RIASSUNTO

Titolo tesi: "*Analisi e sviluppo di un crawler per la creazione di una base di conoscenza semantica del personale universitario*"

Autore: Martina Stefano (`stefano.martina@gmail.com`)
Relatore: Barcucci Elena (`barcucci@dsi.unifi.it`)
Corelatore: Nesi Paolo (`nesi@dsi.unifi.it`)

Il testo della tesi è il risultato del lavoro di *stage* svolto presso il laboratorio *DISIT* del Dipartimento di Sistemi e Informatica dell'Università degli Studi di Firenze.

Il lavoro effettuato si inserisce nel progetto *OSIM* che ha l'obiettivo di creare una base di conoscenza semantica riguardante il personale universitario. In particolare è stata sviluppata la parte dell'applicazione che riguarda il *data mining* delle informazioni riguardanti le persone, i corsi associati a queste, e le competenze di ogni persona.

Il progetto fa uso delle tecnologie *RDF*, *RDFS* e *OWL* per l'organizzazione della base di conoscenza semantica. Per quanto riguarda l'applicazione, questa è sviluppata in *Java* facendo uso delle *Servlet* e delle *JSP* per fornire l'interfaccia *web* con l'utente.

Il testo è diviso in due parti: una parte consiste in una sostanziosa lettura sullo stato dell'arte delle tecnologie usate, dalle basi dell'*XML* e delle tecnologie correlate sino ad una visione eterogenea sulle tecnologie del web semantico; l'altra parte concerne la descrizione del lavoro effettuato e un'analisi dettagliata dell'applicazione nelle sue parti.

Il *crawler* è costituito da due parti, una fase di estrazione delle parole chiave ed una fase di estrazione di competenze composte da più parole chiave. Entrambe le fasi vengono eseguite sulle pagine del portale *cercachi* del *web* dell'università.

Per le parti di *NLP* (*Natural Language Processing*) l'applicazione si appoggia al *framework GATE* sviluppato dall'Università di Sheffield.

Gli aspetti del lavoro svolto nell'arco della durata dello *stage* sono stati molteplici, vi sono state fasi di *debug* del codice preesistente, fasi di studio e ricerca in diversi campi e fasi di sviluppo di nuove funzionalità.

Per quanto riguarda lo studio sono state considerate principalmente le tecnologie legate a *XML*, al web semantico, e al *NLP*. Lo sviluppo si è concentrato sul linguaggio *Java*, sulle pagine *JSP*, e sulle *pipe* di *GATE*.

Lo *stage* è stato svolto principalmente nei locali del laboratorio *DISIT* ed è stato coadiuvato dai ricercatori presenti in un ambiente familiare e costruttivo di cooperazione.