



Spring 2011 Colloquium

Temple University

Computer and Information Sciences

CiteSeerX and Friends: The Open Source SeerSuite

Lee Giles

Penn State University

Wednesday, 5 / 4, 11am, Wachman Hall 447

Abstract

Cyberinfrastructure or e-science has become crucial in many areas of science as data access often defines scientific progress. Open source systems have greatly facilitated design and implementation and supporting cyberinfrastructure. However, there exists no open source integrated system for building an integrated search engine and digital library that focuses on all phases of information and knowledge extraction, such as citation extraction, automated indexing and ranking, chemical formulae search, table indexing, etc. We propose the open source SeerSuite architecture which is a modular, extensible system built on successful OS projects such as Lucene/Solr and discuss its uses in building enterprise search and cyberinfrastructure for the sciences and academia. We highlight application domains with examples from computer science, CiteSeerX, chemistry, Chem XSeer, and archaeology, ArchSeer. CiteSeerX, the successor to CiteSeer, currently offers or intends to offer some unique aspects of search not yet present in other scientific search services or engines, such as table, figure, algorithm and author search. In addition, CiteSeerX continuously crawls the web and author submissions and now has nearly 1.5 million documents, close to 30 million citations, a million authors and comparable database tables. It has nearly 1 million unique users with several million hits a day. In chemistry, the growth of data has been explosive and timely, and effective information and data access is critical. The ChemXSeer (funded by NSF Chemistry) system is a portal and search engine for academic researchers in environmental chemistry, which integrates the scientific literature with experimental, analytical and simulation datasets. Chem XSeer consists of information crawled from the web, manual submission of scientific documents and user submitted datasets, as well as scientific documents and metadata provided by major publishers. Information gathered from the web is publicly accessible whereas access to restricted resources such as user submitted data will be determined by those users. Thus, instead of being a fully open search engine and repository, Chem XSeer will be a hybrid one, limiting access to some resources. Because such enterprise systems require unique information extraction approaches, several different machine learning methods, such as conditional random fields, support vector machines, mutual information based

feature selection, sequence mining, etc. are critical for performance. We draw lessons for other e- science and cyberinfrastructure systems in terms of design, implementation and research and discuss future directions and systems.

Bio:

Dr. C. Lee Giles is the David Reese Professor at the College of Information Sciences and Technology at the Pennsylvania State University, University Park, PA. He is also graduate college Professor of Computer Science and Engineering, courtesy Professor of Supply Chain and Information Systems, and Director of the Intelligent Systems Research Laboratory. He directs the Next Generation CiteSeer, CiteSeerX project and codirects the ChemXSeer project at Penn State. He has been associated with Columbia University, the University of Maryland, University of Pennsylvania, Princeton University, and the University of Trento.

His current research and research group interests are in intelligent information processing systems such as:

- Intelligent cyberinfrastructure and portals with a special interest in computer and information science, chemistry, biology and archaeology.
- Novel web tools, search engines, scientometrics, web search and measurement.
- Large scale knowledge and information management and extraction, information retrieval, information and data mining, machine learning, digital libraries and web databases, web services and social networks.
- Novel applications and architectures of intelligent information systems.
- Computational issues in e-commerce, the e-world, markets, and betting.
- Business and economic models for search and search engines.

His research is or has been supported by NSF, NASA, DARPA, Microsoft, FAST Search and Transfer, Ford, IBM, Internet Archive, Lockheed-Martin, Alcatel/Lucent, NEC, Raytheon, Smithsonian, US Department of Treasury, and Yahoo. He has consulted for or been on advisory boards of NEC, FAST Search and Transfer, PJM, KXEN, US Department of Treasury, and the US Department of Defense.

He has published over 300 journal and conference papers, book chapters, edited books and proceedings. He has been involved in the creation and development of various novel search engines and digital libraries.

He has served or is currently serving on the editorial boards of IEEE Intelligent Systems, IEEE Transactions on Knowledge and Data Engineering, Machine Learning Journal, Computational Intelligence and Applications, IEEE Transactions on Neural Networks, Journal of Computational Intelligence in Finance, Journal of Parallel and Distributed Computing, Neural Networks, Neural Computation, and Academic Press.

He is a Fellow of the ACM, a Fellow of the IEEE and a Fellow of the International Neural Network Society, and a member of AAI and AAAS. He has twice received the IBM Distinguished Faculty Award. His graduate degrees are from the University of Michigan and the University of Arizona and his undergraduate degrees are from Rhodes College and the University of Tennessee.