Searching in the “Real World”

(Abstract of Invited Plenary Talk)

Ophir Frieder

Information Retrieval Laboratory
Department of Computer Science
Illinois Institute of Technology
ophir@ir.iit.edu

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, logos, water-marks, stamps, 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 (CDIP) prototype. This prototype integrates "point solution" (mature) technologies, 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".

To evaluate our CDIP prototype as well as to provide an evaluation platform for future CDIP systems, we also introduced a complex document benchmark. This benchmark is currently in use by the National Institute of Standards and Technology (NIST) Text REtrieval Conference (TREC) Legal Track. The details of our complex document benchmark are similarly presented.

Having described the global approach, we describe some additional point solutions developed in the IIT Information Retrieval Laboratory. 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 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 designed, implemented, 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.

A summary of the efforts discussed is found in [1].

Reference

1. Frieder, O.: On Searching in the 'Real World'. In: Argamon, S., Howard, N. (eds.) Computational Methods for Counterterrorism, ch. 1. Springer, Heidelberg (2009) ISBN: 978-3-642-01140-5