Knowledge Base Augmentation

SIGIR 2019 tutorial - Part III

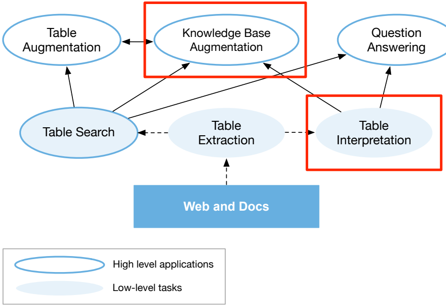
Shuo Zhang and Krisztian Balog

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

Tables for Knowledge Exploration

Definition

The knowledge contained in web tables can be harnessed for knowledge exploration, which explores the knowledge such as relationships.

Knowledge Carousels (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 *knowledge carousels*

Knowledge Carousels (Chirigati et al., 2016)

Knowledge Carousels (Chirigati et al., 2016) is the first system addressing this, by providing support for exploring "is-A" and "has-A" relationships.

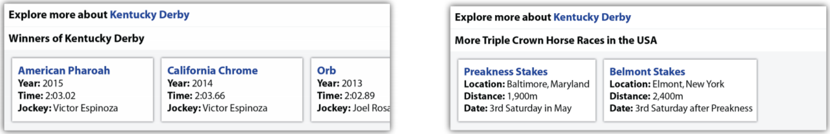


Figure: Illustration of Knowledge Carousels, showing an example of knowledge exploration for the query of "kentucky derby" through Knowledge Carousels: (a) a downward showing the winners of Kentucky Derby; (b) a sideway representing the famous Triple Crown horse races in the US, of which Kentucky Derby 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

Source	Tables	KB	Tasks
Sekhavat et al. (2014)	Spreadsheet	YAGO	Slot filling
Cannaviccio et al. (2018)	Wikipedia	DBpedia	Slot filling
T2K (Ritze et al., 2015)	Web	DBpedia	Entity linking Schema Matching
Ritze et al. (2016)	Web	DBpedia	Slot filling
Hassanzadeh et al. (2015)	Web	DBpedia,Schema.org YAGO, Wikidata, and Freebase	Entity linking Schema matching

Sekhavat et al. (2014)

Ronaldinho	Brazil	Barcelona FC
Fabio Cannavaro	Italy	Juventus
Kaka	Brazil	AC Milan
Lionel Messi	Argentina	Barcelona FC

It focuses on identifying plausible relations between pair of entities that appear in the same row of a table.

Approaches of (Sekhavat et al., 2014)

1. 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
2. Extracting the patterns with the help of PATTY patterns and NELLTriples
3. Estimate the probability of possible relations that can be added to the Linked Data repository

Towards Knowledge Augmentation

1. Evaluation on **spreadsheets**
2. 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
3. In the experiment on 100 NBA <player, team> pairs, YAGO had 92 of them in the *is-affiliated-to* relation

Cannaviccio et al. (2018)

...	Title	Directed by	Written by	...
	"Homer the Whopper"	Lance Kramer	Seth Rogen	
	"Bart Gets a 'Z'"	Mark Kirkland	Matt Selman	
	"The Great Wife Hope"	Matthew Faughnan	Carolyn Omine	
	"Boy Meets Curl"	Chuck Sheetz	Rob LaZebnik	
	"The Color Yellow"	Raymond S. Persi	Billy Kimball	

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

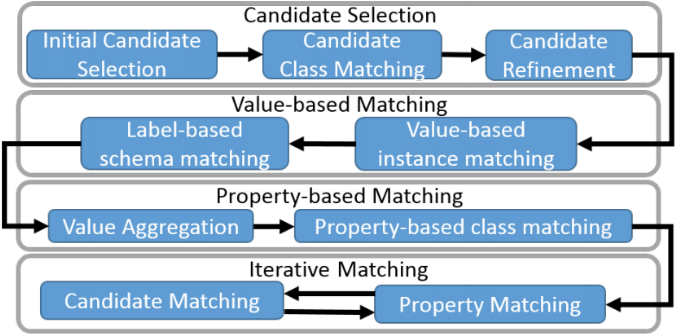
T2K (Ritze et al., 2015)

Matching problems include:

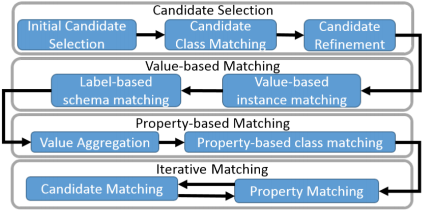
1. *Table-to-class matching* (Table type identificatioin)
2. *Attribute-to-property matching* (Schema matching)
3. *Row-to-instance matching* (Entity linking)

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

Matching steps of T2K

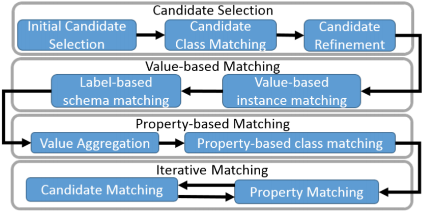


Candidate Selection of T2K



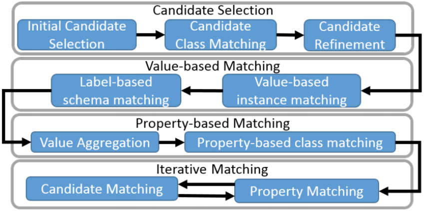
1. Candidate Selection:
   - Search for the entity label in DBpedia, and Top-k candidates are kept
   - Determine the distribution of each entity and choose the most frequent class as candidates for schema matching
   - Candidates not belonging to a chose class are removed

Value-based Matching of T2K



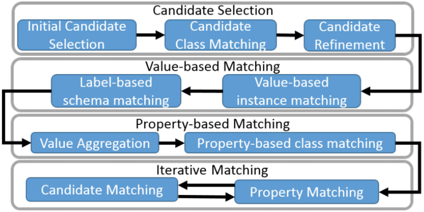
1. Candidate Selection
2. Value-based Matching:
   - The values of each entity are compared to the values of the candidates
   - Only values with the same type are compared
   - Calculate all combination similarities and choose the maximum if multi-values exist

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

Task	Precision	Recall	F1	F1 (opt.)
Entities	0.90	0.76	0.82	0.86
Properties	0.77	0.65	0.70	0.73
Classes	0.94	0.94	0.94	0.97

- 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

Ritze et al. (2016)

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

DBpedia Class	Instances
+ Person	1 445 104
|– Athlete	280 976
+ Organisation	241 286
|– EducationalInstitution	35 190
Place	725 546
|– Country	1 694
Work	396 046
+ MusicalWork	162 397
+ Software	25 649
Species	283 341

Ritze et al. (2016)

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

Strategy	F_o	F_{no}	Precision	Recall	F1
MM	691 622	237 548	.369	.823	.509
KBT	378 892	64 237	.639	.785	.705
PR	691 622	237 548	.365	.814	.504

Fusion Results

Causes of incorrect fusion results:

1. Conversion issues: e.g., date format (6/9/1987 VS 9/6/1987)
2. Ambiguous entities: e.g., common names
3. Performance varies with Classes and Properties

Match Tables to Multiple KBs

(a) DBpedia Ontology Tags

(b) DBpedia YAGO Classes Tags

(c) Freebase Type Tags

(d) Wikidata Type Tags

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**

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 schema:
Area/Nation: Location
Athlete: Person {M, F}
```

```
t2 schema:
Date: Date
Country: Location
Athlete: Person {F}
```

```
t3 schema:
Date: Date
Country: Location
Athlete: Person {M}
```

```
t4 schema:
Date: Date
Country: Location
Athlete: Person {M, age<20}
```

t1: Continental records

Area	Men			Women		
	Time	Athlete	Nation	Time	Athlete	Nation
Africa	**9.85**	Olusoji Fasuba	Nigeria	**10.78**	Murielle Ahoure	Ivory Coast
Asia	**9.91**	Femi Ogunode	Qatar	**10.79**	Li Xuemei	China
Europe	**9.86**	Francis Obikwelu	Portugal	**10.73**	Christine Arron	France
South America	**10.00**	Robson da Silva	Brazil	**11.01**	An Cláudia Lemos	Brazil

```
Table Relations:

(t1,t2): rel_1 = genderRestriction(t1,t2)
         rel_2 = topWomanRecords(t2,t1)

(t1,t3): rel_1 = topMenRecords(t3,t1)
         rel_2 = genderRestriction(t1,t2)

(t1,t4): rel_1 = genderRestriction(t1,t4)
         rel_2 = ageRestriction(t4,t1)

(t3,t4): rel_1 = ageRestriction(t4,t3)

(t2,t4): rel_1 = equivalentTopics(t2,t4)
```

top women records
gender restriction

top men records
gender restriction

gender restriction
age restriction

t2: All-time top 25 women

Rank	Time	Athlete	Country	Date
1	**10.49**	Florence G.-Joyner	United States	16.07.1988
2	**10.64**	Carmelita Jeter	United States	20.09.2009
3	**10.65**	Marion Jones	United States	12.09.1998
4	**10.70**	Shelly-Ann F.-Pryce	Jamaica	29.06.2012

t3: All-time top 25 men

Rank	Time	Athlete	Country	Date
1	**9.58**	Usain Bolt	Jamaica	16.08.2009
2	**9.69**	Tyson Gay	United States	20.09.2009
		Yohan Blake	Jamaica	23.08.2012
4	**9.72**	Asafa Powell	Jamaica	23.08.2012

age restriction

t4: Top 10 Junior (under-20) men

Rank	Time	Athlete	Country	Date
1	9.97	Trayvon Bromell	United States	13.06.2014
2	10.00	Trentavis Friday	United States	05.07.2014
3	10.01	Darrel Brown	Trinidad and Tobago	24.08.2003
		Jeff Demps	Jamaica	28.06.2008
		Yoshiihide Kiryu	Japan	29.0.4.2013

equivalent

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

Candidate Selection in TableNet (Fetahu et al., 2019)

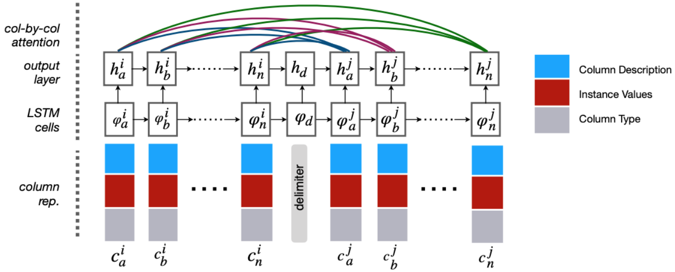
Features for candidate finding (predict if a pair of tables are related).

Feature	Description
TFIDF	TFIDF similarity between abstracts
d2v	Doc2vec similarity between abstracts
w2c	Avg. word2vec abstract vectors similarity
c2v	Category embeddings similarity
category overlap	Direct and parent categories overlap
article sim	Embedding similarity of the article pair
type overlap	Type overlap
column sim	Column title and distance between table headings category representation sim

Candidate Selection in TableNet (Fetahu et al., 2019)

- 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

Classification in TableNet (Fetahu et al., 2019)



- Fetahu et al. (2019) represent tables by joining column description, instance-values and column-type
- Classification is based on an RNN with LSTM cells

TableNet Results (Fetahu et al., 2019)

- 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

Bibliography I

Matteo Cannaviccio, Lorenzo Ariemma, Denilson Barbosa, and Paolo Merialdo. Leveraging wikipedia table schemas for knowledge graph augmentation. In *Proc. of WebDB'18*, pages 1–6, 2018.

Fernando Chirigati, Jialu Liu, Flip Korn, You (Will) Wu, Cong Yu, and Hao Zhang. Knowledge exploration using tables on the web. *Proc. VLDB Endow.*, 10(3):193–204, November 2016. ISSN 2150-8097.

Besnik Fetahu, Avishek Anand, and Maria Koutraki. Tablenet: An approach for determining fine-grained relations for wikipedia tables. In *Proc. of WWW '19*, pages 2736–2742, 2019.

Oktie Hassanzadeh, Michael J. Ward, Mariano Rodriguez-Muro, and Kavitha Srinivas. Understanding a large corpus of web tables through matching with knowledge bases: an empirical study. volume 1545 of *CEUR Workshop Proceedings*, pages 25–34. CEUR-WS.org, 2015.

Dominique Ritze and Christian Bizer. Matching web tables to dbpedia - A feature utility study. In *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017.*, pages 210–221, 2017.

Dominique Ritze, Oliver Lehmberg, and Christian Bizer. Matching html tables to dbpedia. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, WIMS '15, pages 10:1–10:6, New York, NY, USA, 2015. ACM.

Bibliography II

Dominique Ritze, Oliver Lehmberg, Yaser Oulabi, and Christian Bizer. Profiling the potential of web tables for augmenting cross-domain knowledge bases. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 251–261, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

Yoones A. Sekhavat, Francesco Di Paolo, Denilson Barbosa, and Paolo Merialdo. Knowledge base augmentation using tabular data. In *Prof. of WWW '14*, 2014.