Measurement of Similarity in Academic Contexts

Omid Mahian, Marius Treutwein, Patrice Estellé, Somchai Wongwises, Dongsheng Wen, Giulio Lorenzini, Ahmet Selim Dalkilic, Wei-Mon Yan, Ahmet Sahin

publications

MDPI

Communication

Measurement of Similarity in Academic Contexts

Omid Mahian [1], Marius Treutwein [2], Patrice Estellé [3,*], Somchai Wongwises [4], Dongsheng Wen [5], Giulio Lorenzini [6], Ahmet Selim Dalkilic [7], Wei-Mon Yan [8] and Ahmet Z. Sahin [9]

1 Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University, Mashhad 9187147578, Iran; omid.mahian@mshdiau.ac.ir
2 Department of Radiation Oncology, Regensburg University Medical Center, D93053 Regensburg, Germany; Marius.Treutwein@klinik.uni-regensburg.de
3 LGCGM, Equipe Matériaux et Thermo-Rhéologie, Université Rennes 1, 35000 Rennes, France
4 Fluid Mechanics, Thermal Engineering and Multiphase Flow Research Laboratory (FUTURE Lab.), Department of Mechanical Engineering, King Mongkut's University of Technology Thonburi, Bangmod, Bangkok 10140, Thailand; somchai.won@kmutt.ac.th
5 School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK; d.wen@leeds.ac.uk
6 Department of Industrial Engineering, University of Parma, 43124 Parma, Italy; giulio.lorenzini@unipr.it
7 Heat and Thermodynamics Division, Department of Mechanical Engineering, Yildiz Technical University, Yildiz, Besiktas, Istanbul 34349, Turkey; ahmet_selim_dalkilic@hotmail.com
8 National Taipei University of Technology, Taipei 10608, Taiwan; wmyan1234@gmail.com
9 Mechanical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia; azsahin@gmail.com
* Correspondence: patrice.estelle@univ-rennes1.fr; Tel.: +33-223-234-200

Academic Editor: Alan Singleton
Received: 2 March 2017; Accepted: 19 June 2017; Published: 22 June 2017

Abstract: We propose some reflections, comments and suggestions about the measurement of similar and matched content in scientific papers and documents, and the need to develop appropriate tools and standards for an ethically fair and equitable treatment of authors.

Keywords: similarity; plagiarism; standards

Currently, most famous publishers as well as many universities worldwide use software to detect the rate of similarity between scientific papers, dissertations and the available literature. This is a common step in the current editing process. However, not all journals use such software, thus contributing to an overlap between papers in their databases. Specific criteria used in these software tools are also unknown to the authors.

In addition, to date, there is neither a measuring unit nor a general standard in the scientific community for the allowable rate of similarity. For example, based on the knowledge of the presenting authors, some editors reject research papers with a similarity percentage higher than 10% or sometimes up to 35% and review papers with a much higher similarity percentage. Similarity measurement can also be performed at different steps in the review or editing processes, i.e., when a paper is submitted or accepted. Moreover, universities and even general governmental organizations in science from different countries may also have different rules and limits regarding similarity, which has been described with some examples by Maurer et al. [1]. Consequently, some researchers may be accused of bad citing or even plagiarism, brought about by the lack of existing similarity standards and missing knowledge of the applied software tool to measure it.

As an example of our concerns, the preceding two original paragraphs of this letter were checked with available free versions of plagiarism software: PlagScan [2], CheckText [3] and Plagiarism Checker X [4] respectively. The first one reveals that "no strongly similar text sources were found on the internet". The second one indicates 4% of plagiarized words corresponding to common words such

as "currently" or "moreover". The last one reports 31 plagiarized words out of 201 with nine sources identified with an estimated plagiarism percentage of 14%, recommending optional improvement. Although these tools only check text similarity and are probably less powerful than iThenticate [5] used by Crossref Similarity Check and which is subject to a charge, this example demonstrates that the detected similarity depends on the applied software.

Although there exist academic and national boards, e.g., JISC [6] (Launch of Jisc Plagiarism Advisory Service) in the United Kingdom [1], a universal standard would be desirable in the scientific community, for the rate of allowable similar and matched content—referring especially to so-called "self-plagiarism" [7,8] or copying of general phrases in the introduction section [9,10]—for an ethically fair and equitable treatment of authors. This mainly implies a wide consensus between journals, editors, authors and institutions. With this goal, we recommend the creation of a representative committee to propose appropriate common tools and standards for measuring matched content and similarity rate in scientific documents. In a sophisticated version, the standard might be adjusted by considering the field of research, as using technical words in science is unavoidable, and, hence, the rate of similarity might be automatically higher. This is also the reason why author names, affiliations and the reference list are obviously excluded from the similarity analysis. Furthermore, the similarity in some parts of articles such as introductions or methods could be weighed differently from results and discussions as well as conclusions. As Brumfiel discussed [9], open archives or preprint servers such as arXiv [11] are often misused for plagiarism and authors with poor English knowledge tend to copy phrases from their own earlier work or the work of others. Therefore, the content of previously published paper(s) by the same (group of) author(s) also has to be evaluated to differentiate between self-plagiarism and the correct re-use of previous published works. While similar and matched content detection by software is very quick and useful, this could be coupled to human analysis for better efficiency as evidenced in [12]. The more sophisticated the results of text analysis software are, the more solid is the basis on which the editor makes his/her decision. As Glänzel et al. [13] and also *the Nature journal* [8] stated, a careful human cannot be replaced. An automatic rejection based on a simple similarity value therefore should not occur.

Before potentially undertaking this large-scale work, our main suggestions to improve this issue and make the process more transparent are that each journal should systematically reveal in the author guidelines which similarity checker software is used. In addition, concrete similarity limits must also be mentioned.

Author Contributions: All the authors contributed equally.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Maurer, H.; Kappe, F.; Zaka, B. Plagiarism—A Survey. *J. Univers. Comp. Sci.* **2006**, *12*, 1050–1084. [CrossRef]
2. PlagScan. Available online: https://www.plagscan.com/seesources/ (accessed on 2 March 2017).
3. CheckText. Available online: http://www.checktext.org/ (accessed on 2 March 2017).
4. Plagiarism Checker X. Available online: http://plagiarismcheckerx.com/ (accessed on 2 March 2017).
5. iThenticate. Available online: http://www.ithenticate.com/ (accessed on 28 April 2017).
6. Jisc. Available online: https://www.jisc.ac.uk/news/launch-of-jisc-plagiarism-advisory-service-25-sep-2002 (accessed on 28 April 2017).
7. Bird, S.J. Self-plagiarism and dual and redundant publications. What is the problem? *Sci. Eng. Ethics* **2002**, *8*, 543–544. [CrossRef] [PubMed]
8. Plagiarism pinioned. *Nature* **2010**, *466*, 159–160.
9. Brumfiel, G. Turkish physicists face accusations of plagiarism. *Nature* **2007**, *449*, 8. [CrossRef] [PubMed]
10. Li, Y. Text-based plagiarism in scientific publishing: Issues, developments and education. *Sci. Eng. Ethics* **2013**, *19*, 1241–1254. [CrossRef] [PubMed]
11. ArXiv. Available online: https://arxiv.org/ (accessed on 28 April 2017).

12. Bretag, T.; Mahmud, S. Self-plagiarism or appropriate textual re-use. *J. Acad. Ethics* **2009**, *7*, 193–205. [CrossRef]
13. Glänzel, W.; Braun, T.; Schubert, A.; Zosimo-Landolfo, G. Coping with copying. *Scientometrics* **2015**, *102*, 1–3. [CrossRef]