WebDB

General Information

- Betreuer: Sebastian Maneth
- als BS und MS Projekt geplant
- Studiengänge: Inf & WInf
- Masterprofilvorbereitung: SQ & KIKR
- Begleitende LVs: keine

Topics

- Deep Web
- Web Data Extraction
- Wrapper Induction

Description

The web provides access to gigantic amounts of structured information. Very often, however, the access to such databases is given only through a custom web interface. Such web interfaces let us formulate only very simple queries. For instance, we are *not* able to ask at `imdb.com`:

```
find the average length of all movies made by "Werner Herzog".
```

Or, using the web interfaces of property pages such as `immobilienscout24.de` we are *not* able to ask:

```
find the the average price of apartments in "Bremen".
```

The aim of this project is to build a tool that is able to **extract** structured information from web pages. For instance, we would like to extract from a movie web page of `imdb.com` the director of the movie. Such extraction *cannot* be done on the level of the visual (textual) representation of the web browser. Instead, we must tap into the "deep web", i.e., we must extract from the HTML source. Interestingly, the information about the director of a movie can be at a very deep location of the HTML source (tree): at level of 50 or even deeper. The reason for this is that a large part of the design of the web page is captured on the path from the root of the HTML tree to the node that carries the information about the director of the movie. The big challenge in building "web wrappers" (i.e., pieces of

software that extract information from web pages) is that on the HTML source of another movie, the director information may be at a completely different level. This is due to the fact that the web page contains advertisements which change from page to page. Thus, the wrapper must ignore such advertisements and needs to focus on the immediate vicinity of the node that needs to be extracted.

Within the Diadem project [1] that was carried out at the University of Oxford, a tool was built that is able to generate (from web pages that are annotated with the information that shall be extracted) wrappers of very high quality. We would like to reimplement and improve this tool [2]. The software of this tool is currently owned by Meltwater[1] but will most probably be made available for this project. Using this tool, we would like to generate rich databases from web pages, by systematically extracting information from their web pages.

PROs:

- very clear and precise project goal
- demanding programming tasks
- at the forefront of current research about web extraction
- knowledge gained in this project will be highly interesting for industry

CONs:

- possibly too challenging for the limited time and manpower
- outcome heavily depends on programming skills of participants.

References

[1] Tim Furche, Georg Gottlob, Giovanni Grasso, Xiaonan Guo, Giorgio Orsi, Christian Schallhart, and Cheng Wang. DIADEM: thousands of websites to a single database. *PVLDB*, 7(14):1845–1856, 2014.

[2] Tim Furche, Jinsong Guo, Sebastian Maneth, and Christian Schallhart. Robust and noise resistant wrapper induction. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 773–784, 2016.

[1] `https://www.meltwater.com/de/`