UNIVERSITY OF TURKU
Department of Information Technology

JYRI LEHTONEN: Characterizing the deep Web

Master's thesis, 91 p., 18 supplementary p.
Computer Science
December 2011

The deep Web is a section of the Web. It contains over 99% of the Web, yet the major search engines cannot access it efficiently. Defining the difficulties of automatically accessing the deep Web is the main research topic of this thesis. Two other thesis statements are also taken into account. First, we create an estimation of the current deep Web scale concluded from the previous scale-related research. Second, we address the problem of querying multiple deep Web resources "on the fly".

The thesis is written in a pattern of first explaining what is currently considered as part of the deep Web content (i.e. dynamic pages, limited content and databases). This content is the part of the Web that the search engines do not index well, therefore users cannot query a general search engine index to find this data.

After the deep Web content has been defined, the thesis illustrates the data gathering techniques of the Web (i.e. screen scraping and APIs). These techniques are used when one creates a system that collects information from other servers (i.e. a Web crawler).

The final part of this thesis concentrates on how to query a deep Web resource, using theoretical and practical examples. The research is concluded with a project: the Scavenger Crawling System. The system was written for this thesis using the methods and techniques explained throughout this research. The project was a success: The system can query a deep web resource that is not predefined for it.

Keywords: Deep Web, Deep Web Navigation, Deep Web Analysis, Deep Web Crawling, Metasearch Engine, HTML Form Analysis, Web Crawler