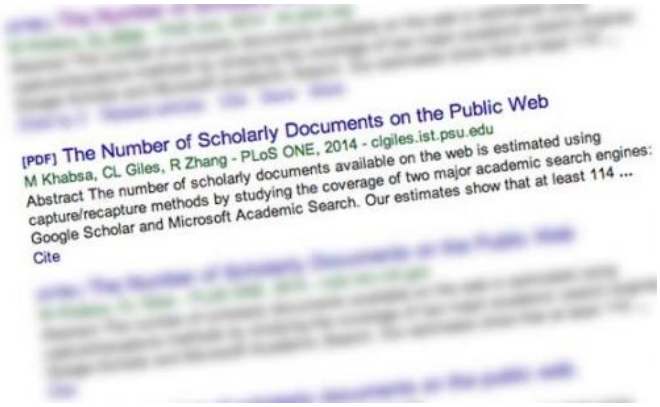
How many scholarly papers are on the Web? At least 114 million, professor finds

10 October 2014, by Stephanie Koons



Google Scholar boasts nearly 100 million English-language scholarly documents. Credit: Penn State

(Phys.org) —Lee Giles, a professor at Penn State's College of Information Sciences and Technology (IST), has devoted a large portion of his career to developing search engines and digital libraries that make it easier for researchers to access scholarly articles. While numerous databases and search engines track scholarly documents and thus facilitate research, many researchers and academics are concerned about the extent to which academic and scientific documents are available on the Web as well as their ability to access them. As part of an effort to make the process of accessing documents more efficient, Giles recently conducted a study of two major academic search engines to estimate the number of scholarly documents available on the Web.

"How many scholarly papers are out there?" said Giles, who is also a professor of computer science and engineering (CSE), a professor of supply chain and information systems, and director of the Intelligent Systems Research Laboratory. "How many are freely available?"

Giles and his advisee, Madian Khabsa, a doctoral candidate in CSE, presented their findings in "The Number of Scholarly Documents on the Public Web," which was published in the May 2014 edition of *PLOS ONE*, a peer-reviewed scientific journal published by the *Public Library of Science*. The paper was also mentioned twice in *Nature*, a prominent interdisciplinary scientific journal, as well as various blogs and websites.

In their paper, Giles and Khabsa report that they estimated the number of scholarly documents available on the Web by studying the overlap in coverage of two major academic search engines: Google Scholar and Microsoft Academic Search. By scholarly documents, they refer to journal and conference papers, dissertations and master's degree theses, books, technical reports and working papers. Google Scholar is a freely accessible Web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. Microsoft Academic Search is a free public search engine for academic papers and literature, developed by Microsoft Research for the purpose of algorithms research in object-level vertical search, data mining, entity linking and data visualization. Using statistical methods, Giles and Khabsa estimated that at least 114 million English-language scholarly documents are accessible on the Web, of which Google Scholar has nearly 100 million. They estimate that at least 27 million (24 percent) are freely available since they do not require a subscription or payment of any kind. The estimates are limited to English documents only.

Giles' and Khabsa's study, Giles said, is the "first to use statistical, rigorous techniques in doing these estimations." The researchers conducted their study using capture-recapture methods, which were pioneered in ecology and derive their name from censuses of wildlife in which several animals are captured, marked, released and subject to recapture. The technique examines the degree of overlap between two or more methods of

ascertainment and uses a simple formula to estimate the total size of the population. Since their study was not longitudinal, Giles said, he and Khabsa plan to do another capture in the future to verify their results.

Giles' interest in determining the number of scholarly documents on the Web was inspired by more than just curiosity—as a developer of various novel search engines and digital libraries, there are practical implications for his research. CiteSeer, a public search engine and digital library for scientific and academic papers, primarily in the fields of computer and information science, was created by Giles, Kurt Bollacker and Steve Lawrence in 1997 while they were at the NEC Research Institute (now NEC Labs), in Princeton, New Jersey, CiteSeer's goal was to actively crawl and harvest academic and scientific documents on the Web and use autonomous citation indexing to permit querying by citation or by document, ranking them by citation impact. CiteSeer, which is often considered to be the first automated citation indexing system, was considered a predecessor of academic search tools such as Google Scholar and Microsoft Academic Search. Released in 2008, CiteSeerX was loosely based on the previous CiteSeer search engine and digital library and is built with a new open source infrastructure, SeerSuite, and new algorithms and their implementations. While CiteSeerX has retained CiteSeer's focus on computer and information science, it has recently been expanding into other scholarly domains such as economics, medicine and physics. One of the motivations for determining the number of scholarly documents on the Web, Giles said, is to increase the number of papers in CiteSeerX.

A significant finding in their study, Giles and Khabsa wrote in their paper, is that almost one in four of Web accessible scholarly documents are freely and publicly available. The researchers used Google Scholar to estimate this percentage because Scholar provides a direct link to the publicly available document next to each search result where a link is available. The findings are important, Giles said, because publicly available documents carry more weight in the research community. Governments, especially those in Europe, fund a lot of scientific research and don't want papers not to be freely available. In addition, he said, it's been shown that freely available papers are much more likely to be cited than those that are not.

By having an idea of how many scholarly documents are on the Web as well as how many are freely available, Giles said, researchers can be better equipped to manage scholarly document research and related projects.

"It was surprising to see how many scholarly documents were digitized and how many were freely available," Giles said. "But keep in mind, these estimates were only for those written in English. How many are there in other languages, more or less than English?"

More information: Khabsa M, Giles CL (2014) "The Number of Scholarly Documents on the Public Web." *PLoS ONE* 9(5): e93949. DOI: 10.1371/journal.pone.0093949

Provided by Pennsylvania State University

APA citation: How many scholarly papers are on the Web? At least 114 million, professor finds (2014, October 10) retrieved 25 January 2021 from https://phys.org/news/2014-10-scholarly-papers-web-million-professor.html