Synthesis

International Scientific Conference of IT and Business-Related Research

# ANTI-PLAGIARISM SOFTWARE: USAGE, EFFECTIVENESS AND ISSUES

## SOFTVER ZA ANTIPLAGIJARIZAM: UPOTREBA, DELOTVORNOST I MOGUĆI PROBLEMI

Marina Marjanović, Violeta Tomašević, Dejan Živković
Singidunum University, Danijelova 32, Belgrade, Serbia

**Abstract:**
In today's digital age, with never ending advances of information technology, plagiarism of textual documents is becoming one of leading issues that causes wide concern in higher education and science. The ubiquity of Internet has created a plethora of possibilities for unacknowledged copying and paraphrasing of other people's work. This has grave legal and moral repercussions for the society and seriously undermines its system of values. This paper discusses different types of document plagiarism and examines methods for their detection. It also presents software solutions that implement particular plagiarism detection techniques. Following the overview in the first part, the paper focuses on the analysis of difficulties arising in the plagiarism detection process and points out to open questions that need to be solved. Moreover, it offers some principal suggestions for possible improvements.

**Key words:**
plagiarism, plagiarism types, plagiarism detection methods, plagiarism detection issues, plagiarism detection software.

**Apstrakt:**
U današnje digitalno doba, koje odlikuju konstantne inovacije u oblasti informacionih tehnologija, pitanje plagijarizma u tekstualnim dokumentima postaje jedno od glavnih problema u domenu nauke i visokog obrazovanja. Sveprisutnost Interneta pruža mnoštvo mogućnosti za nelegalno kopiranje i parafraziranje radova drugih autora. To sa sobom nosi ozbiljne zakonske i moralne posledice za celokupno društvo i ozbiljno podriva njegov sistem vrednosti. U radu se razmatraju različite vrste plagijarizama u tekstualnim dokumentima kao i metode za njihovo otkrivanje. Takođe, pominju se softverska rešenja koja implementiraju određene tehnike za otkrivanje plagijarizama. Nakon kratkog pregleda u prvom odeljku, rad stavlja naglasak na analizu poteškoća koje mogu nastati u procesu otkrivanja plagijarizma i ukazuje na pitanja kojima bi se trebalo pozabaviti. Štaviše, autori nude načelne predloge i sugestije u cilju unapređenja softvera za otkrivanje plagijarizma.

**Ključne reči:**
plagijarizam, tipovi plagijarizma, metode za otkrivanje plagijarizma, problemi vezani za otkrivanje plagijarizma, softver za otkrivanje plagijarizma.

**Acknowledgments:**
The second and the third author wish to express their sincere gratitude to the Serbian Ministry of Education, Science and Technological Development for its support to paper presentation (Research Project III44006).

## 1. INTRODUCTION

In very broad terms, plagiarism can be defined as the act of uncritical use of other people's work (writings, thoughts, ideas, inventions, *etc.*) without acknowledging the source. While plagiarism can be traced back to almost the beginning of human civilization, the Internet has opened up numerous new possibilities for plagiarism, thus making it a very tempting endeavor.

The growing trend of plagiarism has become a grave global issue that seriously undermines the society's value system. This is why numerous countries have intensified their efforts to cope with the rise of plagiarism based on the combination of plagiarism prevention and detection. Plagiarism prevention refers to raising the society's awareness about the plague of plagiarism to a higher level, along with the implementation of a range of measures that include media campaigns and development of deterring strategies, honesty policies and sanctions. Plagiarism detection implies identification of unacceptable similarity between documents, usually by means of some sort of software systems.

In the battle against plagiarism, experience has shown that prevention is more effective than detection in the long run. This is the case mainly due to the fact that only plagiarism prevention measures can fully or to a great extent eliminate plagiarism, albeit consistently applied within a long period of time. On the other hand, plagiarism detection methods can only decrease the amount of plagiarism, even though they may achieve such positive results in the short term.

Not considering moral and ethical issues, it seems that there are two main reasons why confronting plagiarism is so difficult. Firstly, plagiarism itself eludes a clear universal definition because the borderline between the plagiarized and authentic work can be surprisingly blurred. Namely, except for obvious cases of the cut-and-paste kind, one can use many plagiarism techniques to disguise genuine scholarly work. Secondly, all plagiarism detection methods usually rely on software tools, while plagiarism is practiced by humans. Assuming that the plagiarist's goal is to go undetected with the "intellectual theft", the plagiarism issue can be considered an artificial intelligence problem of how well the computer can simulate the human thinking process.

The remainder of this paper is organized as follows: Section 2 reviews the existing types of plagiarism, while Section 3 elaborates on the methods and software tools used for plagiarism detection. The main goal of the paper is to analyze plagiarism issues and reasons for their emergence, point out open questions, and suggest possible improvements, shall be further discussed in Section 4. The final section includes a short summary with conclusions.

E-mail: mmarjanovic@singidunum.ac.rs

DOI: 10.15308/Synthesis-2015-119-122

## 2. PLAGIARISM TYPES

Plagiarism can be divided into about 15 types (Park, 2003; Park, 2004; Hiremath & Otari, 2014; Kashkur *et al.*, 2010; Turnitin white paper). They are listed below based on their detection difficulty.

1. Clone plagiarism – Taking someone else's work entirely.
2. Copy-and-paste plagiarism – Copying large parts of someone else's work.
3. Re-tweet plagiarism – Contains correctly quoted text, but relies too much on someone else's work.
4. Recycle or auto plagiarism – Publishing the same work many times.
5. Find-replace or word-switch plagiarism – Using synonyms for words in someone else's work.
6. Hybrid plagiarism – Combining judiciously quoted and unquoted parts of someone else's text.
7. Mashup plagiarism – Copying material from different sources.
8. Error plagiarism – Using incorrect citation.
9. Aggregator plagiarism – No originality in the work, although it contains references to the original work.
10. Style plagiarism – Paraphrasing to the extent that the original text is unrecognizable, but the structure of both documents is similar (essential schemes, main arguments, or examples coincide).
11. Translation plagiarism – Translating someone else's work into another language.
12. Idea plagiarism – The main idea of the work is not original, but it is masked by the plagiarist's knowledge.
13. Graphics plagiarism – Using a figure or a picture without permission.
14. Source-code plagiarism – Taking the source code in computer programming.
15. Ghostwrite plagiarism – Contracting another person or website to produce the work for someone.

## 3. PLAGIARISM DETECTION

The extraordinary popularity of the Internet has enabled easy access to useful and credible information for use by everyone. At the same time, the Internet has taken the plagiarism issue to a higher level by making it extremely easy for uncritical use of other people's work and even for finding numerous services on the Web that will do scholarly work for someone else. Thus, the continued growth of plagiarism cases has drawn the increased attention to the anti-plagiarism tools.

In today's digital marketplace, one can find many software products that offer defensive solutions against plagiarism based on various techniques. The following list summarizes the most commonly used methods for detecting different types of plagiarism (Hiremath & Otari, 2014).

***Text-based plagiarism detection methods.*** Kashkur *et al.* (2010), provide a classification of techniques for plagiarism detection of textual documents. Furthermore, Meyer and Stein (2006) described a heuristic method for style plagiarism detection. The method is based on finding stylistic inconsistencies in the document being checked for plagiarism. However, this approach can give false positive results in case the document represents a joint work with multiple authors. Anzelmi *et al.*, elaborate on a detection algorithm that uses SCAM (Standard Copy Analysis Mechanism) formula for the so-called bag of words analysis (Anzelmi *et al.*, 2011). Moreover, Hoad and Zobel (2013) suggested that one can use the fingerprinting method to estimate the likelihood of similarity when two or more documents are compared.

**Citation-based plagiarism detection methods.** Hiremath and Otari (2014) described the method that uses citations and references for plagiarism detection. The method is based on an estimate of the degree of similarity in citations and the order of the documents being compared. This method can give good results for detecting idea plagiarism, but not for the hybrid and error types of plagiarism.

**Shape-based plagiarism detection methods.** Meyer and Stein (2006) derived a formula for detection of improper use of a figure without permission. The method is based on figure shape recognition, but it is very sensitive to even small changes in figure shape.

**Source-code plagiarism detection methods.** Kashkur *et al.*(2010), present many algorithms for the source-code plagiarism detection based on Kolmogorov complexity and fingerprinting method. Lukashenko *et al.* (2007), described various methods based on finding patterns of the same variable names in programs and identifying similarities in the syntax complexity of programs.

**Translation plagiarism detection methods.** Gipp describes a method for detection of the translation plagiarism based on the citation pattern analysis (Gipp, 2014).

Moreover, Urbina *et al.* (2010), provide an extensive list of commonly used software tools to detect plagiarism. Here, we shall reproduce the list by dividing it into two tables as given below: the first one is free software, while the other one is commercial plagiarism detection software. For each software tool in both tables, the second column specifies whether the tool is Web-based or a desktop application (web, desktop). The third column represents the corpus of documents that is searched over when plagiarism detection is performed (the Internet, database, files). The fourth column gives the acceptable file format of the document submitted for check (txt, pdf, img, ppt, and html). Finally, the fifth column shows the form of the report obtained as a result of plagiarism detection. The results can be sent as a percentage probability that the submitted document is plagiarized (%), as a website link to the report (link), or as a list of suspicious documents similar to the submitted document (list).

| Software | App | Corpus | File format | Report |
|---|---|---|---|---|
| Approbo | web | internet | txt, pdf, doc | %, link, list |
| Image Stamper | web | internet | img | link |
| DocCop | web | internet, files | txt, pdf, doc | %, link, list |
| Plagiarism Checker | web | internet | txt | link |
| WCopyfind | desktop | files | txt, doc, html | %, link |
| Jplag | desktop | files | txt | %, list |

Table 1. Free plagiarism detection software

DOI: 10.15308/Synthesis-2015-119-122

| Software | App | Corpus | File format | Report |
|---|---|---|---|---|
| iThenticate | web | internet, database | txt, pdf, doc, html | %, link |
| Turnitin | web | internet, database | txt, pdf, doc, html | %, link |
| Plagiarism Detect | web, desktop | internet | txt, pdf, doc, html | link |
| Docoloc | web | internet | txt, pdf, doc, html | %, link, list |
| EVE2 | desktop | internet, database | txt, doc | %, link |
| Scriptum | web | internet, database | txt, doc | %, link |

Table 2. Commercial plagiarism detection software

## 4. PLAGIARISM ISSUES

The issue of plagiarism is inherently associated with the manner in which creative work is produced. There are dishonest authors who intentionally try to steal other persons' work. This, of course, represents a blatant case of plagiarism. However, the authors frequently create their own work by following and imitating others. There is a great risk that the work created may turn into a non-authentic piece, which is why the authors must be aware of the fact that good intentions are not an excuse. The only way the authors can be sure that their work is authentic is by doing it entirely on their own and by giving proper credit to other people's ideas.

The use of software systems for plagiarism detection has both positive and negative aspects. The advantages of software plagiarism detection come from the mere assets of the technology itself such as speed, reliability, easy reporting, *etc.* Negative effects result from misunderstanding of the role of software in plagiarism detection process. This can be manifested in the following ways:

- Software is accepted to definitely decide on whether some work is plagiarism or not, but it only serves the purpose to detect similarity of the document contents. This is exacerbated by the fact that software plagiarism detection tools often use catchy names to associate their purpose to something which is a way out of their league;
- Consequently, the authors often check their work for plagiarism and after obtaining a negative result, they deny that their work is plagiarism. This argument may not be in agreement with other people's judgment on (non-) authenticity of the work. The possible solution for such conflicting situations is to inform the authors in advance about the proper role of the software.

Software tools determine similarity of the document contents by using different methods and produce appropriate reports. The results of these reports need to be taken with great care and human judgment is necessary to decide whether something is plagiarism or not. However, in doing so, the question of the correctness of the obtained results must be clarified. Firstly, different software can give diverse results for the same type of document plagiarism, and thus, it is not clear how to interpret the results. Secondly, different software can analyze a document for different types of plagiarism, which raises the question of what is more relevant and how to reach a final decision.

Some authors have exploited the wide availability of cheap, even free, software tools for plagiarism detection by checking their work against plagiarism prior to making it public. Such bad practice can lead to an absurd situation in which the authors desperately try to revise a non-authentic work until it passes the plagiarism check, not paying much attention to the quality that usually deteriorates. Furthermore, for those who know how particular software works and which methods to implement, it is tempting to adapt their work by trying to circumvent software methods and pass the plagiarism check. These problems stem from the easy availability of software plagiarism detection tools to both authors and referees. If the plagiarism check was only in the hand of referees, the authors would be paying much more attention to their creative process and the quality of their work. The very idea of using software to check one's own work arouses suspicion that the work is not authentic and that it is nothing more than a lame attempt to soothe one's conscience. There is no need at all to use software to self-check an independent work for plagiarism, even though it is possible for software plagiarism detection tools to show similarities with other works.

Given the fact that human brain is much more complex than the computer, the role of human judgment in the process of plagiarism detection is indispensable. Namely, when detecting plagiarism, humans use semantic and statistical methods to apply them to all kinds of information. Human intuition, hunch and experience are also very important. Unfortunately, transferring these main features of human intelligence into software is far from possible today, and it may never be. On the other hand, the main advantage of computers is reflected in the computers' ability to access and process a huge amount of data with astonishing speed (in the plagiarism detection case, the data is a corpus of documents for comparison). This intrinsic characteristic of computers should certainly be of great assistance to humans in detecting whether something is plagiarism or not.

The complex questions regarding the plagiarism issue might be better answered by introducing more order in the process of plagiarism detection. One possibility may be the establishment of certified organizations with exclusive authority to check for plagiarism. The organizations would obtain certificates based on their activity, which would entail a certain level of responsibility. The certified entities could include publishing houses, universities, schools, and agencies providing plagiarism detection services. In the process of plagiarism detection, certified organizations should use software tools that are not freely available. The results of preliminary plagiarism detection process would be then made available to a committee which would be responsible for making the final decision on whether something is plagiarism or not. This would be an effective plagiarism deterrent and would boost the authors' morale and their self-confidence, while reducing the damage caused by plagiarism.

## 5. SUMMARY

Even though the incidents of plagiarism can be found since ancient times, plagiarism has never been as widespread as today. The rapid development of the Internet has significantly contributed to the proliferation of plagiarism cases. In fact, new digital technologies have triggered more opportunities for uncritical use of other people's work, thus making such new forms of plagiarism harder to detect and control. This unethical practice has become so serious that its erosive and corruptive effects are felt in all spheres of society. That is why the efforts against plagiarism have been intensified through implementation of a range of measures that usually involve software systems. Thus, one can ironically note that the information technology represents both the cause and the solution to the plagiarism problem.

DOI: 10.15308/Synthesis-2015-119-122

The lack of a clear and universal definition of the concept of plagiarism makes it more difficult to effectively prevent the old problem. As described in Section 2, it is possible to list 15 types of plagiarism, but this number is constantly increasing with the advent of new technologies. In order to detect new types of plagiarism, methods for determining the similarity of the document contents are also adapted (Section 3). A growing number of anti-plagiarism software tools are also available today, and some of the most popular ones are compared in Section 3. However, it appears that the anti-plagiarism software per se cannot solve the plagiarism problem. Moreover, the software tools are often misused and their results are misinterpreted, and thus, new problems emerge as a result of uncontrolled and incorrect use of the software. Some of these issues shall be discussed in Section 4.

In this paper, we have argued that the existing practice aimed to cure the plagiarism problem relying mostly on software tools is questionable. It should be thus revised by keeping its positive elements (for example, that humans make a final decision on plagiarism or that emphasis is placed on preventive measures to raise social awareness), as well as by getting rid of those elements that may cause new problems (for example, that anyone can verify the document contents). Greater level of discipline, stricter deterrent rules and more responsibility in the process of solving the plagiarism issue would probably give better results in the future.

## REFERENCES

Anzelmi, D., Carlone, D., Fabio, R., Thomsen, R., & Hussain, D.M.A. (2011). Plagiarism Detection Based on SCAM Algorithm. Proceedings of the International MultiConference on Engineers and Computer Scientists, pp. 272-277.

Asim, M., El Tahir, A., Hussam, M.D.A., & Snasel, V. (2015). Overview and Comparison of Plagiarism Detection Tools. Department of Computer Science, VSB-Technical University of Ostrava, 17., Ostrava - Poruba, Czech Republic.

Gipp, B. (2014). Citation-based Plagiarism Detection – Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis. Berlin: Springer Vieweg Research.

Hoad, T., Zobel, J. (2003). Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology*, 54(3), 203–215.

Kashkur, M., Parshutin, S., Borisov, A. (2010). Research into Plagiarism Cases and Plagiarism Detection Methods. *Scientific Journal of Riga Technical University Computer Science, Information T and Management Science*. 44, 139-143.

Lukashenko, R., Graudina, V., & Grundspenkis, J. (2007). Computer-Based Plagiarism Detection Methods and Tools: An Overview. International Conference on Computer Systems and Technologies - CompSysTech'07.

Meyer, S., & Stein, B. (2006). Intrinsic Plagiarism Detection. 28th European Conference on IR Research, ECIR 2006 London, pp. 565-569, Springer.

Park, C. (2003). In other (people's) words: Plagiarism by university students – literature and lessons learned. Assessment & Evaluation in Higher Education, 28, 471-488.

Park, C. (2004). Rebels without a clause: Towards an institutional framework for dealing with plagiarism by students. *Journal of Further and Higher Education*, 28, 291-306.

Senosy, A., Fadhil, N., Maidorawa, A., & Salim, N. (2014). Shape-Based Plagiarism Detection for Flowchart Figures in Texts. *International Journal of Computer Science & Information Technology (IJCSIT)*. 6(1). DOI: 10.5121/ijcsit.2014.6108

Shiremath, S.A., & Otari, M.S. (2014). Plagiarism Detection-Different Methods and Their Analysis: Review. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*. 1(7), 41-47.

Si, H., Leong, V., & Rynson, W.H. (1997). CHECK: A Document Plagiarism Detection System. In ACM symposium on Applied computing SAC'97, pp. 70-77, DOI: 10.1145/331697.335176.

Urbina, S., Ozollo, R., Gallardo, J.M., & Aina, C.M. (2010). Analisis de Herramientas para la Deteccion de Ciberplagio. XIII International conference EDUTEC 2010.

The Plagiarism Spectrum: Instructor Insights into the 10 Types of Plagiarism. Retrieved 15.02.2015 from https://www2.nau.edu/d-elearn/support/tutorials/academicintegrity/pdf/Turnitin_WhitePaper_PlagiarismSpectrum.pdf

DOI: 10.15308/Synthesis-2015-119-122