

CiteSeer: Past, Present, and Future

Dr. C. Lee Giles

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

Abstract. CiteSeer, a computer science digital library, has been a radical departure for scientific document access and analysis. With nearly 600,000 documents, it has over a million page views a day making it one of the most popular document access engines in computer and information science. CiteSeer is also portable, having been extended to ebusiness (eBizSearch) and more recently to academic business documents (SMEALSearch). CiteSeer is really based on two features: actively acquiring new documents and automatic tagging and linking of metadata information inherent in an academic document's syntactic structure. We discuss methods for providing new tagged metadata and other data resources such as institutions and acknowledgements.