Motivation

How do I get the soccer result as soon as possible?

“A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.” (Kobayashi and Takeda, 2000)

Evaluation Process

Tasks
- Creation of Web Scrapper Description
- Creation of Functional Specification
- Creation of Benchmark Catalogue
- Comparison of Existing Web Scrapers
- Validation which Scrapper is applicable in which areas
- Evaluation which Scrapper fits the need the best
- Definition of further research assignments

Output
- Web Scrapper Description
- Functional Specification
- Benchmark Catalogue
- Scrapper Comparison
- Benchmark Results
- Topics for further research assignments

General Strengths
- Good for gathering lots of information
- Various output formats of the crawled information
- Good support for different forms of navigation
- Quite easy to use
- Good support for proxy servers

General Weaknesses
- No quality assurance
- No intelligent adjustment
- Problems with extracting information from flash pages
- No prevention algorithm against crawler traps
- No support for CAPTCHAs
- Not possible to respect the robot exclusion standard

Comparison of the evaluated Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Good for gathering lots of information</th>
<th>Various output formats of the crawled information</th>
<th>Good support for different forms of navigation</th>
<th>Quite easy to use</th>
<th>Good support for proxy servers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tool 1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Tool 2</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tool 3</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Results based on the Evaluation

Unstructured information can be scraped and saved in an appropriate file format like XML.

Case Study

“A more recent variant of the web crawler is the web scraper, which looks for certain kinds of information - prices of particular goods from various online stores for instance - and then aggregates it into new web pages, often using a database.” (Adams and McCrindle, 2007)

Is it possible to use a web scraper to gather information about sport fixtures and results?

Conclusion and Further Work

General
- Scrapers are good for collecting data – but for nothing else
- Works smoothly as long as the structure of the website does not change
- Minimization of the available risks by configuring different web sites
- Data processing and quality assurance system needs to be programmed
- Configuration/Maintenance effort vs. benefit?
- No assurance that quality is better than from feed providers

Open Questions
- How often do big sporting websites change their structure?
- How many sports and leagues should be supported?
- Is there a data processing and quality assurance system which could be used

Further Application Areas
- Fixtures/Results
- Names of Players
- Live bet analysis of competitors (coverage, availability,..)
- Odds comparison

Kontakt: a.mehlfuehrer@gmail.com