Plagiarism: A Threat to Intellectual Property Rights

Dr. Ruchi Tuli

Assistant professor, Department of Computer Science & Engineering, Jubail University College, Jubail, Kingdom of Saudi Arabia

Abstract- **Plagiarism, in general sense, is making one's original idea as your own. One can say that it is a form of cheating as it steals and implicitly attack one's intellectual property rights. Plagiarism is generally defined as "Academic Dishonesty "and can be considered as one of the electronic crimes. Thus it is important to develop effective approaches to detect plagiarism. This paper presents some common software tools for detecting the plagiarism. The tools are then compared qualitatively on various parameters. At the end, the pros and cons of the different tools are also described.**

Keywords: - Plagiarism detection, plagiarism types, plagiarism algorithms

I. INTRODUCTION

The rapid advancement in technology has made the internet as primary media for information access and almost every piece of information can be found on internet. Today, text can easily be copied and pasted from internet. So, it becomes very easy to use another author's work from internet without giving proper reference. The term "Plagiarize" is defined as imitation of the methods or ideas of some another author and representing it as one's own original work without citation.Plagiarism has become one of the greatest educational challenge because most of the students or researchers are cheating when they do the assigned tasks and projects. This is because a lot of resources can be found on the internet. It is so easy to them to use one of the search engines to search for any topic and to cheat from it without citing the owner of the document. Therefore, it becomes necessary to use plagiarism detection soft-wares to stop or to eliminate students cheating, copying and modifying documents when they know that they will be found. Some types of plagiarism acts can be detected easily by using some of the recent plagiarism detection software available on the market or over the internet. Plagiarism can occur in many forms such as reusing the ideas of others without properly citing, translated plagiarism in which the original content is translated, plagiarism is the source code where whole or a part of the code written by someone else is copied and many others. Plagiarism of computer programs is quite common among the students where they tend to copy the source programs and modify them with little changes in the appearance. In both the textual document plagiarism and source code plagiarism, detection can be either: Manual detection or automatic detection [1].

– Manual detection: done manually by human, it's suitable for lectures and teachers in checking student's assignments but it is not effective and impractical for a large number of documents and not economical also need highly effort and wasting time.

– Automatic detection(Computer assisted detection): there are many software and tools used in automatic plagiarism detection, likeTurintin, Edutie, PlagiServe, Glatt Plagiarism Self-Detection program (GPSD) and many more.This paper compares the various software tools for detecting the plagiarism and then pros and cons of each tool are also mentioned. In this paper, ten software tools namely Copycatch, Glatt (GPSP), Turnitin, Eve2, PlagiServe, Moss, Sim, Jplag, Gplagand Google are compared on various parameters. Section II of the paper classifies the plagiarism detection software's in different categories. In section III different plagiarism detection software are discussed. Section IV provides a comparative study of the features and also discusses pros and cons of each.

II. CLASSIFICATION OF PLAGIARSIM DETECTION SOFTWARE

Plagiarism detection software can be classified into four main categories namely [2] – online or remote search tools, stand-alone desktop software, web search engines and subscription databases.

A. Online or remote search tools

Developing web systems for plagiarism detection overcomes machine capability problems, facilitate the availability of the system to many users and extend the search of plagiarized resources to the World Wide Web easily. Turnitin [3][4]is the most well-known commercial plagiarism detection system to which many universities from UK and USA subscribe. It uses

an enormous database from the Internet and previous student works to be compared with the query document. Plagiserve is another online plagiarism detection software

B. Stand-alone desktop software

Stand-alone software is developed to be installed on computers. EVE [4][3] (The Essay Verification Engine) is a desktop application but it has the capability to make large number of searches on the Internet to locate matches between sentences in the query document and suspected websites. Thus, in order for EVE to work, the machine should be connected to the Internet. Other examples include CopyCatch Gold, Glatt Plagiarism Screening program (GPSP) and Word Check Keyword DP

C. Web search engines

Internet search engines such as Google, AltaVista, and Yahoo can be used an alternate method to detect suspected plagiarism without the need to download software or register for a detection service. Examples of such systems include Google, Alta vista, Looksmart, Amazon and many more.

D. Subscription databases

Finally, subscription databases of scholarly and popular literature which include abstracts or full texts of articles may be searched, particularly if assignments relate to the location of such material. These are specialized databases which we subscribe to which are not available on Google or other common search engines. They include online encyclopaedias, periodical indexes, business directories and other resources. They are broken down by subject to help you find exactly what you need.

III. SOFTWARE PLAGIARSIM DETECTION TOOLS

A. TURNITIN

Turnitin.com[3] [4][5] uses digital fingerprinting to match submitted papers against internet resources and them against an in-house database of previously submitted papers. The website provides online tutorials in the use of the service for both lecturers and students, with a free trial period of one month. Turnitin.com has the highest rate of detection amongst subscription detection tools. Papers can be submitted individually by either student or lecturer. All papers are archived for future checking – a feature which is particularly useful if copying of previous students' papers is suspected. Reports are provided within 24-48 hours and show similarities of the submitted text to other sources.

B. PLAGISERVE

Plagiserve.com[2][6] is a free service which searches the internet for duplicates of submitted papers, analyses them, and provides evidence of plagiarism to the lecturer. It has an extensive database of 90,000 papers, essays and Cliff Notes study guides, and papers from all known paper mills. Reports are generated in 12 hours. The service is only available through its website, and papers must be submitted in one batch.

C. GLATT PLAGIARISM SCREENING PROGRAM (GPSP)

The Screening Program (GPSP)[2][7] evaluates a student's knowledge of their own writing by producing a test whereby every fifth word of a student's paper is eliminated and replaced with blanks which the student has to replace. Accuracy and speed in replacing the blanks is evaluated against a proprietary database, and a probability score returned immediately. Useful for detecting plagiarism where the original source cannot be located through other sources such as internet search engines and other plagiarism detection services, its limitations lie in not being able to identify the source of the suspect text and the requirement for students to sit a test.

D. COPYCATCH GOLD

CopyCatch Gold[8][9]is stand-alone desktop software which can be either installed on a single PC or on a network. It detects collusion between students by checking similarities between words and phrases within work submitted by one group of students.

E. EVE2 - ESSAY VERIFICATION ENGINE

EVE2[3] is a windows based system, installed on individual workstations. It is not easily installed on servers. Papers are submitted by cutting and pasting plain text, Microsoft Word, or Word Perfect documents into a text box. The program then searches internet resources for matching text. Reports are provided within a few minutes, highlighting suspect text, and indicating the percentage of the paper that is plagiarized.

F. GPLAG

GPlag[3]was developed by Chao LIU, Chen Chen, Jiawei Han at the University of Illinois-UC, Urban in 2006. GPlag, which detects plagiarism by mining program dependence, graphs (PDGs). A PDG is a graphic representation of the data and control dependencies within a procedure. The PDG thus developed from original program and modified program are checked whether it is copied or not by

graph isomorphism. In order to make GPlag scalable to large programs, a statistical lossy filter is proposed to prune the plagiarism search space.

G. JPLAG

JPlag[10][11]was developed by Guido Malpohl at the University of Karlsruhe. In 1996 it started out as a student research project and a few months later it evolved into a first online system. In 2005 JPlag was turned into a web service by EmericKwemou and Moritz Kroll. Jplag converts programs into token strings that represent the structure of the program, and can therefore be considered as using a structure-based approach. For comparing two token strings JPlag uses the Greedy String Tiling" algorithm as proposed by Michael Wise but with different optimizations for better efficiency. JPlag is a system that finds similarities among multiple sets of source code files. JPlag currently supports Java, C#, C, C++, Scheme and natural language text. Jplag has a powerful graphical interface for presenting its results. It takes input as set of programs, compares these programs pair wise (computing for each pair a total similarity value and a set of similarity regions), and provides as output a set of HTML pages that allow for exploring and understanding the similarities found in detail.

H. MOSS

Moss[12] is an acronym for Measure Of Software Similarity. Moss was developed in 1994 at Stanford University by Aiken et al. It is being provided as a web service that can be accessed using a script. The MOSS submission script works for Unix/Linux platforms and may work under Windows with Cygwin. To measure similarity between documents, moss compares the standardized versions of the documents: moss uses a document fingerprinting algorithm called winnowing. Document fingerprinting [13]is a technique that divides a document into contiguous substrings, called k-grams, with k being picked by the user. Every k-gram is hashed, and a subset of all the k-gram hashes is selected as the document's fingerprint. Moss is an automatic system for determining the similarity of programs. Moss can currently analyse code written in the following languages: C, C++, Java, C#, Python, Visual Basic, JavaScript, FORTRAN, ML, Haskell, Lisp, Scheme, Pascal, Modula2, Ada, Perl, TCL, Matlab, VHDL, Verilog, Spice, MIPS assembly, a8086 assembly, a8086 assembly, MIPS assembly, HCL2.Moss is also being provided as an Internet service.

I. SIM

SIM[14] is a software similarity tester for programs written in C, Java, Pascal, Modula-2, Lisp, Miranda, and for natural language. It was developed in 1989 by Dick Grune at the VU University Amsterdam. The process SIM uses to detect similarities is to tokenize the source code first, then to build a forward reference table that can be used to detect the best matches between newly submitted files, and the text they need to be compared to. SIM detects similarities between programs by evaluating their correctness, style, and uniqueness.

J. GOOGLE

Google [15] is an American multinational corporation specializing in Internet-related services and products. These include online advertising technologies, search, cloud computing, and software. It can be used as an alternate method to detect suspected

IV. RESULTS AND DISCUSSIONS

Comparisons of all the above listed plagiarism detection tools are done based on their features. Feature comparisons are qualitative comparisons; they describe the properties of a tool, like whether it is a local or a web-based application, whether it available freely, response time etc. The criteria that is used for qualitative comparison are:-

1. Local or web based:- Some tools are provided as web services. This requires a lecturer to send the student assignments over the network. Here you take a risk of exposing confidential information to the outside world. Other tools have to be downloaded and run locally.
2. Open Source :- An advantage of open source is of course the possibility of extending or improving the program to better suit the situation you intend to use it for
3. Database checking: -Database checking can be local or global. A large number of resources can be compared if the tool check the databases all over the internet apart from local database checking
4. Internet checking: - Refers to searching of live and cached links to websites and database to have extensive internet checking to all submitted documents.
5. Publicationchecking: -Extensive internet checking to all submitted documents checking most types of submitted publication like documents, including, books, articles, magazines, journals, newspapers, PDFs etc. online only.
6. Free Text :- The Tool checks only documents
7. Source Code :- The tool operates on the structured text i.e. source code
8. Designed for students :- The tool is designed for the students to check their documents

9. Designed for Teachers :- The tool is designed for teachers to check the plagiarism in submitted assignments
10. Instant response:- The results are available instantly or takes time to publish the results.

A table (see Table 1) was produced to report the criteria defined above.

TABLE 1 : - COMPARISON OF THE PLAGIARISM DETECTION TOOLS

Feature	Turnitin	Plagiserve	GPSP	Copy catch Gold	EVE2	Gplag	Jplag	MOSS	SIM	Google
Local or web based	Web	Web	Local	Local	Local	Web	Web	Web	Local	Web
Open Source	No	Yes	No	No	No	Yes	No	No	Yes	Yes
Database checking	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	No
Internet Checking	Yes	Yes	No	No	Yes	No	No	No	No	Yes
Publication Checking	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes	No
Free Text	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
Source Code	No	No	No	No	No	Yes	Yes	Yes	Yes	No
Designed for students	Yes	No	No	No	No	No	No	No	No	Yes
Designed for teachers	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Instant Response	No	No	No	No	Yes	No	No	No	Yes	Yes

Operation of plagiarism detection tools is based on statistical or semantical methods or both to get better results. In[2] it is stressed that "although plagiarism detection tools provide excellent service in detecting matching text between documents, care needs to be taken in their use". Plagiarism detection tools inability to distinguish correctly cited text from plagiarised text is one of the serious drawbacks of these tools[10][2]. That is why human interposition is necessary before a paper is declare plagiarised – manual checking and human judgment are still needed [8][10]. Table 2 discusses the pros and cons of each tool

TABLE 2 : PROS AND CONS OF PLAGIARISM DETECTION TOOLS

Plagiarism Detection Tool	Pros	Cons
Turnitin	Covers a huge range of sources. Offers a digital portfolio service. in which students' work is archived.	The user has to check the report carefully because the software detects correctly-cited material as well as plagiarised material. As in similar programs, formatting is lost in the checking procedure, so essays for marking have to be submitted separately from essays for checking.
Plagiserve	Tests against extensive in-house database and internet searches	Formatting is lost during the checking process, so material has to be handed in separately. Has been claimed to be associated with cheat sites.
GPSP	Useful for detecting plagiarism where the original source material cannot be located.	Students actually have to sit down to a test to fulfil the requirements.
CopyCatch Gold	The JISC (Joint Information Systems Committee - HEFCE-funded UK organization) gave this software five stars for detection, clarity, value, user-friendliness, speed and reliability.	Detects only collusion among students, and cannot detect material downloaded from the Web
EVE2	Tests against wide area of internet.	Each piece of work has to be individually loaded and checked by the lecturer.
GPlag	GPLAG is both effective and efficient: It detects plagiarism that easily slips over existing tools, and it usually takes a few seconds to find (simulated) plagiarism in programs having thousands of lines of code.	Limited and less effective in its use with ordinary text.
JPlag	The only software that can deal with programming-type work as well as ordinary text.	Limited and less effective in its use with ordinary text.
MOSS	Designed with a special focus on computer programming code rather than text.	Limited in scope.
SIM	Designed with a special focus on computer programming code rather than text.	Limited in scope.
Google	Quick and free. Google extracts from pdf files, which many search engines cannot do.	Unsystematic, and involves manual entry of strings. Labour-intensive.

V. CONCLUSION

In this paper we have compared ten plagiarism detection tools with respect to ten tool features. Pros and cons of each tool are also discussed. Performance was compared by a sensitivity analysis on a collection of intentionally plagiarized programs and on a set of real life submissions. The results of the comparison give good insight into the strong and weak points of the different tools. In the age of information technologies plagiarism has become more actual and turned into a serious problem.

Education institutions need to focus on plagiarism detection methods. Analysis of the known plagiarism detection tools shows that although these tools provide excellent service in detecting matching text between documents, even advanced plagiarism detection software can't detect plagiarism as good as human does. They have several drawbacks and, so manual checking and human judgment is still needed. Human brain is universal plagiarism detection tool, which is able to analyse document using statistical and semantical methods, is able to operate with textual and non-textual information. At the present such abilities are not available for plagiarism detection software tools. In accordance with[2] "…at least for now – nothing can completely replace the watchful eye of human beings". But nevertheless computer – based plagiarism detection tools can considerably help to find plagiarized documents.

REFERENCES

[1] Asim M., Hussan M., Vaclav S., "Overview and comparison of Plagiarism detection tools," *Dateso*, pp. 161-172, 2011.

[2] R. Symons, "Teaching and Learning Committee Plagiarism Detection software Report," University of Sydney, 2003.

[3] L. Chao, "GPLAG : detection of software plagiariasm by program dependence graph analysis," in *The 12th ACM SIGKDD international conference on knowledge discovery and data mining*, Philadelphia, PA, USA,pp 872-881 2006.

[4] C. J. Neill, G. Shanmuganthan, "A Web-enabled plagiarism detection tool," *IEEE IT Professional*, Vol. 6 No. 5, pp 19-23, 2004.

[5] Jasper P, "Turnitin.com," *Library Journal 126(8)*, p. 138, May, 2001.

[6] [Online]. Available: http://www.plagiserve.com.

[7] [Online]. Available: http://www.plagiarism.com.

[8] [Online]. Available: http://www.copycatch.freeserve.co.uk.

[9] University of Birmingham, "Report on Plagiarism workshops: show and tell - tackling the causes of collusion and Plagiarism," University of Birmingham, Birmingham, UK, 2002.

[10] Delvin M., "Plagiarism Detection Software : How effective is it?," *Assessing Learning in Australian Universities, Elvines*, 2002.

[11] [Online]. Available: http://wwwipd.ira.uka.de/jplag/..

[12] [Online]. Available: http://theory.stanford.edu/~aiken/moss/..

[13] S. Burrows., S. M. M. Tahaghoghi, and J. Zobel, "Efficient Palgiarsim detection for large code repositiories," *Softw. Pract. Expert*, pp. 37(2): 151-175, 2007.

[14] D. Gitchell and N. Tran, "Sim : A utility for detecting similarity in computer programs," in *Proceedings of Tech. Symp. Comput. Sci. Ed.*, pp 266-270, 1999.

[15] [Online]. Available: http://en.wikipedia.org/wiki/Google.

