Extracting and Archiving Rich Content from the Web: the ARCOMEM approach

Mots clés :

- Directeur de thèse : BOGDAN CAUTIS
- Co-encadrant(s) :
- Unité de recherche : Laboratoire Traitement et Communication de l'Information
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

The goal of this thesis is to develop new algorithms and advanced practical solutions for the extraction and archiving of rich content from the Web. In the context of the ARCOMEM project, archivists will specify what information should be targeted in Web pages (the goal). A goal-driven crawling and extraction process will then harvest the most relevant information, at the level of semantic objects. This will require for instance effective techniques for prioritizing the sources to be crawled, as well as powerful facilities for the targeted extraction of complex objects from the Web. We plan to apply such techniques on both deep Web (information behind forms, Web services) – based on result page analysis - and surface Web pages. Ultimately, the goal is to decompose the relevant content into semantic units, and store the information about individual Web objects, allowing them to be searched and analyzed independently. To this end, we also intend to organize and enrich the archive with semantic information in order to facilitate access. Integrating information extraction in real world archiving applications, based on a crawl and content specification, via various parameters (e.g., search strings) that give a semantic focus/context, requires a significant departure from existing techniques, which mostly consider archiving and extraction in separation and rarely use external semantic information (such as ontologies).

Résumé du projet de recherche (Langue 2)

- archivage - extraction - passage à l'échelle