Knowledge Base Augmentation
SIGIR 2019 tutorial - Part III

Shuo Zhang and Krisztian Balog

University of Stavanger
Outline for this Part

1. Tables for knowledge exploration
2. Knowledge base augmentation
3. Knowledge base construction
Knowledge Base Augmentation vs Table Interpretation

KBA:
1. Table type identification
2. Entity linking
3. Schema matching
4. Slot filling

Table Interpretation:
1. Column type identification
2. Entity linking
3. Relation extraction
Definition

The knowledge contained in web tables can be harnessed for knowledge exploration, which explores the knowledge such as relationships.
Knowledge Carousels \textit{(Chirigati et al., 2016)}

- Knowledge bases tend to be geared towards understanding single entities.
- Web tables contain groups of related entities and require less assembly to produce downwards or sideways from them.
- Chirigati et al. (2016) propose a method for using web tables for generating \textit{knowledge carousels}.
Knowledge Carousels (Chirigati et al., 2016) is the first system addressing this, by providing support for exploring “is-A” and “has-A” relationships.

Figure: Illustration of Knowledge Carousels, showing an example of knowledge exploration for the query of “kentucky derby” through Knowledge Carousels: (a) a downward showing the winners of Kentucky Derby; (b) a sideway representing the famous Triple Crown horse races in the US, of which Kentucky Derby is a member.
Take-away Points for Tables for Knowledge Exploration

1. It is important to know what knowledge is contained in tables.
2. Tables are highly structured and related entities are easy to find, e.g., member entities.
3. Tables are often curated with explicit contextual information and they are important to understand the concepts of entities.
4. Table structure allows for inferring implicit features by reasoning across columns.
Outline for this Part

1. Tables for knowledge exploration
2. Knowledge base augmentation
3. Knowledge base construction
Knowledge Base Augmentation

**Definition**

*Knowledge base augmentation*, also known as *knowledge base population*, is concerned with generating new instances of relations using tabular data and updating knowledge bases with the extracted information.
## Comparison of the Existing Studies

<table>
<thead>
<tr>
<th>Source</th>
<th>Tables</th>
<th>KB</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sekhavat et al. (2014)</td>
<td>Spreadsheet</td>
<td>YAGO</td>
<td>Slot filling</td>
</tr>
<tr>
<td>Cannaviccio et al. (2018)</td>
<td>Wikipedia</td>
<td>DBpedia</td>
<td>Slot filling</td>
</tr>
<tr>
<td>T2K (Ritze et al., 2015)</td>
<td>Web</td>
<td>DBpedia</td>
<td>Entity linking</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Schema Matching</td>
</tr>
<tr>
<td>Ritze et al. (2016)</td>
<td>Web</td>
<td>DBpedia</td>
<td>Slot filling</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YAGO, Wikidata, and Freebase</td>
<td>Schema matching</td>
</tr>
</tbody>
</table>
It focuses on identifying plausible relations between pair of entities that appear in the same row of a table.
To match under-explored tabular data to a Linked Data repository, Sekhavat et al. (2014) propose a probabilistic method by collecting sentences containing pairs of entities in the same row in a table.

Extracting the patterns with the help of PATTY patterns and NELLTriples.

Estimate the probability of possible relations that can be added to the Linked Data repository.
Evaluation on **spreadsheets**

Sekhavat et al. (2014) looked at 48 <singer, song> pairs from Frank Sinatra, manually verified 48 facts and found only 31 were already in YAGO.

In the experiment on 100 NBA <player, team> pairs, YAGO had 92 of them in the *is-affiliated-to* relation.
Cannaviccio et al. (2018) leverage the patterns that occur in the schemas of a large corpus of Wikipedia tables.
Methods in (Cannaviccio et al., 2018)

1. Use the facts already in DBpedia to associate a bi-column with a relation
2. Associate schemas to relations
3. Associate relations to Bi-columns
Take-away Points from (Cannaviccio et al., 2018)

1. Headings are useful, especially for Wikipedia tables
2. Find 1.7M facts
3. Resources: http://dx.doi.org/10.7939/DVN/F36TGC
Matching problems include:

1. **Table-to-class matching** (Table type identification)
2. **Attribute-to-property matching** (Schema matching)
3. **Row-to-instance matching** (Entity linking)

Ritze et al. (2015) propose an iterative matching method, T2K, to match web tables to DBpedia for augmenting knowledge bases.
Matching steps of T2K

Candidate Selection
- Initial Candidate Selection
- Candidate Class Matching
- Candidate Refinement

Value-based Matching
- Label-based schema matching
- Value-based instance matching

Property-based Matching
- Value Aggregation
- Property-based class matching

Iterative Matching
- Candidate Matching
- Property Matching
Candidate Selection:

- Search for the entity label in DBpedia, and Top-k candidates are kept.
- Determine the distribution of each entity and choose the most frequent class as candidates for schema matching.
- Candidates not belonging to a chose class are removed.
Value-based Matching of T2K

1. Candidate Selection
2. Value-based Matching:
   - The values of each entity are compared to the values of the candidates
   - Only values with the same type are compared
   - Calculate all combination similarities and choose the maximum if multi-values exist
Property-based Matching of T2K

1. Candidate Selection
2. Value-based Matching
3. Property-based Matching:
   - Aggregate the value similarities per attribute for schema matching
   - Votes from all values are summed up and the attribute property pair with the highest value is chosen (a similar attribute property pair has many similar values)
   - Heading labels are not considered
Iterative Matching of T2K

1. Candidate Selection
2. Value-based Matching
3. Property-based Matching
4. Iterative Matching: Value-based Matching and Property-based Matching are refining each other until the similarities do not change
Take-away Points of T2K

- Low recall of entities. *Solution*: soft constrain...
- Low recall of properties. *Solution*: include heading...
- T2K works well for large tables
- Feature study *(Ritze and Bizer, 2017)* (Part-2)
- This work focuses on table to DBpedia matching
- T2D golden collection is made public available
Facts about Web tables and DBpedia when matching:

1. **Entity:** 949970 of 33.3M (English relational) tables have row-to-entity correspondence. A total of 361 different classes from DBpedia ontology.

2. **Schema:** 301450 tables match 274 different DBpedia classes.

3. **Table type:** Almost 50% describe Persons and Organizations.

<table>
<thead>
<tr>
<th>DBpedia Class</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Person</td>
<td>1445104</td>
</tr>
<tr>
<td></td>
<td>- Athlete</td>
</tr>
<tr>
<td>+ Organisation</td>
<td>241286</td>
</tr>
<tr>
<td></td>
<td>- EducationalInstitution</td>
</tr>
<tr>
<td>Place</td>
<td>725546</td>
</tr>
<tr>
<td></td>
<td>- Country</td>
</tr>
<tr>
<td>Work</td>
<td>396046</td>
</tr>
<tr>
<td>+ MusicalWork</td>
<td>162397</td>
</tr>
<tr>
<td>+ Software</td>
<td>25649</td>
</tr>
<tr>
<td>Species</td>
<td>283341</td>
</tr>
</tbody>
</table>
Facts about Web tables and DBpedia when matching:

1. **Data type:** String > Numerical > Date
2. Only 2.85% of all Web tables can be matched to DBpedia
   - Cover 15.6% DBpedia entities and 3% of the entities are described in more than 100 tables
   - Cover 721 unique properties
   - Coverage can be enhanced in many manners
Manual Evaluation

Shortcomings of the method:

1. Temporal facts: objects are changing over time
2. Different granularity and conflicting values: the city of the *Emroy university* is *Druid Hills Georgia* in DBpedia. In tables, it is *Atlanta*. *Druid Hills Georgia* is a community in *Atlanta*
3. Missing objects in lists: novel entities and concept population
Data Fusion

Data fusion aims to select the triples of a group with the same subject/predicate and used as slot filling. Strategies of data fusion for generating new facts:

1. Majority/Median Fusion: voting for strings, and median for numeric and date
2. Knowledge-based Trust: assign a trust score by calculating the overlap
3. PageRank-based Trust: PageRank scores for assessing the tables

<table>
<thead>
<tr>
<th>Strategy</th>
<th>$F_o$</th>
<th>$F_{no}$</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>691 622</td>
<td>237 548</td>
<td>.369</td>
<td>.823</td>
<td>.509</td>
</tr>
<tr>
<td>KBT</td>
<td>378 892</td>
<td>64 237</td>
<td>.639</td>
<td>.785</td>
<td>.705</td>
</tr>
<tr>
<td>PR</td>
<td>691 622</td>
<td>237 548</td>
<td>.365</td>
<td>.814</td>
<td>.504</td>
</tr>
</tbody>
</table>
Fusion Results

Causes of incorrect fusion results:

2. Ambiguous entities: e.g., common names
3. Performance varies with Classes and Properties
Match Tables to Multiple KBs

Figure: Most frequent column headings. (Illustration from (Hassanzadeh et al., 2015))
Take-away Points of Knowledge Base Augmentation

1. Table matching is a key step towards knowledge base augmentation
2. Only a small portion of tables can be matched to the knowledge bases
3. The unmatched tabular data remains under exploration
Outline

1. Tables for knowledge exploration
2. Knowledge base augmentation
3. Knowledge base construction
### Definition

Instead of augmenting existing knowledge bases, web tables contain abundant information to be turned into knowledge bases themselves.
TableNet (Fetahu et al., 2019)

### t1: Continental records

<table>
<thead>
<tr>
<th>Area</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>9.85 Olusoji Fasuba, Nigeria</td>
<td>10.78 Murielle Ahoure, Ivory Coast</td>
</tr>
<tr>
<td>Asia</td>
<td>9.91 Femi Ogundode, Qatar</td>
<td>10.79 Li Xuemei, China</td>
</tr>
<tr>
<td>Europe</td>
<td>9.86 Francis Obikwelu, Portugal</td>
<td>10.73 Christine Arron, France</td>
</tr>
<tr>
<td>South America</td>
<td>10.00 Robson da Silva, Brazil</td>
<td>11.01 An Cláudia Lemos, Brazil</td>
</tr>
</tbody>
</table>

**Table Relations:**

- \((t1,t2): rel_1 = \text{genderRestriction}(t1,t2)\)
- \((t2,t4): rel_1 = \text{equivalentTopics}(t2,t4)\)
- \((t1,t3): rel_1 = \text{topMenRecords}(t3,t1)\)
- \(rel_2 = \text{genderRestriction}(t1,t2)\)
- \((t1,t4): rel_1 = \text{genderRestriction}(t1,t4)\)
- \(rel_2 = \text{ageRestriction}(t4,t1)\)
- \((t3,t4): rel_1 = \text{ageRestriction}(t4,t3)\)

### t2: All-time top 25 women

<table>
<thead>
<tr>
<th>Rank</th>
<th>Time</th>
<th>Athlete</th>
<th>Country</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.49</td>
<td>Florence G.-Joyner</td>
<td>United States</td>
<td>16.07.1985</td>
</tr>
<tr>
<td>2</td>
<td>10.64</td>
<td>Carmelita Jeter</td>
<td>United States</td>
<td>20.09.2009</td>
</tr>
<tr>
<td>3</td>
<td>10.65</td>
<td>Marion Jones</td>
<td>United States</td>
<td>12.09.1996</td>
</tr>
<tr>
<td>4</td>
<td>10.70</td>
<td>Shelly-Ann F.-Prayce</td>
<td>Jamaica</td>
<td>29.06.2012</td>
</tr>
</tbody>
</table>

### t3: All-time top 25 men

<table>
<thead>
<tr>
<th>Rank</th>
<th>Time</th>
<th>Athlete</th>
<th>Country</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.58</td>
<td>Usain Bolt</td>
<td>Jamaica</td>
<td>16.08.2009</td>
</tr>
<tr>
<td>3</td>
<td>9.72</td>
<td>Asafa Powell</td>
<td>Jamaica</td>
<td>23.08.2012</td>
</tr>
</tbody>
</table>

### t4: Top 10 Junior (under-20) men

<table>
<thead>
<tr>
<th>Rank</th>
<th>Time</th>
<th>Athlete</th>
<th>Country</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.97</td>
<td>Trayvon Bromell</td>
<td>United States</td>
<td>13.06.2014</td>
</tr>
<tr>
<td>2</td>
<td>10.00</td>
<td>Trentavis Friday</td>
<td>United States</td>
<td>05.07.2014</td>
</tr>
<tr>
<td>3</td>
<td>10.01</td>
<td>Darrel Brown</td>
<td>Trinidad and Tobago</td>
<td>24.06.2003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jeff Dims</td>
<td>Jamaica</td>
<td>28.06.2005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yoshihide Kiryu</td>
<td>Japan</td>
<td>29.04.2013</td>
</tr>
</tbody>
</table>
TableNet (Fetahu et al., 2019)

- TableNet is an approach to construct a knowledge graph of interlinked tables with *has-a* and *is-a* relations.
- It has two main steps:
  1. Given a input table, it finds all candidate tables with high coverage.
  2. A neural approach that takes the columns and decides the type of relations.
Features for candidate finding (predict if a pair of tables are related).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFIDF</td>
<td>TFIDF similarity between abstracts</td>
</tr>
<tr>
<td>d2v</td>
<td>Doc2vec similarity between abstracts</td>
</tr>
<tr>
<td>w2c</td>
<td>Avg. word2vec abstract vectors similarity</td>
</tr>
<tr>
<td>c2v</td>
<td>Category embeddings similarity</td>
</tr>
<tr>
<td>category overlap</td>
<td>Direct and parent categories overlap</td>
</tr>
<tr>
<td>article sim</td>
<td>Embedding similarity of the article pair</td>
</tr>
<tr>
<td>type overlap</td>
<td>Type overlap</td>
</tr>
<tr>
<td>column sim</td>
<td>Column title and distance between table headings</td>
</tr>
<tr>
<td>category rep sim</td>
<td>Category representation similarity</td>
</tr>
</tbody>
</table>
Features in the previous slide are used to remove irrelevant article pairs

In terms of recall, in most of the cases the individual features have over 0.8 coverage

*Doc2vec* provides a high reduction of 0.91
Fetahu et al. (2019) represent tables by joining column description, instance-values and column-type.

Classification is based on an RNN with LSTM cells.
LSTM and BiLSTM are able to capture the sequence information in the table schemas.

TableNet can provide the means to capture the contextual similarity between the column description, type and instance cell-values.

TableNet+type outperforms on all classes in terms of F1.

Resources: https://github.com/bfetahu/wiki_tables

Need to match to the knowledge bases before complementing the existing KBs.
Summary of this Part

1. Knowledge exploration is important for knowledge base augmentation.
2. More efficient methods are needed for table-to-KB match.
3. The unmatched tabular data deserves exploration.
4. KBs can be constructed based on tables.


