Table Extraction and Understanding for Scientific and Enterprise Applications

Douglas Burdick, Marina Danilevsky, Alexandre V Evfimievski, Yannis Katsis, Nancy Wang
IBM Research - Almaden
650 Harry Rd
San Jose, CA 95120
{drburdic,mdanile,evfimi}@us.ibm.com, {yannis.katsis, wangnxr}@ibm.com

ABSTRACT

Valuable high-precision data are often published in the form of tables in both scientific and business documents. While humans can easily identify, interpret and contextualize tables, developing general-purpose automated techniques for extraction of information from tables is difficult due to the wide variety of table formats employed across corpora.

To extract useful data from tables, data cells must be correctly extracted and linked to all relevant headers, units of measure and in-text references. *Table extraction* involves identifying the border and cell structure for each document table, while *table understanding* provides context by linking cells with semantic information inside and outside the table, such as row and column headers, footnotes, titles, and references in surrounding text.

The objective of this tutorial is to provide a detailed synopsis of existing approaches for table extraction and understanding, highlight open research problems, and provide an overview of potential applications.

PVLDB Reference Format:
Douglas Burdick, Marina Danilevsky, Alexandre V Evfimievski, Yannis Katsis, and Nancy Wang. Table Extraction and Understanding for Scientific and Enterprise Applications. *PVLDB*, 13(12): 3433-3436, 2020.
DOI: https://doi.org/10.14778/3415478.3415563

1. INTRODUCTION

Automatic identification, separation, parsing, and interpretation of tables that appear within documents are critical tasks in both enterprise and scientific applications, as valuable high-precision data in documents are often publicized in the form of tables. These tasks are especially challenging due to the significant diversity of both tables and documents, which are usually formatted with human consumption in mind. Table appearance - both in terms of formatting and in the placement of important semantic information, such as units and headers within the table - differs greatly across subject matters, publisher regulations, source institutions, localities, and typesetting tools. Authors adhere to different sets of conventions, sometimes breaking them if they feel that data or aesthetics call for it. Tables are elevated to a form of speech and even art, seeking to combine the expressive power of natural communication with very high accuracy requirements. To make matters even more challenging, table-specific machine readable markup is lacking in many popular document formats, such as PDF or (scanned) images, which leaves automated extractors to tackle boundless human creativity unaided.

To acquire useful information from tables (which can then be used for downstream applications, such as database population, knowledge base creation, question answering, data integration, etc.), two high-level tasks must be performed: *Table extraction*, which involves identifying the borders of tables and extracting their cell structure (i.e., their grid) and *table understanding*, which provides context by linking each cell to semantic information pertaining to it appearing within or outside the table, such as row and column headers, footnotes, units, and references in surrounding text.

The scientific community has amassed multiple decades of research on both table extraction and understanding, across multiple sub-communities. These include the document analysis community, which has established benchmarks on the table extraction task [16]. The data management community has also worked extensively on table understanding in the context of web tables (including the seminal work on web tables [5], which received the VLDB Test of Time Award in 2018 [4], and the significant amount of follow-up work in the area). Multiple other communities, such as natural language processing, machine learning, and human computer interaction, also have ongoing research on table extraction and understanding.

The goal of this tutorial is to consolidate the significant amount of work spread across many sub-communities. It presents a unified view of existing work and open challenges in extraction and understanding with a focus on tables that appear in scientific and business documents. For the data management community, with its extensive work on table understanding for web tables, the tutorial will help place the work in web tables in the general context of table processing for more complex document types. In particular, it introduces the table extraction work of accurately inferring table border and structure (usually assumed to be correct in web tables), and explain how inaccuracies may affect the downstream task of table understanding. It also motivates the

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415563

I. Introduction: Table Processing Pipeline (20 min)
1. End-to-end Example
2. Problem Definition: Table Extraction & Understanding
3. Motivation and Challenges

II. Table Extraction (70 min)
1. Overview
2. Document Feature Analysis
3. Table and Cell Detection
4. Refinement Process
5. Machine Learning Infrastructure

III. Table Understanding (70 min)
1. Overview
2. Row and Column Header Identification
3. Decoding Hierarchical Relationships
4. Linking Context to Tabular Data
5. Resources

IV. Integration and Applications (20 min)
1. Putting it All Together
2. Applications

Figure 1: Tutorial Outline

need for table understanding support for the complex structures of tables that appear in scientific and business documents, including multiple nested headers and additional complex semantic information (issues not as prevalent in the case of simpler web tables).

We believe that such cross-pollination of ideas can both help the scientific community tackle the open challenges in table extraction and understanding, and contribute towards the greater goal of leveraging the high-accuracy information found within tables. This is an important problem, as evidenced by the significant amount of activity in developing new research prototypes as well as commercial offerings for table processing (e.g., see recent table processing offerings from Amazon [1], Google [17], IBM [21], and Microsoft [26]).

Target Audience and Prerequisites. The tutorial should be of interest to researchers and practitioners from the document analysis, data mining, and data management communities interested in either contributing improved approaches for table extraction and table understanding or leveraging the table processing output for other downstream tasks such as question answering, knowledge-base population, and table search and retrieval. No prior knowledge of table processing is required as the tutorial includes a detailed background section. Basic familiarity with machine learning and deep learning would be helpful, but is not required.

2. TUTORIAL OUTLINE

Figure 1 outlines the tutorial and provides time estimates for each topic, described in more detail below.

Part I, Introduction: Table Processing Pipeline, introduces and motivates the problem of table extraction and understanding. It uses an end-to-end example to illustrate the different tasks to be solved and the corresponding steps in the table processing pipeline. It then provides definitions of these steps and describes the main challenges associated with them. The first challenge is the extraordinary diversity and complexity of tables as a form of human communication. The second is the lack of table metadata in common document formats such as PDF, images, and plain text. Even formats with metadata (Word, Excel, HTML) frequently leave out vital information such as header hierarchies. The third challenge is understanding text elements in table cells and their relationships to semantic clues inside and outside the table. Each challenge is illustrated and motivated by examples from real-life scenarios.

Part II, Table Extraction, dives deeper into table detection and cell structure recognition. It starts by giving a historical perspective on how the table extraction technology evolved over the past 30 years. It outlines a set of tasks commonly addressed in prior work, organizing them into four categories: Analysis, Detection, Refinement, and Learning. Each category is then discussed in turn. *Analysis* comprises the pre-processing tasks that prepare for table extraction, such as character and font recognition [27], text alignment and page layout analysis [34], grouping text into larger semantic units [29], and identifying ruled and white-space separator lines [18]. They also provide shortcuts to bypass expensive table detection in most cases with no tables on the page. *Detection* performs initial table detection and cell structure recognition using a variety of techniques, from line clustering [22], row classification [30], and tracing text alignment [34] to modern deep learning methods [14, 35, 31]. *Refinement* filters out false positives [38], adjusts candidate tables and cells [8], addresses special cases, and resolves conflicts and ambiguities. Finally, *Learning* comprises the tasks and resources required for model and parameter training. This includes common accuracy metrics [15], challenges related to labeling ground truth data [11], popular benchmarks [16], and a discussion of the performance of published systems. While most of the work on table extraction has focused on unstructured formats (such as PDF documents), the tutorial also compares and contrasts this to work done on extracting tables from semi-structured formats (such as HTML pages) [25, 13].

Part III, Table Understanding, overviews the conversion of tables and cells into data records. Using an example, it walks through the semantic constituents of the knowledge in tabular form. As before, the conversion is presented as a set of common tasks grouped into categories. The first task is *row and column header identification*, important to both table extraction and understanding. In extraction, headers help tell apart a single table from multiple co-aligned tables; in understanding, headers provide critical semantics to the data values. Many real-life tables have multiple header rows or columns, whose header cells, in turn, span multiple columns or rows. The tutorial surveys existing approaches to header identification, including unsupervised and rule-based methods [10, 33], "traditional" machine learning [30, 32], and deep learning methods [40, 28]. The next task is *capturing hierarchical relationships* across headers and table sections, required to correctly assign headers to data cells. Methods to extract header hierarchies include graphical models [7] and nested rectangle models [6]. The third task is *linking table cells to additional context* beyond headers, which includes context from other cells within the table, from the surrounding text outside the table [23, 19], and from knowledge bases outside the document [42]. For each of the aforementioned tasks, in addition to reviewing the existing work, the tutorial also explains the challenges posed by the complex structure of tables found in documents and

compares and contrasts them to the work done on web tables. Finally, it outlines the table understanding resources that are available to the scientific community, such as the two large-scale corpora of web tables; the WDC Web Table Corpus [24] and the Dresden Web Table Corpus [9].

Part IV, Integration and Applications, starts by showing how the tasks and solutions from Parts II and III fit together into a single pipeline. It explains the interaction between the two parts, with errors made by table extraction potentially affecting, and/or corrected by, table understanding. The tutorial then discusses the considerable value of table extraction and understanding by presenting a range of enterprise and scientific applications. These include knowledge base construction [39, 3], table search and retrieval [5, 37, 2], question answering [36], and leaderboard construction [20].

3. PRESENTERS

Douglas Burdick is a Research Staff Member at IBM Research - Almaden currently working on the application of AI and machine learning to document understanding, which includes table extraction and understanding in addition to inferring document structure. His document understanding work is incorporated into the IBM Watson Compare & Comply and IBM Watson Discovery products. His other research focuses on creation of financial knowledge graphs from unstructured data sources such as regulatory filings and analyst reports, which includes interpretation of tabular data from these documents [3]. He has contributed to Apache SystemML and OpenII data integration toolkit, and co-organizes the DSMM workshop series (co-located with SIGMOD). He presented a tutorial on table processing at ICDM 2019. He received his PhD in Computer Science from the University of Wisconsin - Madison.

Marina Danilevsky is a Research Staff Member at IBM Research - Almaden whose work centers on understanding structured and unstructured text data. An important research direction for her has been in creating and querying knowledge bases from unstructured data sources, with particular focus on interpreting text and tabular data from the financial domain, including multiple published research efforts [6, 3]. She received her PhD in Computer Science from the University of Illinois at Urbana-Champaign.

Alexandre Evfimievski is a Research Staff Member at IBM Research - Almaden who implemented core components of the table extraction pipeline in the IBM Watson Compare and Comply product. He also contributed to table understanding research [3, 6]. Prior to that, he contributed to the Apache SystemML project, implemented algorithms for statistical data analysis, and did research in data mining and data privacy. He received his PhD in Computer Science from Cornell University in 2004.

Yannis Katsis is a Research Staff Member at IBM Research - Almaden with experience in management, integration and knowledge extraction from both structured and unstructured data. His current work focuses on improving the state-of-the-art in table understanding by enriching the extracted tables with extended semantic information that can be leveraged by downstream tasks, such as question answering, knowledge base generation, and others. His work in the area has been already incorporated into both the IBM Compare & Comply and IBM Discovery products. He received his PhD in Computer Science from UC San Diego in 2009.

Nancy Wang is a Researcher with IBM Research - Almaden who is currently working on applying deep learning and computer vision methods for table extraction and table understanding from documents. She graduated from the University of Washington with her PhD in Computer Science in 2018 in the area of computer vision for computational neuroscience. Her new table extraction work is under review at top-level AI conferences and is in the process of being incorporated into Watson Discovery. She was also one of the presenting tutors for the Table Extraction and Understanding Tutorial at ICDM 2019.

4. RELATED WORK

While this tutorial builds upon the tutorial on table processing presented by the authors at ICDM 2019, it extends it with additional information to make it more relevant to the database community. This includes the latest developments in the area of deep learning for table identification, additional material on downstream applications, such as table retrieval and search, and discussions on how the extensive work done by the database community on web tables relates to the broader context of document processing.

This tutorial is also related to a recent tutorial on web table processing that appeared in SIGIR 2019 [41]. However, [41] focuses on web tables and thus covers only the subset of the topics presented in this tutorial related to table understanding (i.e., interpretation of table values) and downstream applications (i.e., table search, knowledge base augmentation and question answering). Since table border and structure of web tables are often apparent, [41] does not contain material related to table extraction or its interaction with table understanding, which is particularly important for scientific and business documents that are often provided in PDF or image formats.

Finally, this is also related to several benchmarks and competitions on table extraction (such as the ICDAR 2013 and 2019 table competitions [16, 12]), which have attracted primarily the interest of the document analysis community. The ICDAR 2013 competition is especially known for helping establish a baseline for PDF-based table boundary and structure extraction that many subsequent works compared against. Our tutorial discusses both the results from these competitions, as well as the approaches followed by the submitted solutions, as outlined in Section 2.

5. REFERENCES

[1] Amazon. Amazon Textract. https://aws.amazon.com/textract/features/.

[2] C. S. Bhagavatula, T. Noraset, and D. Downey. Methods for exploring and mining tables on wikipedia. In *ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*, 2013.

[3] S. Bharadwaj, L. Chiticariu, M. Danilevsky, S. Dhingra, S. Divekar, A. Carreno-Fuentes, H. Gupta, N. Gupta, and et al. Creation and interaction with large-scale domain-specific knowledge bases. *PVLDB*, 10(12):1965–1968, 2017.

[4] M. Cafarella, A. Halevy, H. Lee, J. Madhavan, C. Yu, D. Z. Wang, and E. Wu. Ten years of WebTables. *PVLDB*, 11(12):2140–9, 2018.

[5] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. WebTables: Exploring the power of tables on the web. *PVLDB*, 1(1):538–549, 2008.

[6] X. Chen, L. Chiticariu, M. Danilevsky, A. Evfimievski, and P. Sen. A rectangle mining method for understanding the semantics of financial tables. In *ICDAR*, pages 268–273, 2017.

[7] Z. Chen and M. Cafarella. Integrating spreadsheet data via accurate and low-effort extraction. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1126–1135, 2014.

[8] F. Deckert, B. Seidler, M. Ebbecke, and M. Gillmann. Table content understanding in smartFIX. In *ICDAR*, pages 488–492, 2011.

[9] J. Eberius, K. Braunschweig, M. Hentsch, M. Thiele, A. Ahmadov, and W. Lehner. Building the Dresden web table corpus: A classification approach. In *2015 IEEE/ACM International Symposium on Big Data Computing (BDC)*, pages 41–50, 2015.

[10] J. Fang, P. Mitra, Z. Tang, and C. L. Giles. Table header detection and classification. In *AAAI*, 2012.

[11] M. Frey and R. Kern. Efficient table annotation for digital articles. *D-Lib Magazine*, 21(11/12), Nov 2015.

[12] L. Gao, Y. Huang, H. Déjean, J.-L. Meunier, Q. Yan, Y. Fang, F. Kleber, and E. Lang. ICDAR 2019 competition on table detection and recognition (ctdar). In *ICDAR*, pages 1510–1515. IEEE, 2019.

[13] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krüpl, and B. Pollak. Towards domain-independent information extraction from web tables. In *WWW*, pages 71–80, 2007.

[14] A. Gilani, S. R. Qasim, I. Malik, and F. Shafait. Table detection using deep learning. In *ICDAR*, 2017.

[15] M. Göbel, T. Hassan, E. Oro, and G. Orsi. A methodology for evaluating algorithms for table understanding in PDF documents. In *ACM SIGWEB*, pages 45–48, 2012.

[16] M. Göbel, T. Hassan, E. Oro, and G. Orsi. ICDAR 2013 table competition. In *ICDAR*, 2013.

[17] Google. Google Document AI. https://cloud.google.com/solutions/document-ai.

[18] J. Ha, R. M. Haralick, and I. T. Phillips. Recursive X-Y cut using bounding boxes of connected components. In *ICDAR*, volume 2, 1995.

[19] B. Hancock, H. Lee, and C. Yu. Generating titles for web tables. In *WWW*, pages 638–647. ACM, 2019.

[20] Y. Hou, C. Jochim, M. Gleize, F. Bonin, and D. Ganguly. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *ACL*, 2019.

[21] IBM. IBM Watson Discovery. https://www.ibm.com/cloud/watson-discovery.

[22] K. Itonori. Table structure recognition based on textblock arrangement and ruled line position. In *ICDAR*, pages 765–768, October 20–22 1993.

[23] D. H. Kim, E. Hoque, J. Kim, and M. Agrawala. Facilitating document reading by linking text and tables. In *ACM UIST*, pages 423–434, 2018.

[24] O. Lehmberg, D. Ritze, R. Meusel, and C. Bizer. A large public corpus of web tables containing time and context metadata. In *WWW*, pages 75–76, 2016.

[25] K. Lerman, L. Getoor, S. Minton, and C. Knoblock. Using the structure of web sites for automatic segmentation of tables. In *ACM SIGMOD*, pages 119–130, 2004.

[26] Microsoft. Microsoft Form Recognizer. https://azure.microsoft.com/en-us/services/cognitive-services/form-recognizer/.

[27] S. Mujumdar, N. Gupta, A. Jain, and D. Burdick. Simultaneous optimisation of image quality improvement and text content extraction from scanned documents. In *ICDAR*, 2019.

[28] K. Nishida, K. Sadamitsu, R. Higashinaka, and Y. Matsuo. Understanding the semantic structures of tables with a hybrid deep neural network architecture. In *AAAI*, pages 168–174, 2017.

[29] E. Oro and M. Ruffolo. PDF-TREX: An approach for recognizing and extracting tables from PDF documents. In *ICDAR*, pages 906–910, 2009.

[30] D. Pinto, A. McCallum, X. Wei, and W. B. Croft. Table extraction using conditional random fields. In *ACM SIGIR*, pages 235–242, 2003.

[31] S. R. Qasim, H. Mahmood, and F. Shafait. Rethinking table parsing using graph neural networks. In *ICDAR*, 2019.

[32] R. Rastan, H.-Y. Paik, and J. Shepherd. TEXUS: A unified framework for extracting and understanding tables in PDF documents. *Information Processing and Management*, 56(3):895–918, May 2019.

[33] S. Seth and G. Nagy. Segmenting tables via indexing of value cells by table headers. In *2013 12th International Conference on Document Analysis and Recognition*, pages 887–891. IEEE, 2013.

[34] F. Shafait and R. Smith. Table detection in heterogeneous documents. In *International Workshop on Document Analysis Systems (DAS)*, 2010.

[35] S. A. Siddiqui, M. I. Malik, S. Agne, A. Dengel, and S. Ahmed. DeCNT: Deep deformable CNN for table detection. *IEEE Access*, 6, 2018.

[36] H. Sun, H. Ma, X. He, W.-T. Yih, Y. Su, and X. Yan. Table cell search for question answering. In *WWW*, pages 771–782, 2016.

[37] P. Venetis, A. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. *PVLDB*, 4(9), 2011.

[38] Y. Wang, I. T. Phillips, and R. M. Haralick. Table structure understanding and its performance evaluation. *Pattern Recognition*, 37(7), July 2004.

[39] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, P. Levis, and C. Re. Fonduer: Knowledge base construction from richly formatted data. In *ACM SIGMOD*, pages 1301–1316, 2018.

[40] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *NAACL'16*, pages 1480–1489, 2016.

[41] S. Zhang and K. Balog. Web table extraction, retrieval and augmentation. In *ACM SIGIR*, 2019.

[42] S. Zhang and K. Balog. Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(2), January 2020. Article 13.