Ontology Matching
- using -
Outlier Detection

Master Thesis

presented by
Alexander Müller
Matriculation Number 1376818

submitted to the
Chair of Information Systems V
Prof. .Dr. Heiko Paulheim
University Mannheim

Mai 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overcoming the Disparate Data Space

Maybe use Linked Open Data The Story so far as a motivating example, or the smart data initiative of the federal government of Germany

Defining an ontology is not a deterministic task, so different authors will produce different ontologies that capture the same real life task

1.2 Contributions

Chapter 2

The Ontology Matching Problem

- First define basic terms, like ontology, ontology matching process, ontology alignment
- Give a simple motivating example
- Shortly review the state of the art in ontology matching, mostly ensembles of multiple matchers
- Express challenges, and focus on matcher selection and matcher cominbation

2.1 Defintions

2.1.1 Ontologies

What is an Ontology

In philosophy the term ontology describes the study of being and existence, trying to define categories of things and to discover relationships among them. Computer Science adopted this term for their own needs and consequently for artificial intelligence and web researchers an ontology is a formal model of a domain.([2], [9]).

In literature there exist various definitions for ontologies on different levels, some of which are discussed in [7]. Nevertheless one of the most cited definitions is the one by [6]: "An ontology is an explicit specification of a conceptualization". But probably as often as its cited its often extended by other definitions like: "An ontology is an explicit formal specification of a shared conceptualization of a domain of interest"[11] and "a logical theory which gives and explicit, partial account of a conceptualization" [8]. These two definitions extends the understanding of an ontology in three points. First of all an ontology needs to be formal, so for instance a textual description is not sufficient.[11] Moreover it needs to cover a specific domain, so that there exist more than one ontology (a key difference to philosophy).

And finally a ontology can only partial conceptualize the facts of world, so it's a simplified abstraction of the reality.[2]

This rather textual definition will be in the following precised by the introduction of the Web Ontology Language (OWL) which is heavily used in the semantic web (TODO cite linked open data the story to far) and most of the datasets provided by the OAEI are in OWL format. [1]

OWL a Ontology Language

- Show main elements of owl and relate them to the main elements of a an ontology based on [5]
- stress uniqueness of URI
- Then define it formally

Definition base on [5]

Definition 2.1 (Ontology) *An ontology is a tuple* $o = (C, I, R, T, V, \leq, \perp, \in, =)$ *such that:*

C *is the set of classes,*
I *is the set of individuals,*
R *is the set of relations,*
V *is the set of datatypes,*
with C, I, R, V *being pairwise disjoint,*
$\leq$ *is a relation on* $(C \times C) \cup (R \times R) \cup (T \times T)$ *called specialization,*
$\perp$ *is a relation on* $(C \times C) \cup (R \times R) \cup (T \times T)$ *called exclusion,*
$\in$ *is a relation over* $(I \times C) \cup (V \times T)$ *called instantiation,*
$=$ *is a relation over* $I \times R \times (I \cup U)$

2.1.2 Ontology Matching

In general ontology matching aims to reduce heterogeneity between different ontologies, to overcome disparate data spaces. As mentioned in the introduction designing an ontology in not a deterministic process (TODO Cite), as heterogeneity may occur due to different Reasons: For instance because of a different usage terms to described the same real world concept (e.g.: car vs. automobile) or different perspectives on the modeling domain (TODO cite).

Despite that common ground it needs to be said that in literature there exist various terms and definitions for ontology matching. Terms like ontology alignment, ontology mapping, integration of ontologies (TODO cite) are referring to the same concepts and techniques. To have a common understanding in the work, the term ontology matching is used when to the process of matching ontologies is referred. Complementary ontology alignment or simply alignment is used when to

the resulting output of a matching process is referred. Therefore in the following sections these two questions will be answered:

1. What is a matching process?
2. What is an alignment?

Ontology Matching Process

The foundation for each ontology matching system is a process that matches two or more ontologies in order to produce an alignment between those ontologies. That is called in general a ontology matching process. In the definition of a process is given by considering two input ontologies that should be aligned.

In the database oriented research on ontology matching this process is often called match operator ([10]), which refers in its core to the same abstraction as made in [5] and [2].

Nevermind the fact that matters is that these definitions have in common that the process has as an input two ontologies that needs to be matched , and some parameters to control the internal behavior of the underlying algorithm, usually threshold parameters. In some definitions ([4]) furthermore other resources are also included into the input domain, those can include initial alignments that will bootstrap the alignment process or background knowledge items. The output is consequently an alignment, which consists out of correspondences between entities of the given ontologies. In [5] and [2] the functional character of the matching process is stressed leading to the definition of the matching process.

Definition 2.2 (Matching Process) *An ontology matching process is a function, based on two ontologies to match O_1 and o_2, a set of parameters p and a set of resources r*

$$A = f(o_1, o_2, p, r)$$

TODO Think about adding figure

Considering this function, the expected alignment is a real subset of $A \subseteq o_1 \times o_2 \times \Theta$ with Θ being the set of possible relation types (see Section 2.1.2). This shows that the possible search space can become very big. Assuming that each Ontology contains 1000 entities meaning classes, object properties and data type properties and possibly 4 type of relations in Θ, the search space is $1000 \times 1000 \times 4 = 4,000,000$. These sizes are rather small especially in the medicine domain exists ontologies much larger than this. Nevermind this shows that the process to find an alignment for ontologies is not trivial. [2] Which properties an alignment has is illustrated in the following section.

Ontology Alignment

The definition of alignments in literature have the following criteria, to express the correspondence between entities, that belong two different ontologies:

1. How to correspond to entities of the underlying ontology
2. Which types of relationships between entities are distinguished
3. How to address the allowed or wished multiplicity of alignments

To tackle the first feature in [5] an entity language is introduced, that defines a separate ontology-format-independent language to address entities and perform operations with them. The basic advantage of this language is that it is a abstraction over the underlying ontology description language and by this allows matching entities of ontology across multiple formats. Since this is not in scope for this work, entities are assumed to have a unique identifier, which they are referenced with, without the use of an entity language.

Another important aspect of a correspondence between two entities is the type of relationship they stand to each other. Those types are often inferred from data modeling techniques([10]) or have set-theoretic background ([5]) and can define for example when two entities are equal to each other ($=$) or one entities more general than another one ($>$). The set of possible relations is called Θ. The predominant relationship in this work is the equality between two entities.

Furthermore in contrast to [5] - but in agreement with [2] and [10]- it is considered that each correspondence has a degree of confidence , expresses the likelihood that the correspondence holds. This confidence is real valued number in the interval $[0, 1]$, where 1 expresses the highest confidence and 0 the lowest. The reason why [5] is not considering a degree of confidence as part of the correspondence definition is that they define a separate meta-data element for each correspondence, which can contain arbitrary information, including the confidence. This work does not follow this definition, because of the high importance of the confidence value for the presented approach to ontology matching.

From all this, the definition of a correspondence and an alignment can be expressed as follows:

Definition 2.3 (Correspondence) *Given two ontologies o_1, o_2 and a set of relations Θ a correspondence is a 4-Tuple:*

$$(e_1, e_2, r, c)$$

with

$$\begin{aligned} e_1 &\in o_1 \wedge e_2 \in o_2 \\ r &\in \Theta \\ c &\in [0, 1] \end{aligned}$$

Definition 2.4 (Alignment) *Given two ontologies o_1, o_2 an alignment between them is defined as a set of correspondences of entities e_1 and e_2, where $e_1 \in o_1 \wedge e_2 \in o_2$.*

Ontology Alignment Multiplicity

The previous defined alignment does not consider the allowed multiplicities. But for a lot of problems an alignments with a specific cardinality are necessary. For instance that for each entity of a ontology o_1 needs to be exactly one entity of ontology o_2 mapped, or more likely at most one entity of 0_2.

In [10] and [2] those cardinalities are considered in a way that is known from data modeling languages like the Entity Relationship Diagrams or the Unified Modeling Language in the form of 1:1, 1:N, N:1 and N:M cardinalities between entities of two sets. In [5] however the multiplicities are defined more formally and in a more granular way. Thus we will follow this definition, since this definitions are used in the remainder of this thesis.

In order to define the multiplicity between entities of an alignment, the alignment between two ontologies is treated as a function f where ontology 0_1 is the domain X of f and o_2 is the co-domain Y. Thus we can use the mathematical terms surjective, injective and bijective to define the multiplicity of a function.

A function is surjective when every element $y \in Y$ of f has a corresponding value $x \in X$. This means that the function may map more than one x-value to a y value.

In contrast to that a function is injective when it preserves the uniqueness of a value $x \in X$ when it's mapped to a value $y \in Y$. Therefore intuitively speaking a function is injective if for each value $y \in Y$ it maps at most one value $x \in X$. Combining those two properties a function is bijective if it is surjective and bijective, so for each value of Y has exactly one value in X. In practice this type of function is often called one-to-one mapping.

In [5] however they define one more property of an alignment which is called total. It is an inversion of the surjective properties so that a function is total if for each value $x \in X$ at least one value $y \in Y$ is mapped. Analog to [5] the properties total and injective of an alignment can be defined:

Definition 2.5 (Total and Injective Alignment) *Given two ontologies o_1 and o_2, an alignment A is called total iff:*

$$\forall e_1 \in o_1, \exists e_2 \in o_2 : (e_1, e_2, =) \in A$$

And an alignment A, with the domain o_1 and the codomain 0_2 is called injective from 0_1 to 0_2 iff:

$$\forall e_2 \in o_2, \exists e_{2'}, e_1 \in o_2 : (e_1, e_2, =) \in A \wedge (e_{2'}, e_2, =) \in A \Rightarrow e_1 = e_{2'}$$

In [3] a expressive notation for the definition of multiplicities of alignments in ontology matching is introduced. There the an total and injective alignment is represented by 1, ? represents an injective alignment, + for total and * for an alignment that does not hold the definitions above. These properties are sensitive to the direction they are seen, for instance and A alignment of o_1 and o_2 can be injective from 0_1 to o_2 and total from o_2 to o_1. Thus this results in the following

different combinations: ?:?, ?:1, 1:?, 1:1, ?:+, +:?, 1:+, +:1, +:+, ?:*, *:?, 1:*, *:1, +:*, *:+, *:* .

An example for this can be seen in Section 2.2.

2.1.3 Matching Representation

Think about if necessary

RDF Format [4]

2.2 Motivating Example

2.3 State-of-the Art

What is the current state in research, e.g. current advances at OAEI Data sets:

- Conference, Benchmark, Library, Anatomy

2.4 Challenges

Basicalle the Euzenant Paper

Chapter 3

Ontology Matching Approaches (Related Work)

This chapter presents the results of a conducted literature survey in the area of ontology matching.

3.1 Classification of Approaches

Tailored classification of approaches

3.2 Base Matcher

3.2.1 Label Based

TODO

3.2.2 Instance Based

TODO

3.2.3 Structure Based

TODO

3.3 Hybrid Matching Approaches

3.4 Analysis of Hybrid Matching Approaches

Show weaknesses of current approaches, supervised, often weighted average based, so not flexible (transferable to other domains)

Chapter 4

Selected Approaches to Outlier Analysis

4.1 Definition Outlier Analysis

4.2 Approaches

Chapter 5

Hybrid Ontology Matching using Outlier Analysis

5.1 Motivation for using Outlier Analysis for Ontology Matching

Flexibility towards changing data domains No need to train weights upfront

5.2 Ontology Matching as an Outlier Detection Problem

5.2.1 Creating the Feature Vector

5.2.2 Significance of Outliers for Ontology Matching

5.2.3 Transforming the Outlier Analysis Result to a Matching

Chapter 6

A Matching Pipeline using Outlier Detection

6.1 Overview

Presents the implemented Pipeline

6.2 Base Matcher used

What base matcher survived the selection process

6.3 Methods used to combine Matcher

6.4 Feature Selection

6.5 Outlier Analysis

Chapter 7

Evaluation

7.1 Datasets

7.2 Experimental Setup

7.3 Used Baselines

7.4 Results

Chapter 8

Discussion

8.1 Flexibility towards changing Data Domains

8.2 Runtime Considerations

8.3 Comparison with current OAEI Participants

Chapter 9

Conclusion

Bibliography

[1] Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Daniel Faria12, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, et al. Results of the ontology alignment evaluation initiative 2014.

[2] M. Ehrig. *Ontology Alignment: Bridging the Semantic Gap*. Semantic Web and Beyond. Springer US, 2006.

[3] Jérôme Euzenat. Towards composing and benchmarking ontology alignments. In *Proc. ISWC-2003 workshop on semantic information integration*, pages 165–166, 2003.

[4] Jérôme Euzenat. An api for ontology alignment. In *The Semantic Web–ISWC 2004*, pages 698–712. Springer, 2004.

[5] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.

[6] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, December 1995.

[7] Nicola Guarino. Understanding, building and using ontologies. *Int. J. Hum.-Comput. Stud.*, 46(2-3):293–310, March 1997.

[8] Nicola Guarino and Pierdaniele Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In *Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing*, pages 25–32. IOS Press, 1995.

[9] H. Paulheim. *Ontology-based Application Integration*. SpringerLink : Bücher. Springer, 2011.

[10] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, December 2001.

[11] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: Principles and methods. *Data Knowl. Eng.*, 25(1-2):161–197, March 1998.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Mannheim, den 31.5.2015 Unterschrift