Bringing Magic to Web Archive Creation

The L3S Research Center, under supervision of Prof. Dr. Nejdl, offers a M.Sc. thesis “A Wizard for Web Crawl Specifications” in the scope of the project iCrawl: The Integrated Focused Crawling ToolBox.


Focused Web Crawlers allow the creation of targeted Web datasets for future Web Science research. A major obstacle to their widespread use is however the low usability of current systems. Existing Web crawlers are designed to be configured and controlled by expert users. This requires extensive configuration using text files and manual control, often using command line interfaces. This causes a larger barrier of entry for end users interested in collecting relevant Web content such as researchers or journalists. The goal of this thesis is therefore to provide an easy-to-use graphical interface to specify the scope and focus of a crawl, as well as an automatic scheduling of crawls and crawling sub-tasks.

The aim of the M.Sc. thesis is the development of a wizard-like interface that allows the user to specify a crawl based on their expected outcome. To achieve this goal the following tasks need to be addressed:

- **Analysis of typical Web crawls** to extract the relevant dimensions, such as exhaustive vs. focused, complete Web pages vs. only textual content, or prioritized vs. diversified. The result of this task will be a model of the configuration settings.

- **Identifying good starting points**: As a high-quality Web crawl depends on having good starting points (so called seeds) for the crawl, different ways to bootstrap the initial seed list need to be investigated. Possible methods are to query a Web search engine, to use curated topical lists of Web sites such as DMOZ.org, or to monitor social media platforms such as Twitter to find current links for given keywords.

- **Wizard-like interface creation**: Based on the created model a user interface will be created that allows the user to specify a crawl based on their expected outcome. This crawl specification is then automatically translated into the configuration format of the crawler.

You should have:

- Programming skills in Java
- Basic knowledge of text mining or machine learning methods
- Interest in user interface design
- Independent thinking and willingness to learn

Are you interested or have questions? Contact us!

Gerhard Gossen, email: gossen@L3S.de, or Thomas Risse, email: risse@L3S.de

Forschungszentrum L3S, Appelstr. 9a, 30167 Hannover