Identifying Web Tables
Supporting a Neglected Type of Content on the Web

Michael Galkin, Dmitry Mouromtsev, Sören Auer
Outline

• Motivation
• Neglected?
• Pipeline
• Machine Learning
• Evaluation
• Applications
Motivation

Motivation

the DATA

http://commoncrawl.org/

http://stats.lod2.eu/
Motivation

the DATA

• 9 tables on an average HTML page

Motivation

the DATA

• 9 tables on an average HTML page

• 12 B tables were extracted

Motivation

the DATA

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts

Motivation

the DATA

- 9 tables on an average HTML page
- 12 B tables were extracted
- A table contains ~50 facts
- 600 B facts

Motivation

Tables are a natural way how people interact with structured data (for humans)

<table>
<thead>
<tr>
<th>Thing1</th>
<th>Thing2</th>
<th>Thing3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Param1</td>
<td>val11</td>
<td>val21</td>
</tr>
<tr>
<td>Param2</td>
<td>val12</td>
<td>val22</td>
</tr>
<tr>
<td>Param3</td>
<td>val13</td>
<td>val23</td>
</tr>
</tbody>
</table>
## Motivation

Tables are a natural way how people interact with structured data (for humans)

<table>
<thead>
<tr>
<th>Thing1</th>
<th>Thing2</th>
<th>Thing3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Param1</td>
<td>val11</td>
<td>val21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Param2</td>
<td>val12</td>
<td>val22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Param3</td>
<td>val13</td>
<td>val23</td>
</tr>
</tbody>
</table>
Neglected?

• Processing - search engines index anything except tables

• Annotation - web tables annotation tools are on the edge of extinction

• Retrieval - exists for pictures, audio, video data
Goals

Bring the tables knowledge to the LOD Cloud

• Search & Ingestion

• Explore characteristics of a table

• Extract knowledge & map with the existing knowledge graph
Pipeline

Search for tables → Extract → Genuineness assessment → Generate an ontology → Enrich the KB
Machine Learning

Did we find a data table or not?

- Genuineness assessment
  a table contains some data, not HTML formatting

both are represented via `<table>`
Machine Learning

Which data table did we find?

- **Orientation assessment**
  it is essential to distinguish concepts from their properties

<table>
<thead>
<tr>
<th></th>
<th>Header 1</th>
<th>Header 2</th>
<th>Header 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obj1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

horizontal orientation

<table>
<thead>
<tr>
<th></th>
<th>Obj1</th>
<th>Obj2</th>
<th>Obj3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Header 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Header 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Header 3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

vertical orientation
Machine Learning

String Metric → Heuristics → ML

Levenshtein

Horizontal cell similarity

Naive Bayes

Jaro-Winkler

Vertical cell similarity

kNN

n-grams

Max/avg

J48

SVM

Improvements
1. All numbers are the same
2. Fixed similarity for long sentences
Evaluation

Genuineness assessment, F-Measure
Evaluation

Orientation assessment, Precision

On Fig. 4 the results of tables recognition are presented. All the tables that are marked in HTML code of web-pages as tables are coloured in red and blue. The tables were extracted from the websites of Associated Press, Sports.ru and Saint Petersburg Government Portal. According to the theory those tables might be divided in genuine and non-genuine (relevant or irrelevant) groups. It might be easily noted that the tables coloured in red use the tag for formatting reasons and do not contain appropriate table data. In contrast, the tables coloured in blue are relevant tables which data might be parsed and processed. ScraperWiki was able to extract all the red and blue tables. The user therefore should choose relevant tables for a further processing. As a counter to ScraperWiki the developed system was able to extract and process only blue genuine tables using appropriate heuristics and machine learning algorithms. Taking into account the achieved results we consider the hypothesis suggested in Section 4 demonstrated. Indeed, unstructured data contains semantics. Hence,
Applications

• Enrich enterprise or web Knowledge Graphs with billions of facts from the most neglected content type

• Business domain - refine openly published tables with financial KPIs for the data analysis; Maintain connection with the Enterprise Knowledge Graph; Use semantic technologies for business processes re-engineering.