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. Michael Cafarella, Alon Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. Uncovering the Relational Web. In *Proceedings of the 11th International Workshop on Web and Databases*, 2008.
2. Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proc. of the 20th SIGKDD*, pages 601–610, 2014.
3. Julian Eberius, Katrin Braunschweig, Markus Hentsch, Maik Thiele, Ahmad Ahmadov, and Wolfgang Lehner. Building the dresden web table corpus: A classification approach. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, pages 41–50, 2015.

[1] http://webdatacommons.org/webtables/

4. Julian Eberius, Maik Thiele, Katrin Braunschweig, and Wolfgang Lehner. Top-k Entity Augmentation Using Consistent Set Covering. In *Proc. of the 27th Int. Conf. on Scientific and Statistical Database Mgmt*, 2015.
5. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, 6:167–195, 2015.
6. Oliver Lehmberg and Christian Bizer. Web table column categorisation and profiling. In *Proceedings of the 19th International Workshop on Web and Databases*, page 4, 2016.
7. Oliver Lehmberg, Dominique Ritze, Robert Meusel, and Christian Bizer. A large public corpus of web tables containing time and context metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 75–76, 2016.
8. Oliver Lehmberg, Dominique Ritze, Petar Ristoski, Robert Meusel, Heiko Paulheim, and Christian Bizer. The Mannheim Search Join Engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:159–166, 2015.
9. John Morcos, Ziawasch Abedjan, Ihab Francis Ilyas, Mourad Ouzzani, Paolo Papotti, and Michael Stonebraker. Dataxformer: An interactive data transformation tool. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 883–888, New York, NY, USA, 2015. ACM.
10. Yaser Oulabi, Robert Meusel, and Christian Bizer. Fusing time-dependent web table data. In *Proceedings of the 19th International Workshop on Web and Databases*, page 3, 2016.
11. Dominique Ritze, Oliver Lehmberg, and Christian Bizer. Matching HTML Tables to DBpedia. In *Proc. of the 5th Int. Conf. on Web Intelligence, Mining and Semantics*, 2015.
12. Dominique Ritze, Oliver Lehmberg, Yaser Oulabi, and Christian Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 251–261, 2016.
13. Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012.