Web Table Extraction, Retrieval and Augmentation

Shuo Zhang*
University of Stavanger
shuo.zhang@uis.no

Krisztian Balog
University of Stavanger
krisztian.balog@uis.no

ABSTRACT

This tutorial synthesizes and presents research on web tables over the past two decades. We organize the work into six main categories of information access tasks: (i) table extraction, (ii) table interpretation, (iii) table search, (iv) question answering on tables, (v) knowledge base augmentation, and (vi) table completion. For each category, we identify and introduce seminal approaches, present relevant resources, and point out interdependencies among the different tasks.

CCS CONCEPTS

• **Information systems → Environment-specific retrieval**; Search in structured data; Data extraction and integration;

ACM Reference Format:
Shuo Zhang and Krisztian Balog. 2019. Web Table Extraction, Retrieval and Augmentation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3331184.3331385

Motivation

Tables are a practical and useful tool in many application scenarios. Tables can be effectively utilized for collecting and organizing information from multiple sources. With the help of additional operations, such as sorting, filtering, and joins, this information can be turned into knowledge and, ultimately, can be used to support decision-making. Thanks to their convenience and utility, a large number of tables are being produced and are made available on the Web. These tables represent a valuable resource and have been a focus of research for over two decades now. In this tutorial, we provide a systematic overview of this body of research.

Tables on the web, referred to as *web tables* henceforth, differ from traditional tables (that is, tables in relational databases and tables created in spreadsheet programs) in a number of ways. First, web tables are embedded in webpages. There is a lot of contextual information, such as the embedding page's title and link structure, the surrounding text, etc. that can be utilized. Second, web tables are rather heterogeneous regarding their quality, organization, and content. For example, tables on the Web are often used for layout and navigation purposes. Among the different table types, *relational tables* (also referred to as *genuine tables*) are of special interest. These describe a set of entities (such as people, organizations, locations, etc.) along with their attributes [8, 10, 13, 27]. Relational tables are considered to be of high-quality, because of the relational knowledge contained in them. However, unlike from tables in relational databases, these relationships are not made explicit in web tables; uncovering them is one of the main research challenges. The uncovered semantics can be leveraged in various applications, including table search, question answering, knowledge base augmentation, and table completion. For each of these tasks we identify seminal work, describe the key ideas behind the proposed approaches, discuss relevant resources, and point out interdependencies among the different tasks.

The tutorial is organized into six session, half an hour each. Below is a brief outline of the contents.

(1) Introduction
  - Motivating scenarios
  - Table types
  - Table extraction and table corpora
(2) Table interpretation
  - Column type identification
  - Entity linking in tables
  - Relation extraction
(3) Table search
  - Keyword query search
  - Search by table
(4) Question answering on tables
  - QA using a single table
  - QA using multiple tables
(5) Knowledge base augmentation
  - Tables for knowledge exploration
  - Knowledge base augmentation and construction
(6) Table augmentation (and wrap-up)
  - Row extension
  - Column extension
  - Data completion

Table extraction. A vast number of tables can be found on the Web, produced for various purposes and storing an abundance of information. These tables are available in heterogenous format, from HTML tables embedded in webpages to files created by spreadsheet programs (e.g., Microsoft Excel). To conveniently utilize these resources, tabular data should be extracted, classified, and stored in a consistent format, resulting ultimately in a table corpus. This process is referred to as *table extraction*. In this tutorial, we present approaches for the table extraction task, organized around three main types of tables: web tables [7, 8, 13], Wikipedia tables [5], and spreadsheets [9].

Table Interpretation. Table interpretation encompasses methods that aim to make tabular data processable by machines. Specifically, it focuses on interpreting tables with the help of existing knowledge bases. Bhagavatula et al. [5] identify three main tasks

*Main contact

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331385

aimed at uncovering table semantics: (i) *column type identification* [25], that is, associating a table column with the type of entities or relations it contains, (ii) *entity linking* [5], which is the task of identifying mentions of entities in cells and linking them to entries in a reference knowledge base, and (iii) *relation extraction* [25], which is about associating a pair of columns in a table with the relation that holds between their contents.

Table Search. Table search is the task of returning a ranked list of tables in response to a query. It is an important task on its own and is regarded as a fundamental step in many other table mining and extraction tasks as well, like table integration or data completion. Table search functionality is also available in commercial products; e.g., Microsoft Power Query provides smart assistance features based on table search. Depending on the type of the query, table search may be classified as *keyword query search* [7, 32] and *table query search* [1, 11]. We also introduce methods that generate tables "on the fly" in response to keyword queries [33].

Question Answering on Tables. Tables are a rich source of knowledge that can be utilized for answering natural language questions. This problem has been investigated in two main flavors: (i) where the table, which contains the answer to the input question, is given beforehand [21], and (ii) where a collection of tables are to be considered [24]. Question answering on tables is closely related to work on natural language interfaces to databases, where the idea is that users can issue natural language queries, instead of using formal structured query languages (like SQL), for accessing databases [2, 17, 18, 22]. *Semantic parsing* is the task of parsing natural language queries into a formal representation. Semantic parsing is used for answering natural language questions, by generating logical expressions that are executable on knowledge bases [3, 14].

Knowledge Base Augmentation. *Knowledge base augmentation* leverages tabular data for exploring, constructing, and augmenting knowledge bases. Knowledge bases need to be complete, correct, and up-to-date. A precondition of extending knowledge bases using web tables is matching table content to entities, classes, and attributes already existing in those knowledge bases. Specifically, matching problems include *table-to-class matching*, *row-to-instance matching*, and *attribute-to-property matching* [4, 12, 23].

Table Augmentation. *Table augmentation* refers to the task of extending a seed table with more data. Specifically, we discuss three tasks in this section: row extension [11, 26, 31], column extension [11, 16, 31], and data completion [28, 29, 31]. Row extension is similar to the problems of *concept expansion*, also known as *entity set expansion*, where a given set of seed entities is to be completed with additional entities [6, 15, 19, 20]. One might envisage these functionalities being offered by an intelligent agent that aims to provide assistance for people working with tables [30].

REFERENCES

[1] Ahmad Ahmadov, Maik Thiele, Julian Eberius, Wolfgang Lehner, and Robert Wrembel. 2015. Towards a Hybrid Imputation Approach Using Web Tables. In *Proc. of BDC '15*. 21–30.

[2] Ion Androutsopoulos, Graeme D. Ritchie, and Peter Thanisch. 1995. Natural Language Interfaces to Databases - An Introduction. *CoRR* cmp-lg/9503016 (1995).

[3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proc. of EMNLP '13*. 1533–1544.

[4] Avishek Anand Besnik Fetahu and Maria Koutraki. 2019. TableNet: An Approach for Determining Fine-grained Relations for Wikipedia Tables. In *Proc. of WWW '19*.

[5] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2015. TabEL: Entity Linking in Web Tables. In *Proc. ISWC '15*. 425–441.

[6] Marc Bron, Krisztian Balog, and Maarten de Rijke. 2013. Example Based Entity Search in the Web of Data. In *Proc. of ECIR'13*. 392–403.

[7] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: Exploring the Power of Tables on the Web. *Proc. of VLDB Endow.* 1, 1 (Aug. 2008), 538–549.

[8] Michael J. Cafarella, Alon Y. Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu 0002. 2008. Uncovering the Relational Web. In *Proc. of WebDB '08*.

[9] Zhe Chen and Michael Cafarella. 2013. Automatic Web Spreadsheet Data Extraction. In *Proc. of SS@ '13*. 1–8.

[10] Eric Crestan and Patrick Pantel. 2011. Web-scale Table Census and Classification. In *Proc. of WSDM '11*. 545–554.

[11] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. 2012. Finding Related Tables. In *Proc.of SIGMOD '12*. 817–828.

[12] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proc. of KDD '14*. 601–610.

[13] Julian Eberius, Katrin Braunschweig, Markus Hentsch, Maik Thiele, Ahmad Ahmadov, and Wolfgang Lehner. 2015. Building the Dresden Web Table Corpus: A Classification Approach. In *2nd IEEE/ACM International Symposium on Big Data Computing, BDC 2015, Limassol, Cyprus, December 7-10, 2015*. 41–50.

[14] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open Question Answering over Curated and Extracted Knowledge Bases. In *Proc. of KDD '14*. 1156–1165.

[15] Yeye He and Dong Xin. 2011. SEISA: Set Expansion by Iterative Similarity Aggregation. In *Proc. of WWW '11*. 427–436.

[16] Oliver Lehmberg, Dominique Ritze, Petar Ristoski, Robert Meusel, Heiko Paulheim, and Christian Bizer. 2015. The Mannheim Search Join Engine. *Web Semant.* 35, P3 (Dec. 2015), 159–166.

[17] Fei Li and H. V. Jagadish. 2014. Constructing an Interactive Natural Language Interface for Relational Databases. *Proc. VLDB Endow.* 8, 1 (Sept. 2014), 73–84.

[18] Yunyao Li, Huahai Yang, and H. V. Jagadish. 2005. NaLIX: An Interactive Natural Language Interface for Querying XML. In *Proc. of SIGMOD '05*. 900–902.

[19] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. 2013. QBEES: query by entity examples. In *Proc. of CIKM '13*. 1829–1832.

[20] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. 2014. Aspect-Based Similar Entity Search in Semantic Knowledge Graphs with Diversity-Awareness and Relaxation. In *Proc. of WI-IAT '14*. 60–69.

[21] Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Proc. of ACL '15*. 1470–1480.

[22] Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a Theory of Natural Language Interfaces to Databases. In *Proc.of IUI '03*. 149–157.

[23] Dominique Ritze, Oliver Lehmberg, and Christian Bizer. 2015. Matching HTML Tables to DBpedia. In *Proc. of WIMS '15*. Article 10, 10:1–10:6 pages.

[24] Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table Cell Search for Question Answering. In *Proc. of WWW '16*. 771–782.

[25] Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering Semantics of Tables on the Web. *Proc. VLDB Endow.* 4, 9 (June 2011), 528–538.

[26] Chi Wang, Kaushik Chakrabarti, Yeye He, Kris Ganjam, Zhimin Chen, and Philip A. Bernstein. 2015. Concept Expansion Using Web Tables. In *Proc. of WWW '15*. 1198–1208.

[27] Yalin Wang and Jianying Hu. 2002. Detecting Tables in HTML Documents. In *Proc. of DAS '02*. 249–260.

[28] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *Proc. of SIGMOD '12*. 97–108.

[29] Shuo Zhang. 2018. SmartTable: Equipping Spreadsheets with Intelligent AssistanceFunctionalities. In *Proc. of SIGIR '18*. 1447–1447.

[30] Shuo Zhang, Vugar Abdulzada, and Krisztian Balog. 2018. SmartTable: A Spreadsheet Program with Intelligent Assistance. In *Proc. of SIGIR '18*.

[31] Shuo Zhang and Krisztian Balog. 2017. EntiTables: Smart Assistance for Entity-Focused Tables. In *Proc. of SIGIR '17*. 255–264.

[32] Shuo Zhang and Krisztian Balog. 2018. Ad Hoc Table Retrieval using Semantic Similarity. In *Proceedings of The Web Conference 2018 (WWW '18)*.

[33] Shuo Zhang and Krisztian Balog. 2018. On-the-fly Table Generation. In *Proc. of SIGIR '18*.