Comparison of Table Type Taxonomies

A. Shigarov[1,2], V. Paramonov[1,2]

[1]Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of Russian Academy of Sciences, 134 Lermontov str., Irkutsk, Russia

[2]Institute of Mathematics and Information Technologies, Irkutsk State University, 20 Gagarina ave., Irkutsk, Russia

Abstract

Tables are widely used to present related data in various formats and media. In particular, an immense number of HTML tables can be found on web pages. These tables serve as a valuable source of data for applications such as web mining, question-answering systems, and knowledge base development. However, not all HTML tables contain actual relational data. Many of them are used primarily for layout and navigation purposes. On the other hand, genuine tables may have different layouts, formatting, and content. One of the challenges in extracting data from web tables is to identify the main functional and layout types and properties of the tables that are commonly found on the internet. Currently, we have a wide range of taxonomies that classify table types and their properties. These taxonomies can be used to categorize tables and select the appropriate method for processing tabular data based on their specific type. Although the existing taxonomies provide a variety of table types, they often use confusing terminology. This paper aims to provide an overview of these taxonomies by aligning their terminology and comparing them in a qualitative manner.

Keywords

table understanding, table extraction, type classification, table layout

1. Introduction

Tables are a one way to organise and present multiple sets data efficiently. They are widely used on the web and often embedded in HTML pages. These tables can contain various types of data, including unstructured, semi-structured, and relational data. Approachers to automated tables processing depend on their type and the way of data organisation.

According to one of the first studies on table classification, presented by Wang et al. [1], all tables are divided into two categories: *genuine* and *non-genuine*. A *genuine* table includes relational data, where the table structure helps convey the relationships between the data in the cells. In contrast, *non-genuine* tables are used for layout purposes.

Extracting, integrating, and analysing information from *genuine* tables can help uncover new knowledges [2, 3]. All genuine tables can be classified into different types based on how the data is organized. Understanding these types can improve the quality of data processing. Currently, there are some studies discussing table taxonomies [4, 5, 6, 7, 8, 9, 10, 11, 12]. Exploring these taxonomies can lead to better data extraction from HTML tables and understanding them.

6^{th} International Workshop on Information, Computation, and Control Systems for Distributed Environments (ICCS-DE 2024), July 01–05, 2024, Irkutsk, Russia

shigarov@icc.ru (A. Shigarov); slv@icc.ru (V. Paramonov)

0000-0001-5572-5349 (A. Shigarov); 0000-0002-4662-3612 (V. Paramonov)

ICCS-DE 2024 Workshop Proceedings (iccs-de.icc.ru)

DOI: 10.47350/ICCS-DE.2024.08

HTML tables are commonly classified on their layout. Defining the type of a table can simplify data extraction and analysis [1, 13, 4]. In this analysis, we aim to establish which cells contain meta-data or semantic hints that help understand the meaning of the data, and which cells hold actual data. Further, in all figures 1, 2, 3, 4, 5, 6 cells with meta-data have a blue background, and cells with data have a white background.

This paper focuses on analysing existing taxonomies for web tables in HTML. The main goal is to identify relationships between these taxonomies and understand how they can be applied to improve data extraction and analysis from HTML tables.

2. Existing Table Type Taxonomies

In the study [13] the goal of Cafarella et al. research is to extract a set of high-quality relations [13] from HTML documents. In the research, authors present the fundamental division of whole tables into two types: *relational* and *non-relational*, which in the other classification [1] are designated as *genuine* and *non-genuine*. The results of Crestan's research, which became a basis of further research, presented in the paper [4]. During the analysis of these tables, a broad classification and set of problems in extracting tables of certain types were established. This study divides the entire set of tables into *relational knowledge* tables and *layout* tables, which are equivalent to *genuine* and *non-genuine* accordingly in [1]. *Layout tables* are further divided into *navigational* and *formatting*. However, these table types are not relevant in our study. Below we consider different types of tables according to the Crestan's taxonomy [4] and give examples to illustrate table reliability to a certain type.

- *Vertical listings*. A table that lists the attributes of several entities that are similar in meaning (Fig. 1, *a*).
- *Horizontal listings*. Tables are similar to a *vertical listings*, but with a different orientation. This type of tables often is used for comparison (Fig. 1, *b*).
- *Attribute/value*. A special case of vertical and *horizontal listings*, that does not contain a subject (Fig. 1, *c*).
- *Matrix*. Tables that have values at the junctions of rows and columns, while having headers on the left and top of the table (Fig. 1, *d*).
- *Calendar*. A special case of *matrix tables* in which the subject is a date and the predicate is some kind of event (Fig. 1, *e*).
- *Enumeration*. The tables list a number of objects that have a similar meaning (Fig. 1, *f*).
- *Form*. Tables consist of input fields that the user can fill in or select (Fig. 1, *g*).
- *Other*. This type includes tables that cannot be attributed to the types listed above.

In 2013, Lautert [5] presents a classification of 631K tables which were extracted from Wikipedia, news sites, etc. All amount of tables were divided into two classes – *main* and *secondary*. The main classification includes *horizontal web table* (Fig. 2, *a*), *vertical web table* (Fig. 2, *b*) and *matrix web table* (Fig. 2, *c*). The secondary classification is represented by following types (generally, tables can belong to several types at the same time):

- *Concise web table* contains merged cells (Fig. 2, *d*).

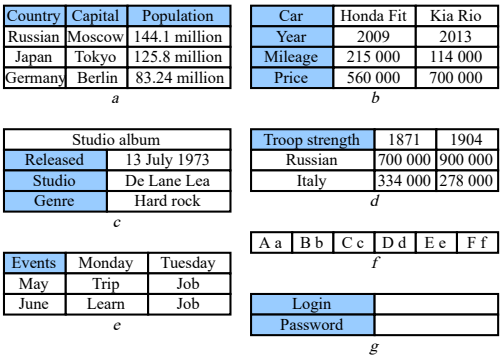

Figure 1: Examples of tables layout according Crestan classification: vertical listing *(a)*; horizontal listing *(b)*; attribute/value *(c)*; matrix *(d)*; calendar *(e)*; enumeration *(f)*; form *(g)*

- *Splitted web table* has sequential ordered repeating labels (Fig. 2, *e*).
- A table is classified as *nested web table* if it forms a part of another table (Fig. 2, *f*).
- *Multivalued web table*. Table in which cells have multiple values.
- *Simple multivalued web table* has multivalued data related to one specific type (Fig. 2, *g*).
- *Composed multivalued web table* contains multivalued data related to different types (Fig. 2, *h*).

In the paper [6] the classifying on six types for tables — *layout* (*non-genuine*), *genuine* which are divided into *vertical listings*, *horizontal listings*, *matrix*, *others* (all tables that do not belong to the above types) — was suggested. This classifier was tested on a July 2014 version of the Common Crawl.

In 2015, Braunschweig [7] presented several types of tables based on the study of taxonomies in [4, 5], such as *vertical listing* (Fig. 1, *a*), *horizontal listing* (Fig. 1, *b*), matrix (Fig. 1, *d*), and special case including *attribute/value* (Fig. 1, *c*) and *nested* (Fig. 2, *f*). In turn, Lehmberg et al. using the WebDataCommons framework [8] extracted and classified tables from the July 2015 version of the Common Crawl based on [7]. As a result, *relational tables* (Fig. 3, *a*) that contain a certain number of entities were extracted. In turn, these entities can be described by attributes, which are represented by rows or columns. Another version of tables in which a certain set of attributes describes one entity only (Fig. 3, *b*) is named *entity tables*. There are also *matrix tables* (Fig. 3, *c*).

The approach of Nishida [9] uses on combines recurrent neural network (RNN) and convolutional neural network (CNN) for table extraction from a subset of the April 2016 Common Crawl. The extracted tables were classified as: *relational tables* (Fig. 4, *a*), *entity tables* (Fig. 4, *b*), *matrix tables* (Fig. 4, *c*) and *other tables*. Tables are represented by semantic triples: subject, property, and object. It is noteworthy that in this taxonomy, *matrix tables* do not contain properties in their structure.

Gasemi-Gol and Szekely presented TabVec [10] — an unsupervised method of embedding tables into a vector space in order to perform classification of tables into categories. Its taxonomy includes *relational* (Fig. 5, *a*), *entity* (Fig. 5, *b*), *matrix* (Fig. 5, *d*) tables. *List* (fig. 5, *c*) and *non-data*.

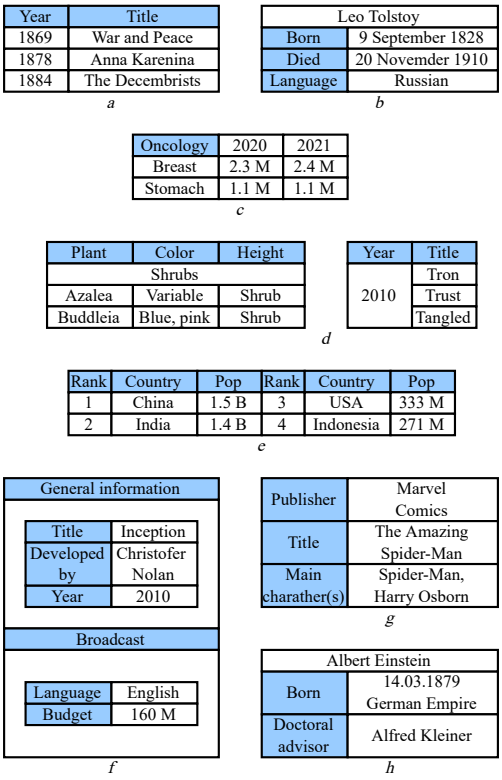

Figure 2: Examples of tables layout according Lautert classification: horizontal web table *(a)*; vertical web table *(b)*; matrix web table *(c)*; concise web table *(d)*; splitted web table *(e)*; nested web table *(f)*; simple multivalued web table *(g)*; composed multivalued web table *(h)*

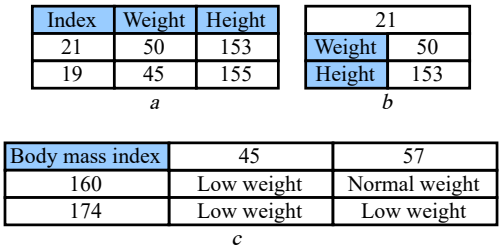

Figure 3: Examples of tables layout according Lehmberg classification: relational tables *(a)*; entity tables *(b)*; matrix tables *(c)*

The classifier was evaluated on a July 2015 version of the Common Crawl. The same types of tables are also used in the paper [12].

In the [14] Roldan introduced classification of tables types as the following: *horizontal listings* (Fig. 6, *a*), *vertical listings* (Fig. 6, *b*), form (Fig. 6, *c*), *matrix* (Fig. 6, *d*). The sources of the tables were Wikipedia (1496 tables) and Dresden Web Table Corpus (1513 tables).

Name	Age	Sex
Anna	25	Fermale
Mike	35	Male

a

Anna	
Age	25
Sex	Fermale

b

Weight

	2020	2021
Anna	57	60
Mike	75	75

c

Figure 4: Examples of tables layout according Nishida classification: relational tables *(a)*: contain complete semantic triples (may contain multiple subjects). Anna, Mike – subject, Age, Sex – properties, other – objects; entity tables *(b)*: describe the properties of one specific subject only (Anna - subject); matrix tables *(c)*: have the same property for each cell object at the junction of a row and a column. Anna, Mike, 2020,2021 – subject, Weight - properties, other – objects

Price	Dish
200	Soup
400	Steak

a

Steak	
Weight	500
Price	400

b

Idaho
Iowa
Alabama

c

Dish	Soup	Steak
Price	200	400
Weight	400	500

d

Figure 5: Examples of tables layout according Gasemi-Gol classification: relational tables *(a)*; entity tables *(b)*; list tables *(c)*; matrix tables *(d)*

SKU	Item	Price
AU-12	Bread	$0.90
PZ-18	Butter	$5.00
XX-99	Water	$1.00
WI-09	Milk	$3.95

a

Maker	Apple	Xiaomi
Model	8 Plus	MI6
Screen	LED	LED
RAM	64 Gb	64 Gb
Cores	8	8

b

Name	Pedro
Surname	Lopez
Age	47
Birthplace	Spain

c

	2018	2019
DINER	95B	100B
CHASE	98B	103B
VISA	78B	82B

d

Figure 6: Examples of tables layout according Roldan classification: horizontal listing *(a)*; vertical listing *(b)*; vertical form *(c)*; matrix *(d)*

Table 1
Matching Terminology of Table Type Taxonomies

Crestan	*Lautert*	*Eberius*	*Braunschweig*	*Lehmberg*	*Nishida*	*Gasemi-Gol*	*Wang*	*Roldan*
Vertical listing	Horizontal web table	Vertical listing	Vertical listing	(Horizontal) Relational	(Vertical) Relational	Relational	(Vertical) Relational	Horizontal listing
Horizontal listing	Vertical web table	Horizontal listing	Horizontal listing	(Vertical) Relational	(Horizontal) Relational	-	(Horizontal) Relational	Vertical listing
Matrix	Matrix web table	Matrix	Matrix	Matrix	Matrix	Matrix	Matrix	Matrix
Attribute/ value	Vertical web table	-	Attribute/value	Entity	Entity	Entity	Entity	Form
Enumeration	-	-	-	-	-	List	List	-

Table 2
Comparison of Table Type Taxonomies

Taxonomy of	*Number of types*	*Levels*	*Classifier*	*Classifier availability*	*Corpora*	*Corpora availability*
Crestan	13	Multi	TabEx (ML-based)	-	(8.2B tables from high quality English pages on the Web — Proprietary crawl)	-
Lautert (main)	7	Multi	WTClassifier (DL-based)	-	(0.6M HTML tables extracted from Wikipedia, e-commerce, news and university sites)	-
Lautert (secondary)	6	Multi	see above	-	see above	-
Eberius	6	Multi	DWTC-Tools[1] (ML-based)	+	Dresden Web Table Corpus (DWTC)[1] (125M tables from July 2014 version of the Common Crawl[2])	+
Lehmberg	4	Multi	WDC Extraction Framework[3] (ML-based)	+	Web Data Commons — Web Table Corpora[4] (233M tables from July 2015 version of the Common Crawl[2])	+
Nishida	5	Single	TabNet[5] (DL-based)	+	(0.19M tables from Top 500 websites from April 2016 version of the Common Crawl[2])	-
Gasemi-Gol	5	Single	TabVec[6] (DL-based)	+	(Web pages from unusual domains: human tracking advertisements, fire arms trading, microcap stock market; Random samples from July 2015 Common Crawl[2])	-
Roldan	4	Single	TOMATE[7] (cluster analysis)	+	Wikipedia, DWTC[1]	+
Wang	5	Single	TUTA[8] (DL-based)	+	(tables extracted from July 2015 version of the Common Crawl[2])	-

3. Results of the Comparison

As can be seen, some of the taxonomies considered above are similar but have different names. It should be noted that there is no any agreement in the literature regarding the types of web-tables. The comparison of web table types of different authors is presented in table 1. Each row of the table contains names for one type of table (if one is presented by the researcher).

The characteristics of the taxonomy kinds and classifiers, such as the number of types, the presence of a hierarchy, the classifier the data set on which the experiments were conducted and their availability are shown in table 2. The column "Number of Types" contains information about the quantity of genuine web-table types proposed by the author. The "Taxonomy levels" column indicates the way for table types definition. Two cases are possible:

[1]https://wwwdb.inf.tu-dresden.de/research-projects/dresden-web-table-corpus/
[2]https://commoncrawl.org/
[3]http://webdatacommons.org/framework/
[4]http://webdatacommons.org/webtables/#results-2015
[5]https://github.com/dreamquark-ai/tabnet
[6]https://github.com/majidghgol/tabvec
[7]https://data.mendeley.com/datasets/zcn6h2fvz7/2
[8]https://github.com/microsoft/TUTA_table_understanding

- Multi-Level — table types are defined hierarchically. Firstly, the set of tables divides into genuine and non-genuine. Next, the genuine tables divide into some types. The obtained types can be divided into other types and so on;
- Single-Level — table types are represented a plain structure. Each table type has no further splitting on other types.

The "Classifier" column shows what classifier and what kind of model are used for types classification. The general, method of classification (ML — "traditional" Machine Learning; DL — Deep Learning; cluster analysis) is given in parentheses. The availability of the classifier is specified in the corresponding column. "Corpora" column contains data about corpora where the tables were extracted from. Information about corpora availability is given in the rightmost column.

4. Summary

In this paper, we explored the various taxonomies of table types that have been proposed in studies of web tables since the groundbreaking work by Cafarella in 2008. Building upon this work, several other research have been conducted, offering different classifications of tables found in HTML documents.

The most comprehensive taxonomy, which includes 12 types, was suggested by Crestan. All other taxonomies we found are based on this one. These taxonomies usually focus on the most common types, such as *listing*, *matrix*, and *attribute/value*. Taking Crestan's taxonomy as a foundation, Lautert proposed two classes — the main one (containing the basic types of table layouts) and the secondary one (more complex cases of layouts). However, a single table may relate to types in both classes.

Crestan's primary taxonomy was based on a large corpus of 15 billion HTML tables. The other taxonomies were studied on open corpora with smaller tables amount. Other researchers proposed classifiers based on supervised learning methods.

We have noticed that some of the existing taxonomies are similar but have different terminology. In this study, we have made an attempt to unify the existing terminology by matching different names for common types. Also, the analysis of the taxonomies by comparing their key features was made.

Acknowledgment

The research was supported by the Program of the Fundamental Research of the Siberian Branch of the Russian Academy of Sciences, project no. 121030500071-2

References

[1] Y. Wang, J. Hu, Detecting tables in html documents, in: International Workshop on Document Analysis Systems, Springer, 2002, pp. 249–260.

[2] D. Ritze, O. Lehmberg, C. Bizer, Matching html tables to dbpedia, in: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, 2015, pp. 1–6.
[3] M. Yakout, K. Ganjam, K. Chakrabarti, S. 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, pp. 97–108.
[4] E. Crestan, P. Pantel, Web-scale table census and classification, in: Proceedings of the fourth ACM international conference on Web search and data mining, 2011, pp. 545–554.
[5] L. R. Lautert, M. M. Scheidt, C. F. Dorneles, Web table taxonomy and formalization, ACM SIGMOD Record 42 (2013) 28–33.
[6] J. Eberius, K. Braunschweig, M. Hentsch, M. Thiele, A. Ahmadov, W. Lehner, Building the dresden web table corpus: A classification approach, in: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), IEEE, 2015. URL: https://doi.org/10.1109/bdc.2015.30. doi:10.1109/bdc.2015.30.
[7] K. Braunschweig, Recovering the semantics of tabular web data, Ph.D. thesis, Saechsische Landesbibliothek-Staats-und Universitaetsbibliothek Dresden, 2015.
[8] O. Lehmberg, D. Ritze, R. Meusel, C. 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, 2016, pp. 75–76.
[9] K. Nishida, K. Sadamitsu, R. Higashinaka, Y. Matsuo, Understanding the semantic structures of tables with a hybrid deep neural network architecture, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[10] M. Ghasemi-Gol, P. Szekely, Tabvec: Table vectors for classification of web tables, arXiv preprint arXiv:1802.06290 (2018).
[11] J. C. Roldán, P. Jiménez, P. Szekely, R. Corchuelo, Tomate: A heuristic-based approach to extract data from html tables, Information Sciences 577 (2021) 49–68.
[12] Z. Wang, H. Dong, R. Jia, J. Li, Z. Fu, S. Han, D. Zhang, Tuta: Tree-based transformers for generally structured table pre-training, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1780–1790.
[13] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, Y. Zhang, Webtables: exploring the power of tables on the web, Proceedings of the VLDB Endowment 1 (2008) 538–549.
[14] J. C. Roldán Salvador, Enterprise Data Integration: On Extracting Data from HTML Tables, Ph.D. thesis, 2020.