How Accurate is Item-Level Search for Digital Records?

DOUGLAS OARD AND JASON BARON

Abstract: Rapidly expanding digital record collections pose daunting challenges for item-level search, as is required for many eDiscovery and Freedom of Information requests. Digital records also offer opportunities for enhancing search, however, including content-based "full-text" search and statistical modeling of language use. Over the last five years, we have helped lead a global effort to develop ways of answering the question of "how well are we doing." This question is key both because archivists, the judiciary, and other interested parties need that answer now, and because those who seek to build better search technology, or better processes for using technology that we already have, need to know what "better" means. We are working with an international group of experts in technology design, process design, and evaluation design to address this need. We call this the "Legal Track" of the Text Retrieval Conference. To date, that effort has resulted in the creation of two benchmark "test collections" for which evaluation results from dozens of academic and commercial search systems are now publicly available. In this brief talk we will describe the key details of our evaluation process, we will illustrate what we have learned by presenting two key results from participating research teams, we will explain where to find additional details that would be of interest to both researchers and practitioners, and we will explain the relationship between this work and other initiatives such as the Sedona Conference and the Electronic Discovery Reference Model (EDRM).

About the authors:

Douglas W. Oard is a Professor at the University of Maryland, College Park, with joint appointments in the College of Information Studies and the Institute for Advanced Computer Studies. Dr. Oard earned his Ph.D. in Electrical Engineering from the University of Maryland. His research interests center around the use of emerging technologies to support information seeking by end users. His recent work has focused on automated tools for supporting search and sense-making in conversational media (e.g., email, text chat and speech), interactive techniques for cross-language information retrieval, and development of test collections for evaluation of search technology in e-discovery applications in the TREC Legal Track. Additional information is available at http://terpconnect.umd.edu/~oard/.

Jason R. Baron has served as the National Archives' Director of Litigation since May 2000. Prior to his appointment as Director of Litigation, Mr. Baron held successive positions as trial attorney and senior counsel with the Federal Programs Branch of the Civil Division of the Department of Justice, where he represented the Archivist and various Executive Office of the President components in Armstrong v. Executive Office of the President (the PROFS case) and Public Citizen v. Carlin (the GRS 20 case). As NARA's representative to The Sedona Conference®, Mr. Baron currently serves as Co-Chair of the Working Group on Electronic Document Retention and Production. Mr. Baron was a founding coordinator of the TREC Legal Track, and he currently serves as an Adjunct Professor in the University of Maryland's College of Information Studies.