EXHIBIT 5

WIKIPEDIA

CiteSeer^{x}

CiteSeer^{x} (originally called **CiteSeer**) is a public search engine and digital library for scientific and academic papers, primarily in the fields of computer and information science. CiteSeer is considered as a predecessor of academic search tools such as Google Scholar and Microsoft Academic Search. CiteSeer-like engines and archives usually only harvest documents from publicly available websites and do not crawl publisher websites. For this reason, authors whose documents are freely available are more likely to be represented in the index.

CiteSeer's goal is to improve the dissemination and access of academic and scientific literature. As a non-profit service that can be freely used by anyone, it has been considered as part of the open access movement that is attempting to change academic and scientific publishing to allow greater access to scientific literature. CiteSeer freely provided Open Archives Initiative metadata of all indexed documents and links indexed documents when possible to other sources of metadata such as DBLP and the ACM Portal. To promote open data, **CiteSeer^{x}** shares its data for non-commercial purposes under a Creative Commons license.[1]

CiteSeer changed its name to ResearchIndex at one point and then changed it back.

CiteSeer^{x}

Type of site	Bibliographic database
Owner	Pennsylvania State University College of Information Sciences and Technology
URL	citeseerx.ist.psu.edu (https://citeseerx.ist.psu.edu/)
Registration	Optional
Launched	2008 / 1997
Current status	Active
Content license	Creative Commons BY-NC-SA license[1]

Contents

Exhibit #
Hall-Ellis 8
3-01-2021 JB
exhibitsticker.com

History

CiteSeer and CiteSeer.IST

CiteSeer was created by researchers Lee Giles, Kurt Bollacker and Steve Lawrence in 1997 while they were at the NEC Research Institute (now NEC Labs), Princeton, New Jersey, USA. 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. At one point, it was called ResearchIndex.

CiteSeer became public in 1998 and had many new features unavailable in academic search engines at that time. These included:

- Autonomous Citation Indexing automatically created a citation index that can be used for literature search and evaluation.
- Citation statistics and related documents were computed for all articles cited in the database, not just the indexed articles.
- Reference linking allowing browsing of the database using citation links.
- Citation context showed the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest.
- Related documents were shown using citation and word based measures and an active and continuously updated bibliography is shown for each document.

CiteSeer was granted a United States patent # 6289342, titled "*Autonomous citation indexing and literature browsing using citation context*", on September 11, 2001. The patent was filed on May 20, 1998, and has priority to January 5, 1998. A continuation patent (US Patent # 6738780) was filed on May 16, 2001 and granted on May 18, 2004.

After NEC, in 2004 it was hosted as CiteSeer.IST on the World Wide Web at the College of Information Sciences and Technology, The Pennsylvania State University, and had over 700,000 documents. For enhanced access, performance and research, similar versions of CiteSeer were supported at universities such as the Massachusetts Institute of Technology, University of Zürich and the National University of Singapore. However, these versions of CiteSeer proved difficult to maintain and are no longer available. Because CiteSeer only indexes freely available papers on the web and does not have access to publisher metadata, it returns fewer citation counts than sites, such as Google Scholar, that have publisher metadata.

CiteSeer had not been comprehensively updated since 2005 due to limitations in its architecture design. It had a representative sampling of research documents in computer and information science but was limited in coverage because it was limited to papers that are publicly available, usually at an author's homepage, or those submitted by an author. To overcome some of these limitations, a modular and open source architecture for CiteSeer was designed – CiteSeer^x.

CiteSeer^x

CiteSeer^x replaced CiteSeer and all queries to CiteSeer were redirected. CiteSeer^x[2] is a public search engine and digital library and repository for scientific and academic papers primarily with a focus on computer and information science.[2] However, recently CiteSeer^x has been expanding into other scholarly domains such as economics, physics and others. Released in 2008, it 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. It was developed by researchers Dr. Isaac Councill and Dr. C. Lee Giles at the College of Information Sciences and Technology, Pennsylvania State University. It continues to support the goals outlined by CiteSeer to actively crawl and harvest academic and scientific documents on the public web and to use a citation inquiry by citations and ranking of documents by the impact of citations. Currently, Lee Giles, Prasenjit Mitra, Susan Gauch, Min-Yen Kan, Pradeep Teregowda, Juan Pablo Fernández Ramírez, Pucktada Treeratpituk, Jian Wu, Douglas Jordan, Steve Carman, Jack Carroll, Jim Jansen, and Shuyi Zheng are or have been actively involved in its development. Recently, a table search feature was introduced.[3] It has been funded by the National Science Foundation, NASA, and Microsoft Research.

CiteSeer^x continues to be rated as one of the world's top repositories and was rated number 1 in July 2010.[4] It currently has over 6 million documents with nearly 6 million unique authors and 120 million citations.

CiteSeer^x also shares its software, data, databases and metadata with other researchers, currently by Amazon S3 and by rsync.[5] Its new modular open source architecture and software (available previously on SourceForge but now on GitHub) is built on Apache Solr and other Apache and open source tools which allows it to be a testbed for new algorithms in document harvesting, ranking, indexing, and information extraction.

CiteSeer^x caches some PDF files that it has scanned. As such, each page include a DMCA link which can be used to report copyright violations.[6]

Current features

Automated information extraction

CiteSeer^x uses automated information extraction tools, usually built on machine learning methods such ParsCit, to extract scholarly document metadata such as title, authors, abstract, citations, etc. As such, there are sometime errors in authors and titles. Other academic search engines have similar errors.

Focused crawling

CiteSeer^x crawls publicly available scholarly documents primarily from author webpages and other open resources, and does not have access to publisher metadata. As such citation counts in CiteSeer^x are usually less than those in Google Scholar and Microsoft Academic Search who have access to publisher metadata.

Usage

CiteSeer^x has nearly 1 million users worldwide based on unique IP addresses and has millions of hits daily. Annual downloads of document PDFs was nearly 200 million for 2015.

Data

CiteSeer^x data is regularly shared under a Creative Commons BY-NC-SA license with researchers worldwide and has been and is used in many experiments and competitions.

Thanks to its OAI-PMH endpoint,[7] CiteSeerX is an open archive and its content is indexed like an institutional repository in academic search engines, for instance BASE and Unpaywall consumers.

Other SeerSuite-based search engines

The CiteSeer model had been extended to cover academic documents in business with SmealSearch and in e-business with eBizSearch. However, these were not maintained by their sponsors. An older version of both of these could be once found at BizSeer.IST but is no longer in service.

Other Seer-like search and repository systems have been built for chemistry, Chem$_X$Seer and for archaeology, ArchSeer. Another had been built for robots.txt file search, BotSeer. All of these are built on the open source tool SeerSuite, which uses the open source indexer Lucene.

See also

- Arnetminer
- arXiv
- Collection of Computer Science Bibliographies
- DBLP (Digital Bibliography & Library Project)
- Disciplinary repository
- Google Scholar
- List of academic databases and search engines
- Microsoft Academic
- Research Papers in Economics (RePEc)
- Semantic Scholar

References

1. "CiteSeerX Data Policy" (https://web.archive.org/web/20120105193216/http://csxstatic.ist.psu.edu/about/data). Archived from the original (http://csxstatic.ist.psu.edu/about/data) on 2012-01-05. Retrieved 2015-11-10.
2. "About CiteSeerX" (http://citeseerx.ist.psu.edu/about/site). Retrieved 2010-05-07.
3. "The CiteSeerX Team" (https://web.archive.org/web/20180726034438/http://csxstatic.ist.psu.edu/about/team). Pennsylvania State University. Archived from the original (http://csxstatic.ist.psu.edu:80/about/team) on 2018-07-26.

4. "Ranking Web of World Repositories: Top 800 Repositories" (https://web.archive.org/web/20100724004342/http://repositories.webometrics.info/top800_rep.asp). Cybermetrics Lab. July 2010. Archived from the original (http://repositories.webometrics.info/top800_rep.asp) on 2010-07-24. Retrieved 2010-07-24.
5. "About CiteSeerX Data" (https://web.archive.org/web/20120105193216/http://csxstatic.ist.psu.edu/about/data). Pennsylvania State University. Archived from the original (http://csxstatic.ist.psu.edu/about/data) on 2012-01-05. Retrieved 2012-01-25.
6. For example, "CiteSeerx – DMCA Notice". CiteSeerX 10.1.1.604.4916 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.604.4916). "The document with the identifier "10.1.1.604.4916" has been removed due to a DMCA takedown notice. If you believe the removal has been in error, please contact us through the feedback page, along with the identifier mentioned in this page."
7. Hirst, Author Tony (2011-12-08). "Using OAI-PMH as a Single Record Level Query Interface to Citeseer" (https://blog.ouseful.info/2011/12/08/using-oai-pmh-as-a-query-interface-to-citeseer/). Retrieved 2020-04-25.

Further reading

- Giles, C. Lee; Bollacker, Kurt D.; Lawrence, Steve (1998). "CiteSeer: an automatic citation indexing system". *Proceedings of the Third ACM Conference on Digital Libraries*. pp. 89–98. CiteSeerX 10.1.1.30.6847 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.6847). doi:10.1145/276675.276685 (https://doi.org/10.1145%2F276675.276685). ISBN 978-0-89791-965-4. S2CID 514080 (https://api.semanticscholar.org/CorpusID:514080).

External links

- Official website of CiteSeer^{x} (https://citeseerx.ist.psu.edu/)
- CiteSeerX (https://github.com/SeerLabs/CiteSeerX) on GitHub
- SeerSuite (https://sourceforge.net/projects/citeseerx/) on SourceForge.net (historic)

Retrieved from "https://en.wikipedia.org/w/index.php?title=CiteSeerX&oldid=1001491063"

This page was last edited on 19 January 2021, at 22:49 (UTC).