Extraction of Symbolic Patterns from the Web

Geisberger, Bernhard

Web scraping has become increasingly important in recent years. It is part of the branch of data analysis, search engines and data collection programs. The need for web scraping programs is rising in a variety of industries. This project treats web scraping in a very special way. With experimental prototyping individual algorithms are developed to learn structures within web documents by machine learning. These experiments should lead to one resulting, comprehensive algorithm which is able to extract data out of any SGML-based document by identifying its particular structure. Specifically, the algorithms in this project are based on a theoretical paper by Dana Angluin about minimal, zero-reversible finite deterministic automata. To improve this process, heuristic measures are used prior to the identification of a structure. These measures will alter the contents of the input documents to optimize further processing. Already after a few experiments the project was bearing its first fruit. With continuous improvements, the algorithm was refined. Finally, it was possible to extract data out of about 80 per cent of the test websites without any human interaction. In conclusion, the project was quite successful and has shown several new and promising ways of web scraping.