Plagiarism Culprit: Citations

The information scientist Prof. Dr. Bela Gipp develops new methods of plagiarism detection - even beyond language boundaries

The plagiarism detection software, CitePlag, checks literature for semantically similar content. Even if passages have been paraphrased or translated into another language, colored text highlights and lines connecting shared citations can indicate similarity between the texts.

"Verbatim" or word by word! This is the mantra followed by today's most widely used plagiarism detection software. Potentially suspicious documents are automatically compared with many other documents from large document repositories. The software determines whether any text sections match one-to-one. However, these types of text-based plagiarism detection systems quickly reach their limit when plagiarists do not copy text verbatim. When it comes to disguising their misconduct, plagiarists are known for their creativity: Texts are paraphrased to change the wording, or they are restructured, or disguised by translation. Plagiarism detection software fails to detect such reinvented text as plagiarism. "Plagiarism detection software should not only check for matching text strings" this is the recommendation of the Computer Science Prof. Dr. Bela Gipp from Konstanz. In his project CitePlag, he makes use of semantically similar content to use for plagiarism detection.

Citations - a reliable and distinctive feature of academic text

One of the most telling features used by CitePlag are a text's citations. With 'citations' we don't mean any quotes cited in the text, or the content of footnotes; Rather, we describe the in-text references to other works that were simply copied from another source. No academic literature can get by without citing other academic work on which the author bases their claims. However, some plagiarist simply copy citations from other academics and include them in their own text and reference list without even having read them. After all, it would be very time-intensive for a plagiarist to research their own citations and to replace them with similar citations or re-arrange all citations from the document that is being plagiarized. To do this, a plagiarist would have to be so familiar with the material that plagiarizing citations from others would hardly be time-saving. "Citations are a reliable distinctive feature", according to Bela Gipp, who explains that it is "extremely unlikely for two academic articles to share the same citations in the same order." Citation analysis is just one of several approaches that are combined in CitePlag to examine text for semantic similarity, or similarity of content. The software allows a side-by-side visual comparison of two documents highlighting their similarities using colors and lines to connect shared citations. Even if the plagiarized text passages have been paraphrased or generously rearranged within the document, their content can still be identified as semantically similar. Even plagiarism that has been translated from another language can be detected by Bela Gipp's software.

CitePlag is in the testing phase

"I want to point out that I'm not a 'plagiarist-hunter'", says Bela Gipp. "My goal is to increase the effort of plagiarizing until it is no longer worth it." Bela Gipp's CitePlag project is currently in the testing phase. In the future, the software will be offered free of charge as a service of the University. The Computer Science professor furthermore thinks it is important that users do not rely exclusively on automated plagiarism detection: "a human expert will always be needed to check the documents being compared and to make the final decision on whether it is a case of plagiarism". Plagiarism detection software is just one tool that makes plagiarizing more difficult. Even more importantly, says Gipp is the public discussion that must take place around proper scientific conduct and academic honesty: "the problem of plagiarism will not be solved by technical means alone. Plagiarizing of ideas remains the largest problem. How society chooses to deal with plagiarism will be the crucial factor."

Translation by CB

Isg.uni-konstanz.de

"My goal is to increase the effort of plagiarizing until it is no longer worth it."

Prof. Dr. Bela Gipp

Prof. Dr. Bela Gipp is a Junior Professor of Information Science at the University of Konstanz, in Germany. Previously, he researched at the University of California, Berkeley and at the National Institute of Informatics in Tokyo. His research focus includes methods for extracting and visualizing information, information management systems, as well as developing semantic analysis methods for plagiarism detection software and for recommender systems.

This is a translation of the article "*Am Zitat überführt*" (in German), which appeared in *issue 59 of uni'kon*, the campus magazine of the University of Konstanz.

You can read and download the original article at:
www.uni-konstanz.de/unikon