



Spring 2012 Colloquium

Temple University

Computer and Information Sciences

Searching in the "Real World"

Distinguished Speaker

Ophir Frieder

Georgetown University

Wednesday 2/15, 11am, Wachman 447

Abstract:

For many, "searching" is considered a mostly solved problem. In fact, for text processing, this belief is factually based. The problem is that most "real world" search applications involve "complex documents", and such applications are far from solved. Complex documents, or less formally, "real world documents", comprise of a mixture of images, text, signatures, tables, etc, and are often available only in scanned hardcopy formats. Search systems for such document collections are currently unavailable.

We describe our efforts at building a complex document information processing prototype. This prototype integrates "point solution" (mature) technologies, such as document readability enhancement, OCR capability, signature matching and handwritten word spotting techniques, search and mining approaches, among others, to yield a system capable of searching "real world documents". The described prototype demonstrates the adage that "the whole is greater than the sum of its parts". Our complex document benchmark development efforts are likewise presented.

Having described the global approach, we describe some point solutions which we developed over the years. These include an image enhancer, an Arabic stemmer, and a natural language source integration fabric called the Intranet Mediator.

Bio:

Ophir Frieder holds the Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt L.C.H.S. Chair in Computer Science and Information Processing and is Chair of the Department of Computer Science at Georgetown University. He holds a courtesy membership in the Georgetown University Lombardi Comprehensive Cancer Center. He is a Fellow of the AAAS, ACM, and IEEE.