A Large Public Corpus of Web Tables containing Time and Context Metadata

Oliver Lehmberg, Dominique Ritze, Robert Meusel, Christian Bizer
Data and Web Science Group, University of Mannheim, Mannheim, Germany
{oli,dominique,robert,chris}@informatik.uni-mannheim.de

1. INTRODUCTION

The Web contains vast 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 [2]. As these relational Web tables cover a very wide range of different topics, there is a growing body of research investigating the utility of Web table data for completing cross-domain knowledge bases [6], for extending arbitrary tables with additional attributes [7, 4], as well as for translating data values [5]. The existing research shows the potentials of Web tables. However, comparing the performance of the different systems is difficult as up till now each system is evaluated using a different corpus of Web tables and as most of the corpora are owned by large search engine companies and are thus not accessible to the public.

In this poster, we present a large public corpus of Web tables[1] which contains over 233 million tables and has been extracted from the July 2015 version of the CommonCrawl. By publishing the corpus as well as all tools that we used to extract it from the crawled data, we intend to provide a common ground for evaluating Web table systems.

The main difference of the corpus compared to an earlier corpus[2] that we extracted from the 2012 version of the CommonCrawl as well as the corpus extracted by Eberius et al. [3] from the 2014 version of the CommonCrawl is that the current corpus contains a richer set of metadata for each table. This metadata includes table-specific information such as table orientation, table caption, header row, and key column, but also context information such as the text before and after the table, the title of the HTML page, as well as timestamp information that was found before and after the table. The context information can be useful for recovering the semantics of a table [7]. The timestamp information is crucial for fusing time-depended data, such as alternative population numbers for a city [8].

[1] `http://webdatacommons.org/webtables/#results-2015`
[2] `http://webdatacommons.org/webtables/#results-2012`

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889386 .

2. EXTRACTION PROCESS

We extract our Web tables corpus from the July 2015 version of the CommonCrawl. This public web crawl contains more than 1.78 billion HTML pages originating from over 15 million websites. The extraction is performed using the *WebDataCommons* (WDC) extraction framework.[3] The current version is an extension of the version used in [3], which was itself based on an earlier version of the WDC framework. Its main purpose is the efficient extraction of tables from HTML pages and their classification as *layout* table or specific type of *content* table [1]. For the current release, we extended the framework with methods for determining the orientation of tables, the header row, and the key column, as well as methods for extracting timestamp and context information from the HTML page.[4]

The extraction process results in 233 million content tables (2.25% of all tables) which are classified as either *relational* (90 million), *entity* (139 million), or *matrix* (3 million). Relational tables contain entities which are described by several attributes, both of which can either be represented by rows or columns. In entity tables, attributes characterize one single entity where the name of the entity is usually not contained in the table but may be found in the table context. Matrix tables are usually used for statistics and contain numbers that relate to the dimensions given in the column and row headers.

3. CORPUS STATISTICS

In the following, we provide statistics about the tables that have been classified as *relational*. A relational table is called *horizontal* if entities are represented in rows and attributes in columns, and it is called *vertical* if it is the other way around. We did not directly transpose vertical tables during the extraction to not exclude any kind of use case, e.g. for a specific application only vertical tables might be interesting. As out of the 90 million relational tables 84 million are horizontal, we focus in the following on this type of table. On average, horizontal tables have 5.2 columns (attributes) and 14.45 rows (entities) with a median of 4, respectively 6. Figure 1 shows the distribution of tables having a certain number of rows and columns. Only a very small fraction of tables consists of a larger number of columns (max. 18,106) or rows (max. 17,033).

[3] `http://webdatacommons.org/framework/`
[4] Due to space constraints we refer the interested reader to our webpage for details: `http://webdatacommons.org/webtables/index.html`.

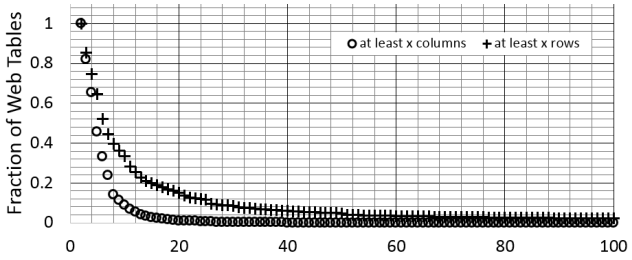

Figure 1: Distribution of # rows & columns

Regarding the origin of the tables, we find that over 91% of the tables are from webpages registered under the five top-level domains (TLD): `com`, `org`, `gov`, `edu`, and `net`, with the majority (68%) originating from the `com` TLD. Language-wise we find a large fraction (over 80%) of English tables.

In Table 1 we list the top 10 pay-level domains (PLDs) ordered descending by the number of tables they contain. The topical domains of the PLDs span from search (*Google*, *Cappex*, and *healthgrades*) to sport and gaming (*gtaforums* and *3dfootball*) to shopping pages (*Hollister & Co*) which indicates a broad topical coverage of the whole corpus. The table also shows the ten most common headers (attribute names), ordered descending by their frequency. Beside general-purpose headers like `date`, `name`, `description` and `title` we also find topic specific headers describing products (`price`), and likely sport teams (`team`).[5]

Table 1: 10 most frequent PLDs and headers

PLD	Header
cappex.com	date
hollisterco.com	name
ucm.es	comments
wikipedia.org	categories
google.com	title
d3football.com	description
heatlthgrades.com	time
reef.org	team
seatgeek.com	price
gtaforums.com	forum

We find almost equal fractions of numeric and string attributes which together form the majority of attributes. Other data types that were detected but do not account for large fractions are date and boolean.

In addition to former extractions, our corpus also includes contextual metadata. For each table, we extract the URL, the page title, the table title as well as 200 words before and after the table. Further more, we extract text that covers timestamp information and include the last modified property of the HTTP response. For almost half of the tables we find a timestamp located after the table, which in most cases is the imprint of the page.

In order to gain further insights into the topical contents of the tables we applied the *T2K* matcher [6], which matches Web tables to DBpedia. Table 2 lists the most frequent table-to-class correspondences in the resulting mapping. Altogether, we were able to match 5.6 million tables to DBpedia classes. This relatively low number might be explained by the limited topical coverage of DBpedia, for example, products or events are hardly covered.

[5]We provide more comprehensive statistics on our webpage: `http://webdatacommons.org/webtables/2015/relationalStatistics.html`.

Table 2: Most frequent topics in the Web tables

DBpedia class	# Tables
dbo:Magazine	598,175
dbo:Protein	409,752
dbo:Country	349,220
dbo:Single	291,099
dbo:PopulatedPlace	208,772
dbo:TelevisionShow	195,811
dbo:City	160,658
dbo:Region	156,701
dbo:SoccerClub	155,101
dbo:MusicGenre	134,719
dbo:BaseballPlayer	130,868
dbo:Company	113,024
dbo:University	109,346
dbo:Album	100,099
dbo:Film	96,378
dbo:MusicalArtist	92,922
dbo:AmericanFootballPlayer	87,102
dbo:AnatomicalStructure	79,230
dbo:Software	68,113
dbo:Cricketer	66,519

4. SUMMARY

This poster presents the largest and most up-to-date corpus of Web tables that is currently available to the public outside the large search engine companies. In addition to previously published corpora we include time and context metadata. Based on our first analysis, the topical coverage is broad and allows for a variety of application scenarios. We hence believe that this corpus can serve as a common ground further research and comparability in this area.

5. REFERENCES

[1] K. Braunschweig, M. Thiele, J. Eberius, and W. Lehner. Column-specific Context Extraction for Web Tables. In *Proc. of the 30th ACM Symp. on Appl. Computing*, 2015.

[2] M. Cafarella, Y. Halevy, Alonand Zhang, D. Z. Wang, and E. Wu. Uncovering the Relational Web. In *Proc. of the WebDB Workshop*, 2008.

[3] J. Eberius, M. Thiele, K. Braunschweig, and W. Lehner. Top-k Entity Augmentation Using Consistent Set Covering. In *Proc. of the 27th Int. Conf. on Scientific and Statistical Database Mgmt*, 2015.

[4] O. Lehmberg, D. Ritze, P. Ristoski, R. Meusel, H. Paulheim, and C. Bizer. The Mannheim Search Join Engine. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:159–166, 2015.

[5] J. Morcos, Z. Abedjan, I. F. Ilyas, M. Ouzzani, P. Papotti, and M. Stonebraker. Dataxformer: An interactive data transformation tool. In *Proc. of the 2015 SIGMOD*, 2015.

[6] D. Ritze, O. Lehmberg, and C. Bizer. Matching HTML Tables to DBpedia. In *Proc. of the 5th Int. Conf. on Web Intelligence, Mining and Semantics*, 2015.

[7] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *Proc. of the 2012 SIGMOD*, 2012.

[8] M. Zhang and K. Chakrabarti. InfoGather+: Semantic Matching and Annotation of Numeric and Time-varying Attributes in Web Tables. In *Proc. of the 2013 ACM SIGMOD Int. Con. on Mgmt. of Data*, pages 145–156, 2013.