PROBLEM

Tables on web sites or in scientific papers represent a valuable source of information for the human reader. The information itself is meaningless to machines — unless it is enriched with semantic information. The Semantic Web, and specifically the Linked Open Data initiative provide means for representing any kind of knowledge semantically.

If tables were enriched semantically a variety of new applications could evolve, as is the idea of Google Fusion Tables [1] where the annotation is done by humans.

CONTRIBUTIONS

We formulated column type inference as the majority search of cell types. The method is based on the majority vote algorithm proposed in [2]. Our main contributions are:

1. Simple algorithm to annotate table headers with semantic types based on the types cell entities
2. Investigating the influence of the number of cells on the accuracy of the header-type inference

DATA SET

We used DbPedia as knowledge base with type relations rdf:type and dcterms:subject from the Dublin Core Metadata Ontology. We evaluated our system on 50 tables extracted from Wikipedia including tables mentioned in [2].

RESULT

As a result our work contributes:

⇒ Algorithm to annotate table headers
⇒ Similar accuracy as previous work with more complex methods
⇒ Reasonable to use only a small number of cells for annotating the header

METHOD

We assessed the overall performance on the complete data set. The table below shows the results for three different type vocabularies (using Rdf-Type relations only, using DublinCore subjects only, or using both), whereby macro-averaged precision is denoted as $\pi$ and recall is denoted as $\rho$. In terms of precision the combined vocabulary performs best (0.64), however only slightly better than using DublinCore subjects only (0.59), whereas Rdf-type annotations are worst (0.23). For the combined approach, $F_1$ is low due to the low recall, which is because we have more correct results in the ground-truth but consider only the best result in the evaluation.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>$\pi^M$</th>
<th>$\rho^M$</th>
<th>$F^M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rdf-Type</td>
<td>0.24</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>DublinCore</td>
<td>0.59</td>
<td>0.51</td>
<td>0.55</td>
</tr>
<tr>
<td>Rdf-Type + DublinCore</td>
<td>0.64</td>
<td>0.27</td>
<td>0.38</td>
</tr>
</tbody>
</table>

We assessed the influence of the number of cells on the accuracy of table header disambiguation. From all 192 columns we randomly selected $k$ entity annotation step and assessed the header disambiguation accuracy using the DublinCore vocabulary. We repeated the experiment 10 times with different randomly selected cells for each $k \in \{1, 2, \ldots, 8, 10, 12, 15, 20\}$. As expected for small numbers of cells the performance increases significantly when adding one more cell (e.g. from 3 to 4 cells the $F_1$ measure increases from 0.27 to 0.35 a growth of 26%). For larger numbers of cells there is less information gain by adding one more cell resulting in smaller increases in performance (all below 10%). Using 20 cells results in $F_1$ of 0.514, which is 94% of the $F_1$ achieved with all cells (0.547).

REFERENCES


COLUMN TYPE INFERENCE

We assessed the influence of the number of cells on the accuracy of table header disambiguation. From all 192 columns we randomly selected $k$ entity annotation step and assessed the header disambiguation accuracy using the DublinCore vocabulary. We repeated the experiment 10 times with different randomly selected cells for each $k \in \{1, 2, \ldots, 8, 10, 12, 15, 20\}$. As expected for small numbers of cells the performance increases significantly when adding one more cell (e.g. from 3 to 4 cells the $F_1$ measure increases from 0.27 to 0.35 a growth of 26%). For larger numbers of cells there is less information gain by adding one more cell resulting in smaller increases in performance (all below 10%). Using 20 cells results in $F_1$ of 0.514, which is 94% of the $F_1$ achieved with all cells (0.547).