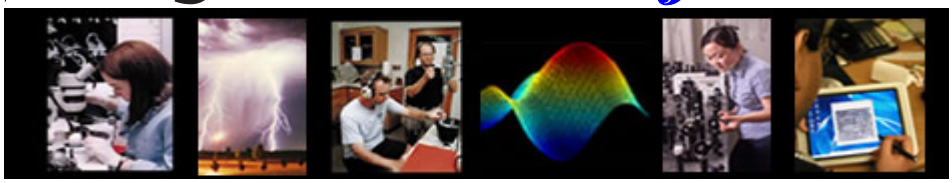
Discovery Park

Cyber Center



DISTINGUISHED LECTURE

April 19, 2010

12:00- 1:30PM

Lunch Provided

Lawson 3102

Ophir Frieder

Professor and Chair of the Department of Computer Science at Georgetown University

'Searching in the "Real World"'

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) echnologies, such as 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 potential future point solutions which we have developed over the years. These include an Arabic stemmer and a natural language source integration fabric called the Intranet Mediator. In terms of stemming, we developed and commercially licensed an Arabic stemmer and search system. Our approach was evaluated using the benchmark Arabic collections and favorably compared against the state of the art.

We also focused on source integration and ease of user interaction. By integrating structured and unstructured sources, we developed and commercially licensed our mediator technology that provides a single, natural language interface to querying distributed sources. Rather than providing a set of links as possible answers, the described approach actually answers the posed question. Both the Arabic stemmer and the mediator efforts are likewise discussed.

BIO: Ophir Frieder is Professor and Chair of the Department of Computer Science at Georgetown University. His research interests focus on scalable information retrieval systems spanning search and retrieval and communications issues. He frequently consults for industry and government and for key intellectual property litigation; his systems are deployed in commercial and governmental production environments worldwide. In 2007, Springer Science and Business Media designated his co-authored book entitled "Information Retrieval: Algorithms and Heuristics" with the "Top Selling Title" award. He is the recipient of the 2007 ASIS&T Research in Information Science Award and a recipient of the 2008 IEEE Technical Achievement Award. He is a Fellow of the AAAS, ACM, and IEEE.

Please Distribute to Faculty