'Text overlap' clutters scientific papers, arXiv analysis finds

December 23 2014, by Bill Steele

Credit: Petr Kratochvil/public domain

Computer text analysis of a huge database of scientific papers shows a large amount of "text overlap," where authors use text from previous papers of their own and others, not always with attribution. This is not necessarily good or bad, Cornell researchers say.

"Our first goal was to characterize the accepted practice, not to be

judgmental," said Paul Ginsparg, professor of physics and information science and founder of the online arXiv collection of scientific papers, now maintained by Cornell University Library. The analysis was conducted on thousands of papers in the *arXiv*. Ginsparg and Cornell graduate student Daniel Citron reported their study in the Dec. 8 online edition of the *Proceedings of National Academy of Sciences*.

"While it is technically plagiarism, which more generally is stealing of ideas," Ginsparg said, "it's a benign form in the sense that most of it cites the source (at least somewhere in the article), and many authors have rationales for the practice." Many readers find the reuse of text "an annoyance and a distraction," he added, and some worry that it wastes space online and in print journals.

Sometimes the copied text is a description of experimental apparatus or procedures. Mathematicians often repeat well-known theorems that underlie their arguments. Many authors reuse material from their own previous papers about the same line of research, and a Ph.D. candidate's final thesis may pull in material from papers the student had published along the way. An amusing sidelight is the copying of acknowledgements with new names inserted, possibly because authors think what they are using is a standard format.

Sometimes the roots of the practice may be cultural. In China and some other Eastern cultures it's considered disrespectful to change someone else's words. And if English is not their native language, authors may not feel comfortable rewriting their source material. Some may just have a different view of what is acceptable and may have been trained that way by their mentors. "It's not so much malfeasance as ignorance of the standards," Ginsparg explained.

In what may seem like the Law of Karma in action, the analysis showed that papers with the most text overlap were the least cited. Partly, the

researchers said, this is because many papers with large overlap come from a group of countries in eastern Europe and the Middle East, and papers from these countries are less often cited anyway.

Scientists can deposit their papers as electronic "preprints" in the online arXiv to make them promptly available to colleagues around the world. By the end of 2014 the arXiv will pass a million submissions in mathematics, physics, computer science and a growing list of other disciplines.

The study compared text in 757,000 articles deposited from 1991 to 2012. Each text was broken into overlapping phrases seven words long, and other texts were then scanned to see how many matching phrases they had in common. The system skips common usages like "The remainder of this article is organized as follows," as well as direct quotations with proper attribution.

The entire database was preprocessed to create an index of all the seven-word patterns in the arXiv. The analysis only became possible, Ginsparg said, with the low-cost availability of computers with enough random-access memory (RAM) to hold the entire 12.5 GB index. Given that, a single paper can be checked in less than a second, he said.

Since June 2011, every new article submitted to the arXiv has been scanned against the entire database, and papers with significant text overlap have been flagged with a notice to that effect. The current study grew partly from the need to establish a threshold for aberrant practice, permitting a response to indignant authors saying in effect, "Everybody does it."

That's what the study is about, Ginsparg said. "This is the first systematic assay of what's out there on a large scale, permitting assessment of community standards."

An example of copied text

"I cannot describe how indebted I am to my wonderful girlfriend, Amanda, whose love and encouragement will always motivate me to achieve all that I can. I could not have written this thesis without her support; in particular, my peculiar working hours and erratic behaviour towards the end could not have been easy to deal with!"

"I cannot describe how indebted I am to my wonderful wife, Renata, whose love and encouragement will always motivate me to achieve all that I can. I could not have written this thesis without her support; in particular, my peculiar working hours and erratic behaviour towards the end could not have been easy to deal with!"

Provided by Cornell University

Citation: 'Text overlap' clutters scientific papers, arXiv analysis finds (2014, December 23) retrieved 27 April 2024 from https://phys.org/news/2014-12-text-overlap-clutters-scientific-papers.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.