DWTC reloaded: Feature Selection for Classification of HTML Web-Tables

Julius Gonsior

October 23, 2017

Final presentation for INF-D-960 Analyse eines Forschungsthemas

DWTC

Dresden Web Table Corpus

- Contains 125 millions web data tables
- Lastly updated in July 2014
- Extracted from Common Crawl
    - Freely available web crawl
    - 3.6 billions web pages (July ’14)
    - 266 TB in total (July ’14)

Classification

- Phase 1: filtering out layout tables
- Phase 2: classification into Relational, Entity, Matrix and Other tables
- Accuracy of Phase 1 was around 90%
- Accuracy of Phase 2 was around 80%, especially bad accuracy for matrix tables
  → improvement possible with a better Gold Standard which contains better training samples?

Table layout classes

Entity

- Describes (mostly) one object
- First row contains attribute names, second row attribute values
- Transposed form possible too

Country	State	Mayor	Population	Elevation
Germany	Saxony	Dirk Hilbert	545.000	113m

Table layout classes

Relation

- Each column represents one attribute, each row one object
- Transposed form possible too

City	Country	Population
Mexico City	Mexico	20,116,842
Shanghai	China	19,210,000
Peking	China	15,796,450
Istanbul	Turkey	14,160,467

Table layout classes

Matrix

- First row or columns contains attribute names
- Other row or columns represents multiple dimensions for the attributes
- Datatype is mostly consistent inside of Matrix table

Means of transport	2013	2008	2003
Car	38%	41%	43%
Bicycle	17%	16%	12%
Pedestrian	24%	22%	24%
Public transport	21%	21%	20%

Table layout classes

Other

- Everything else
- Often garbage tables which contain no information

Used Features

Features

- Global Features are calculated for whole table
- Local Features only for the following rows and columns:
    - First and second row and column
    - Two middlemost rows and columns
    - Last and second to last row and column
- Feature Selection using Wekas implementation of WrapperSubsetEval for RandomForest Classifier

Global Features

- Cumulative content consistency ✓
- Average cell length ✗
- Average number of rows ✓
- Average number of cols ✓
- Ratio alphabetical cells ✗
- Total amount of rows ✗
- Total amount of columns ✗
- Area size ✗
- Ratio empty cells ✗
- Standard deviation columns ✓
- Standard deviation rows ✗

Ratio Empty Cells

$$RATIO\_EMPTY\_CELLS = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ where } X_i = \begin{cases} 1, \text{ if cell is empty} \\ 0, \text{ else} \end{cases}$$

- n denotes the total amount of cells

Area size

$$AREA\_SIZE = t_w * t_h$$

- t_w denotes the width of the table
- t_h denotes the height of the table

Max rows

$$MAX\_ROWS = \max_{\forall c_i \in C} rows(c_i)$$

- C denotes the set of all rows
- $rows()$ denotes a function counting the amount of cells in a row

Local Features

- Ratio empty ✓
- Empty variance ✗
- Amount of digits variance ✓
- Average length ✓
- Length Variance ✓
- Ratio anchor ✓
- Ratio image ✓
- Ratio input ✗
- Ratio select ✓
- Ratio colon ✓
- Ratio comma ✓
- Ratio numbers ✓
- Ratio header ✓
- Ratio whitespace ✓
- Ratio special char (non alphanumeric) ✓
- Ratio percentage ✗
- Ratio year ✗
- Ratio only number ✓

Local Ratio Empty

$$RATIO\_EMPTY$$
$$= \frac{1}{n} \sum_{i=1}^{n} X_i, \text{ where } X_i = \begin{cases} 1, \text{ if cell is empty} \\ 0, \text{ else} \end{cases}$$

- *n* denotes the total amount of cells in the respective row/column

Local Empty Variance

$$EMPTY\_VARIANCE = \frac{1}{n}\sum_{i=1}^{n}(c_{ei} - \overline{c_e})^2 \text{ where } \overline{c_e} = \frac{1}{n}\sum_{i=1}^{n} c_{ei}$$

- n denotes the total amount of cells in the respective row/column
- c_{ei} is one if cell i is empty, zero otherwise

Local Amount of Digits Variance

$AMOUNT\_OF\_DIGITS\_VARIANCE$
$= \frac{1}{n} \sum_{i=1}^{n} (c_{di} - \overline{c_d})^2$ where $\overline{c_d} = \frac{1}{n} \sum_{i=1}^{n} c_{di}$

- n denotes the total amount of cells in the respective row/column
- c_{di} denotes the count of digits in the cell i

Results

2017 Gold Standard

- Relabeled tables from DWTC-2014
- Distribution over classes:

	Entity	Matrix	Other	Relation
Original classification result	1999	1309	470	1999
New assigned layout class	1798	846	831	2302

- Previous Gold Standard contained less Matrix and Other tables → hopefully the accuracy for those two classes can be improved

Evaluation of Phase 2: Layout Identification task

Metric	Entity	Relational	Matrix	Other	Weight. Avg.
			2014		
Precision	71.22	90.02	35.70	80.89	80.18
Recall	86.87	89.24	17.93	56.90	80.71
F1	77.98	89.50	21.69	65.87	79.35
			2017		
Precision	86.7	81.6	87.1	83.6	84.3
Recall	88.6	90.1	79.0	63.8	84.2
F1	87.6	85.6	82.8	72.4	83.9

Evaluation of Phase 2: Layout Identification task

Relation	Entity	Matrix	Other	← classified as
		2017		
2073	106	83	40	Relation
135	**1593**	8	62	Entity
171	5	**668**	2	Matrix
160	133	8	**530**	Other

Conclusion

Conclusion

- Improved precision and recall for almost every class, especially for matrix tables
- Without changing the Classifier the results could be improved by using a Gold Standard which contained more data points from problematic classes

Questions?