How to avoid plagiarism?

Tiina Niklander

Department of Computer Science, University of Helsinki

P.O.Box 68, FI-00014 University of Helsinki, Finland

E-mail: {tiina.niklander@cs.helsinki.fi

Abstract

Plagiarism and its avoidance have gained for attention in educating students. A lot of institutes define it using examples. General agreement is that plain copying of existing material is definitely plagiarism. The current information flow in the Internet and the use of computers in the writing process makes it too easy for students to avoid the actual learning-by-writing and just to perform cut-and-paste for their essay or report. At the same time the information flow makes it almost impossible for teachers to detect all of these flaws. Automatic detection tools have been developed to help teachers in this process. However, the detection is just one part of the story. It is much more important that the authors (students as well as researchers) do not copy from others, but say it using their own voice and words. We should also remember that some students have copied material even before the current technology.

1 Introduction

Everybody agrees that plagiarism in general is unethical. The key problem with plagiarism is that it is not black and white. There is a huge gray area, in which one person would consider the text fragment as plagiarism while another one could interpret the issue just as missing reference or poorly formulated sentences.

Plagiarism is defined as "the use of another's thoughts, or work, without acknowledgement or permission. In plagiarism, one author takes another's idea and presents it as his/her own" [9]. Similar definitions can be found from all papers discussing this issue. For example, L. Bilic-Zulle et.al. [1] give the definition as "misappropriation of another person's ideas, methods, results or words, i.e. using the intellectual property of another person without giving appropriate credit".

The word plagiarism is clearly latin based, but there are at least two latin words, that have been given as the origin. Gu and Brooks [3] claim that the origin word is plagiaries, which means "the theft of words as well as slaves". On the other hand Naveed Imran [5] gives the word plagium as the latin origin. According to him the meaning of plagium is kidnapping. Either way, both words contain the idea of taking something that is not legally mine.

The modern definitions include the idea that the usage is OK if you have permission or you give the credits. For a student this might look like that the citations is all what the teachers require. While it is the beginning it is not enough. This paper will mainly discuss about the tools that will help teachers in this first step. The proper writing skills would be topic of its own.

2 Aspects of plagiarism

Naveed Imran [5] explains the different aspects of plagiarism using taxonomy. He has divided the aspects into three major groups. The method covers aspects that deal with the actual implementation. The cast covers all kind of forms for plagiarism. Finally, the purpose explains the different motivations behind the plagiarism. The purpose is simply either intentional or unintentional. Naturally the intentional behaviour is more problematic from the ethical perspective.

The methods used in plagiarism start with the most straightforward copy-and-paste. It is the simplest to do and easiest to detect, because no words have been changed. This is often due time limits or laziness, or lack of writing skills. It is always an easy way out.

The inappropriate paraphrasing is not as easy to detect as copy-and-paste, because in the paraphrasing some of the original wordings have been changed. The key problem here is the amount of partially copied text and the lack of the author's own voice. A lot of scientific papers use paraphrasing properly in small scale with quotation marks, if direct copying, and citation information.

Omitting quotation marks and/or citations, automatically causes otherwise properly paraphrased text to be classified as plagiarism. Had the author marked the direct loans and given the credits to the original material, it would have been accepted. Forgetting the citation is undistinguishable from the purposeful omission and both are thus considered as the missing was purposeful. Worse case would be faked citations, where the cited article does not even contain the material, citation was just added to mislead

the reader.

Plagiarism can even cover the stealing of ideas [5]. In such a case the text is presented as if the idea was the author's own and there is no citation available. Even when there are no textual similarities, but the original source of the idea has not been given credits, the author is guilty of improper behaviour and violates the ethics of academic writing. This could naturally happen with missing citations, but often the author has not even thought about adding the citation there.

3 Copy detection tools

The automated detections tools should be able to find both copy-and-paste and paraphrasing kinds of plagiarism. If they are capable of doing it, they would find most missing citations also.

According to [8] there were several tools available already in 2007. The paper compares tools, like Eve2, CopyCatchGold, WordCheck, Glatt, Moss, JPlag, with the de facto standard Turnitin [1]. However, in 2007, the detection tools were not that known and the Turnitin did not have such a strong position as it has today with its easy-to-use web-based user interface. In 2002 it was claimed [4] that "the available software tends to come and go: new software and websites surface and then disappear".

Most of the tools worked with plain text and tried to estimate the similarity between the submitted text and some other texts. For the comparison they used different metrics, such as N-gram, Euclidean distance, Jaccard measure. A nice collection of different metrics is presented in [8].

Most tools simply compare the document with its database. They do not try to analyse the content of the text, but just do some sort of character-based comparison. To catch also translations, the tool must be able to support some kind of semantical matching in addition to plain text comparison. This would be an issue, if and when a Finnish university wants to systematically detect the possible plagiarism in the bachelor thesis, which are written in Finnish using mainly English references.

4 Turnitin

Turnitin has gained the position of de facto standard over the years. It uses its own database to store the texts to be compared. It has also been criticized about the coverage and cheating possibilities. Kaner and Fiedler

[1] http://www.turnitin.com

[7] in 2008 claimed that the tool was missing articles from ACM and IEEE. If that is still true, the usability of the tool for computer science is much reduced.

Turnitin is a web-based tool, which has its own database of articles and pages. All submitted papers are compared against the existing material in the database. The tool works relatively fast and the comparation result is available for the teacher within minutes after the submission. According to [6] "the software looks for matches of strings of eight to ten words", while "ignoring the commonly used words". Basically the tool counts as similarity a sequence of identical words in identical order. The words do not have to be in consecutive sequence, there can be some additional words in between.

Turnitin works nicely when used exactly as the company has planned. The tool has been designed for identical submission deadlines for the whole class. It cannot handle nicely a situation where students have individual deadlines, and the others should be able to see the submitted papers before their own submission. Our department uses this model in some seminars, where students give their presentations one by one during the weeks. They must submit their paper a give time before their individual presentation. The other participants are expected to read and comment the paper before the presentation.

Using Turnitin through the web-interface is relatively easy. Teacher creates a course and the tasks for the students. She also adds students to the course and the students submit their papers themselves. There is also a possibility for teacher submission, but the tool is designed for student submission.

It is announced to be possible to integrate the Turnitin with course platforms like blackboard or moodle. From a teacher's perspective the integration may mean giving up some of the features that are available through the web-interface while gaining the more familiar environment of the platform. Integrated version, the main interface is the course platform and Turnitin is used only to check the originality of the submitted papers.

The author tried the web-interface based service in two separate courses. One of them was a seminar where students wrote their papers in English. The other one was a Bachelor thesis writing group, where the papers were written in Finnish. With Finnish papers, the main benefit came from the possibility of doing electronic peer reviews of the submitted papers by the students.

The tool shows the results of the originality analyses using traffic lights (blue, green, yellow, orange, red) with similarity percentage. The percent-

age tells how much of the report matches with existing material. The tool compares the submitted paper with its own database. The database contains articles collected from web and directly from publishers.

For the author, best feature of the tool is the visual report of the similarities. The tool shows the submitted student paper with colour-marked similarity locations. Each colour represents different original paper. The simple similarity percentage, as such, does not give any information about what kind of text is found to be similar counterpart in the tool's database stored papers.

Because the similarity percentage just shows how much of the student paper has identical short word sequences with other papers, it should not be used automatically by bureaucrats [3] to classify papers as copied vs. not copied. A native speaker is able to use rephrasing easier to avoid the detection of identical word sequences. A non-native speaker has smaller vocabulary, which makes e.g. using synonyms in rephrasing more difficult for them.

The low-similarity-percentage papers indicated with blue or green colour may still contain parts that fulfil the plagiarism definition. There can be one or two paragraphs that are identical with an other paper in the tool's database. It is also possible to have a high-percentage student paper with yellow or orange that does not fulfil the plagiarism definitions. Such a paper may contain directly copied formulas and definitions with proper citations. It is not custom to use quotation marks with formulas. Of course such a paper might not have high quality because the amount of student's original text is low. Usually the red-marked papers contain too much copied material in all cases.

5 Avoiding unintentional plagiarism

Most of the plagiarism in the student papers is not fully intentional, but could be due to the missing skills of students. To solve such issues teaching is more important than punishment. According to Smith and Wren [9] "avoiding plagiarism does not need to be difficult or require an in-depth knowledge of copyright law". Very simple mechanisms are enough to avoid unintentional plagiarism. It is enough to learn proper referencing and paraphrasing techniques. Naturally they mean that the writer (student or researcher) should summarize the referenced article content in own voice and mark it with citation. To be able to do that she has to actually understand the source information [5]. If the student does not understand what she is writing the direct copying might be the only (but wrong) solution within

the deadline.

When students were asked [2] about explanations to plagiarism, they had quite normal explanations from bad time management and simple opportunity to uninterested teachers and eminently theoretical subject. The time management included explanations like, too many assignments to be handed in short time or personal shortcomings in preparation. A lot of these can be solved by the institutions, when these issues gain the attention of the faculty and teachers.

6 Conclusion

The plagiarism avoidance can be done with just common sense. Ethical behaviour is the key here. Teachers must give a good example and instructions for the students about scientific writing practises. However, it should be clear for everyone that you are not allowed to steal ideas and you should give credit to the right persons and papers. The credit giving is done by using a proper citation technique.

The automatic plagiarism detection tools make the life of the teachers a bit easier. They save time in the evaluation process. Instead of having to manually search for excerpts of the student paper from the existing, published papers, the tool does the similarity check automatically in seconds or minutes. Teachers are still very much needed to evaluate the quality of the student paper, as well as, possible similarity findings of the tool. To improve the student's writing skills teachers also need to explain to the student how the writing process should be improved.

Bibliography

[1] Bilic-Zulle, L., Azman, J., Frkovic, V., and Petrovecki, M. *Is There an Effective Approach to Deterring Students from Plagiarizing?* Sci Eng Ethics (2008) 14:139-147

[2] Comas-Forgas, R. and J. Sureda-Negre, J. *Academics Plagiarism: Explanatory Factors from Students' Perspective.* J Acad Ethics (2010) 8:217-232. Springer.

[3] Gu. Q. and Brooks, J. *Beyond the accusation of plagiarism.* System 36 (2008) pp 337-352, Elsevier.

[4] Delvin, M. *Plagiarism detection software: how effective is it? Assessing Learning in Australian Universities, 2002.* Available

at: http://www.cshe.unimelb.edu.au/assessinglearning/docs/ Plag-Software.pdf

[5] Imran, N. *Electronic Media, Creativity and Plagiarism.* SIGCAS Computers and Society, Volume 40, No. 4, December 2010

[6] Jones. K. *Practical Issues for Academics Using the Turnitin Plagiarism Detection Software.* International Conference on Computer Systems and Technologies - CompSysTech'08, ACM, 2008, IV.1-1 - IV.1-5

[7] Kaner, C. and Fiedler, R. *A Cautionary Note on Checking Software Engineering Papers for Plagiarism.* IEEE Transactions on, Volume 51, No. 2, May 2008, pp. 184-188

[8] Lukashenko, R., Graudina, V., and Grundspenkis, J. *Computer-Based Plagiarism Detection Methods and Tools: An Overview.* International Conference on Computer Systems and Technologies - CompSysTech'07, ACM, 2007, IIIA.18-1 - IIIA.18-6

[9] Smith, N. and Wren. K. *Ethical and Legal Aspects Part 2: Plagiarism – "What Is It and How Do I Avoid It?".* Journal of PeriAnesthesia Nursing, Vol 25, No 5 (October), 2010: pp 327-330