International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
(Vol. 23, Issue 01) and (Publishing Month: August 2015)
(An Indexed, Referred and Impact Factor Journal)
ISSN (Online): 2319-6564
www.ijesonline.com

A Review of Figure Plagiarism Detection Techniques

Mohammed Altyab Mohammed Ali[1] and Professor Dr. Izzeldin Mohamed Osman Elamin[2]

[1]Computer Science, University of The Holy Qura'n And Taseel Of Sciences
Rufaa, Sudan
moh_altayb@yahoo.com

[2]Sudan University of Science and Technology
Khartoum, Sudan
izzeldinosman@hotmail.com

Publishing Date: 5th August 2015

Abstract

Plagiarism detection and prevention became one of the research and educational challenges. Many students and researchers tend to copy other works and idea when doing researches, projects, and assigned tasks. Most of the papers that are meant for Plagiarism detection are focusing on the Plagiarism from text point of view. So far there is no significant work in figure Plagiarism detection. This paper is a review paper that point out the main methods and techniques that are used for plagiarism detection.

Keywords: *Plagiarism, plagiarism detection, plagiarism prevention, figure plagiarism.*

1. Introduction

Currently there is an increasing in the amount of materials available in the electronic form and the ease of accessing to the internet has increased plagiarism. Plagiarism is an unethical act and must be eradicated from the researcher's studies and mind. The result of it is on the students and can also stain the good reputation of an institution. The most interesting definitions are given by the IEEE (2008): "plagiarism is the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source". Another definition of the plagiarism is: "Plagiarism, the 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" [2].directly. The document you are reading is written in the format that should be used in your paper.

Studies have shown that plagiarism must be dealt with in serious way, because the understanding of the concept of plagiarism using the ICT is still unsatisfactory. An ongoing effort must be undertaken to arise the understanding between the students on plagiarism to avoid doing this act in the future. Usually the plagiarism detection is based on comparison of two or more documents. In order to compare these two or more documents and to reason about degree of similarity between them, it is needed to assign numeric value, so called, similarity score to each document [1]. This score can be based on different metrics. There are many parameters and aspects in the document which can be used as metrics.

Manual detection of plagiarism is very complicated as well as time consuming due to the vast amount of contents available, therefore many researches are conducted to invent automated tools to deal with plagiarism. Most of existing works are focusing only on text plagiarism; therefore, works that focus on figure plagiarism detection is needed due to the increase of using graphics and images as document content especially in academic areas. This focus of this research is to invent a novel methodology and techniques for figures and image plagiarism detection**.**

2. Classification of Plagiarism

Plagiarism can be classified in different ways:[3]

The first possible way is by the type of person who is committing the plagiarism.

Student plagiarism is where a student is submitting plagiarized work for academic credit. *Academic plagiarism* is where an academic uses plagiarized materials for professional development, primarily by submitting plagiarized papers to conferences or journals. *Professional plagiarism*, such as copying a report from a competitor, refers to plagiarism within the workplace. Academic plagiarism could be considered as one example of professional plagiarism.

Plagiarism can also be classified according to the source of the plagiarism. These definitions are concerned especially with student plagiarism but could be applicable to elsewhere. Traditionally a student will be asked to complete a piece of work to a given *assignment specification*. The piece of work they hand in is then known as a *student submission*. The set of all student submissions for a given assignment specification is then known as a *corpus* (a corpus is a standard term used for a set of documents in linguistics). *Corpora* (the plural of corpus) can also be made in different ways so long as all the documents inside are linked in some way. A number of corpora could be produced from a single assignment specification if the same specification was presented in multiple years. Then a corpus could be produced for every new batch of students studying the material, or a single corpus containing every submission could be produced. Another possible linkage might be a corpus containing all the work produced by a given individual student. Submissions within such a corpus could be expected to have similar linguistic properties.

Plagiarism within a corpus is known as *intra-corpal plagiarism*. For a corpus containing the work of a single group of students this would represent the case where one student was copying from another. In such a case the *source submission* and the *copy submission* would not be immediately identifiable, although there may be clues.

When the plagiarism source is outside the corpus, such as in a journal, or in a submission from another institution, this is known as *extra-corpal plagiarism*. One type of extra-carpal plagiarism worth classifying further is that of *Web plagiarism* where some or all of a submission is sourced from the World Wide Web, a problem that has sprung up over the last few years. Many cases of Web plagiarism are *multiply sourced*, where material has been copied from more than one place. Intra-carpal plagiarism is more likely *singularly sourced*, committed with a one-to-one correspondence. A set of submissions containing similar material within a corpus is known as a *cluster*.

It is important to differentiate between plagiarism and *similarity* in this context.

Similarity refers to two documents, or part of two documents, containing material that has been judged alike in some way. This similarity can only be referred to as plagiarism once it has been examined and verified in some way by a tutor. Otherwise there might be legitimate reasons for the similarity, such as the two students using the same correctly cited materials.

Major headings are to be column centered in a bold font without underline. They need be numbered. "2. Headings and Footnotes" at the top of this paragraph is a major heading.

3. Forms of plagiarism

Plagiarism can take several distinct forms, including the following [4]:

(1) **Word-For-Word Plagiarism:** direct copying of phrases or passages from a published text without quotation or acknowledgement.

(2) **Paraphrasing Plagiarism:** when words or syntax are changed (rewritten), but the source text can still be recognized.

(3) **Plagiarism of Secondary Sources*:*** when original sources are referenced or quoted, but obtained from a secondary source text without looking up the original.

(4) **Plagiarism of The Form of A Source*:*** the structure of an argument in a source is copied (verbatim or rewritten).

(5) **Plagiarism of Ideas*:*** the reuse of an original thought from a source text without dependence on the words or form of the source.

(6) ***Plagiarism of Authorship:*** the direct case of putting your own name to someone else's work.

The easiest form of plagiarism to detect and prove is verbatim or word-for-word text reuse (given a possible source text to compare with). This can often be detected using the simplest of automatic methods, but occurrences by students

are often due to the fact that they are uncertain as to how to reuse source texts legitimately.
Other forms, such as paraphrasing and the reuse of structure can also be identified relatively easily, but get progressively harder as the plagiarist uses more complex rewrites or to hide the original text, or reuses only ideas and not the content. The extreme is ghost-writing: getting someone else to write the text for you. These forms of plagiarism are not just harder to detect, but also harder to prove [5].
Another form of plagiarism is a figure or graph plagiarism where plagiarist uses someone else figure or graph without making a citation to the original source.

4. Plagiarism Detection Systems

Most existing plagiarism detection tools are specially designed to process natural language text or program source code. Systems designed for finding similarities in natural language texts mainly search the Internet for the possible matches. Text comparisons use simple comparison methods aiming mostly at processing speed and wide coverage. The program source code usually performs a pair wise comparison between single submissions only. Though sophisticated procedures are being developed which compares with multiple source code programs simultaneously [6].

4.1 Text Based Detection systems

Most of the relevant research focus on finding plagiarism in free text. As such the research and techniques that can be reported are limited and the methods used differ widely, with their being little evidence available about how well they work and which are best. The literature on free text detection is substantially limited compared to that on source code detection [3].
It is worth noting that there are many sites and articles on the Internet that cover plagiarism. Many institutions around the world and the departments within them have one. The content on the sites is fairly standard. Instructions for students about what plagiarism is and how to cite properly may be provided. Details of electronic tools, mainly the Web-based plagiarism detection services are given. Links to other sites and articles of interest are common. They might also contain advice for tutors on how to assess students whilst reducing the chance of cheating. These sites have been deliberately excluded for the most part since they are both of no technical interest and cover no original ground. In some cases it could be argued that the shared similarity on them is rather a suspect [3].
Shivakumar & Garcia-Molina [7] describes the SCAM engine that registers documents for copy protection purposes, the same authors also describe a clustering method that allows a database of Web documents to be produced, that may subsequently be checked for plagiarism [8].
Ribler & Abrams [9] presents two visualizations that show commonality in documents. Primarily used with source code but may be applicable to free text.
Hoad & Zobel [10] investigates how documents can be identified that they are derived from another.

4.2 Source Code based plagiarism

Source-code plagiarism detection in programming assignments is a task many higher education academics carry out. Source-code plagiarism in programming assignments occurs when students reuse source-code authored by someone else, either intentionally or unintentionally, and fail to adequately acknowledge the fact that the particular source-code is not their own [11]. Once similarity between students work is detected, the academic proceeds with the task of investigating this similarity. The investigation process involves comparing the detected source-code files for plagiarism by examining their similar source-code fragments.
Many different plagiarism detection tools exist and these can be categorized depending on their algorithms. Mozgovoy [12] identified two categories; *fingerprint based systems*, and *content comparison techniques*. Various other classifications exist in the literature [13, 14, and 15].

4.3 Fingerprint based systems

Tools based on the fingerprint approach create *fingerprints* for each file, which contain statistical information about the file, such as average number of terms per line, number of unique terms, and number of keywords. Files are considered similar if their fingerprints are close

to each other. This closeness is usually determined by measuring the distance (using a distance function in a mathematical sense) between them [12].

In fingerprint based system the first known plagiarism detection system was an attribute counting program developed by Ottenstein for detecting identical and nearly-identical student work [16, 17].

Robinson and Soffa developed a plagiarism detection program that combined new metrics with Halsteads metrics in order to improve plagiarism detection [18].

Rambally and Sage [19] created an attribute counting system, which accepts student's programs, parses them and then creates a *knowledge system* which contains knowledge vectors, where each vector holds information about the attributes in a student's program.

More recent plagiarism detection tools such as MOSS (Measure of Software Similarity) [20] combine the fingerprinting approach with the structure metric approach.

4.4 Content comparison techniques

Content Comparison techniques are often referred to as *structure-metric* systems in the literature.

These systems convert programs into tokens and then search for matching contiguous sequence of substrings within the programs. Similarity between programs depends on the percentage of the text matched. Mozgovoy [12] classified content comparison techniques into string matching-based algorithms, parameterized matching algorithms and parse trees comparison algorithms.

5. Conclusion

Plagiarism is awfully more successful in the academic world than other types because academicians may not have sufficient time to track their own ideas, and publishers may not be well-equipped to check where the contributions and results come from the authors of the material to be published. From the literature it's obvious that current antiplagiarism tools for educational institutions, academicians, and publishers mainly concentrate on text. On the other hand some idea and materials can be expressed in figures, thus copying figures is a serious plagiarism issue. So developing tools for figure plagiarism is very essentials. In our future research, we will consider the development of novel tools and techniques for figure plagiarism.

References

[1] Lukashenko, Romans, Vita Graudina, and Janis Grundspenkis. "Computer-based plagiarism detection methods and tools: an overview." Proceedings of the 2007 international conference on Computer systems and technologies. Vol. 285. ACM, 2007.

[2] Encyclopedia Britannica, http://www.britannica.com/EBchecked/topic/462640/plagiarism 3428.

[3] Lancaster, Thomas. Effective and efficient plagiarism detection. Diss. South Bank University, 2003.

[4] Martin, Brian. "Plagiarism: a misplaced emphasis." Journal of Information Ethics 3.2 (1994): 36-47.

[5] Old and new challenges in automatic plagiarism detection Paul Clough, Department of Information Studies. February 2003.

[6] Jamal, Sangeetha. "Plagiarism Detection Techniques" (2010).

[7] N. Shivakumar & H.Garcia-Molina. SCAM: A Copy Detection Mechanism for Digital Documents, Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries, Austin, Texas, June 1995.

[8] N. Shivakumar & H.Garcia-Molina. Finding Near Replicas of Documents on the Web, Proceedings of Workshop on Web Databases, in conjunction with EDBT '98, March 1998.

[9] R. L. Ribler & M. Abrams. Using Visualization to Detect Plagiarism in Computer Science Classes, Information Visualization 2000, pp173-177.

[10] T. Hoad & J. Zobel .Methods for Identifying Versioned and Plagiarized Documents, to appear in Journal of the American Society for Information Science and Technology.

[11] G. Cosma and M. Joy. Towards a definition of source-code plagiarism. IEEE

Transactions On Education, 51:195–200, 2008.

[12] M. Mozgovoy. Desktop tools for offline plagiarism detection in computer programs. Informatics in Education, 5(1):97–112, 2006.

[13] T. Lancaster and F. Culwin. Classifications of plagiarism detection engines. Italics, 4(2), 2005.

[14] K. Verco and M. Wise. Plagiarism `a la mode: A comparison of automated systems for detecting suspected plagiarism. The Computer Journal, 39(9):741–750, 1996.

[15] K. Verco and M. Wise. Software for detecting suspected plagiarism: Comparing structure and attribute-counting systems. In John Rosenberg, editor, Proceedings 287 of the First Australian Conference on Computer Science Education, pages 81–88, Sydney, Australia, July 3–5 1996. SIGCSE, ACM.

[16] K. Ottenstein. A program to count operators and operands for ansi fortran modules.Computer Sciences Report TR 196, Purdue University, 1976.

[17] K. Ottenstein. An algorithmic approach to the detection and prevention of plagiarism. ACM SIGCSE Bulletin, 8(4):30–41, 1976.

[18] S. Robinson and M. Soffa. An instructional aid for student programs. In SIGCSE ’80: Proceedings of the Eleventh SIGCSE Technical Symposium on Computer Science Education, pages 118–129, New York, NY, USA, 1980. ACM.

[19] G. Rambally and M. Sage. An inductive inference approach to plagiarism detection in computer program. In In Proceedings of the National Educational Computing Conference, pages 22–29, Nashville, Tennesse, 1990.

[20] A. Aiken. Moss: A system for detecting software plagiarism. Software: www.cs.berkeley.edu/ aiken/moss.html, accessed: July 2008.

[21] K. R. Rao, “Plagiarism, a scourge,” Current Sci., vol. 94, pp. 581–586, 2008.

[22] L. Stenflo, “Intelligent plagiarists are the most dangerous,” Nature, vol. 427, p. 777, 2004.