

DICE Project Groups SS-2021

Data Science Group (DICE)

Tutors: *Michael Röder, Mohamed Sherif and Stefan Heindorf*



PADERBORN UNIVERSITY
The University for the Information Society

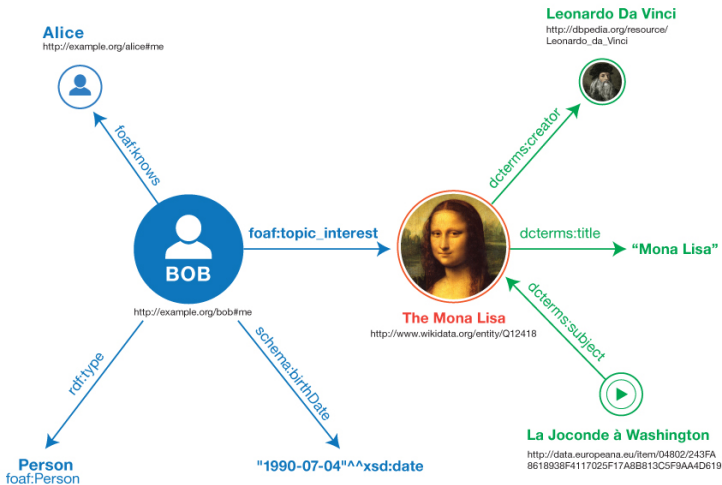
DICE – Data Science Group, Paderborn University, Germany

February 15, 2021

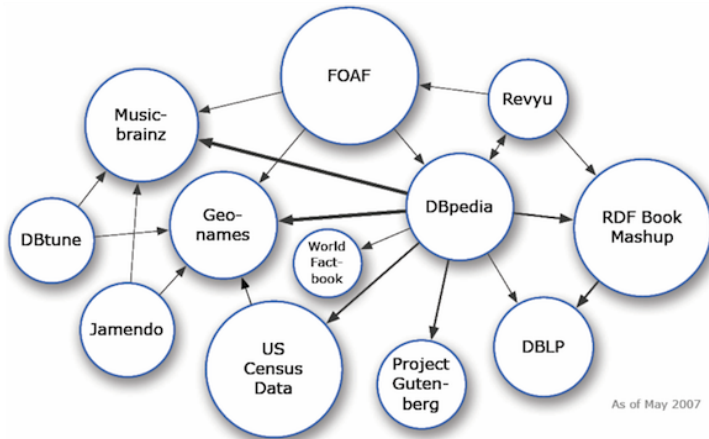
- DICE – Data Science Group
- [LEMMING](#) is an example mimicking graph generator
- [ORCA](#): a crawler analysis benchmark
- Knowledge Graph Fusion ([KGFusion](#))
- Explainable Artificial Intelligence ([XAI](#))

Section 1

DICE – Data Science Group



<https://www.w3.org/TR/rdf11-primer/>



<http://lod-cloud.net>

- Knowledge reasoning
- Explainable AI
- Never ending learning
- Natural language processing
- Data integration
- Intelligent Question answering
- Fact checking
- Digital assistants
- ...

We want you

Create new software: Develop new software and research prototype

Enhance code: Improve existing solutions.

Participate: Bring your own ideas in.

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

Search

Question answering

Intelligent assistants

Machine learning

⋮

Search

Question answering

Intelligent assistants

Machine learning

⋮

LEMMING

Data sharing problem

LEMMING

Data sharing problem

LEMMING

Data sharing problem

LEMMING

Data sharing problem

LEMMING

Growth of Knowledge Graphs

Conrads et al. \IGUANA : a generic framework for benchmarking the read- write performance of triple stores". ISWC 2017.

LEMMING

Growth of Knowledge Graphs

Conrads et al. \IGUANA : a generic framework for benchmarking the read- write performance of triple stores". ISWC 2017.

- ! Predict the future performance of storage solutions given existing versions of a dataset.

LEMMING

Goal

LEMMING

Goal

LEMMING

Overview

Summary

Problem: LEMMING is slow and its functionality is limited

Solution: Enhance the existing LEMMING implementation

Goal: Improved efficiency and effectiveness

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

...

LEMMING

Summary

Technologies:

Java / Maven

RDF (helpful)

Graph theory (helpful)

Further information:

<https://dice-research.org/teaching/LemmingPG/>

Section 3

ORCA

Search

Question answering

Intelligent assistants

Machine learning

...

Search

Question answering

Intelligent assistants

Machine learning

⋮

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

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?

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



ORCA

Why a new benchmark?

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



ORCA

Why a new benchmark?

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



ORCA

Why a new benchmark?

Problems

Repeatability

Unknown ground truth

LOD cloud gures from <https://www.lod-cloud.net/>



Generate a synthetic Data Web

Repeatable

Scalable

Configurable

Ground truth is known

	RDF Serialisations									Comp.			HTML					
	RDF/XML	RDF/JSON	Turtle	N-Triples	N-Quads	Notation 3	JSON-LD	TriG	TriX	HDT	ZIP	Gzip	bzip2	RDFa	Microdata	Microformat	SPARQL	CKAN
ORCA	X	(X)	X	X	(X)	X	(X)	(X)	(X)	{	X	X	X	X	{	}	X	X

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

...

Technologies:

Java / Maven

RDF (helpful)

Docker (helpful)

Further information:

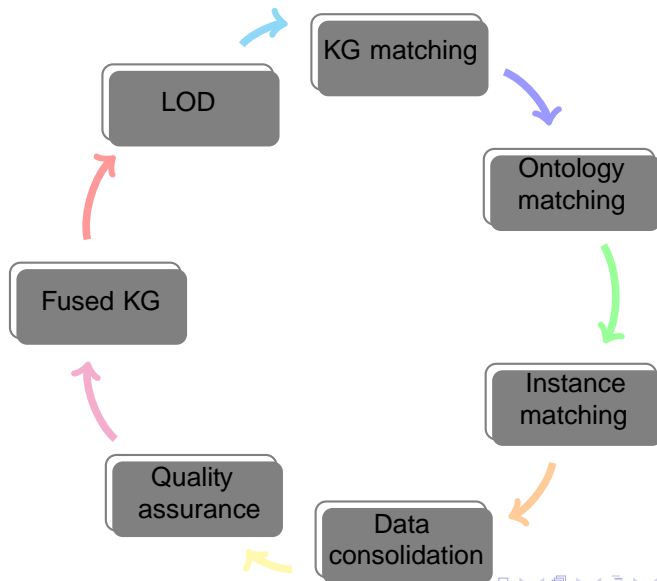
<https://dice-research.org/teaching/LemmingPG/>

Section 4

Knowledge Graph Fusion (KG Fusion)

KG Fusion: Overview

What is KG fusion?



KG Fusion: Motivation

Why we need KG fusion?

Fused KG should be

more complete

more accurate

non redundant

richer

cleaner and

as universal description for the
respective resources

KG Fusion: KG matching

Linked Open Data 2020 (<http://lod-cloud.net>)

KG Fusion

Example: FusingDBpedia and Yago

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

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

KG Fusion

Example: Fusing DBpedia and Yago

Data consolidation

Paderborn location in DBpedia is defined using `geo: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 values` strategy

Summary

Problem: KG instances have different properties labels and values

Solution: Implement automatic fusion strategies

Goal: Generate fused KG

Quality 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)

XAI: Motivation

Explainable Machine Learning 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

XAI: Motivation

Evolutionary Algorithms for Rule Mining

Summary

Problem: Many (bad) candidates generated

Solution: Guide crossover and mutation with neural network

Goal: Generate promising candidates as soon as possible

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

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

Thank you!



Topics:

- Knowledge Graphs
- Machine Learning
- Explainability

The topics are subject to change.

More information at

<https://dice-research.org>