Intelligent Content Acquisition in Web Archiving

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

Résumé du projet de recherche (Langue 1)

Background ---------- Since its beginnings in the 1990s, the set of hyperlinked resources known as the World Wide Web has become an extremely precious source of information. For easy access to this content, search engines such as Google or Bing crawl the Web and index pages. But crawling technologies also have an important application in the field of Web archiving [1]: the storage of (versioned) archives of the Web, to ensure continuous access to its resources despite data deletion, to carry out temporal analyzes of its content, or to provide a user the possibility to see a Web page or a Web site as it was at a given date in the past. This PhD topic is part of the ARCOMEM European project (2011-2013). This project, involving 12 institutions around Europe, aims at leveraging the wisdom of the crowds for intelligent preservation of Web content. PhD Topic Description

The purpose of this PhD is to develop intelligent and adaptive models, methods, algorithms, and tools for making the content acquisition process in Web archiving more effective and efficient. The objective is to leverage existing work (in particular inside the ARCOMEM consortium) on the extraction and analysis of events, entities, topics, opinions, perceptions, etc., to select and prioritize sources to be crawled. This will be combined with techniques that go beyond traditional page-level crawling, allowing object-level, goal-driven crawling of the Web. The focus is on unsupervised methods, that can scale to the whole Web. In particular, the following aspects will be covered: - assessing the relevance [2], importance, and coverage of available content with respect to a Web archiving task; - combining evidence to select or prioritize the crawling process [3]; - accessing content at the level of objects inside Web pages [4] and hidden behind deep Web forms [5] or Web 2.0 applications. Depending on the content to be archived (social networks, structured Web, deep Web, etc.), different solutions for assessing relevance, prioritizing the crawling, and extracting Web objects can be proposed.