Generating Linked Data by Inferring the Semantics of Tables

Varish Mulwad, Ph.D. 2015

http://ebiq.org/j/96
<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Position</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jordan</td>
<td>Chicago</td>
<td>Shooting guard</td>
<td>1.98</td>
</tr>
<tr>
<td>Allen Iverson</td>
<td>Philadelphia</td>
<td>Point guard</td>
<td>1.83</td>
</tr>
<tr>
<td>Yao Ming</td>
<td>Houston</td>
<td>Center</td>
<td>2.29</td>
</tr>
<tr>
<td>Tim Duncan</td>
<td>San Antonio</td>
<td>Power forward</td>
<td>2.11</td>
</tr>
</tbody>
</table>
Goal: Table => LOD*

<table>
<thead>
<tr>
<th>Name</th>
<th>Team</th>
<th>Position</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael Jordan</td>
<td>Chicago</td>
<td>Shooting guard</td>
<td>1.98</td>
</tr>
<tr>
<td>Allen Iverson</td>
<td>Philadelphia</td>
<td>Point guard</td>
<td>1.83</td>
</tr>
<tr>
<td>Yao Ming</td>
<td>Houston</td>
<td>Center</td>
<td>2.29</td>
</tr>
<tr>
<td>Tim Duncan</td>
<td>San Antonio</td>
<td>Power forward</td>
<td>2.11</td>
</tr>
</tbody>
</table>

@prefix dbpedia: <http://dbpedia.org/resource/>.
@prefix dbo: <http://dbpedia.org/ontology/>.
@prefix yago: <http://dbpedia.org/class/yago/>.

"Name"@en is rdfs:label of dbo:BasketballPlayer.
"Team"@en is rdfs:label of yago:NationalBasketballAssociationTeams.

"Michael Jordan"@en is rdfs:label of dbpedia:Michael Jordan.
dbpedia:Michael Jordan a dbo:BasketballPlayer.

"Chicago Bulls"@en is rdfs:label of dbpedia:Chicago Bulls.
dbpedia:Chicago Bulls a yago:NationalBasketballAssociationTeams.

All this in a completely automated way

* DBpedia
Tables are everywhere!! ... yet ...

The web – 154 million

high quality relational tables

Table 1—Characteristics and fasting lipid profiles of African-American and Caucasian patients with type 2 diabetes

Table 2. Results in the intent-to-treat population

Clinically significant bleeding, n (%) 7 (3.9) 34 (19.1) 32 (18.0)
Any overt bleeding, n (%) 34 (19.1) 34 (19.1) 32 (18.0)
Inadequate pH control, n (%) 19 (10.5) 58 (32.0) 105 (58.0)
Evidence–based medicine

Evidence-based medicine judges the efficacy of treatments or tests by meta-analyses of clinical trials. Key information is often found in tables in articles.

Figure: Evidence-Based Medicine - the Essential Role of Systematic Reviews and the Need for Automated Text Mining Tools, IHI 2010
~ 400,000 datasets 🤗😊

~ < 1 % in RDF 😞
2010 Preliminary System

T2LD Framework

Predict Class for Columns → Linking the table cells → Identify and Discover relations

Class prediction
Entity Linking

Examples of class labels:
Column – Nationality
Prediction – MilitaryConflict
Column – Birth Place
Prediction – PopulatedPlace

Predict Class for Columns

Linking the table cells

Identify and Discover relations

<table>
<thead>
<tr>
<th>Categories</th>
<th>Person</th>
<th>Place</th>
<th>Organization</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>16.95%</td>
<td>19.57%</td>
<td>38.10%</td>
<td>29.22%</td>
</tr>
<tr>
<td>Correct</td>
<td>83.05%</td>
<td>80.43%</td>
<td>61.90%</td>
<td>70.78%</td>
</tr>
</tbody>
</table>

% of correct and incorrect instances linked
Sources of Errors

• The *sequential* approach let errors percolate from one phase to the next
• The system was biased toward predicting overly general classes over more appropriate specific ones
• *Heuristics* largely drive the system
• Although we consider multiple sources of evidence, we did not *joint assignment*
A Domain Independent Framework

Pre-processing modules

Sampling → Acronym detection

Query and generate initial mappings

Joint Inference/Assignment

Generate Linked RDF → Verify (optional) → Store in a knowledge base & publish as LOD

<table>
<thead>
<tr>
<th>City</th>
<th>Mayor</th>
<th>State</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>T. Menino</td>
<td>MA</td>
<td>610,000</td>
</tr>
<tr>
<td>New York</td>
<td>M. Bloomberg</td>
<td>NY</td>
<td>8,400,000</td>
</tr>
<tr>
<td>Philadelphia</td>
<td>M. Nutter</td>
<td>PA</td>
<td>1,500,000</td>
</tr>
<tr>
<td>Baltimore</td>
<td>S. Dixon</td>
<td>MD</td>
<td>640,000</td>
</tr>
<tr>
<td>Washington</td>
<td>A. Fenty</td>
<td>DC</td>
<td>595,000</td>
</tr>
</tbody>
</table>
Query Mechanism

Team

Michael Jordan | Chicago Bulls | Shooting Guard | 1.98

{dbo:Place, dbo:City, yago:WomenArtist, yago:LivingPeople, yago:NationalBasketballAssociationTeams...}

possible types

possible entities

Chicago Bulls, Chicago, Judy Chicago ...

.........
Ranking the candidates

- $C_i = \text{“State” ; } L_{Ci} = \text{AdministrativeRegion}$

- $f_1 = \left[ \text{Levenshtein distance}(C_i, L_{Ci}), \right.$
  $\left. \text{Dice Score } (C_i, L_{Ci}), \right.$
  $\left. \text{Semantic Similarity } (C_i, L_{Ci}), \right.$
  $\text{InformationGain}(L_{Ci}) \right]$  

- $\psi_1 = \exp(w_1^T f_1(C_i, L_{Ci}))$
Ranking the candidates

- \( R_{ij} = \text{“Baltimore”} ; E_{ij} = \text{Baltimore}_\text{Maryland} \)

- \( f_2 = \left[ \text{Levenshtein distance}(R_{ij}, E_{ij}), \right. \)
  \( \text{Dice Score } (R_{ij}, E_{ij}), \)
  \( \text{PageRank } (E_{ij}), \)
  \( \text{KBScore } (E_{ij}), \)
  \( \text{PageLength } (E_{ij}) \left. \right] \)

\[
\psi_2 = \exp(w_2^T f_2(R_{ij}, E_{ij}))
\]
Joint Inference over evidence in a table

✓ Probabilistic Graphical Models
A graphical model for tables
Joint inference over evidence in a table
Parameterized graphical model

Captures interaction between row values

Captures interaction between column headers

Function that captures the affinity between the column headers and row values

Variable Node: Column header

Factor Node
Challenge: Interpreting Literals
Many columns have literals, e.g., numbers

- Predict properties based on cell values
- Cyc had hand coded rules: *humans don’t live past 120*
- We extract *value distributions* from LOD resources
  - Differ for subclasses: age of *people* vs. *political leaders* vs. *athletes*
  - Represent as *measurements*: value + units
- Metric: possibility/probability of values given distribution
Other Challenges

- Using table captions and other text is associated documents to provide context
- **Size** of some data.gov tables (> 400K rows!) makes using full graphical model impractical
  - **Sample** table and run model on the subset
- Achieving acceptable accuracy may require human input
  - 100% accuracy unattainable automatically
  - How best to let humans offer advice and/or correct interpretations?