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 Definitions

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.([5], [15]).

In literature there exist various definitions for ontologies on different levels, some of which are discussed in [13]. Nevertheless one of the most cited definitions is the one by [12]: "An ontology is an explicit specification of a conceptualization". But probably as often as its cited its extended by other definitions like: "An ontology is an explicit formal specification of a shared conceptualization of a domain of interest"[19] and "a logical theory which gives and explicit, partial account of a conceptualization" [14]. 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.[19] 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 facts of the real world, so it's a simplified abstraction of the reality.[5]

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. [3]

OWL a Ontology Language

There exist a variety of ontology languages, but in this work the Web Ontology Language (OWL) will be used. OWL is based upon the eXtensinble Markup Language, a succressor of the ontology languages DAMl and OIL, extending RDF and RDFS. In this thesis the OWL 2 Standard is used, which is the second iteration of OWL and became an W3C recommendation 2009 [20]. [1]

OWL is an essential part of the linked open data stack, used as a format to express knowledge about the world. OWL exists in two flavors: OWL DL and OWL Full. The main difference lies in the decidability with reasoning with those ontologies. Since reasoning is not in scope of this thesis, the only thing worth mentioning for completeness it that OWL DL is decidable, because it is based on description logic, where OWL Full not decidable. Despite this fact, the main focus at this point lies in the concepts of OWL to conceptualize facts from the real world, by modeling an ontology. Thus is in the following an introduction of basic OWL elements is given, which is based on [1].

TODO Ontology Listing

The key element of OWL is the definition of classes. They represent an abstraction of concepts from the real world, modeled in the ontology. For instance in Listing XXX an example ontology with 4 classes can be seen. Those classes do have properties and are related to each other. The class Person for example defines the set of person, so each instanziation of the class person is a member of a set containing all persons, these members are usually called individuals or instances.

In order to model complex ontologies classes can also have a subclass relationship. So here we can see that in Listing XXX the class Celebrity is a subclass of Person, so it specializes the Person class. In OWL every class existing is a subclass of owl:Thing the most general class, so each member or instance of a class is also member of the set of instances of owl:Thing. For classes further build-in properties can be used, so for instance classes can be said to be equivalent(owl:equivalentClass) or disjoint(owl:disjointClass). More complex constructs are also possible but are not in scope of this short introduction.

Class may need to have properties to capture real world properties correctly. In OWL there exists two different types of properties: Object properties, which relate individuals with each other, so e.g. the properties lives_in of the Person class, and

[1] In the following the term OWL is used instead of OWL 2, when to some specialties of OWL 2 is referred, it will be explicitly mentioned

datatype properties that relate individuals to literal values of a certain datatype, e.g name of the Person class. For those properties a specific domain and range can be defined. So for the property lives_in the domain is that a person can only live in a city, but not in another Person, which kind of makes sense.

In addition to those key properties there exits annotation property, which contain more information about a class. A typical annotation property is rdfs:label, which often gives textual description of the given class. Those properties are ignored in OWL DL and thus are part of OWL Full. Properties itself can have some settings. For instance one can define whether there exists other properties with the same meaning or if they are functional, or symmetric and much more. Those properties will not further be mentioned because they are not relevant to the problem solved in this thesis.

The last component in OWL are individuals, as already mentioned they are instances of a specific class and there for belong to the set of all members of a class. An ontology contains individuals when some base facts needs to be modeled which can be a starting point for inferring new knowledge with reasoners, applied to the ontology. Since reasoning is out of scope in this thesis, individuals are simply treating as parts of an ontology.

Out of this summary of the definition above, now a more formal definition based on [10] of an ontology can be inferred, which will be the basis of the following chapters.

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,*
T *is the set of datatypes,*
V *is the set of values,*
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 pf terms to describe 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 this thesis, 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 ([17]), which refers in its core to the same abstraction as made in [10] and [5].

Nevermind the fact that matters is that these definitions have in common that the process has as an input of 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 ([9]) 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 [10] and [5] 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. [5] 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 [10] 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([17]) or have set-theoretic background ([10]) 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 [10] - but in agreement with [5] and [17]- 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 [10] 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 [17] and [5] 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 [10] 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 [10] 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 [10] 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 [8] 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 [9]

2.2 Motivating Example

2.3 State-of-the Art

There exits several state-of-the-art of the ontology matching systems, developed all over the world. To evaluate the performance of different systems and approaches the Ontology Alignment Evaluation Initiative was started, to asses weaknesses and strength of ontology matching algorithms and by this give developers and researches and developers in this are a platform for knowledge transfer. [7] There the OAEI publishes each year a report on an assesment of current ontology matching systems. In order to validate the state-of-the-art, analyzing the techniques used by participants of OAEI is a good starting point. In this section this finding are shortly summarized, based on those reports. [6] Add other reports
Mostly all systems that attended the challenge have in common that they rely on multiple ontology matching functions. In order to overcome weakness of different types of functions, state-of-the art ontology matching systems use different measures. They exploit the structural properties, the name of elements and sometimes individuals of ontologies in order to match entities. The main drawback of this multiple strategy or hybrid matching is that there is a need of a combination function of the results of the so called base matchers. This can be a weighted average sum (TODO cite YAM) or a majority vote of the base matchers ([4]). [18] Another similarity is modern matchers try to be as efficient as possible and therefore align ontologies faster and in addition be able to align big ontologies.[16] Furthermore

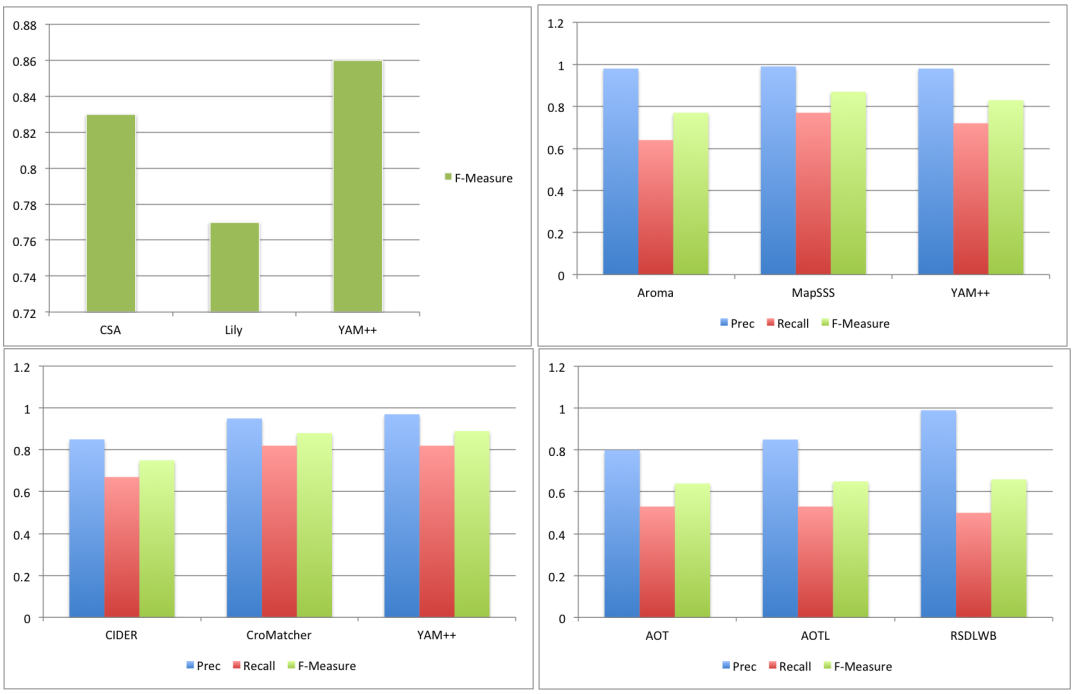

Figure 2.1: Top performing Ontology Matching Systems for the OAEI Bibliography Benchmark Dataset

most matching approaches use base matchers that rely on background knowledge, which is used to overcome a terminological mismatch, between the matched ontologies. Such resources can be WordNet [11] or UMLS[2]. [7]

Of particular interest is consequently the performance of matchings systems, which is usually estimated using the measures $rrecision$, $recall$ and $f_1-measure$ [2] from information retrieval. The OAEI offers different tracks that contain different ontologies to be matched. To illustrate the performance over those different datasets in the following the results of matching systems for four different datasets over last four years will be presented, which is an update of the work done in [18].

2.3.1 Bibliography Benchmark

2.3.2 Conference

2.3.3 Anatomy

2.3.4 Library

[2]TODO add reference to section

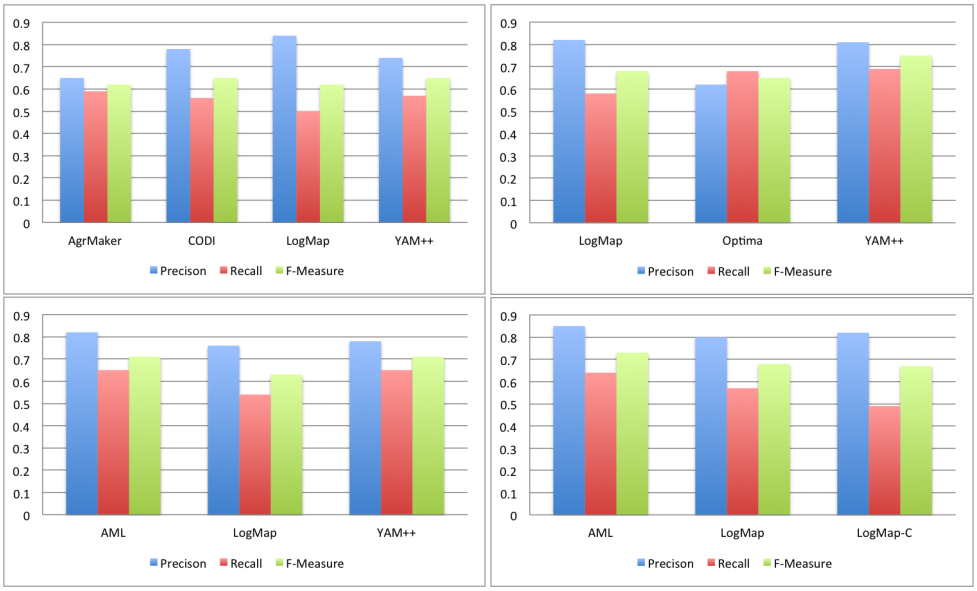

Figure 2.2: Top performing Ontology Matching Systems for the OAEI Conference Dataset

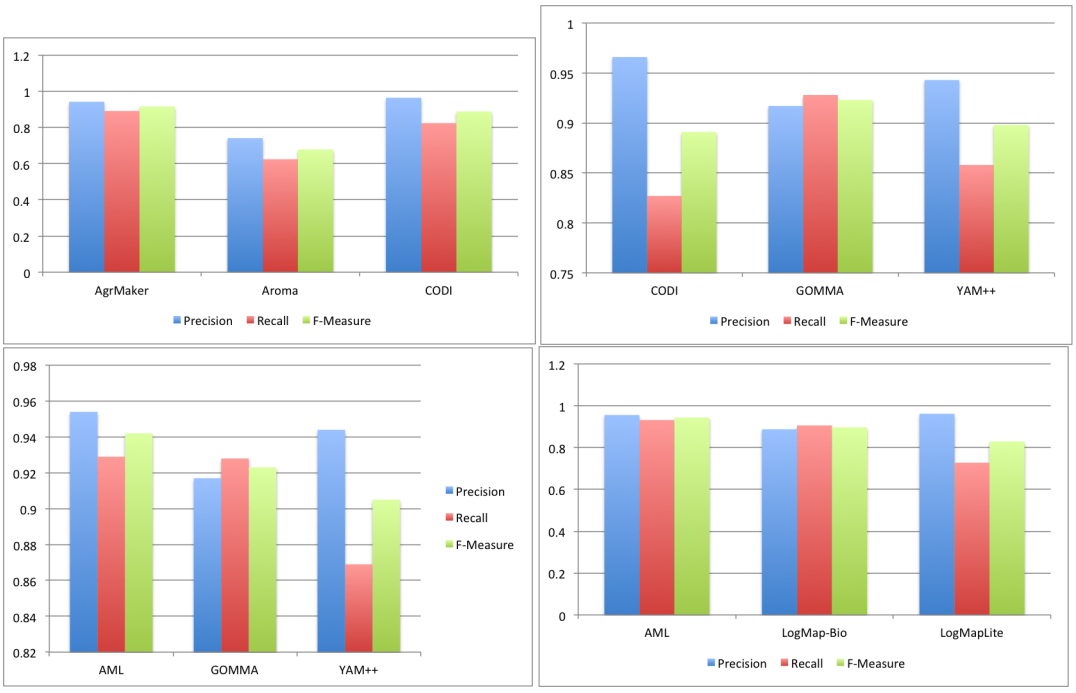

Figure 2.3: Top performing Ontology Matching Systems for the OAEI Anatomy Dataset

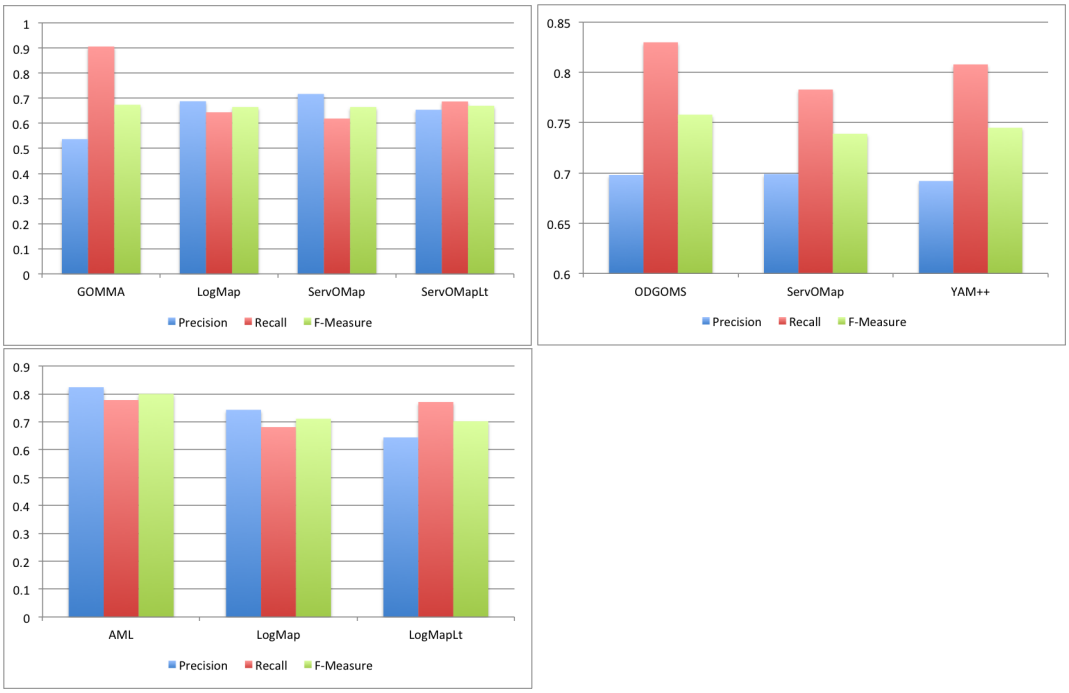

Figure 2.4: Top performing Ontology Matching Systems for the OAEI Library Dataset

2.4 Challenges

The previous section demonstrated that the ontology matching community is improving their systems in terms of $f_1 - measure$. Another way of looking at this is that there still exist various challenges that need be solved, to overcome the disparate dataspace. In [18] a research agenda is presented, that list a lot of open issues of ontology matching, This now summarized.

- Name all challenges named in Euzenat Paper
- Focus on Matcher selection, Combination and Tuning

Need for finding a function that combines different matcher, into one final matching. No Training required Unsupervised

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

Correlation between Outlier score and label

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 Matchers used

What base matchers survived the selection process

6.3 Methods Used to combine Matchers

6.4 Feature Selection

6.5 Outlier Analysis

Chapter 7

Evaluation

7.1 Datasets

7.1.1 Conference

7.1.2 Benchmark

7.1.3 Anatomy

7.1.4 Library

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] Grigoris Antoniou, Paul Groth, Frank vanHarmelen, and Rinke Hoekstra. *A Semantic Web Primer*. MIT Press, Cambridge, MA, USA, 2012.

[2] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[3] 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.

[4] Kai Eckert, Christian Meilicke, and Heiner Stuckenschmidt. Improving ontology matching using meta-level learning. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 158–172, Berlin, Heidelberg, 2009. Springer-Verlag.

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

[6] J. Euzenat. Results of the results of the ontology alignment evaluation initiative 2014. In *Ontology Matching*, 2014.

[7] JÃ©rÃŽme Euzenat, Christian Meilicke, Heiner Stuckenschmidt, Pavel Shvaiko, and CÃ!'ssia Trojahn. Ontology alignment evaluation initiative: Six years of experience. In Stefano Spaccapietra, editor, *Journal on Data Semantics XV*, volume 6720 of *Lecture Notes in Computer Science*, pages 158–192. Springer Berlin Heidelberg, 2011.

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

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

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

[11] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[12] 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.

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

[14] 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.

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

[16] Erhard Rahm. Towards large-scale schema and ontology matching. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors, *Schema Matching and Mapping*, Data-Centric Systems and Applications, pages 3–27. Springer Berlin Heidelberg, 2011.

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

[18] P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):158–176, Jan 2013.

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

[20] World Wide Web Consortium (W3C). Owl 2 web ontology language. structural specification and functional-style syntax, 2009.

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