Knowledge Graph Construction, Curation and Its Machine Learning Applications

Jiaoyan Chen (陈矫彦)
Department of Computer Science, University of Oxford
Outline

• **Background**
  • Knowledge Graph and Semantic Web

• **Knowledge Graph Construction and Curation**
  • Knowledge Construction (Semantic Table Annotation)
  • Knowledge Integration
  • Knowledge Refinement

• **Knowledge Graph Applications**
  • Urban Computation
  • Transfer Learning
  • Zero-shot Learning
The Term of “Knowledge Graph” (KG)

- The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with knowledge gathered from a variety of sources.
  - Proposed around 2012
- Knowledge ≈ Instances + Facts
- KG ≈ Linked Structured Data (can be regarded as a multi-relational graph)
KG in The Semantic Web

- RDF (Resource Description Framework)
  - Triple (三元组): <Subject, Predicate, Object>
  - Widely used for representing facts
    - E.g., <University of Oxford, locatedIn, Oxford>

- RDF Schema
  - Meta data (schema) of facts and instances
    - E.g., classes, properties, property domain and range

- Web Ontology Language (OWL)
  - Schema and description logics
    - E.g., if Tom has a colleague who works for University of Oxford, then Tom is involved in University of Oxford
  - Widely used for representing schemas, domain taxonomies and vocabularies
    - E.g., GeoName Ontology, The Orphanet Rare Disease Ontology, FoodOn

What is a KG?
(1) Facts in RDF alone, (2) Facts in RDF + RDF Schema or Ontology, (3) Ontology alone
Why use a Knowledge Graph?

- Intuitive (e.g., no “foreign keys”)
- Data + schema (ontology)
- IRI/URI not strings
- Flexible & extensible
- Rule language
  - Location + capital → location
  - Parent + brother → uncle
- Other kinds of query
  - Navigation
  - Similarity & Locality

(This slide is from Prof. Ian Horrocks)
KG Construction

- **Crowdsourcing (Encyclopedias) & Domain Experts**
  - DBpedia, Wikidata, Zhishi.me (中文), LinkedGeoData, GeoName
- **The Web Mining**
  - NELL (Never-ending Language Learning)
- **Natural Language Text**
  - Open Information Extraction
- **Tabular data**
  - DBs, Web Tables, Excel Sheets, CSV files, etc.
- **KG Alignment (Integration)**
Transform Tables to KG (Example #1)

<table>
<thead>
<tr>
<th>StockTicker</th>
<th>Name</th>
<th>GICS Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMZN</td>
<td>Amazon</td>
<td></td>
</tr>
<tr>
<td>GOOG</td>
<td>Alphabet</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basic_Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stock</td>
</tr>
<tr>
<td>AMZN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Finance_Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ticker</td>
</tr>
<tr>
<td>AMZN</td>
</tr>
</tbody>
</table>

Shortcomings: rely on table headers, often with incomplete semantics!

- **Table to RDF Mapping:**
  - Table and column to class
  - Inter-column relation to data or object property
  - Cell to entity

Knowledge Graph Construction, Curation and Its Machine Learning Applications

2021年研究生地理大数据与空间智能暑期学校
Transform Tables to KG (Example #1)

**Basic_Info**

<table>
<thead>
<tr>
<th>StockTicker</th>
<th>Name</th>
<th>GICS Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMZN</td>
<td>Amazon</td>
<td>...</td>
</tr>
<tr>
<td>GOOG</td>
<td>Alphabet</td>
<td>...</td>
</tr>
</tbody>
</table>

**Background_Info**

<table>
<thead>
<tr>
<th>Stock</th>
<th>URL</th>
<th>CEO Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMZN</td>
<td><a href="http://www.amazon.com">www.amazon.com</a></td>
<td>Jeff Bezos</td>
<td>USA</td>
</tr>
</tbody>
</table>

**Finance_Result**

<table>
<thead>
<tr>
<th>Ticker</th>
<th>Date</th>
<th>Income ($)</th>
<th>Liabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMZN</td>
<td>06/01/2018</td>
<td>177.86 billion</td>
<td>...</td>
</tr>
</tbody>
</table>

**Table to RDF Mapping:**
- Table and column to class
- Inter-column relation to data or object property
- Cell to entity

**Semantic annotation:**
- Name of Basic_Info by class **Company** and object property **hasCompany**;
- Lead to more semantics: Company is a sub-class of Organization; the definition of hasCompany ...

**External KGs (ontologies)**

- DBpedia
- Wikidata

Knowledge Graph Construction, Curation and Its Machine Learning Applications
Table for KG Population (Example #2)

- **Semantic annotation (cell by entity):**
  - Sebastian Ferrari = dbp:Ferrari
  - Germany = dbp:Germany

- **New knowledge extraction and population**
  - Hamilton races-for Mercedes?
  - Hamilton lives-in England?
  - Hamilton rdf:type Racing Driver?
  - ..... 

<table>
<thead>
<tr>
<th>Alonso</th>
<th>McLaren</th>
<th>Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamilton</td>
<td>Mercedes</td>
<td>England</td>
</tr>
<tr>
<td>Sebastian Vettel</td>
<td>Ferrari</td>
<td>Germany</td>
</tr>
</tbody>
</table>

Table on F1

Existing KG on F1

Knowledge Graph Construction, Curation and Its Machine Learning Applications

17/08/2021
CoNet: Embedding the Semantics of Web Tables for Column Type Prediction

Jiaoyan Chen,1 Ernesto Jiménez-Ruiz,1,2,4 Ian Horrocks,1,2 Charles Sutton1,3
1Department of Computer Science, University of Oxford, UK
2School of Informatics, The University of Edinburgh, UK
3Department of Informatics, University of Oslo, Norway

Abstract
Automatically annotating column types with knowledge base (KB) concepts is a critical task to gain a basic understanding of web tables. Current methods rely on either table metadata like column name or entity correspondences of cells in the KB, and may fail to deal with growing web tables with incomplete meta information. In this paper we propose a neural network based column type annotation framework named CoNet which is able to integrate KB reasoning and lookup with machine learning and can automatically train Convolutional Neural Networks for prediction. The prediction model not only considers the contextual semantics within a cell using word representation, but also embeds the semantics of a column by learning locally features from multiple cells. The method is evaluated with DBPedia and two different web table datasets, T2Dv2 from the general Web and LinEase from Wikipedia pages, and achieves higher performance than the state-of-the-art approaches.

Learning Semantic Annotations for Tabular Data

Jiaoyan Chen1, Ernesto Jiménez-Ruiz1,2,4 Ian Horrocks1,2 and Charles Sutton1,3
1Department of Computer Science, University of Oxford, UK
2School of Informatics, The University of Edinburgh, UK
3Department of Informatics, University of Oslo, Norway

Abstract
The usefulness of tabular data such as web tables critically depends on understanding their semantics. This study focuses on column type prediction for tables without any meta data. Unlike traditional lexical matching-based methods, we propose a deep prediction model that can fully exploit a table’s contextual semantics, including table locally features learned by a Hybrid Neural Network (HNN), and inter-column semantics features learned by a knowledge base (KB) lookup and query answering algorithm. It exhibits good performance not only on individual table sets, but also when transferring from one table set to another.

Read more ...

- Our papers on semantic column type annotation (https://github.com/alan-turing-institute/SemAIDA)
- “Ten years of WebTables”, VLDB 2018 (a review on Web table and KG, especially in industry)
SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However, a lack of understanding of the semantic structure and meaning of the content may hinder this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what transformation are appropriate on the data.

Tables on the Web may also be the source of highly valuable data. The addition of semantic information to Web tables may enhance a wide range of applications, such as web search, query construction.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the table. This task however (e.g., table and column names) being missing, incomplete or ambiguous.

The SemTab challenge aims at benchmarking systems dealing with the tabular data to KG matching problem, so as to facilitate their comparison on the same basis and the reproducibility of the results.

The 2021 edition of this challenge will be collocated with the 20th International Semantic Web Conference and the 16th International Workshop on Ontology Matching.

Participation: forum and registration

We have a discussion group for the challenge where we share the latest news with the participants and we discuss issues risen during the evaluation rounds.

Please register your system using this google form.

Note that participants can join SemTab at any Round for any of the tasks/tracks.

Challenge Tasks

Accuracy Track

As in previous editions, SemTab includes the following tasks organised into several evaluation rounds:

- CTA Task: Assigning a semantic type (a DBpedia class as fine-grained as possible) to a column.
- CEA Task: Matching a cell to a Wikidata entity.
- CPA Task: Assigning a KG property to the relationship between two columns.

Ontology Alignment

- Discover **equivalence** or **subsumption** mappings between classes across two ontologies (often taxonomies)
  - E.g., Canned Mushroom in HeLis vs mushroom (canned) in FoodOn

- **Traditional Solutions**
  - Lexical index, lexical matching
  - Structure matching
  - Logical reasoning (post-checking and repair)

- **Emerging Solutions**
  - Machine learning
    - Features (name similarity, neighbor similarity, etc.)
    - Embedding (deep learning)
LogMap and its Extensions

- LogMap: lexical index/matching, structure matching, logics-based repair
- Online service: http://krrwebtools.cs.ox.ac.uk/logmap/
- Software: https://github.com/ernestojimenezruiz/logmap-matcher
- Extensions: LogMap-ML (ESWC’21), Distributed LogMap (ECAI’20), etc.
LogMap and its Extensions

- LogMap: lexical index/matching, structure matching, logics, based repair
- Extensions: LogMap-ML (ESWC’21), Distributed LogMap (ECAI’20), etc.
- Online service: http://krrwebtools.cs.ox.ac.uk/logmap/
- Software: https://github.com/ernestojimenezruiz/logmap-
matcher
Entity Alignment

- KG entity alignment
  - Equivalent entities/instances across KGs composed of large-scale facts
  - E.g., Victoria vs Queen Victoria in the right figure
- Traditional solutions
  - Lexical index and matching
  - Structure matching
  - Machine learning feature engineering
- Deep learning solutions
  - Learning embeddings in the same vector space
  - See [Su, Zequn et al. VLDB’20] for a survey and benchmarking study
Entity Alignment

- **State-of-the-art (in collaboration with Tencent Jarvis Lab)**
  - **OntoEA**: utilize the ontology to distinguish entity alignment [Findings of ACL’21]

- **PARSE**: combing probabilistic reasoning with KG embeddings [IJCAI’21]
  - E.g., utilizing relation functionality to identify equivalent objects

- **New industrial benchmarks and benchmarking study** with Tencent’s medical KGs [COLING’20]
Entity Alignment

- **OntoEA**: utilize the ontology to distinguish entity alignment [Findings of ACL’21]

- **PARSE**: combing probabilistic reasoning with KG embeddings [IJCAI’21]
  - E.g., utilizing relation functionality to identify equivalent objects

- **New industrial benchmarks and benchmarking study** with Tencent’s medical KGs [COLING’20]
Fact Correction

• Informally represented facts
  • <Yangtze_River, passesArea, “three gorges district”>
  • “tree gorges district”, as a phrase, is short of semantics such as types (remember IRI instead of string)

• Erroneous facts caused by confusions
  • <Sergio_Agüero, playsFor, Manchester_United> to <Sergio_Agüero, playsFor, Manchester_City>

A lexical matching, KG embedding and reasoning-based correction framework [WWW’20]
Applications

• Search engines (e.g., Google KG)
• Search, browse and recommendation in e-Commerce (e.g., Amazon Product Graph)
• Personal assistants (e.g., Apple Siri, Amazon Alex)
• Clinical AI (e.g., *online doctor*)
• Smart City (e.g., Cities Knowledge Graph)
• ...
KG in e-Commerce

Query Recommendation

Faceted Search and Browse
Cities Knowledge Graph

- **Understanding city dynamics**
  - Conflicting demands – housing, business, leisure, energy, ecology, etc.
- **Improve city and data and knowledge systems (manage “big data” via KG and Ontology)**
  - Data: building form, transport flows, underground infrastructure, humidity, or temperature, etc.
- **Artificial Intelligence Agents**
- **Use case: city planning and design**

See more [https://www.cares.cam.ac.uk/research/cities/](https://www.cares.cam.ac.uk/research/cities/)
Transfer Learning Augmentation and Explanation

- Transfer learning
  - Transfer data and model from source domains to a target domain for addressing sample shortage
  - Which source domains to transfer? Why?

Want to train a flight delay forecasting model for flights from LAX to JFK, but have only 20 flight delay records ...

OK. I have a lot of records from ORD to LAX, ORD to JFK, ...

I also have a lot of records from LHR to LAX, ...

If we have a KG with knowledge of all airline companies, airports, flights, etc., ... [Chen Jiaoyan, et al., KR’18]
Knowledge-driven Zero-shot Learning (ZSL)

- What is ZSL?
  - Predict samples with new classes that have never appeared in training
  - Seen classes vs unseen classes

Zero-shot Image Classification

Zero-shot Knowledge Graph Completion

Knowledge Graph

- Basketball Player
- Chicago
- ?
- work as
- locate in
- nationality

Michael Jordan
Chicago Bulls
Michael Jordan
Kobe Bryant

nationality

play for
Knowledge-driven Zero-shot Learning

- **External knowledge** (a.k.a. side information) model the relationship between classes, thus enabling the transfer of the model from seen classes to unseen classes.

- Textual description:

  "**Zebras** are white animals with black stripes, they have larger, rounder ears than horses ..."

- Attribute descriptions, e.g., visual properties of animals

- Logics & rules

- Taxonomy (category)

- Knowledge Graph (Relational Facts + Categories + Literals)
Knowledge-driven Zero-shot Learning

• More readings ...

Knowledge-aware Zero-Shot Learning: Survey and Perspective
Jiaoyan Chen*, Yuxia Geng*, Zhue Chen†, Ian Horrocks‡, Jeff Z. Pan* and Huijuan Chen*†
†Department of Computer Science, University of Oxford
‡College of Computer Science & ASIFT Knowledge Engineering Lab, Zhejiang University
*School of Informatics, The University of Edinburgh

Abstract
Zero-shot learning (ZSL) which aims at predicting classes that have never appeared during the training using external knowledge (e.g., side information) has been widely investigated. In this paper we present a literature review towards ZSL in the perspective of external knowledge, where we categorize the external knowledge, review their methods and compare different external knowledge. With the literature review, we further discuss and outline the role of symbolic knowledge in addressing ZSL and other machine learning sample shortage issues.

1 Introduction
Normal supervised machine learning (ML) classification trains a model with labeled samples and predicts the classes of subsequent samples using classes that were encountered during the training stage. Zero-shot learning (ZSL), however, aims to also predict novel classes that did not occur in the training samples. Such novel classes are known as unseen classes. ZSL has been widely investigated as a means of addressing common ML issues such as emerging

IJCAI’2021 Survey Track

Knowledge Graph Construction, Curation and Its Machine Learning Applications
2021年研究生地理大数据与空间智能暑期学校

https://mp.weixin.qq.com/s/7izRcGcEvOSqM4DF-NHw

17/08/2021
Conclusions

• What are knowledge graphs and why?

• How to construct knowledge graphs?
  • From tabular data
  • Alignment
  • Correction

• Knowledge graph applications
  • Urban computation
  • Low-resource learning
Thanks!

Q&A