Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

LEARNING ARBITRARY RDF DATASET ENRICHMENT GRAPHS USING PRE- & POSTCONDITION BROADCASTING

Leipzig, July 2019

vorgelegt von
Kevin Dreßler
Studiengang Informatik

Betreuender Hochschullehrer:
Prof. Dr. Axel-Cyrille Ngonga Ngomo

ABSTRACT

Short summary of the contents in English...a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beck00PSLA.html

ZUSAMMENFASSUNG

Kurze Zusammenfassung des Inhaltes in deutscher Sprache...

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— **knuth:1974** [**knuth:1974**]

ACKNOWLEDGMENTS

Put your acknowledgments here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, Jörg Weber, Henri Menke, Claus Lahiri, Clemens Niederberger, Stefano Bragaglia, Jörn Hees, Scott Lowe, Dave Howcroft, José M. Alcaide, David Carlisle, Ulrike Fischer, Hugues de Lassus, Csaba Hajdu, Dave Howcroft, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was intially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and for the contributions to the original style.

[1] Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

I APPENDIX

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

LOD	Linked Open Data
W3C	World Wide Web Consortium
DEER	RDF Dataset Enrichment Framework
DEER2	RDF Dataset Enrichment Framework 2
DAG	Directed Acyclic Graph
GP	Genetic Programming

1 INTRODUCTION

1.1 MOTIVATION

The Web of Data, also known as the Semantic Web, is growing from year to year[1], giving leeway to a vast amount of applications to harvest the knowledge cotained within the Linked Open Data (LOD) Cloud.

With growing numbers of datasets, we also see a growing number of domains being represented in the LOD Cloud, leading to the need for a growing number of novel ontologies and vocabularies. While some of these ontologies and vocabularies are well known and standardized by e.g. the World Wide Web Consortium (W3C), most are distributed over the web, hard to find and potentially model the same domain, therefore being redundant vocabularies.

This leads to a retrieval problem of ontologies and vocabularies for dataset curators which the term „Ontology Dowsing"[a] was coined[2] in order to capture the problematic unscientific guessing nature which is most common today when trying to locate a suitable ontology or vocabulary for modeling data in RDF. As a result, applications that comsume the Web of Data also often define their own specific vocabularies as it is not feasible to support, let alone be aware of, all potentially applicable ontologies and vocabularies for the specific application domain.

[a] *Dowsing is the practice of searching for ground water or metal ores using a Y-shaped rod.*

Moreover, limited resources on clients mean that the large datasets in the LOD Cloud, which are often schemaless due to the underlying Open World Assumption, have to be filtered and distilled before they can be used by applications.

We refer to the processes needed to solve the above mentioned problems as RDF Dataset Enrichment. RDF Dataset Enrichment is a quintessential part of Linked Data Integration, which also consists of the *linkage* and *fusion* of RDF Datasets.

While there has been a lot of work on the automatic linkage and fusion, RDF Dataset Enrichment has been paid little atten-

cite

[1] https://lod-cloud.net
[2] https://www.w3.org/wiki/Ontology_Dowsing

tion to, despite there being a critical need for better solutions in order to truly enable Semantic Web powered applications.

1.2 OBJECTIVES

In this thesis we address this shortcoming by extending DEER[1], the only existing approach to automated RDF Dataset Enrichment we are aware of. While DEER implements a fixed set of so called Enrichment Functions and Operators and only allows chaining them lineraly in a pipeline, we argue that this approach is too limited to of use to the very specialized needs of real-world RDF Dataset Enrichment. Therefore, our first objective will be to build an extension of DEER, called DEER2, which should be (1) highly modular, meaning that the framework should be easily extendable by third party developers in order to create specialized enrichment operators and (2) allow to represent the enrichment process as a Directed Acyclic Graph (DAG) of modular operations. These extensions should provide enough flexibility for dataset curators as well as application developers to use DEER2 in real world RDF Dataset Enrichment workflows.

The original DEER publications main contribution was the introduction of a Refinement Operator-based learning algorithm which enabled novice users to define adequate RDF Dataset Enrichment workflows. As highly modular applications in general require a lot of manual configuration of their components and therefore presume expert knowledge to precisely define how the modules operate and interact with each other, a machine learning based approach to automatic configuration will be the second and main objective of this paper.

Since introducing DAG-shaped RDF Dataset Enrichment workflows in DEER2, the complexity of the learning problem is greatly increased in comparison to DEER. We will therefore base our approach on Genetic Programming (GP) instead of Refinement Operators, since GP is known for its ability to find good solutions for hard symbolic regression problems, albeit at the cost of being non-deterministic.

1.3 DESIGN GOALS AND RESEACH QUESTIONS

We set the following goals for the design of DEER2:

(G1) DEER2 should be highly modular

(G2) DEER2 should represent RDF Dataset Enrichment workflows efficiently as DAGs

(G3) DEER2 should include a GP based learning algorithm for automatic configuration of RDF Dataset Enrichment workflows

(G4) DEER2 should improve all the identified shortcomings of DEER.

In order to measure the success of our learning approach we will aim to answer the following research questions:

(Q1) What is the optimal set of hyperparameters?

(Q2) Does our approach generalize well?

(Q3) How does our approach perform on real world datasets?

1.4 STRUCTURE OF THIS THESIS

The remainder of this thesis is structured as follows: In Chapter 2 we explore the State of the Art for fields relevant to this work and introduce some of the basic concepts required to understand this thesis. After that, we present our approach DEER2 in Chapter 3. We evaluate our approach and answer the posed research questions in Chapter 4. Finally, we conclude in Chapter 5.

Extend this

2 STATE OF THE ART

2.1 genetic algorithms

2.1.1 *Genetic Programming*

2.1.2 *Multi Expresssion Programming*

2.1.3 *Semantic Genetic Operators*

2.2 LINKED DATA

2.2.1 *Ontologies*

2.2.2 *Linked Data Quality*

2.3 LINKED DATA INTEGRATION

2.3.1 *Linking*

2.3.2 *Fusion*

2.3.3 *Enrichment*

2.3.4 *DEER*

3 APPROACH

3.1 DEER2! (DEER2!)

3.1.1 *The Enrichment Graph*

3.1.2 *On Modularity*

3.1.3 *Overview of Implemented Enrichment Operators*

3.2 A GENETIC PROGRAMMING APPROACH TO ENRICHMENT GRAPH LEARNING

3.2.1 *The Learning Problem*

3.2.2 *Heuristic Self-Configuration of Enrichment Operators*

3.2.3 *Baseline Algorithm*

3.2.4 *Enrichment Graph Compaction*

3.2.5 *Semantic Genetic Operators*

4 EVALUATION

In this chapter we are going to define our experimental protocol as well as present and discuss our results.

4.1 EXPERIMENTAL SETUP

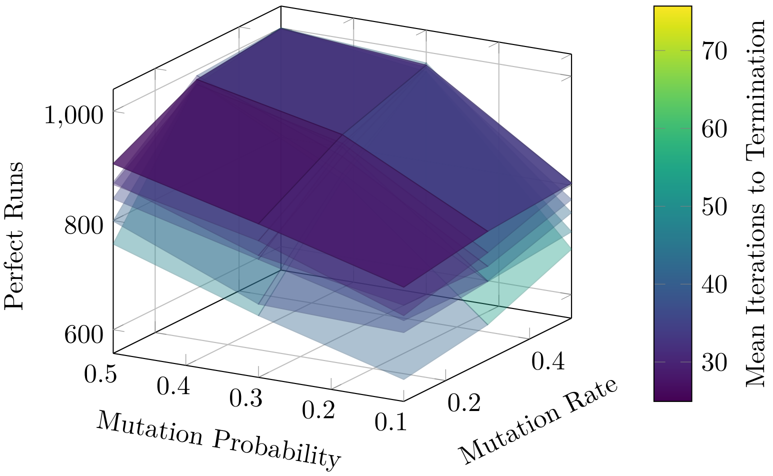
All experiments were carried out on a 64-core 2.3 GHz server running *OpenJDK* 64-Bit Server 1.8.0_151 on *Ubuntu* 16.04.3 LTS. Each experiment was assigned 128 GB RAM.

4.1.1 *Datasets*

4.1.2 *Experiments*

4.2 RESULTS

4.2.1 *Hyperparameter Optimization*



4.2.2 *Performance Evaluation*

4.3 DISCUSSION

5 CONCLUSION & FUTURE WORK

5.1 SUMMARY

5.2 FUTURE WORK

Part I

APPENDIX

A APPENDIX

Algorithm A.1: Integer division.

```
input: int N, int D
output: int
begin
  % a comment about the code
  res ← 0
  while $N \geq D$
    $N \leftarrow N - D$
    res ← res + 1
  end
  return res
end
```

Listing A.1: A floating example (`listings` manual)

```
for i:=maxint downto 0 do
begin
{ do nothing }
end;
```

BIBLIOGRAPHY

[1] Mohamed Ahmed Sherif, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. "Automating RDF Dataset Transformation and Enrichment." In: *12th Extended Semantic Web Conference, Portorož, Slovenia, 31st May - 4th June 2015*. Springer, 2015.

ERKLÄRUNG

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, July 2019

Kevin Dreßler