Web Table Extraction, Retrieval and Augmentation

SIGIR 2019 tutorial

Shuo Zhang and Krisztian Balog

University of Stavanger

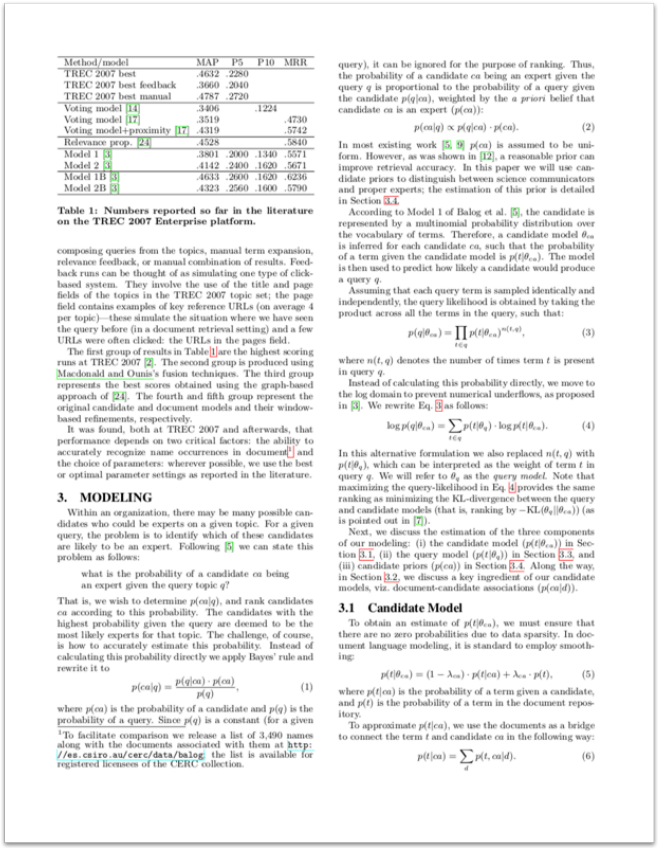
About Us

Shuo Zhang is a final-year PhD student at the University of Stavanger. His PhD research is concerned with developing intelligent tools for tabular data.

Krisztian Balog is a full professor at the University of Stavanger, currently on a sabbatical at Google. He has worked extensively on semi-structured data, entity-oriented and semantic search.

Tables are Everywhere

Banner	
Menu bar	
Navigation	Content
Footer	



Method/model	MAP	P5	P10	MRR
TREC 2007 best	.4632	.2280		
TREC 2007 best feedback	.3660	.2040		
TREC 2007 best manual	.4787	.2720		
Voting model [14]	.3406		.1224	
Voting model [17]	.3519			.4730
Voting model+proximity [17]	.4319			.5742
Relevance prop. [24]	.4528			.5840
Model 1 [3]	.3801	.2000	.1340	.5571
Model 2 [3]	.4142	.2400	.1620	.5671
Model 1B [3]	.4633	.2600	.1620	.6236
Model 2B [3]	.4323	.2560	.1600	.5790

Table 1: Numbers reported so far in the literature on the TREC 2007 Enterprise platform.

composing queries from the topics, manual term expansion, relevance feedback, or manual combination of results. Feedback runs can be thought of as simulating one type of click-based system. They involve the use of the title and page fields of the topics in the TREC 2007 topic set; the page field contains examples of key reference URLs (on average 4 per topic)—these simulate the situation where we have seen the query before (in a document retrieval setting) and a few URLs were often clicked: the URLs in the pages field.

The first group of results in Table 1 are the highest scoring runs at TREC 2007 [2]. The second group is produced using Macdonald and Ounis's fusion techniques. The third group represents the best scores obtained using the graph-based approach of [24]. The fourth and fifth group represent the original candidate and document models and their window-based refinements, respectively.

It was found, both at TREC 2007 and afterwards, that performance depends on two critical factors: the ability to accurately recognize name occurrences in document[1] and the choice of parameters: wherever possible, we use the best or optimal parameter settings as reported in the literature.

3. MODELING

Within an organization, there may be many possible candidates who could be experts on a given topic. For a given query, the problem is to identify which of these candidates are likely to be an expert. Following [5] we can state this problem as follows:

> what is the probability of a candidate ca being an expert given the query topic q?

That is, we wish to determine $p(ca|q)$, and rank candidates ca according to this probability. The candidates with the highest probability given the query are deemed to be the most likely experts for that topic. The challenge, of course, is how to accurately estimate this probability. Instead of calculating this probability directly we apply Bayes' rule and rewrite it to

$$p(ca|q) = \frac{p(q|ca) \cdot p(ca)}{p(q)}, \quad (1)$$

where $p(ca)$ is the probability of a candidate and $p(q)$ is the probability of a query. Since $p(q)$ is a constant (for a given query), it can be ignored for the purpose of ranking. Thus, the probability of a candidate ca being an expert given the query q is proportional to the probability of a query given the candidate $p(q|ca)$, weighted by the *a priori* belief that candidate ca is an expert ($p(ca)$):

$$p(ca|q) \propto p(q|ca) \cdot p(ca). \quad (2)$$

In most existing work [5, 9] $p(ca)$ is assumed to be uniform. However, as was shown in [12], a reasonable prior can improve retrieval accuracy. In this paper we will use candidate priors to distinguish between science communicators and proper experts; the estimation of this prior is detailed in Section 3.4.

According to Model 1 of Balog et al. [5], the candidate is represented by a multinomial probability distribution over the vocabulary of terms. Therefore, a candidate model θ_{ca} is inferred for each candidate ca, such that the probability of a term given the candidate model is $p(t|\theta_{ca})$. The model is then used to predict how likely a candidate would produce a query q.

Assuming that each query term is sampled identically and independently, the query likelihood is obtained by taking the product across all the terms in the query, such that:

$$p(q|\theta_{ca}) = \prod_{t \in q} p(t|\theta_{ca})^{n(t,q)}, \quad (3)$$

where $n(t, q)$ denotes the number of times term t is present in query q.

Instead of calculating this probability directly, we move to the log domain to prevent numerical underflows, as proposed in [3]. We rewrite Eq. 3 as follows:

$$\log p(q|\theta_{ca}) = \sum_{t \in q} p(t|\theta_q) \cdot \log p(t|\theta_{ca}). \quad (4)$$

In this alternative formulation we also replaced $n(t, q)$ with $p(t|\theta_q)$, which can be interpreted as the weight of term t in query q. We will refer to θ_q as the *query model*. Note that maximizing the query-likelihood in Eq. 4 provides the same ranking as minimizing the KL-divergence between the query and candidate models (that is, ranking by $-\mathrm{KL}(\theta_q||\theta_{ca})$) (as is pointed out in [7]).

Next, we discuss the estimation of the three components of our modeling: (i) the candidate model ($p(t|\theta_{ca})$) in Section 3.1, (ii) the query model ($p(t|\theta_q)$) in Section 3.3, and (iii) candidate priors ($p(ca)$) in Section 3.4. Along the way, in Section 3.2, we discuss a key ingredient of our candidate models, viz. document-candidate associations ($p(ca|d)$).

3.1 Candidate Model

To obtain an estimate of $p(t|\theta_{ca})$, we must ensure that there are no zero probabilities due to data sparsity. In document language modeling, it is standard to employ smoothing:

$$p(t|\theta_{ca}) = (1 - \lambda_{ca}) \cdot p(t|ca) + \lambda_{ca} \cdot p(t), \quad (5)$$

where $p(t|ca)$ is the probability of a term given a candidate, and $p(t)$ is the probability of a term in the document repository.

To approximate $p(t|ca)$, we use the documents as a bridge to connect the term t and candidate ca in the following way:

$$p(t|ca) = \sum_d p(t, ca|d). \quad (6)$$

[1] To facilitate comparison we release a list of 3,490 names along with the documents associated with them at `http://es.cairo.au/cerc/data/balog`, the list is available for registered licensees of the CERC collection.

Statistics on Tables

- **Web Tables:** The WebTables systems (Cafarella et al., 2008a) extracts 14.1 billion HTML tables and finds 154 million are high-quality tables (1.1%)
- **Web Tables:** Lehmberg et al. (2016) extract 233 million content tables from Common Crawl 2015 (2.25% of all tables)
- **Wikipedia Tables:** The current snapshot of Wikipedia contains more than 3.23 million tables from 520k articles Fetahu et al. (2019)
- **Spreadsheets:** The number of worldwide spreadsheet users is estimated to exceed 400 million, and about 50 to 80% of businesses use spreadsheets
- ...

Outline

- First half (1.5 hrs):
  - Part I: Introduction
  - Part II: Table Interpretation
  - Part III: Knowledge base augmentation
- 30 mins break
- Second half (1.5 hrs):
  - Part IV: Table Search
  - Part V: Table augmentation
  - Part VI: Question answering on tables and other tasks

Part I: Table Extraction (in Introduction)

Definition

Table extraction is the process of extracting, classifying and storing tabular data in a consistent format, resulting ultimately in a table corpus.

Part II: Table Interpretation

Definition

Table interpretation encompasses methods that aim to make tabular data processable by machines.

Three specific subtasks:

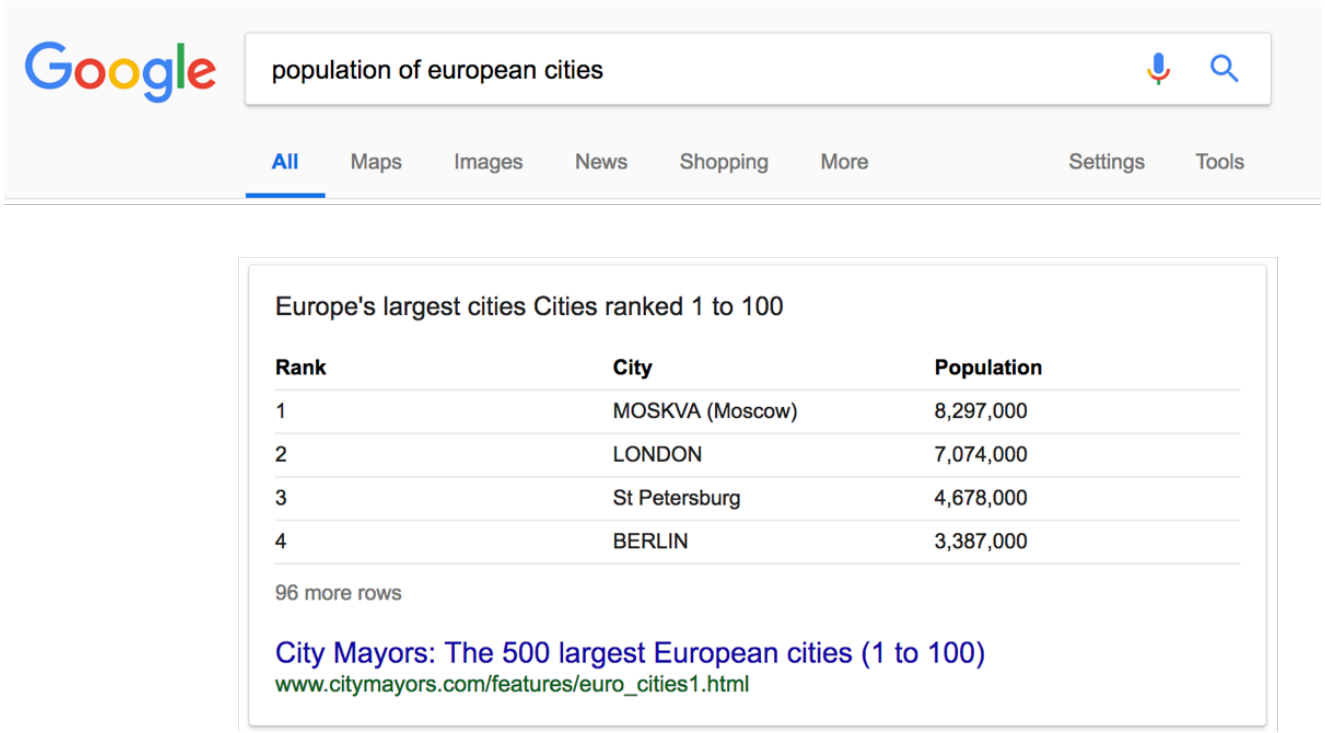
1. Column type identification
2. Entity linking in tables
3. Relation extraction

Part III: Knowledge Base Augmentation

Definition

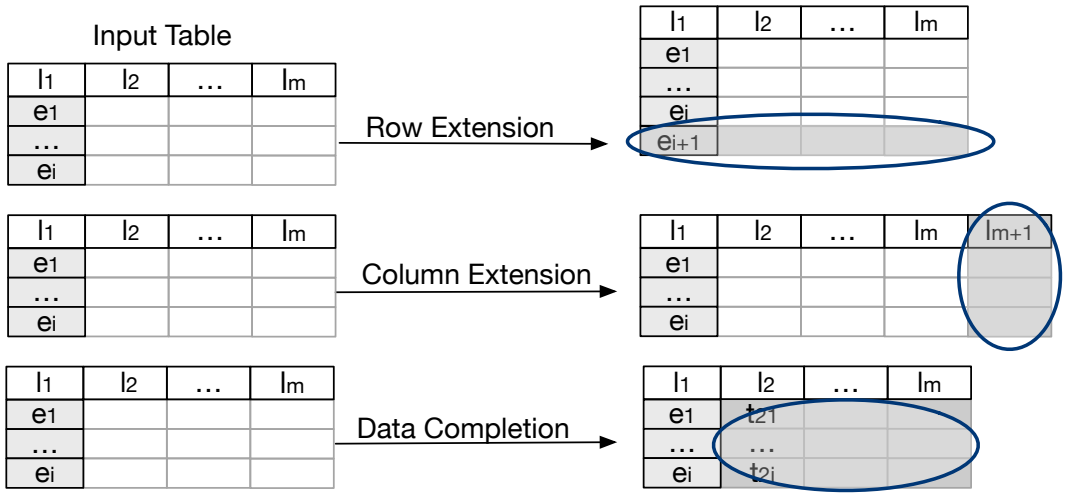
Knowledge base augmentation, also known as *knowledge base population*, is concerned with generating new instances of relations using tabular data and updating knowledge bases with the extracted information.

Part IV: Table Search



Part V: Table Augmentation

1. Row extension
2. Column extension
3. Data completion



Part VI: QA on Tables

Facts/relations in tables can be used for answering questions

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...	...	...	...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

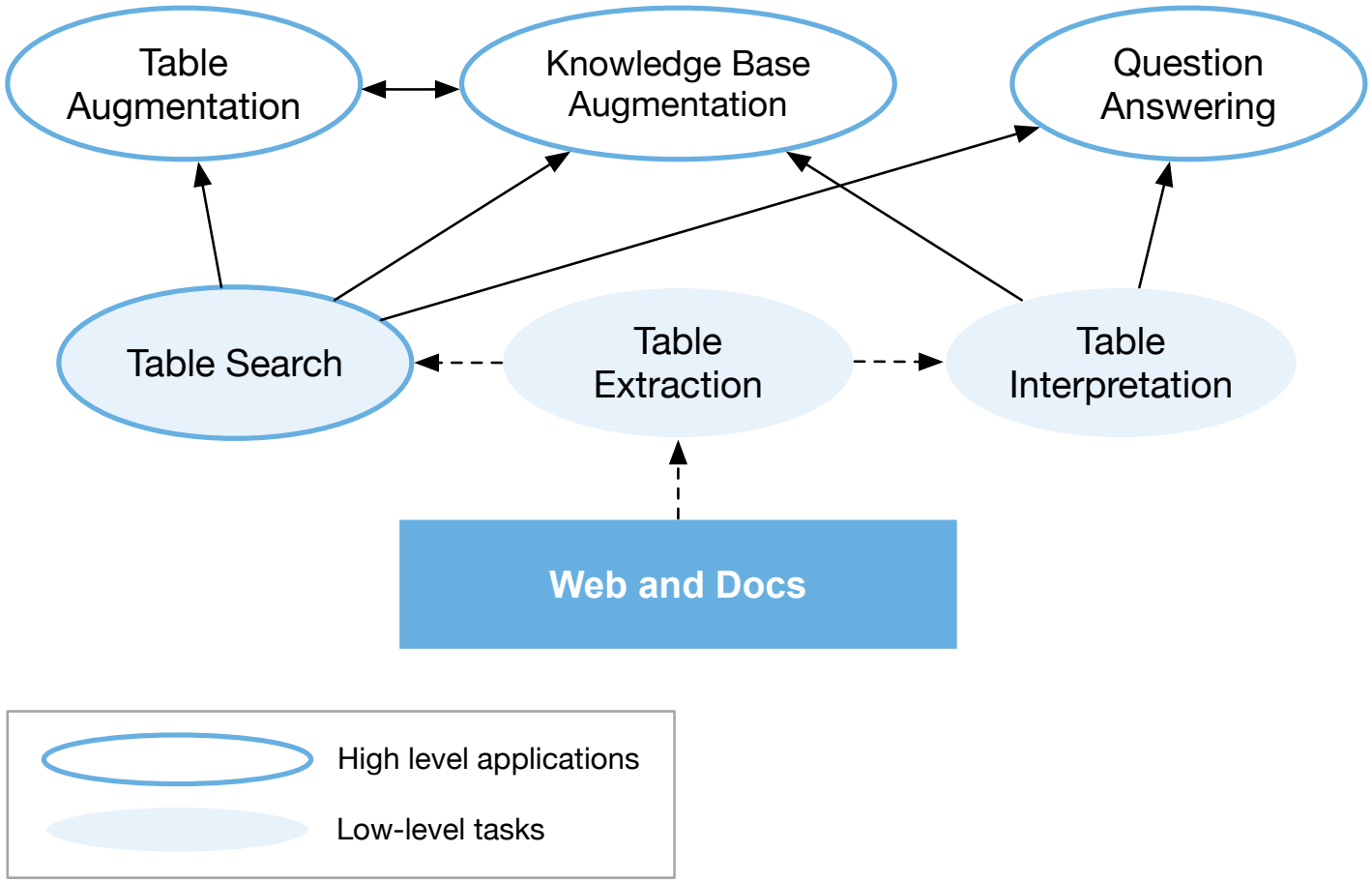
x_1: *"Greece held its last Summer Olympics in which year?"*
y_1: {2004}

x_2: *"In which city's the first time with at least 20 nations?"*
y_2: {Paris}

x_3: *"Which years have the most participating countries?"*
y_3: {2008, 2012}

Figure: Illustration from Pasupat and Liang (2015)

Table-related Information Access Tasks and Their Relationships



About this Tutorial

- Focus is on the breadth of tasks and approaches
- Key ideas are highlighted, without discussing intricate details
- Slides will be distributed after the tutorial

Introduction

SIGIR 2019 tutorial - Part I

Shuo Zhang and Krisztian Balog

University of Stavanger

Outline for this Part

1. **Table types**
2. Table extraction
3. Table corpora

Types of Tables

List of most Olympic gold medals over career

No.	Athlete	Nation	Sport
1	Michael Phelps	United States	Swimming
2	Larisa Latynina	Soviet Union	Gymnastics
3	Nikolai Andrianov	Soviet Union	Gymnastics
4	Ole Einar Bjørndalen	Norway	Biathlon
5	Boris Shakhlin	Soviet Union	Gymnastics
6	Edoardo Mangiarotti	Italy	Fencing
7	Takashi Ono	Japan	Gymnastics
8	Paavo Nurmi	Finland	Athletics

General

Lens Material	chemically strengthened glass
Bezel Material	fiber-reinforced polymer
Case material	fiber-reinforced polymer
QuickFit™ watch band compatible	yes (22 mm)
Strap material	silicone
Physical size	47 x 47 x 13.9 mm
Weight	49 g

Relational tables **Entity tables** **Other**

Development of Type Taxonomy

Reference	#types	Types
Wang and Hu (2002)	2	Genuine, non-genuine
Cafarella et al. (2008b)	5	Extremely small tables, HTML forms, calendars, non-relational tables, relational tables
Crestan and Pantel (2011)	2+7	*Relational*: listings, attribute, matrix, enumeration and form; *Layout*: navigational and formatting tables
Lautert et al. (2013)	6+3	...
Chen and Cafarella (2013)	2+7	...
...		

Type Taxonomy of (Cafarella et al., 2008b)

A number of table classification schemes have been proposed in the literature. For example, Cafarella et al. (2008b) classify web tables into five main types:

1. *Extremely small tables* are those having fewer than two rows or columns
2. *HTML forms* are used for aligning form fields for user input
3. *Calendars* are a specific table type, for rendering calendars
4. *Non-relational tables* are characterized by low quality data, e.g., used only for layout purposes (many blank cells, simple lists, etc.)
5. *Relational tables* contain high-quality relational data

The above categorization systems are quite diverse. We propose a normalized categorization scheme based on the main aspects these share.

Type Taxonomy for this Tutorial

Tables are distinguished along two main dimensions: content and layout.

Dimension	**Type**	**Description**
Content	**Relational***	Describes a set of entities with their attributes
	Entity	Describes a specific entity
	Matrix	A three dimensional data set, with row and column headers
	Other	Special-purpose tables, including lists, calendars, forms, etc.
Layout	Navigational	Tables for navigational purposes
	Formatting	Tables for visual organization of elements

Relational Tables

Definition

A *relational table* describes a set of entities in the core column(s) along with their attributes in the remaining columns.

Among all types of tables, relational tables are regarded as being of the highest quality, and are the main focus of this tutorial.

The Anatomy of a Relational Table

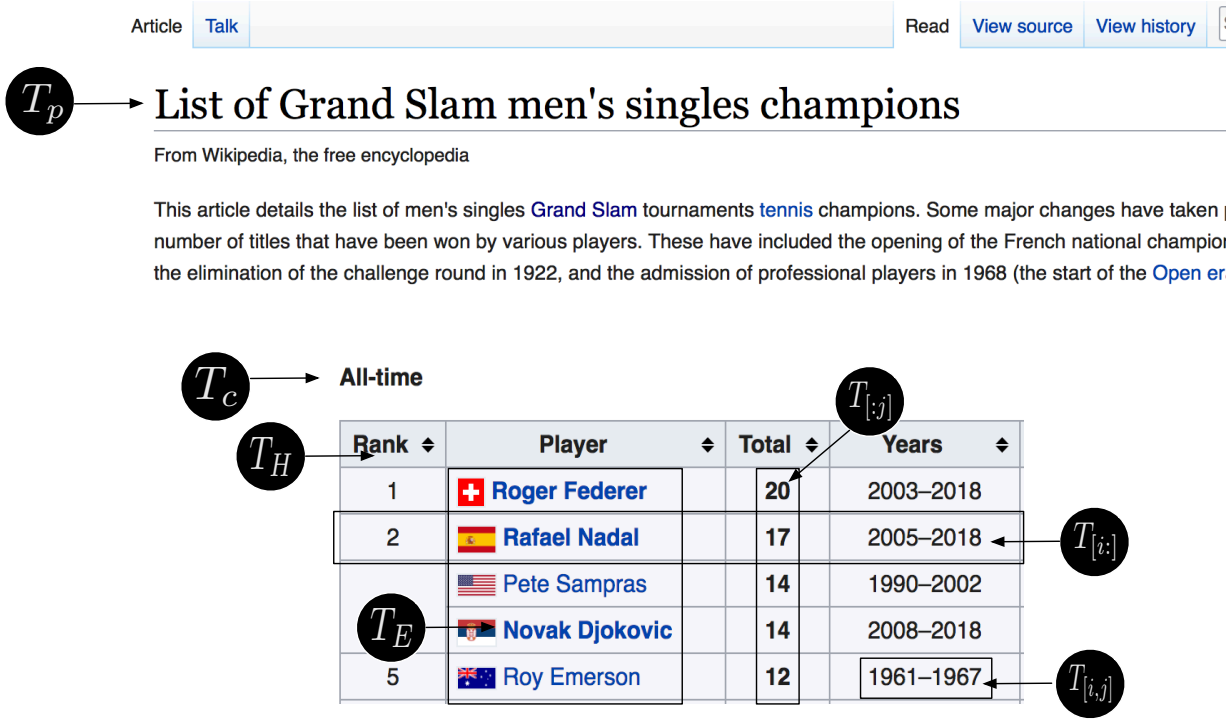
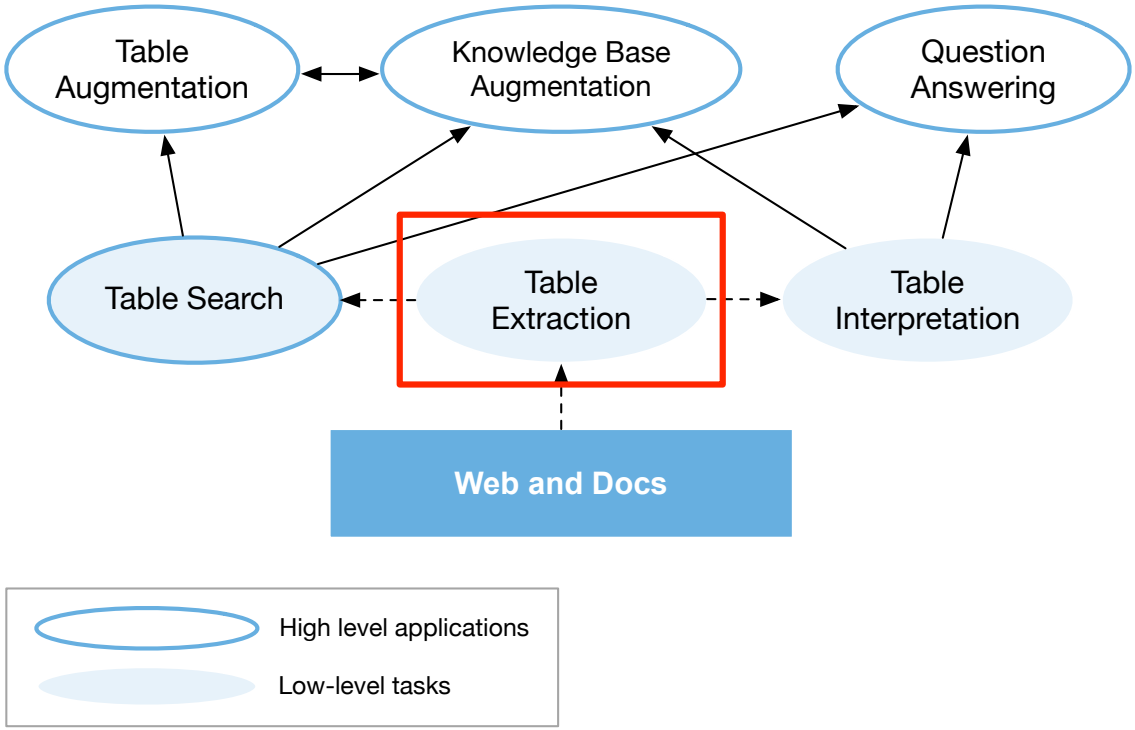


Figure: Illustration of table elements in a web table: table page title (T_p), table caption (T_c), table headings (T_H), table cell ($T_{[i,j]}$), table row ($T_{[i,:]}$), table column ($T_{[:,j]}$), and table entities (T_E).

Outline for this Part

1. ~~Table types~~
2. **Table extraction**
3. Table corpora

Table Extraction



Definition

Table extraction is the process of extracting, classifying and storing tabular data in a consistent format, resulting ultimately in a table corpus.

Table Extraction

Table extraction is concerned with the problem of identifying and classifying tables in web pages, which encompasses a range of more specific tasks:

1. Relational table classification
2. Header detection
3. Table type classification

Relational Table Classification

Definition

Relational table classification (also known as identifying high-quality or genuine tables) refers to the task of predicting whether a web table contains relational data.

Cafarella et al. (2008b)

rows
cols
% rows w/mostly NULLS
cols w/non-string data
cell strlen avg. μ
cell strlen stddev. σ
cell strlen $\frac{\mu}{\sigma}$

- One of the pioneering works
- Relational tables are filtered by training a rule-based classifier
- The classifier uses table characteristics, like table size and table tags, as features. The model is trained on a set of manually annotated tables (as being relational or non-relational) by two human judges
- As a result, they construct a high-quality table corpus, consisting of 154 million tables, filtered from 14.1 billion HTML tables

Results

True class	Precision	Recall
Relational	0.41	0.81
Non-relational	0.98	0.87

- Cafarella et al. (2008b) tuned the training procedure to favor recall over precision, since most downstream applications will need to perform the relevance ranking in any case
- Retain about 125M of the 154M relations they believe exist in the raw crawl, at a cost of sending 271M tables to the WebTables search indexer

Outline for this Part

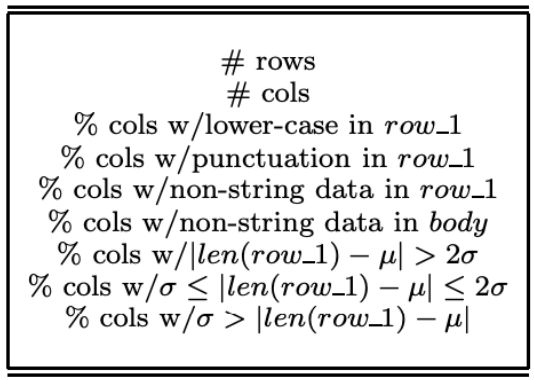
1. ~~Table types~~
2. **Table extraction**
   1. ~~Relational table classification~~
   2. **Header detection**
   3. Table type classification
3. Table corpora

Header Detection

Definition

To extract data in a structured format, the semantics of tables need to be uncovered to some extent, for instance, whether they contain a header row or column. This is known as the task of *header detection*.

Header Detection

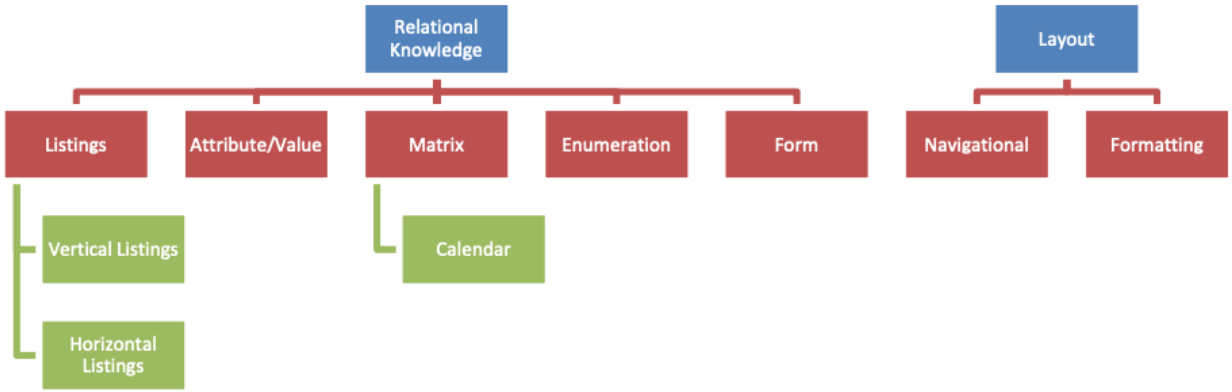


- Headers may be seen as a particular kind of table metadata
- Header detection is commonly addressed along with the other two tasks and uses similar features
- Find resources here (statistic on table schema): https://web.eecs.umich.edu/~michjc/data/acsdb.html

Outline for this Part

1. ~~Table types~~
2. **Table extraction**
   1. ~~Relational table classification~~
   2. ~~Header detection~~
   3. **Table type classification**
3. Table corpora

Table Type Classification



Definition

Table type classification is the task of classifying tables according to a pre-defined type taxonomy.

Balakrishnan et al. (2015)

- They follow a similar approach to (Cafarella et al., 2008b) for relational table classification, but use a richer set of features, which include both syntactic and semantic information
- Syntactic features are related to the structure of the table, as in (Cafarella et al., 2008b) (e.g., number of rows and columns).
- Semantic features are obtained by (detailed in part-2)
  1. Determining whether the table falls into a boilerplate section of the containing page
  2. Detecting core columns
  3. Identifying column types
  4. Detecting binary relationships between columns
- They also developed a table search engine (Google Fusion Tables)

Take-away Points from (Balakrishnan et al., 2015)

1. Most certainly bad tables can be easily excluded using simple rules
2. Semantic features contribute to table type identification

Lehmberg et al. (2016)

- A web table corpus is constructed from the Common Crawl (WDC Web Table Corpus, to be detailed later)
- First, they filter out non-genuine tables (referred to as not innermost tables, i.e., tables that contain other tables in their cells) and tables that contain less than 2 columns or 3 rows
- Then, using the table extraction framework of DWTC [1], the filtered tables are classified as either relational, entity matrix, or layout tables (Eberius et al., 2015)

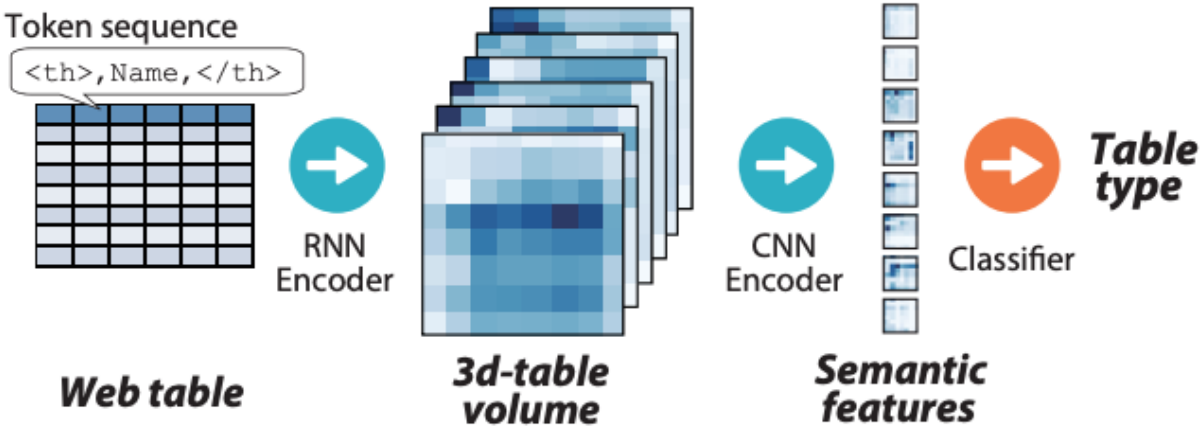
[1] https://wwwdb.inf.tu-dresden.de/misc/dwtc/

Take-away Points from (Lehmberg et al., 2016)

Top level domains	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

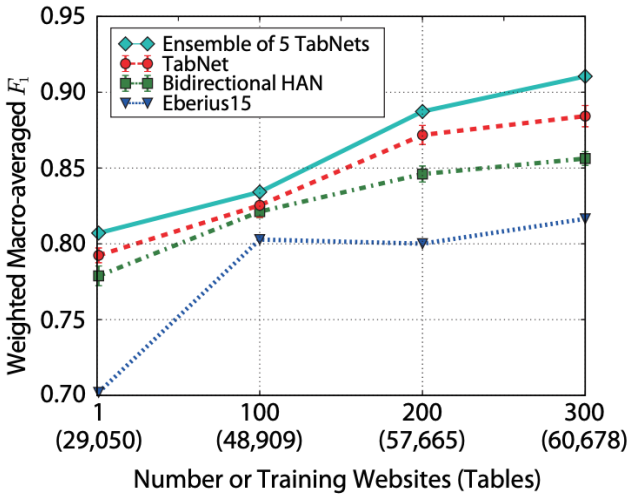
1. Numerical and string attributes share almost equal fractions
2. Contextual metadata and timestamp information are provided

Nishida et al. (2017)



1. They design a framework named TabNet, consisting of RNN Encoder, CNN Encoder, and Classifier
2. The RNN Encoder encodes the input table cells to create a 3D table volume, like image data, in the first step
3. The CNN encoders encode the 3D table volume to capture table semantics, which is used for table type classification by the Classifier

Results



Method	F1
Cafarella08	0.6926
TabNet	0.8842

- Full training dataset result in the best performance
- Learning more tables covering various structures and topics helps to understand the semantics of tables
- Though TabNet is designed to capture table structure, it can be applied to any matrix for type classification

Feature Summary I

Table: Selected features for relational table classification (RTC), header detection (HD), and table type classification (TTC) (Part 1/3).

Features	Explanation	Task
Global layout features		
Max rows	Maximal number of cells per row	RTC, TTC
Max cols	Maximal number of cells per column	RTC, TTC
Max cell length	Maximal number of characters per cell	RTC, TTC
#rows	Number of rows in the table	RTC, HD
#cols	Number of columns in the table	RTC, HD
%rows	Percentage of rows that are mostly NULL	RTC
#cols non-string	Number of columns with non-string data	RTC
μ	Average length of cell strings	RTC
δ	Standard deviation of cell string length	RTC
$\frac{\mu}{\delta}$	Cell string length	RTC
%length one	Percentage of columns with $\lvert len(row_1) - \mu\rvert > 2\delta$	HD
%length two	Percentage of columns with $\delta \leq \lvert len(row_1) - \mu\rvert \leq 2\delta$	HD
%length three	Percentage of columns with $\lvert len(row_1) - \mu\rvert < \delta$	HD
Avg rows	Average number of cells across rows	RTC, TTC
Avg cols	Average number of cells across columns	RTC, TTC
Avg cell length	Average length of characters per cell	RTC, TTC

Feature Summary II

Table: Selected features for relational table classification (RTC), header detection (HD), and table type classification (TTC) (Part 2/3).

Features	Explanation	Task
Layout features		
Std dev rows	Standard dev. of the number of cells per row	RTC
Std dev cols	Standard dev., of the number of cells per column	RTC
Std dev cell length	Standard dev. of the number of characters per cell	RTC
Local length avg	Average size of cells in segment	RTC
Local length variance	Variance of size of cells in segment	RTC

Feature Summary III

Table: Selected features for relational table classification (RTC), header detection (HD), and table type classification (TTC) (Part 3/3).

Features	Explanation	Task
Content features		
%body non-string	Percentage of non-string data in table body	HD
%header non-string	Percentage of non-string data in the first row	HD
%header punctuation	Percentage of columns with punctuation in the first row	HD
Local span ratio	Ratio of cells with a ⟨span⟩ tag	RTC, TTC
Local ratio header	Cells containing a ⟨th⟩ tag	RTC, TTC
Local ratio anchor	Cells containing an ⟨a⟩ tag	RTC, TTC
Local ratio input	Cells containing ⟨input⟩ tag	RTC, TTC
Ratio img	Ratio of cells containing images	RTC, TTC
Ratio form	Ratio of cells containing forms	RTC, TTC
Ratio hyperlink	Ratio of cells containing hyperlinks	RTC, TTC
Ratio alphabetic	Ratio of cells containing alphabetic characters	RTC, TTC
Ratio digit	Ratio of cells containing numeric characters	RTC, TTC
Ratio empty	Ratio of empty cells	RTC, TTC
Ratio other	Ratio of other cells	RTC, TTC

Extracting Other Tables

- Wikipedia Tables (Bhagavatula et al., 2015; Fetahu et al., 2019)
  - Extract tables from the Wikipedia dump based on markup
  - Entity linking is not needed
  - Pay attention to header detection (spanning headings)
- Spreadsheets (Chen and Cafarella, 2013)
  - Spreadsheets are often roughly relational
  - A data frame is defined as a block of numerical data
  - 50.5% of the spreadsheets contain a data frame and 32.5% of them have hierarchical top or left attributes
- Scientific tables
- Tables from PDFs
- ...

Take-away Points for Table Extraction

1. Table extraction aims to extract and store tabular data for convenient utilization
2. Relational table classification is important as they account for only 1% of all tables
3. Table type taxonomy is developed for table type classification
4. For table extraction, feature engineering works well, but neural approaches may also be used

Outline for this Part

1. ~~Table types~~
2. ~~Table extraction~~
    1. ~~Relational table classification~~
    2. ~~Header detection~~
    3. ~~Table type classification~~
3. **Table corpora**

Table Corpora

Table corpora	Type	#tables	Source
WDC 2012 Web Table Corpus	Web tables	147M	Common Crawl
WDC 2015 Web Table Corpus	Web tables	233M	Common Crawl
Dresden Web Tables Corpus	Web tables	174M	Common Crawl
WebTables	Web tables	154M	Web crawl
WikiTables 2013	Wikipedia tables	1.6M	Wikipedia
WikiTables 2017	Wikipedia tables	3.3M	Wikipedia
TableArXiv	Scientific tables	0.34M	arxiv.org
TableBank	Image tables	417K	Documents

Take-away Points for Table Corpora

1. Several large-scale corpora are publicly available (not limited to large search engine companies)
2. Table corpora are a result of one-off extraction efforts, can become outdated

Bibliography I

Sreeram Balakrishnan, Alon Y. Halevy, Boulos Harb, Hongrae Lee, Jayant Madhavan, Afshin Rostamizadeh, Warren Shen, Kenneth Wilder, Fei Wu, and Cong Yu. Applying webtables in practice. In *Proceedings of the Conference on Innovative Data Systems Research*, CIDR '15, 2015.

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Tabel: Entity linking in web tables. In *Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*, pages 425–441, New York, NY, USA, 2015. Springer-Verlag New York, Inc.

Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: Exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549, August 2008a. ISSN 2150-8097.

Michael J. Cafarella, Alon Y. Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu 0002. Uncovering the relational web. In *Proceedings of the Eleventh International Workshop on the Web and Databases*, WebDB '08, 2008b.

Zhe Chen and Michael Cafarella. Automatic web spreadsheet data extraction. In *Proceedings of the 3rd International Workshop on Semantic Search Over the Web*, SS@ '13, pages 1–8, New York, NY, USA, 2013. ACM.

Eric Crestan and Patrick Pantel. Web-scale table census and classification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 545–554, New York, NY, USA, 2011. ACM.

Bibliography II

Julian Eberius, Katrin Braunschweig, Markus Hentsch, Maik Thiele, Ahmad Ahmadov, and Wolfgang Lehner. 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*, pages 41–50, 2015.

Besnik Fetahu, Avishek Anand, and Maria Koutraki. Tablenet: An approach for determining fine-grained relations for wikipedia tables. In *Proc. of WWW '19*, pages 2736–2742, 2019.

Larissa R. Lautert, Marcelo M. Scheidt, and Carina F. Dorneles. Web table taxonomy and formalization. *SIGMOD Rec.*, 42(3):28–33, October 2013. ISSN 0163-5808.

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*, WWW '16 Companion, pages 75–76, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

Kyosuke Nishida, Kugatsu Sadamitsu, Ryuichiro Higashinaka, and Yoshihiro Matsuo. Understanding the semantic structures of tables with a hybrid deep neural network architecture. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 168–174, 2017.

Bibliography III

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL '15, pages 1470–1480, 2015.

Yalin Wang and Jianying Hu. Detecting tables in html documents. In *Proceedings of the 5th International Workshop on Document Analysis Systems V*, DAS '02, pages 249–260, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-44068-2.