The Future of CiteSeer: CiteSeer^x

C. Lee Giles

David Reese Professor, School of Information Sciences and Technology
Professor, Computer Science and Engineering
Professor, Supply Chain and Information Systems
The Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

Abstract. CiteSeer, a public online computer and information science search engine and digital library, was introduced in 1997 and was a radical departure from the traditional methods of academic and scientific document access and analysis. Computer and information scientists quickly became used to and expected immediate access to their literature and CiteSeer provided a popular partial solution. CiteSeer was based on these features: actively acquiring new documents, automatic citation indexing, and automatic linking of citations and documents. CiteSeer, now hosted at the Pennsylvania State University with several mirrors, has over 750,000 documents. The current CiteSeer model is to a limited extent portable and was recently extended to academic business documents (SMEALSearch).

Why has CiteSeer been so popular and how should it progress? What is its role with regards to other similar systems such as the Google Scholar and DBLP? What role should CiteSeer play in the open access movement? We discuss this and the Next Generation CiteSeer project, CiteSeerx, which will emphasize CiteSeer as a research tool, research web service, and researcher facilitator and testbed. In contrast to the current tightly integrated CiteSeer architecture, CiteSeerx will be modular, scalable and self managed. We will discuss how new intelligent data mining and information extraction algorithms will provide improved and new indexes, enhanced document access, expanded and automatic document gathering, collaboratories, new data and metadata resources, active mirroring, and web services. As an example of new features, we point out our new API based acknowledgement index and search. This new feature not only provides insight into the impact of acknowledged individuals, funding agencies and others, but also presents an architectural model for integration and expansion of our legacy system.

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, p. 2, 2006.