Targeted Crawling and Automatic Extraction of Structured Data

Résumé du projet de recherche (Langue 1)

The Web contains a large amount of useful information and resources. It is expanding rapidly, and changing from a pure document collection to a large space of connected public data. The structure and semantic annotations of the web pages make it easier to search, extract and mine the web content. Focused crawling enables to selectively find Web pages relevant to a specific domain, and the aim of automatic wrapper generation is to map, in an unsupervised way, as accurately as possible, Web page content to relational-style tables. These two techniques are essential for automatic data extraction. In this work, we are interested in targeted data extraction, in the sense that the focus is on structured data linked to a target object, may it be simple or complex. In this case, the crawling and the wrapper generation phases have to be rethought based on the input object description. In ObjectRunner, we demonstrated the accuracy of our approach for events wrapper induction. We plan, in this thesis, to (1) concentrate our effort on the crawling step, based on the recent advances on focused crawling, and (2) broaden the scope of our previous work to more categories of real world objects.