WP5: Intelligent Content Acquisition

Pierre Senellart, Télécom ParisTech

January 25, 2011

General Principles

Make the crawling process more intelligent by:

- Isolating parts of Web pages (Web objects) as atomic items to archive
- Crawling more complex forms of Web content (deep Web, Web applications)
- Guiding the crawl by looking at the content of Web pages at crawl time and automatically selecting relevant information

Partners Involved

- Télécom ParisTech (leading)
- Athena RC
- Internet Memory Foundation
- L3S

Crawling at the Level of Web Objects

- Go down into the DOM tree of a page and consider various blocks as candidate for individual archiving
- In parallel with more classical archiving of whole Web pages
- Extract semantic information (timestamp, author, etc.) from these individual Web pages, from the content of the objects or external information (e.g., RSS feeds)
- Store all of this in a semantic archive of objects
- Good examples: news items, blog posts, forum messages, etc.
- Maximize the level of automation; no human-written wrapper!

Crawling the Dynamic Web

- Archive data that is beyond forms (deep Web)
- Archive data that is retrieved through AJAX calls
- Archive Web applications (social networks, Web mail software, forum, etc.) in a structured manner through the design of a specification formalism of what to crawl, and where to store the content

Relevance, Coverage, Importance

Assess the interest of a source (Web page, Web site) for an archiving task defined through examples, keywords, or semantic concepts:

Relevance and Interest Is a source (its entities, its relations, etc., as extracted by other Web packages) relevant to the task? Go beyond IR-based notions of relevance (e.g., distance) and into more formal notions (e.g., minimum-length description)

Importance Is this source important (cf. PageRank)? For instance, is this Twitter account important enough to use crawling resources on?

Coverage Should this source be added to ensure coverage of the archive. Can this be measured? See also coverage of deep Web crawling.

Intelligent Selection of Web Content

- Automatic classification of Web content at crawl time
- Use of this classification to select content to archive, and to select URLs to add to the crawling frontier
- Automatic assessment of relevance, importance, coverage, and use of these to design an adaptive crawler
- Prioritization vs Selection
- cf. focused crawling, topical crawling

Outcomes and Indicators

Outcomes A crawler implementation supporting:

- Ranking of Web sources wrt relevance, importance, coverage
- Adaptive and prioritized crawling strategies
- Web object archiving from the surface Web, deep Web, Web applications

Indicators

- Support of the three features above
- Workability of the implementation
- Scenarios the system work with
- Publications