Exploring the Application Potential of Relational Web Tables

Christian Bizer
University of Mannheim, Germany
Research Group Data and Web Science
chris@informatik.uni-mannheim.de

The Web contains large amounts of HTML tables. Most of these tables are used for layout purposes, but a small subset of the tables is relational, meaning that they contain structured data describing a set of entities [1]. Relational web tables cover a wide range of topics and there is a growing body of research investigating the utility of web table data for applications such as complementing cross-domain knowledge bases [2], extending arbitrary tables with additional attributes [13, 4], and translating data values [9].

Until recently, most of the research around web tables originated from the large search engine companies as they were the only ones having access to large web crawls and thus were able to extract web table corpora from the crawls. This situation has changed in 2012 with the University of Mannheim [7] and in 2014 with the Dresden University of Technology [3] starting to extract web table corpora from the CommonCrawl, a large public web corpus.

In the talk, I will introduce the 2015 version of the Web Data Commons - Web Table Corpus [7]1. Afterward, I will give an overview of the different efforts that are currently conducted by my group on exploring the application potential of relational web tables. These efforts include profiling the content [12, 6] of web tables by matching [11] them to cross-domain knowledge bases such as DBpedia [5], fusing web table data in order to complement cross-domain knowledge bases [10], and performing SearchJoins between a local table and a web table corpus in order to extend the local table with additional attributes [8].

References


1 http://webdatacommons.org/webtables/