

Computer Assisted Appraisal of Contemporary PDF Documents

PETER BAJCSY and SANG-CHUL LEE

Abstract: In our research we are addressing some of the challenges of computer-assisted appraisal of contemporary documents in portable document format (PDF). The motivation of our work is to provide archivists with tools that will assist them in making appraisal decisions when confronted with large collections of documents in the PDF format.

We examined archival appraisal criteria. This examination led us to investigations related to (a) extracting content from PDF documents, (b) finding groups of PDF documents with similar content, (c) ranking documents according to their creation/modification time and digital volume, and (d) detecting inconsistency between ranking and content within a group of related documents. Our focus is on contemporary PDF documents that contain images, text and graphics objects. We have designed a framework for computer assisted appraisal of PDF documents that relates PDF documents based on the similarity of contained text and images, ranks documents chronologically and verifies the integrity of chronological versions.

The novelty of our work is in designing a methodology and a mathematical framework for document appraisals, as well as in prototyping the framework working with image and text components of PDF documents. We will present example results of automated content extraction, grouping, ranking and integrity verification for PDF documents coming from multiple domains.

About the authors:

Peter Bajcsy's research revolves around theoretical modeling and experimental understanding of eScience components including automation and design of computer-assisted systems that deal with large volumes and computational intensive processing of heterogeneous data. Based on my past work, my research could also be described as X-informatics, where the X stands for document, hydro, geo, bio, medical image, or sensor.

Sang-Chul Lee, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign.