An End-User Pipeline for Scraping and Visualizing Semi-Structured Data over the Web

June 12th 2019

G. Bosetti, S. Firmenich, M. Winckler, G. Rossi, U. Cornejo, E. Egyed

LIFIA, Facultad de Informática, UNLP - Argentina
I3S, Université Nice Sophia Antipolis → Université Côte d’Azur - France
LIRIS, Université de Lyon - France
Motivation

- Information visualization support the exploration and analysis of large datasets. Ex. Civil Analysis: http://www.inf.ufrgs.br/~rnmsilva/CivisAnalysis2/
Visualizing data over the web

- The Web is a massive source of public datasets

  - However, not all Web sites embed information visualization techniques. Many Web sites still only display semi-structured data.

  - The operations required for visualizing semi-structured data obtained from the Web is not straight-forward.
Research question

- How to apply visualization techniques over semi-structured data available on Web pages?
Preliminary study

- We analyzed a sample of sites that can match into a table dataset type
- Top 50 popular sites according to Alexa’s ranking for Argentina with:
  - at least five homogeneous elements representing a dataset member with more than a single variable
- Variables were considered just if present in all occurrences of the dataset members
- HTML elements not containing textual raw-data were not taken into account
Preliminary results

- From the 50 sites, we kept only 42 sites for analysis. We discarded:
  - 3 sites with content not suitable for all audiences
  - 2 sites with the same domain
  - 2 sites that were offline at the time of analyzing
  - 1 site with a broken engine
  - 10 did not present any data with a heterogeneous structure.

<table>
<thead>
<tr>
<th>Dataset presentation</th>
<th>HTML Table</th>
<th>HTML list</th>
<th>HTML hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>table</td>
<td>ol/ul</td>
<td>div</td>
</tr>
<tr>
<td>Variables / Columns</td>
<td>thead &gt;tr</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Members / rows</td>
<td>tbody &gt;tr</td>
<td>li</td>
<td>div / article</td>
</tr>
<tr>
<td>Datum / cell</td>
<td>td &gt;*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Occurrences in the sample</td>
<td>2</td>
<td>8</td>
<td>22</td>
</tr>
</tbody>
</table>
How to go further?
Our main contributions

- A pipeline allowing end-users to collect and visualize semi-structured data directly over the web site that publish the datasets
- Web Scraping: new extractors for semi-structured data
- Information Visualization: improving existing tools and visualization pipelines
- Web Augmentation: innovative strategies for web augmentation
- A support tool
- Preliminary evaluation of evaluate the validity and feasibility
Scrapping and visualizing pipeline

data set source

data acquisition

filtering

visual mapping

rendering

final image

new data
import data
enriching and resampling data
map data to visual representations
adjust data to fit in the screen

user interaction
Basic premises

- Users—with no need for knowledge in low-level scraping—can abstract raw data on a Web page into a data model specification (DMS)
- Users choose and apply alternative visualizations for the DMS
- A repository of information visualization and augmentations do exist
- Open source: so that developers can extend the existing visualizations, so users can apply them on any existing and third-party Web page
Visualizing data from a ranking
Process to visualizing raw-data

1. Annotate the DOM
2. Materialize data-items
3. Data transformation
4. Visualization selection
5. Visualization customization
6. Augmentation method selection

Multiple instances in a page
Multiple pages with a single instance each
Multiple instances in a single page over time
https://www.youtube.com/watch?v=tagToUHWx3c&list=PLHuNJBFXxaLBFGtbBCZ7kOUUFd-Z3aaJK&index=2
Validation of the approach

- HTML data structures that can be visualized without making data transformations
  - From the 50 popular sites according to Alexa’s ranking, we discarded:
    - cases with no numerical variables
    - no repeated textual values or dates
    - We keep 22 sites
  - We took the two first sites from the sample matching each type of HTML structure
- The extractors were successfully tested in the 6 sites
Summary of contributions

• Underlying process for visualizing semi-structured data extracted from Web page
• Identification of suitable data structures
• Preliminary analysis of the potential of web sites that can benefit from the approach
• A tool suite including
  – Extractors
  – Basic information visualization tools
  – Interactive tools allowing users to tune the process
Future work

- Further analysis considering:
  - More extensive amount of Web sites
  - Looking for more data structures
  - User testing with the tools
  - Integration of more complex data visualization tools such as graphs and multivariate data
Thanks for your attention

Contact:
Marco Winckler
wincker@unice.fr
SPARKS team
winckler@unice.fr | http://www.i3s.unice.fr/~winckler/ | +33 (0)4.92.96.51.58

17th IFIP TC.13 International Conference on Human-Computer Interaction – INTERACT 2019
September 2-6, 2019, Paphos, Cyprus.