION CALLS , DECLARATION , ASSIGNMENT , ETC . BY MAKING A PDG A PROGRAM CAN BE GRAPHICALLY REPRESENTED , THEREFORE IT BECOMES EASY TO COMPARE THE GRAPH OF ONE PROGRAM WITCH THE GRAPHS OF OTHER PROGRAMS TO BE CHECKED FOR PLAGIARISM .

III. PLAGERISM DETECTION TOOLS

1. **PLAGAWARE:**

PlagAware is a plagiarism search engine . It provides an online utility for colleges , schools , universities and even for the students . It just compares the main topics of the credentials and does not provide any service for inner details like crux of the sentence or similarities between two words . It provides the user with a number of chronicles for checking for plagiarism between the stated document and the original one. The user can make use of all these reports for comparison .

2. **PLAGSCAN:**

PlagScan is a text based plagiarism checker . It is an online software which is easily available to all the its users for checking their documents for plagiarism . It provides the user with convoluted contrivances for getting into the details of the stated document for more precise plagiarism detection . In this software system , rare trademarks are extricated from the stated document and then these trademarks are compared with all the original credentials that are saved .

3. **CheckForPlagiarism.net:**

CheckForPlagiarism.net is a software that is used to detect online plagiarism by using the exact fingerprint matching technique. This tool uses fingerprint of both the stated document and the original credentials and compared them for detecting plagiarism . It is one of the best tools as it can compare multiple documents at a time and also works on details like the anatomy of the sentence etc .

4. **COPYCATCH:**

Copycatch is an online service that looks for plagiarism by checking the complete meaning of the sentence instead of word to word comparisons. Copycatch makes use of Google API . All the users of copycatch are given a choice of two different versions , which can be adopted by the users depending on their needs .

5. **MOSS(MEASURE OF SOFTWARE SIMILARITY):**

As the name suggests MOSS is used to detect plagiarisms in source codes of different languages like Java , C , Pascal , C++ , and many more . To provide more reliability , MOSS works in two stages for plagiarism detection :

- In the first stage , the stated document is divided into keys or tokens.
- In the second stage , these keys or tokens of the stated document are compared with the original credentials .

Thus the time for detection decreases by representation in the form of tokens and also the reliability of the method increases.

6. SAFE-ASSIGN:

Safe-Assign is a plagiarism detector that is mainly built for the students of schools , colleges or universities . It is a free service that helps the students before submitting their assignments or reports to check it for plagiarism . The stated assignment submitted by the student is compared with the original credentials present in the database to check for plagiarism .

7. WCopyFind:

WCopyFind is an open source online application that is used to detect plagiarism . It provides a very fast service . WCopyFind gives the user several options or criterias for checking plagiarism . These criteria inludes punctuation marks , case sensitivity , string length , numbers , basic characters , etc.

8. iThenticate:

iThenticate is one of the earliest online or offline service provided by iParadigms LLC . It is not for academic use but mainly for corporate use . This application is commonly used by professional authors and writers . Unlike CheckForPlagiarism.net , iThenticate does not check the structure of the sentence and the similarity between words for plagiarism detection . It just checks the documents basic details as a whole and compares it with the original saved credentials .

9. PlagiarismDetection.org:

PlagiarismDtetection.org is contemplated as one of the most accurate tool for plagiarism detection . It is widely used by students and teachers .

Merits:

- It has a great database of its own
- Provides high veracity

Demerits:

- Only compares a single credential at a time .
- Does not supports checking of details like anatomy of a sentence and similar words etc .

10. VIPER:

Viper is a software that provides free services for plagiarism detection . It is mainly used in academic field by teachers , students, writers ,etc. It consists of a millions of original credentials in its database . Whenever a text is to be checked for plagiarism the stated text is compared to the original documents present in the database of viper .

11. URKUND:

Urkund is an online paid service for plagiarism detection . It works by comparing the submitted answers or texts with the original text . After comparing them it analyses the percentage of similarity between the two texts and culminates a report of plagiarism in the stated answer or text .

IV. CONCLUSION:

Plagiarism refers to the act of cloning from an already existing publication of a work. Our paper inculcates the current methodologies and tools for plagiarism detection .It also includes list of various online services meant for detecting plagiarism in corporate and academic field . It is basically comparison of frameworks of original credential to the provided one by various tactics.

V. REFERENCES

1. Paul Clough. Old and New Challenges in Automatic Plagiarism Detection. National UK Plagiarism Advisory Service, 2003.
2. Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism - A Survey. Journal of Universal Computer Science, 12(8): 1050–1084, 2006.
3. Martin Potthast ,Benno Steinan, Alberto Barrón-Cedeño and Paolo Rosso. Evaluation Framework for Plagiarism Detection,2010
4. Antonio Si ,Hong Va Leong and Rynson W .H . Lau CHECK:A Document Plagiarism Detection System,feb 1997
5. Ahmed Hamza Osman1,2, Naomie Salim1, and Albaraa Abuobieda1,2 .Survey of Text Plagiarism Detection, June 2012
6. Romans Lukashenko, Vita Graudina, Janis Grundspenkis .Computer-Based Plagiarism Detection Methods and Tools: An Overview ,
7. Lancaster T., F. Culwin. A visual argument for plagiarism detection using word pairs. Paper presented at Plagiarism: Prevention, Practice and Policy Conference 2004.
8. Lancaster, T., F. Culwin. Classifications of Plagiarism Detection Engines. ITALICS Vol. 4 (2), 2005
9. Delvin, M. Plagiarism detection software: how effective is it? Assessing Learning in Australian Universities, 2002.
10. Vani K and Deepa Gupta. Study on Extrinsic Text Plagriarism Detection Techniques and Tools,2016
11. Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and PaoloRosso. Overview of the 1st International Competition on Plagiarism Detection.
12. Bull, J., Collins, C., Coughlin, E. and Sharp, D. (2001), Technical Review of Plagiarism Detection Software Report, ComputerAssisted Assessment Centre, University of Luton.

13. Hannabuss, S. (2001), Contested texts: issues of plagiarism, Library Management, MCB University Press, Vol. 22(6-7), 311318.
14. Ottenstein, K. J. (1976), An algorithmic approach to the detection and prevention of plagiarism, SIGSCI Bulletin, Vol. 8 (Part 4), 30-41.
15. Parker, A. and Hamblen, J. O. (1989), Computer Algorithms for Plagiarism Detection, IEEE Transactions on Education, Vol. 32(2), 94-99.
16. Prechelt, L., Malpohl, G. and Philippsen, M. (2000), JPlag: Finding plagiarisms among a set of programs.

ABOUT US :

ASHITA MAHAJAN:

I am currently pursuing my B.Tech in IT at CHANDIGARH GROUP OF COLLEGES and I am in my 2ND year of education there . My interest lies in using Machine Learning and AI to improve the upcoming technologies.

ARSHIYA:

I am currently pursuing my B.Tech in IT at CHANDIGARH GROUP OF COLLEGES and I am in my 2ND year of education there . My interest lies in analysis of data and its management.

Neeraj Sharma is serving as Assistant Professor in Chandigarh group of colleges, Landran and is having 5 years of teaching experience. He has completed his Masters of Technology from Himachal Pradesh Technical University. He has done his research work in key areas of text mining using back propagation techniques. He has four research papers in international journals and have participated in two international conferences.

JASHANPREET SINGH is serving as Assistant Professor in Chandigarh group of colleges, Landran and is having 5 years of teaching experience. He has completed his Masters of Technology from LPU.