A Survey on Plagiarism Detection

Prasanth.S
PG Scholar
Sri Ramakrishna Engg Coll
Tamilnadu,India

Rajshree.R
PG Scholar
Sri Ramakrishna Engg Coll
Tamilnadu, India

Saravana Balaji.B
Assistant Professor
Sri Ramakrishna Engg Coll
Tamilnadu, India

ABSTRACT

Being a growing problem, plagiarism is generally defined as literary theft and academic dishonesty in the literature, and it is really has to be prevented and stick to the ethical principles. This paper presents a survey on plagiarism detection systems. Common feature of different detection systems are described.

Keywords

Plagiarism detection, plagiarism types, plagiarism techniques.

1. INTRODUCTION

Plagiarism is one of the growing global problems experienced by the publishers, researches and educational institutions which is generally defined to be the literary theft. That is, taking the ideas, documents, codes, images, etc of another person and presenting them as own works. This is meant by plagiarism. This proves an act of dishonesty in academics and literature and hence it has to be prevented.

This paper presents the various plagiarism detection techniques. Plagiarism detection is a good solution to detect the theft of scientific papers, literary works and source codes. Plagiarism occurs at various levels. For example, students make use of programming codes by copying from other's works for their assignments to get more marks with no effort.

The various forms of plagiarism are as follows,

- Using sources without proper citing or reusing the ideas without citing the references
- Other forms of plagiarism include translated plagiarism in which the original content is translated and used in our paper, artistic plagiarism in which the images and videos are used without proper citation
- Another form of plagiarism is cloning the codes ,i.e,reusing the source code without permission and citation of original authors

Plagiarism of computer programs is quite common among the undergraduate classes where students tend to copy the source programs and modify them with little changes in the appearance. This creates a negative impact on learning process by affecting the quality of students and their credits.. Hence plagiarism is one of the major concerns that have to be dealt with.

Synopsis:

Detection of plagiarism

- Plagiarism in documents
- Plagiarism in codes
- Plagiarism techniques
- Plagiarism algorithms

1.1 Plagiarism in documents

Plagiarism in documents is more among the student community for their academic purpose, especially the post graduate who modifies the source document and present as their own. This has to be prevented as it affects the quality of the student from assessing their own capability. Hence plagiarism in documents must be first detected and the systems that can be used for this purpose are,

1. Web enabled systems
2. Stand-alone systems

*1.1.1 W*eb enabled systems

Web enabled systems are more widely used since they extend their search of the plagiarized resources to world wide web easily and is more reliable. The two systems that come under web enabled detection systems are as follows,

- ➢ Turnitin
  Well known commercial detection system in which the submitted document is compared with the previous student works and other international databases to detect whether it is a plagiarized document.
- ➢ Safe Assign
  Checked the submitted paper with the following databases,
  i. Internet databases
  ii. Documents already submitted to Safe Assign database
  iii. Documents submitted by students to Global reference database in order to prevent plagiarism

1.1.2 Stand-alone systems

These are the detection systems that can be installed in the systems. Two types of stand-alone systems are

- ➢ EVE
  EVE refers to Essay Verification Engine. This system works only when Connected to internet. It searches the internet and locate matches for the sentences in the query document (i.e,submitted paper) with suspected websites.
- ➢ WCopyFind
  This system finds plagiarism between two or more assignments.

Various other plagiarism detection tools are

Diff, SCAM, COPS,KOALA ,SSK , CHECK ,MDR ,PP Checker , SNITCH and Ferret.

2. PLAGIARISM IN CODE

Various approaches have been introduced to detect the source codes written with C,C++ or JAVA. These approaches are used to compare the source codes written in different programming languages. They are also used to detect complex modification in the codes but with a longer detection time. The most suitable approach of detecting plagiarism in programming code is the structure based method. This method makes use of tokenization and string matching algorithms to detect the similarities

The existing plagiarism methods that employ structure based methods are Plague[1],YAP[2] ang JPlag[3].

- Plague is one of the earliest method.
  This method creates a structure profile for each source code and compares them with the heckle algorithm that handles the text files.
- YAP has three versions
  All three versions has two phases of processing. First phase is creating token file for each source code and the second phase is the comparison of every token files. The result of each comparison is a value called percent match. The value of percent match should lie between 0 and 100.If the percent value is larger than the minimum value 0 then code is suspected to be a plagiarized code. The detection result in this case is displayed as a text file.
- JPlag is available as a free web service
  In this method a directory containing source programs are fed as input. Every source code in the directory are parsed and transformed to token strings. These token strings are compared with each other by the greedy string tilling algorithm. The detection result in this case is displayed as HTML files and can be opened using a standard browser.

3. PLAGIARISM TECHNIQUES

Plagiarism technique is commonly known as similarity detection technique. A good example for the plagiarism technique is the attribute counting techniques used in the early days. The attribute collection technique creates special finger prints for collection files and also includes file size, average number of commas per line and so on. The files with close finger prints are considered to be similar. But this technique is not reliable.

Hence this has been replaced by the modern plagiarism technique that employs content comparison techniques. Among them widely used techniques are string tilling[4] and parse tree comparison[5] as discussed before. One more technique known as Fast Plagiarism Detection System (FPDS)[6] is used to increase the algorithm performance of the plagiarism detection. This is because of the special indexed data structure to store the collection files. Next is the tokenization technique [7] that prevents renaming of variables. However renaming of variables will not help the plagiarizer in the case of small tokenization algorithms in which elements of the program are substituted as single tokens for comparison.

4. PLAGIARISM ALGORITHMS

A number of algorithms have been proposed to detect plagiarism. The simple algorithm is based on string comparisons and they are implemented as below:

- Remove all comments
- Ignore all blanks and extra lines, except when needed as delimiters
- Perform a character string comparison between the two for all program pairs

For code plagiarism detection, Faidhi and Robinson[8] characterize six levels of program modification in a plagiarism spectrum.

Level 0 - original program without modifications

Level 1 - only comments are changed

Level 2 – changes identifier names

Level 3 – changes position of variables

Level 4 – changes constants and functions

Level 5 – in this level program loops are changed

Level 6 – control structures are changed to an equivalent form using different control structure

Thus this simple algorithm will detect many cases of plagiarism.

5. EXISTING TOOLS OR METHODS TO DETECT PLAGIARISM

5.1 Usual Approach

The usual approach for detecting plagiarism is utilizing fingerprints technique. A bunch of fingerprint contains pieces of text that may overlaps with one another. A fingerprint is then used as a query to search the web or a database, in order to estimate the degree of against synonyms plagiarism. Most currently available software packages make use of this technique. Variations can be occur only in the fingerprints used and search engines employed. The advantage of this method is that it is fairly stable against the re-arranging of text. It is however not stable and translations.

5.2 Stylometry

Stylometry is an attempt to analyze writing styles based on similarities exists in text. A selective text can be matched with the typical writing style of an individual based on his or her existing works. Moreover the text exists in a single paragraph can be compared with the overall style of writing as found in a paper. As we mentioned, stylometry is able to detect plagiarism without the need for any external corpus of the documents. It can be applied to detect essential patterns within documents that capture style parameters that include syntactic formats, structure of the text as well as the key terms. The detection of plagiarism within the document domain or without any external reference is well described as “intrinsic plagiarism detection”.

5.3 Integrating Search Application Programmers Interface (ISAPI)

A home-grown plagiarism detection method built on top of Google’s search API has surprisingly produced superior results as compared to leading software packages in the industry such as Turnitin and Mydropbox. This is mainly due to Google’s indexing of many more web sites as compared to these tools for plagiarism detection. The disadvantages of using Google’s free API have restricted their system to 1000 queries a day. Instead of license their Search engines, Google maintain the exclusive control to numerous potential applications.

Table 1. Different plagiarism methods comparison

Plagiarism methods	**Advantages**	**Disadvantages**
Plagiarism in documents	Extend their search of the plagiarized resources to internet easily and more reliable.	High cost
Plagiarism in code Plague	Each source code is compared them with the heckle algorithm.	Corruption occurs without warning.
Yap	The token file is created for each source code and it is compared with another token file.	It take more time to create the tokens.
J Plag	Every source code in the directory are parsed and transformed to token strings. These token strings are compared with each other by the greedy string tilling algorithm.	It can not handle files which do not parse.
Plagiarism Techniques	Used to increase performance of the plagiarism detection.	This technique is not reliable.
Plagiarism Algorithms	The algorithm is based on string comparisons and simple.	File size increased the processing cost also increased.

6. CONCULSION

A survey on plagiarism detection systems has been introduced. With the evolution of the internet and the need for information the plagiarism continues to be a concern problem to universities, teachers, policy-makers and students. Concluding that, the need of plagiarism detection systems become very important issues and the use of plagiarism detection systems in E-Learning improve the integrity of academic, and also instances of plagiarism can be successfully reduced with the help of plagiarism detection systems.

7. REFERENCES

[1] G. Whale, "Plague : plagiarism detection using program structure," Dept. of Computer Science Technical Report 8805, University of NSW,Kensington, Australia, 2008

[2] M. J. Wise, "Detection of Similarities in Student Programs: YAP'ing may be Preferable to Plague'ing," ACM SIGSCE Bulletin (proc. Of 23rd SIGCSE Technical Symp.), 2002.

[3] P. Lutz, M. Guido, and M. Phlippsen, "JPlag: Finding plagiarisms among a set of programs,"Fakultätfür Informatik Technical Report 2000-1, Universität Kalrsruhe, Karlsruhe, Germany, 2000. International Journal of Computer Theory and Engineering Vol. 4, No. 2, April 2012 187.

[4] M. J. Wise, "YAP3: improved detection of similarities in computer program and other texts," Proc. of SIGCSE'96 Technical Symposium,2006.

[5] L. Prechelt, G. Malpohl, and M. Philippsen, "Finding plagiarisms among a set of programs with JPlag," Journal of Universal Computer Science, 2008.

[6] D. Gitchell and N. Tran, "Sim: a utility for detecting similarity in computer programs," the 30th SIGCSE Technical Symposium on Computer Science Education, 2006.

[7] M. Mozgovoy, K. Fredriksson, and D. White, "Fast plagiarism detection system," Lecture Notes in Computer Science, 2005.

[8] J. A. Faidhi and S. K. Robinson, "An empirical approach for detecting program similarity and plagiarism within a university programming environment," computer education, 2007.