The Information Science & Technology Center

ISTeC.ColoState.edu

Colorado State University's
Information Science and Technology Center (ISTeC)
presents <u>two</u> lectures by

Dr. Ophir Frieder

IITRI Professor of Computer Science
Director, Information Retrieval Laboratory
Illinois Institute of Technology

ISTeC Distinguished Lecture

in conjunction with the
Electrical and Computer Engineering Department and
Computer Science Department Seminar Series

"Searching in the 'Real World'"

Monday, April 20, 2009
Reception: 10:30 a.m.
Lecture: 11:00 – 12:00 noon
Location: Lory Student Center Grey Rock Room

•

Special Electrical and Computer Engineering Seminar

sponsored by ISTeC

"Information Retrieval Technology in Peer-to-Peer Search"

Tuesday April 21, 2009
Lecture: 11:00 a.m. – 12:00 p.m.
Location: Lory Student Center room 214

ABSTRACTS

"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 that integrates "point solution" (mature) technologies, such as optical character recognition signature matching and handwritten word spotting techniques, and 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 also are presented.

Having described the global approach, we describe some potential future point solutions we 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 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.

"On Exploiting Information Retrieval Technology in Peer-to-Peer Search"

Peer-to-peer file sharing applications are highly popular; in fact, their bandwidth demands are the greatest among today's Internet activities. Given their heterogeneous, autonomous node structure, we envision data sharing over a diversity of domains. An existing obstacle, however, is the low sustained search accuracy.

Towards designing meaningful approaches, we analyzed a peer-to-peer file sharing query log obtained by our query analysis tool. We then designed solutions that efficiently improve search accuracy in peer-to-peer file-sharing systems. Our solutions leverage the traditional focus of our Information Retrieval Laboratory. We highlight several of our metadata management and descriptor enrichment solutions in addition to novel applications of traditional information retrieval techniques. Overall, we improve search accuracy by 30% with only a marginal increase of network resources. We also provide a protocol that sustains a 15% accuracy improvement with a similar reduction in network resource demands.

Finally, we conclude with a brief discussion of P2P spam filtration. Initially, we characterize the common types of spam. We then develop an approach based on strictly local information to reduce spam without excessively taxing the network.

SPEAKER BIOGRAPHY

Dr. Ophir Frieder (http://www.ir.iit.edu/~ophir/), a Fellow of the AAAS, ACM, and IEEE, is the IITRI Professor of Computer Science and Director of the Information Retrieval Laboratory at the Illinois Institute of Technology. His research focuses on scalable distributed information systems.

To arrange a meeting with the speaker, please contact MaryAnn Stroub at (970) 491-2708 or mstroub@engr.colostate.edu.

ISTeC (Information Science and Technology Center) is a university-wide organization for promoting, facilitating, and enhancing CSU's research, education, and outreach activities pertaining to the design and innovative application of computer, communication, and information systems. For more information please see ISTeC.ColoState.edu.