DICE Project Groups SS-2021

Data Science Group (DICE)
Tutors: Michael Röder, Mohamed Sherif and Stefan Heindorf

DICE – Data Science Group, Paderborn University, Germany

February 15, 2021
Overview

Which topics will be presented?

- **DICE – Data Science Group**
- **LEMMING** is an example mimicking graph generator
- **ORCA**: a crawler analysis benchmark
- Knowledge Graph Fusion (**KGFusion**)
- Explainable Artificial Intelligence (**XAI**)

DICE Group (Data Science)
Section 1

DICE – Data Science Group
Linked Open Data (2007)

http://lod-cloud.net
Knowledge reasoning
2. Explainable AI
3. Never ending learning
4. Natural language processing
5. Data integration
6. Intelligent Question answering
7. Fact checking
8. Digital assistants
9. . . .
We want you

- **Create new software**: Develop new software and research prototypes.
- **Enhance code**: Improve existing solutions.
- **Participate**: Bring your own ideas in.
We Offer

- **Machine Learning**: State-of-the-art software (PyTorch, DEAP, ...)
- **Real data**: Millions of facts from Wikipedia (Wikidata, DBpedia)
- **Expert tutors**, who developed the core software
- **Master theses**: Topics can be extended accordingly
- **Publications** at top conferences (ISWC, ESWC, WWW)
Section 2

LEMMING
LEMMING

Data-driven applications

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...
Search
Question answering
Intelligent assistants
Machine learning

...
Industry company

- Owns a lot of data
- Wants high performance solutions for their data
LEMMING
Data sharing problem

Industry company
- Owns a lot of data
- Wants high performance solutions for their data

Solution developer
- Offers software solutions
- Can adapt it to the user's situation
Data sharing problem

**Industry company**
- Owns a lot of data
- Wants high performance solutions for their data
- Cannot share the data

**Research institute**
- Wants to research new approaches
- Has a limited set of data sets and generators

**Solution developer**
- Offers software solutions
- Can adapt it to the user’s situation
**LEMMING**

**Data sharing problem**

- **Industry company**
  - Owns a lot of data
  - Wants high performance solutions for their data
  - Cannot share the data

- **Research institute**
  - Wants to research new approaches
  - Has a limited set of data sets and generators

- **Solution Developer**
  - Offers software solutions
  - Can adapt it to the user’s situation

→ Predict the future performance of storage solutions given existing versions of a dataset.
Let \( \mathcal{J} = \{ G_1, G_2, \ldots, G_v \} \).
LEMMING

Goal

\[ G = \{ G_1, G_2, \ldots, G_v \} \]

\[ K(k) = \text{SPARQL} \]

DICE Group (Data Science)

Project Groups SS-2021

February 15, 2021
Overview

Step 1: Graph Analysis

Step 2: Learn graph-specific Invariants

Step 3: Initial Graph Generation

Step 4: Graph Amendment

Step 5: Graph Finalization
Goal

Summary

- **Problem**: LEMMING is slow and its functionality is limited
- **Solution**: Enhance the existing LEMMING implementation
- **Goal**: Improved efficiency and effectiveness
LEMMING

Goal

Summary

- **Problem**: LEMMING is slow and its functionality is limited
- **Solution**: Enhance the existing LEMMING implementation
- **Goal**: Improved efficiency and effectiveness

- Parallelization
- Smarter metrics
- ...  
- Different distribution types
- More metrics
- ...
Technologies:

- Java / Maven
- RDF (helpful)
- Graph theory (helpful)

Further information:
https://dice-research.org/teaching/LemmingPG/
ORCA
Why crawling?

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...
ORCA

Why crawling?

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...

LOD cloud figures from https://www.lod-cloud.net/
Why crawling?

- Search
- Question answering
- Intelligent assistants
- Machine learning
- ...

→ We need a crawler.

LOD cloud figures from https://www.lod-cloud.net/
ORCA
Why a new benchmark?

Crawler A

LOD cloud figures from https://www.lod-cloud.net/
ORCA

Why a new benchmark?

Crawler A

Crawler B

LOD cloud figures from https://www.lod-cloud.net/
ORCA
Why a new benchmark?

Crawler A

Crawler B

LOD cloud figures from https://www.lod-cloud.net/
ORCA
Why a new benchmark?

Problems
- Repeatability
- Unknown ground truth

LOD cloud figures from https://www.lod-cloud.net/
Generate a synthetic Data Web

- Repeatable
- Scalable
- Configurable
- Ground truth is known
## ORCA Formats

<table>
<thead>
<tr>
<th>RDF Serialisations</th>
<th>Comp.</th>
<th>HTML</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF/XML</td>
<td>ZIP</td>
<td>RDFa</td>
</tr>
<tr>
<td>RDF/JSON</td>
<td>Gzip</td>
<td>Microdata</td>
</tr>
<tr>
<td>Turtle</td>
<td>bzip2</td>
<td>Microformat</td>
</tr>
<tr>
<td>N-Triples</td>
<td></td>
<td>SPARQL</td>
</tr>
<tr>
<td>N-Quads</td>
<td></td>
<td>CKAN</td>
</tr>
<tr>
<td>Notation 3</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JSON-LD</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TriG</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TriX</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HDT</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ORCA</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

### Notes:
- ✓: Supported
- (): Partially supported
- -: Not supported
Summary

- **Problem:** ORCA does not reflect all major technologies
- **Solution:** Extend ORCA in various directions
- **Goal:** Evaluation results of a new ORCA version
**Summary**

- **Problem**: ORCA does not reflect all major technologies
- **Solution**: Extend ORCA in various directions
- **Goal**: Evaluation results of a new ORCA version

- More compression algorithms
- Microdata, microformat, ...
- Existing RDF data generators

- More complex graph generators
- ...

---

**ORCA**

**Goal**
Technologies:
- Java / Maven
- RDF (helpful)
- Docker (helpful)

Further information:
https://dice-research.org/teaching/LemmingPG/
Section 4

Knowledge Graph Fusion (KG Fusion)
KG Fusion: Overview

What is KG fusion?

LOD

KG matching

Ontology matching

Instance matching

Data consolidation

Quality assurance

Fused KG
Fused KG should be

1. more complete
2. more accurate
3. non redundant
4. richer
5. cleaner and
6. as universal description for the respective resources
KG Fusion
Example: Fusing DBpedia and Yago

1. KG matching
   - https://dbpedia.org
   - https://yago-knowledge.org

Summary
- **Problem**: KG topic(s) is not explicitly defined
- **Solution**: Apply KG matching techniques
- **Goal**: Limit next steps to deal only with similar KGs
Ontology matching
- https://dbpedia.org/ontology/Town
- https://yago-knowledge.org/resource/schema:City

Summary
- **Problem**: Classes have different labels, structure and ontologies
- **Solution**: Apply ontology matching techniques
- **Goal**: Next step match only instances of similar classes
Example: Fusing DBpedia and Yago

3 Instance matching
- https://dbpedia.org/resource/Paderborn
- https://yago-knowledge.org/resource/Paderborn

Summary
- **Problem**: KG instances have different labels, structure and ontologies
- **Solution**: Apply link discovery techniques
- **Goal**: Next step fuse only similar instances
Data consolidation

- Paderborn location in *DBpedia* is defined using `georss:point` to be
  (51.71805555555556, 8.754166666666666)
- Paderborn location in *Yago* is defined using `schema:geo` to be
  (51.7167701, 8.7666842)
- Fuse using the *keep most precise value* strategy

Summary

- **Problem**: KG instances have different properties labels and values
- **Solution**: Implement automatic fusion strategies
- **Goal**: Generate fused KG
5Quality assurance
- Benchmark the resulted fused KG

Summary
- **Problem:** No benchmark exist for KG fusion
- **Solution:** Generate our own benchmark for KG fusion
- **Goal:** Assure the quality of the fused KG
Section 5

Explainable Artificial Intelligence (XAI)
**Summary**

- **Problem**: Neural networks not explainable, rule mining not accurate
- **Solution**: Combine neural networks and rule mining
- **Goal**: Explainable and accurate predictions

**Knowledge Graph**

- Angela Merkel: Politician
  - position: Chancellor
  - type: Public Office

- Michael Dreier: Mayor
  - position: Public Office
  - type: Mayor

- Roger Federer: Tennis
  - position: Tennis

**Training Examples**

- Angela Merkel: Politician
- Roger Federer: not Politician

**Is Michael Dreier a politician?**
- Neural network: 0.95, no explanation
- Rules: yes
  - \( \exists \) position.public office \( \subseteq \) Politician
XAI: Motivation
Evolutionary Algorithms for Rule Mining

Initialization
create randomly

Selection
select best

Crossover
combine pairs

Mutation
change slightly

 Initialization: create randomly
 Selection: select best
 Crossover: combine pairs
 Mutation: change slightly
**Summary**

- **Problem**: Many (bad) candidates generated
- **Solution**: Guide crossover and mutation with neural network
- **Goal**: Generate promising candidates as soon as possible

**Initialization**
Create randomly

**Selection**
Select best

**Crossover**
Combine pairs

**Mutation**
Change slightly

---

DICE Group (Data Science)
**Summary**

- **Problem**: Evaluation of fitness function takes long time
- **Solution**: Approximate fitness function with surrogate model
- **Goal**: Enable the evaluation of more candidates
Summary

- **Problem**: Existing benchmarking datasets artificial
- **Solution**: Construct realistic datasets for important use cases
- **Goal**: Realistic evaluation of rule miners

Important use cases:
- Type prediction
- Vandalism detection
XAI: Summary

Explainable Artificial Intelligence on Knowledge Graphs

Summary

- **Problem**: Neural networks not explainable, rule mining not accurate
- **Solution**: Combine neural networks and rule mining
- **Goal**: Explainable and accurate predictions

Technical skills to learn
- Neural networks (PyTorch)
- Evolutionary algorithms (DEAP)
- Data analysis (Pandas)

Scientific skills to learn
- Literature review
- Scientific presentation
- Scientific writing
That’s all Folks!

Thank you!

Topics:
- Knowledge Graphs
- Machine Learning
- Explainability

The topics are subject to change.
More information at https://dice-research.org