Relazione scientifica sui risultati dell'attività di ricerca

svolta dal Prof. Ophir Frieder presso ISTI-CNR
Programma STM 2010
Proponente: Raffaele Perego, ISTI-CNR

Il Prof. Ophir Frieder, *Chair* del *Department of Computer Science at Georgetown University*, ha visitato il laboratorio HPC di ISTI-CNR nell'ambito del programma Short Term Mobility 2010, che ha finanziato la sua visita in base alla proposta presentata dal sottoscritto nel settore delle soluzioni efficienti e scalabili per il recupero di informazioni da grandi collezioni di dati testuali.

Durante il suo soggiorno in Italia, il Prof. Ophir Frieder ha collaborato con il proponente e con i ricercatori del gruppo HPC di ISTI-CNR su tematiche riguardanti gli algoritmi efficienti per il recupero di informazioni testuali da grandi collezioni di dati per mezzo di architetture avanzate di tipo multi-core. In questo ambito si sono pianifiche alcune soluzioni promettenti per l'estrazione efficiente e scalabile di *snippet query-based* per la presentazione dei risultati nei motori di ricerca Web, e per la classificazione di dati testuali e multimediali tramite l'impiego di folksonomy. Ci aspettiamo che le idee sviluppate abbiano importanti ricadute scientifiche e conducano a pubblicazioni congiunte di alto livello.

Nei giorni trascorsi presso ISTI il Prof. Ophir Frieder ha avuto incontri singoli di almeno un ora di durata con i giovani ricercatori ed tutti i nove studenti di dottorato afferenti al Laboratorio che coordino. In questi incontri, importantissimi per la loro valenza formativa, i giovani hanno descritto e discusso la loro attività di ricerca, ricevendo utili consigli e spunti di riflessione.

Il giorno 15 Giugno, presso l'aula 27 dell'area della ricerca di Pisa del CNR, il Prof. Frieder ha tenuto un seminario che ha visto la partecipazione di un numeroso pubblico comprendente anche ricercatori di altri Istituti CNR (IIT e ILC) e dei dipartimenti di Informatica e di Ingegneria Informatica dell'Università di Pisa. L'abstract del seminario è allegato al presente documento.

La collaborazione instaurata grazie al finanziamento ottenuto è molto importante per il nostro gruppo. Credo che essa si concretizzerà presto in pubblicazioni congiunte (un primo lavoro a firma congiunta è gia stato sottomesso per la pubblicazione negli atti di una conferenza internazionale prestigiosa), e nella possibilità di inviare nostri dottorati per periodi di specializzazione presso la Georgetown University.

Pisa, lì 30 Giugno 2010

Raffaele Perego

Raffaele Perego

Allegato 1.
Abstract del seminario: **'Searching in the "Real World"'**, tenuto presso l'area della ricerca di Pisa dal Prof Ophir Frieder, Chair del Department of Computer Science at Georgetown University

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 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.

RP