On the effectiveness of anti-plagiarism software

Franco Raimondi

Abstract—Anti-plagiarism tools are currently used in a large number of institutions to perform an initial assessment of students’ essays, enabling an automated approach to the identification of plagiarised work. This paper presents evidence that students are moving from “copy and paste” plagiarism, which can be detected by anti-plagiarism software, to more complex forms of plagiarism using a variety of techniques. Various examples are presented to show how in a number of cases these techniques can bypass an automated evaluation of originality.

Index Terms—Plagiarism, assessment strategies.

1 INTRODUCTION

It is well known that the ready availability of on-line textual information has resulted in a dramatic increase in the cases of plagiarism [1], [2], especially when students are faced with the task of producing reports and essays. Starting with the rise in the availability of Internet connections fifteen years ago, “copy and paste” plagiarism has affected a large number of works.

Manual techniques to detect this form of plagiarism include taking particular care in sudden changes of styles in a written text, a change in the average length of phrases and paragraphs, etc. In such circumstances, lecturers can use a search engine to try to locate the on-line material that may have been copied.

The manual process of Internet search, text comparison, and false positive evaluation can be very time consuming. As a consequence, the increase in the number of students has and the increase in workload for lecturers has created a market for automated anti-plagiarism tools. It is now common to require students to submit their work, and obtain an originality report, through an on-line tool. Typically, the tool performs a comparison between the submitted work and previously indexed documents, which include other students’ essays, on-line articles and documents, etc.

The evidence presented in this paper shows that students are moving from simple “copy and paste” to more complex forms of plagiarism1. More in detail, this paper investigates various techniques that can be (and have been) successfully employed to bypass automatic plagiarism detection tools, presenting a large number of samples and excerpts from submitted student work. Additionally, this paper proposes a set of practices to identify students’ plagiarism in an effective way, and a new structure for assessing the originality of students’ research work. The results presented here are based on material and experience from the final year projects of the School of Engineering and Information Sciences at Middlesex University, but the considerations presented below are likely to apply to other institutions implementing a similar infrastructure for students’ projects.

The rest of the paper is organised as follows. Section 2 describes the techniques that have been identified in a number of students’ essays to cheat anti-plagiarism software. Section 3 presents various methods that can be used to identify the new forms of plagiarism, together with a proposal for new assessment procedures. Section 4 discusses related work.

2 TECHNIQUES

The results presented in this section have been obtained using the following tools:

• http://turnitin.com/: this is the system Middlesex University students are required to use to submit their work.

• http://www.plagiarismdetect.com/: this is a free tool available on-line. A piece of text can be checked for originality by copy-pasting it into an on-line form.

• http://www.scanmyessay.com/: this is another free tool that requires the installation of local software and works by submitting the original document to be checked.

Unfortunately, no information is available on the underlying anti-plagiarism techniques employed by these tools. The results presented below seem to suggest that plagiarism is detected by measuring the syntactic similarity of a submitted text against a database of texts either submitted by other students, or available from on-line databases, some of which may not be public. The tools employ various optimization techniques to reduce the risk of false negatives, such as removal of accents, reordering of words, etc.

Other tools exist, possibly requiring payment for use, but the considerations presented below are likely to apply to all the tools that compute some form of syntactic similarity between texts to assess plagiarism.

F. Raimondi is with the School of Engineering and Information Sciences, Middlesex University, London, UK email: f.raimondi@mdx.ac.uk

1. Research into the motivations that result in students’ plagiarism is beyond the scope of this paper; plagiarised portions of text may appear intentionally, or simply due to a misunderstanding of university regulations. Every course requiring an essay should necessarily include a module on good (and bad) referencing practices and on plagiarism.
2.1 Text substitution and macros

When submitting an essay, students may be required to use an on-line mechanism and to obtain a successful report from the software to rule out plagiarism. The simplest technique to obtain such a report is to modify the plagiarised text, for instance by replacing all “e” characters in a text with another completely different character (as mentioned at the beginning of this section, most software tools remove accents and other decorations). While this technique may be used to obtain a report, it could be easily detected by checking the submitted file (which is available to lecturers).

A more sophisticated approach involves the use of Word macros to automatically replace the additional character every time the document is opened. Using this technique, a student first substitutes all “e” characters with > (using, for instance, the “Find and Replace” command), and then adds a macro to translate > back to “e” every time the document is opened. Anti-plagiarism tools do not apply macros, but markers typically leave them enabled. Figure 1 reports an excerpt from [3] on the top, and the same text as seen by the anti-plagiarism software when all the “e” characters have been replaced by a > symbol. Notice that the process is reversible when the replacing symbol does not appear in the original document. When submitted to the anti-plagiarism tools, the first excerpt returns a 100% match with [3], while the second is accepted as fully original.

2.2 Non-indexed documents

As mentioned above, anti-plagiarism software tools compare text files that have been previously indexed. Thus, another possibility to submit plagiarised material is to search for documents that are not currently indexed, for one of the following reasons:

1) The original text is hosted on a non-indexed repository. As an example, consider the text in Figure 2 extracted from [4]. All the three tools consider this text “original”, as this is not indexed by their databases. Nevertheless, the text could be found by using an Internet search engine. It seems some students are aware of this method: Figure 3 reports part of an email sent by a student to a former student requesting the removal of his work from a web server (forwarded by the recipient).

2) Certain documents cannot be indexed as text due to their format. It seems that postscript files are not indexed by the three tools employed, probably because text cannot be easily extracted from them (unless OCR techniques are employed). Additionally, it seems that password-protected PDF documents are not indexed. Figure 4 depicts a page from an on-line PDF document which has been submitted verbatim by a student in an essay without being detected by the anti-plagiarism software.

3) Images (and text appearing in images) are not indexed. Therefore, if a piece of text is plagiarised from an image, this is not detected by any of the tools. An example of an image which has been reported verbatim in a student’s essay is depicted in Figure 5. In this case the plagiarism has been detected only because other parts of text had been copied verbatim. Other forms of plagiarism that cannot be detected include the use of diagrams and graphs in the form of Jpeg or PNG images. There have been various cases of students submitting images and graphs of experimental results copied from research papers whose text portions are indexed by the anti-plagiarism tools.

2.3 Use of an automatic text modifier

Another possibility to hide plagiarised material in an essay is to employ synonyms and paraphrases. However, the task of modifying a whole document may be too time consuming to be considered effective by students trying to obtain a successful report. Evidence of plagiarism from submitted essays seem to suggest that there may be tools which can automate the paraphrasing process, by modifying and reordering of words and phrases. As an example, consider the pieces of text in Figure 6. The paragraphs on the left-hand side are extracted from a master thesis [5]. The paragraphs on the right-hand side have been submitted in 2010: notice, in particular, the last word: “environments” has been paraphrased with “working atmospheres”, which seems to suggest that there is some form of automation behind this process (the typos may be deliberate, and some of them may be used as indicators of plagiarism, see below).

The anti-plagiarism tools do not detect the text on the right-hand side of the figure as plagiarised (this case was only discovered because the student used the heading “1.1.1 Technical Overview” in chapter 3 of the thesis). Interestingly, if the text on the left-hand side of the figure is submitted to the anti-plagiarism tools, a plagiarism warning is returned. However, the tools do not refer to [5] as the original source (which, apparently, is not indexed), but to a thesis submitted to another institution by another student in 2008!

As mentioned above, there are good reasons to suspect the existence of tools that perform this kind of paraphrasing in an automatic way. However, an Internet search of public available tools has been unsuccessful, but this does not rule out the possibility that they may be available through non-public channels.

2.4 Use of an automatic translator

Paraphrasing and obfuscation of text can also be obtained by employing an automatic translator to perform a cyclic sequence of translation, using for instance the application available at http://translate.google.com. Figure 7 reports the text generated by using the text reported on the top of Figure 1 after a translation to Afrikaans (the first language available) and then back to English.
The ready availability of Internet resources and the case with which they can be searched and information downloaded has contributed to the rise of plagiarism by students (Born, 2003; Hansen 2003). Internet search engines facilitate easy access to many sites containing relevant material for academic essay writing and, in many cases, these sites have been provided by other academic institutions for the benefit of their own students. Plagiarism can take the form of copying an entire essay, or in other cases, transpires when significant portions of the essay have been copied or paraphrased without reference or quotation. The latter represents a form of cut and paste essay writing and occurs when students combine sentences from a number of different sources to complete an essay (Austin & Brown, 1999). Cut and paste plagiarism can be much more difficult to detect than the case of the student submitting another’s essay as his or her own. Cut and paste essay writing is a concern because it is an evasion of learning but one that can deliver satisfactory results for the student. This outcome is unsatisfactory for an educational institution since it enables students to gain credit for significant portions of assessment [...]
hi there. can you please remove this file from your server if you don’t mind for 3 months? please :$? http://[...] /Project.pdf If you can give me the word file, it will be great :$! please it will be a great help :$!

Fig. 3. Email sent by a student requesting the removal of a document

Fig. 4. A non-indexed encrypted PDF document

occurrence of false positives (as described below).

2.6 Submitting essays written by someone else using eBay

In addition to essay writing services which charge a fixed amount of money for a bespoke service, other sources of essays are available. For instance, former students at the same institution may sell their essays when anti-plagiarism software was not used for the original submissions.

A further source of essays is eBay. There is a large number of auctions where former students sell their works for a fee. The screenshot of an auction is reported in Figure 8. The feedback received by this particular seller (see Figure 9) gives an idea of the number of transactions.

In many cases, these auctions explicitly mention one or more anti-plagiarism tools and guarantees that the essay would pass the plagiarism verification.

2.7 False positives

Anti-plagiarism software tools may falsely report legitimate portions of text as being plagiarised. This is caused by the fact that these tools do not consider quotation marks and text formatting (and this choice is appropriate, otherwise students may include all their essays in quotes or present it in italic to avoid checks). Nevertheless, it is perfectly plausible for students to quote portions of text verbatim (as in many figures appearing in this paper), and to use commonly occurring sequences of words.

The problem of false positives is particularly relevant for short technical documents, which use short phrases with sequence of words that may correspond to common acronyms etc. In this case the lecturer has to mark all these as false positives as “acceptable” before an originality report is generated. However, it becomes difficult to distinguish between “real” false positives
1.1.1 Technical overview

The WiMAX standard defines the air interface for the IEEE 802.16-2004 specification working in the frequency band 2-11 GHz. This air interface includes the definition of the medium access control (MAC) and the physical (PHY) layers. **Medium Access Control (MAC) layer**

Some functions are associated with providing service to subscribers. They include transmitting data in frames and controlling the access to the shared wireless medium. The medium access control (MAC) layer, which is situated above the physical layer, groups the mentioned functions. The original MAC is enhanced to accommodate multiple physical layer specifications and services, addressing the needs for different environments.

---

1.1.1 Technical overview

This considered WiMAX specification has the interface to air required IEEE 802.16-2004 specification for the functions for the considered frequency band 2-11 GHz. This interface for the mobile has the physical (PHY) layers as well as the mobile medium access control (MAC) properly defined.

**Medium Access Control (MAC) layer**

Few functionalities are concerned with the provision of the service to subscribers. These have the data transmission in frames as well as the control of the access to the valuable shared wireless medium. The medium access control (MAC) layer is always residing above the physical one, takes together the functions mentioned. Actual MAC is updated to take numerous physical layer data and works, making the needs for different work atmosphere.
The ready availability of Internet resources and the ease with which they can be searched and information downloaded has contributed to the rise of plagiarism by students (Born, 2003; Hansen 2003). Internet search engines facilitate easy access to many sites containing relevant material for academic essay writing and, in many cases, these sites have been provided by other academic institutions for the benefit of their own students. Plagiarism can take the form of copying an entire essay, or in other cases, transpires when significant portions of the essay have been copied or paraphrased without reference or quotation.

Fig. 7. Example of paraphrasing using an automatic translator.

Fig. 8. Example of eBay auction

Fig. 9. Seller's feedbacks from eBay
and automatic paraphrasing. Overall, this process may take more time than using the standard technique of searching the web for “suspicious” portions of text.

3 Guidelines and Recommendations

Current academic guidelines in most institutions allocate a certain workload to each academic supervision. Typically, the final essay constitutes the largest component of the final assessment for the project, and students’ essays have an average length of approximately 50 pages. Taking into account the initial meeting for the definition of the project topic, intermediate meetings, and the assessment of intermediate milestones, lecturers may be left with a limited amount of time to evaluate the quality and the originality of a large piece of text when this is left at the end of a project. This is probably the reason behind the widespread adoption of automatic anti-plagiarism tools. Anti-plagiarism tools may be useful in certain circumstances, but the techniques described above highlight their weaknesses and overall these tools may not be as efficient as expected. In particular, even in the presence of a positive report from one of these tools, lecturers should look for one or more of the following signs of plagiarism:

- Sudden changes in style are usually associated with plagiarism, and checking these changes against an on-line search engine is a often a successful approach to detect “copy-and-paste” kinds of plagiarism. However, the techniques described above show that even un-grammatical phrases may be plagiarised. A discrepancy between the styles of different works by the same student, or a discrepancy between oral and written language skills, should be a warning of possible plagiarism.

- In addition to malformed phrases, missing characters and unexpected symbols may be a symptom of plagiarised material. As an example, consider the text on the right-hand side of Figure 6: the missing characters in the two words “speciication” (specification) and “di_erent” (different) are the result of copying text from a PDF file. Indeed, in many PDF files some sequences of characters (such as “fi” and “ff”) are rendered as a single non-standard character. In turn, these non-standard characters do not have correspondence in standard editing fonts, thus leaving a white space. As a general rule (and to avoid the problem of Word macros), it is good practice to always check the source file of the submitted essay as seen by the tool itself.

- As mentioned above, anti-plagiarism software do not check images and graphs. It is thus appropriate to check consistency of quality and style of graphs (such as experimental results), diagrams (block diagrams, flow diagrams, etc.), and images. In particular, are these images and graphs produced with the same tool? Do they use the same font and colours? Additionally, care must be taken in checking that (mathematical) formulas and equations are typed and not copied as images from another source.

3.1 A proposal: from essays to portfolios

In many institutions, students in their last year and master students are required to work on a project and submit a “thesis” describing their results. The increase in the number of students and the emergence of new forms of plagiarism make the assessment of the originality of these projects an extremely time consuming task. Additionally, due to the increasing number of students, final-year students are often faced with the task of writing a large piece of text with little one-to-one support from staff, and often in parallel with other exams.

It is thus necessary to re-think the structure of students’ projects and what universities want to assess: is it the student’s ability to write a large piece of text? Or is it the ability to produce original results? The allocation of the time to students should be revised accordingly and, if the latter, the assessment procedures can be modified to make the process more efficient.

Given the evidence presented above, it seems necessary to introduce oral assessments to evaluate the originality of a large piece of text. More in detail, instead of allocating a large block of hours to the assessment of a single thesis, it could be more appropriate to introduce evaluations points during the whole life of the projects consisting of both written and oral work. These could be collected in a portfolio and discussed in a final viva voce. In addition to the structure mentioned above, lecturers should try to supervise projects as close as possible to their area of expertise, thus making easier the process of originality evaluation.

A structure similar to the one presented above is currently being used to assess students’ project in the School of Engineering and Information Sciences at Middlesex University, even if the final report still carries a 60% weight for the final evaluation and only one oral examination is scheduled. Nevertheless, the viva voce and the intermediate milestones made it possible to assess the originality of the work with greater accuracy and in a more efficient way than in past experiences where the viva voce was not present and the final report carried 80% of the weight.

4 Related work

Various other papers have investigated the effectiveness of anti-plagiarism software [6], [3]. In particular, [7] investigates two anti-plagiarism software and concludes that “close review of material suspected of plagiarism is still essential for proper identification.”

Another work [8] claims that text matching is not a good metric to assess originality. The focus of this paper is not on the assessment of existing tools, but instead it is a reflection on new forms of collaboration using new media; in essence, this work argues that a review
of plagiarism policies and pedagogy is needed to take into account new forms of work.

The papers mentioned above have investigated the issue of plagiarism in written students’ essays; there are other areas where plagiarism occurs. In particular, plagiarism of mathematical exercises and software code. These kind of work require a different approach; an automatic detection mechanism is investigated, for instance, in [9].

5 CONCLUSIONS

The introduction of automatic anti-plagiarism detection mechanisms has resulted in an evolution of the forms of plagiarisms occurring in students’ reports. This paper has described a set of techniques that can be employed to cheat anti-plagiarism software, and it has presented a number of real instances of plagiarism extracted from students’ submissions. Additionally, this paper has proposed various detection techniques to identify plagiarised portions of text, and it has proposed a different assessment procedure for student research projects.

A number of works deal with the problem of identifying semantic similarities in different pieces of texts [10], [11]. Work in this area is still in its infancy and to the best of the author’s knowledge tools supporting semantic comparison are not currently used for the submission of students essays. Semantic analysis could improve the detection capabilities of anti-plagiarism software, even if some of the techniques mentioned above could affect these tools as well.

REFERENCES


