Testing OWL Axioms against RDF Facts: A Possibilistic Approach

Andrea G. B. Tettamanzi, Catherine Faron Zucker, Fabien Gandon

Testing OWL Axioms Against RDF Facts: A possibilistic approach

Andrea G. B. Tettamanzi[1], Catherine Faron-Zucker[1], and Fabien Gandon[2]

[1] Univ. Nice Sophia Antipolis, I3S, UMR 7271, Sophia Antipolis, France,
andrea.tettamanzi@unice.fr, faron@unice.fr
[2] INRIA, Sophia Antipolis, France,
fabien.gandon@inria.fr

Abstract. Automatic knowledge base enrichment methods rely critically on candidate axiom scoring. The most popular scoring heuristics proposed in the literature are based on statistical inference. We argue that such a probability-based framework is not always completely satisfactory and propose a novel, alternative scoring heuristics expressed in terms of possibility theory, whereby a candidate axiom receives a bipolar score consisting of a degree of possibility and a degree of necessity. We evaluate our proposal by applying it to the problem of testing SubClassOf axioms against the DBpedia RDF dataset.

Keywords: ontology learning, open-world assumption, possibility theory

1 Introduction

A common approach to the semantic Web puts strong emphasis on a principled conceptual analysis of a domain of interest leading to the construction or reuse of ontologies as a prerequisite step for the organization of the Linked Open Data (LOD), much like a database schema must be designed before a database can be populated. However this approach has some limitations: it is aprioristic and dogmatic in the way knowledge should be organized; while it is quite successful when applied to specific domains, it does not scale well to more general settings; it does not lend itself to a collaborative effort; etc. That is why an alternative, bottom-up, *grass-roots* approach to ontology and knowledge base creation better suits many scenarii: instead of postulating an *a priori* conceptualization of reality (i.e., an ontology) and requiring that our knowledge about facts complies with it, one can start from RDF facts and learn OWL 2 axioms.

Recent contributions towards the automatic creation of OWL 2 ontologies from large repositories of RDF facts include FOIL-like algorithms for learning concept definitions [5], statistical schema induction via association rule mining [6], and light-weight schema enrichment methods based on the DL-Learner framework [9,2]. All these methods apply and extend techniques developed within inductive logic programming (ILP) [10].

On a related note, there exists a need for evaluating and validating ontologies, be they the result of an analysis effort or of a semi-automatic learning method. This need is witnessed by general methodological investigations [7,8] and surveys [13] and tools like OOPS! [11] for detecting pitfalls in ontologies.

Ontology learning and validation rely critically on (candidate) axiom scoring. In this paper, we will tackle the problem of testing a single, isolated axiom, which is anyway the first step to solve the problem of validating an entire ontology. Furthermore, to keep things reasonably simple, we will restrict our attention to subsumption axioms of the form $\mathsf{SubClassOf}(C\ D)$.

The most popular scoring heuristics proposed in the literature are based on statistical inference. We argue that such a probability-based framework is not always completely satisfactory. We will propose an axiom scoring heuristics based on a formalization in possibility theory of the notions of logical content of a theory and of falsification, loosely inspired by Karl Popper's approach to epistemology. Our research question is: "Can we apply a possibilistic approach to the task of testing candidate axioms for ontology learning?". In addition, "Could this be beneficial to ontology and knowledge base validation?".

The paper is organized as follows: Section 2 proposes a heuristics based on possibility theory, alternative to probability-based scoring heuristics. Its implementation is detailed in Section 3 and an evaluation is provided in Section 4. Section 5 draws some conclusions.

2 A Possibilistic Candidate Axiom Scoring Heuristics

Let ϕ be a candidate axiom and u_ϕ the support or sample size for ϕ, i.e., the cardinality of the set of its logical consequences that will be tested in the RDF repository. Notice that every formula which logically follows from an axiom is both a potential falsifier (if it is contradicted by facts) and a potential confirmation (if it is verified by facts) for that axiom.

Let u_ϕ^+ the number of such consequences which are true (confirmations), and u_ϕ^- the number of such consequences which are false (counterexamples). A few interesting properties of these three cardinalities are:

$$u_\phi^+ + u_\phi^- \leq u_\phi; \tag{1}$$

$$u_\phi^+ = u_{\neg\phi}^-, \quad u_\phi^- = u_{\neg\phi}^+, \quad u_\phi = u_{\neg\phi}. \tag{2}$$

A statistics-based heuristics for the scoring of candidate axioms used in the framework of knowledge base enrichment [2] may be regarded essentially as scoring an axiom by an estimate of the probability that one of its logical consequences is confirmed (or, alternatively, falsified) by the facts stored in the RDF repository. As Bühmann and Lehmann point out [2], estimating the probability of confirmation of axiom ϕ just by $\hat{p}_\phi = u_\phi^+/u_\phi$ would be too crude and would not take the magnitude of u_ϕ into account. That is why they base their probabilistic score on Agresti and Coull's binomial proportion confidence interval [1].

One problem with such approaches is that they only look for confirmations of ϕ and treat their absence as failures in the calculation of the confidence interval. This is like making an implicit closed-world assumption. An easy fix, in view of the open-world assumption, might be, for example, to use $\hat{p}^* = u_\phi^+/(u_\phi^+ + u_\phi^-)$ as the sample proportion instead of $\hat{p}$.

A more serious problem, however, is that these approaches rely on the assumption that it is possible to estimate the probability that an axiom ϕ is true given some evidence e, where e = "ψ such that $\phi \models \psi$ is in the RDF repository", or e = "ψ such that $\psi \models \neg\phi$ is in the RDF repository", or e = "ψ such that $\phi \models \psi$ is *not* in the RDF repository", etc., which, by Bayes' formula, may be written as

$$\Pr(\phi \mid e) = \frac{\Pr(e \mid \phi)\Pr(e)}{\Pr(e \mid \phi)\Pr(e) + \Pr(e \mid \neg\phi)\Pr(\neg e)}. \tag{3}$$

Therefore, in order to compute (or estimate) such probability, one should be able to estimate the probabilities on the right-hand side of Equation 3. Now, this is possible only under the (strong) assumption that the data at hand are representative.

To capture the basic intuition behind the process of axiom discovery without making unwarranted assumptions, we propose an alternative axiom scoring heuristics based on possibility theory, which is weaker than probability theory.

2.1 Possibility Theory

Possibility theory [14] is a mathematical theory of epistemic uncertainty. Given a finite universe of discourse Ω, whose elements $\omega \in \Omega$ may be regarded as events, values of a variable, possible worlds, or states of affairs, a possibility distribution is a mapping $\pi : \Omega \to [0, 1]$, which assigns to each ω a degree of possibility ranging from 0 (impossible, excluded) to 1 (completely possible, normal). A possibility distribution for which there exists a completely possible state of affairs $(\exists\omega^* : \pi(\omega^*) = 1)$ is said to be *normalized*.

There is a similarity between possibility distribution and probability density. However, it must be stressed that $\pi(\omega) = 1$ just means that ω is a plausible (normal) situation and therefore should not be excluded. A degree of possibility can then be viewed as an upper bound of a degree of probability. Possibility theory is suitable to represent incomplete knowledge while probability is adapted to represent random and observed phenomena. We invite the reader to see [4] for more informations about the relationships between fuzzy sets, possibility, and probability degrees.

A possibility distribution π induces a *possibility measure* and its dual *necessity measure*, denoted by Π and N respectively. Both measures apply to a set $A \subseteq \Omega$ (or to a formula ϕ, by way of the set of its models, $A = \{\omega : \omega \models \phi\}$), and are defined as follows:

$$\Pi(A) = \max_{\omega \in A} \pi(\omega); \tag{4}$$

$$N(A) = 1 - \Pi(\bar{A}) = \min_{\omega \in \bar{A}}\{1 - \pi(\omega)\}. \tag{5}$$

A few properties of possibility and necessity measures induced by a normalized possibility distribution on a finite universe of discourse Ω are the following. For all subsets $A \subseteq \Omega$,

1. $\Pi(\emptyset) = N(\emptyset) = 0, \quad \Pi(\Omega) = N(\Omega) = 1;$
2. $\Pi(A) = 1 - N(\bar{A})$ (duality);
3. $N(A) > 0$ implies $\Pi(A) = 1, \quad \Pi(A) < 1$ implies $N(A) = 0$.

In case of complete ignorance on A, $\Pi(A) = \Pi(\bar{A}) = 1$.

2.2 Support of an Axiom

At the beginning of this section, we have introduced the notion of support or sample for a candidate axiom ϕ as the number of its logical consequences that will be tested in the RDF repository. We shall now define that notion more precisely.

Let BS be a finite set of *basic statements*, i.e., assertions, like the ones contained in an RDF repository, that may be tested by means of a SPARQL `ASK` query. We define the *content* of an axiom ϕ that we wish to evaluate as the set of its logical consequences, but we restrict it to basic statements, to ensure finiteness and testability:

$$\text{content}(\phi) = \{\psi : \phi \models \psi\} \cap \text{BS}. \tag{6}$$

The cardinality of content(ϕ) is finite, because BS is finite, and every formula $\psi \in \text{content}(\phi)$ may be tested, because it is a basic statement. Now we can define the support of ϕ as the cardinality of content(ϕ):

$$u_\phi = |\text{content}(\phi)|. \tag{7}$$

2.3 Possibility and Necessity of an Axiom

The basic principle for establishing the possibility of a formula ϕ should be that the absence of counterexamples to ϕ in the RDF repository means $\Pi(\phi) = 1$, i.e., that ϕ is completely possible.

A hypothesis should be regarded as all the more *necessary* as it is explicitly supported by facts and not contradicted by any fact; and all the more *possible* as it is not contradicted by facts. In other words, given hypothesis ϕ, $\Pi(\phi) = 1$ if no counterexamples are found; as the number of counterexamples increases, $\Pi(\phi) \to 0$ strictly monotonically; $N(\phi) = 0$ if no confirmations are found; as the number of confirmations increases and no counterexamples are found, $N(\phi) \to 1$ strictly monotonically. Notice that a confirmation of ϕ is a counterexample of $\neg\phi$ and that a counterexample of ϕ is a confirmation of $\neg\phi$. Furthermore, the first counterexamples found to an axiom should determine a sharper decrease of the degree to which we regard the axiom as possible than any further counterexamples, because these latter will only confirm our suspicions and, therefore, will provide less and less information and, similarly, in the absence of counterexamples, the first confirmations found to an axiom should determine a sharper increase of the degree to which we regard the axiom as necessary than any further confirmations, because these latter will only add up to our acceptance and, therefore, will provide less and less information.

A definition of Π and N which captures the above intuitions, but by no means the only possible one, is, for $u_\phi > 0$,

$$\Pi(\phi) = 1 - \sqrt{1 - \left(\frac{u_\phi - u_\phi^-}{u_\phi}\right)^2}; \tag{8}$$

$$N(\phi) = \sqrt{1 - \left(\frac{u_\phi - u_\phi^+}{u_\phi}\right)^2} \quad \text{if } \Pi(\phi) = 1, \text{ 0 otherwise.} \tag{9}$$

Notice that this definition satisfies the duality of possibility and necessity, in that $N(\phi) = 1 - \Pi(\neg\phi)$ and $\Pi(\phi) = 1 - N(\neg\phi)$.

We combine the possibility and necessity of an axiom to define a single handy acceptance/rejection index (ARI) as follows:

$$\mathrm{ARI}(\phi) = N(\phi) - N(\neg\phi) = N(\phi) + \Pi(\phi) - 1 \in [-1, 1]. \tag{10}$$

A negative $\mathrm{ARI}(\phi)$ suggests rejection of ϕ ($\Pi(\phi) < 1$), whilst a positive $\mathrm{ARI}(\phi)$ suggests its acceptance ($N(\phi) > 0$), with a strength proportional to its absolute value. A value close to zero reflects ignorance about the status of ϕ.

3 A Framework for Candidate Axiom Testing

We refer to the model-theoretic semantics of OWL 2 (as defined in [3]), which defines an interpretation $\mathcal{I}$ with a valuation function $\cdot^\mathcal{I}$ mapping OWL 2 expressions into elements and sets of elements of an interpretation domain $\Delta^\mathcal{I}$. We take the set of all the resources that occur in a given RDF store as $\Delta^\mathcal{I}$ and checking an axiom amounts to checking whether $\mathcal{I}$ is a model of the axiom. Also, calling linked data search engines like Sindice could virtually extend the interpretation domain to the whole LOD cloud.

However, unlike interpretation domains, RDF stores are incomplete and possibly noisy. The open-world hypothesis must be made; therefore, absence of supporting evidence does not necessarily contradict an axiom, and an axiom might hold even in the face of a few counterexamples (exceptions or possible mistakes). For example, out of 541 axioms of the form SubClassOf(C D) in the DBpedia ontology, 143 have an empty content (i.e., class C is empty) and 28 have at least one counterexample in DBpedia 3.9.[3]

A general algorithm for testing all the possible OWL 2 axioms in a given RDF store is beyond the scope of this paper. Here, we will restrict our attention

[3] And one, namely SubClassOf(dbo:Person dbo:Agent), even has 76 counterexamples!

to atomic class expressions and ObjectComplementOf expressions, needed to test SubClassOf axioms. The model-theoretic semantics of expressions of the form ObjectComplementOf(C) ($\neg C$ in description logics syntax), where C denotes a concept expression (called *class expression* in OWL 2) is $\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$.

Now, let us define a mapping $Q(E, x)$ from OWL 2 expressions to SPARQL graph patterns, where E is an OWL 2 expression, x is a formal parameter which can be replaced by a SPARQL variable or an RDF term, such that the query `SELECT DISTINCT ?x WHERE` $\{\ Q(E, \texttt{?x})\ \}$ returns all the known instances of class expression E, which we will denote by $[Q(E, x)]$, i.e., the equivalent of $E^{\mathcal{I}}$, and the query `ASK` $\{\ Q(E, a)\ \}$ checks whether $E(a)$ is in the RDF base.

For an atomic concept A, $Q(A, \texttt{?x}) = \texttt{?x a}\ A$, where A is a valid IRI. For concept negation, things are slightly more complicated, for RDF does not support negation. The obvious definition

$$Q(\neg C, \texttt{?x}) = \{\ \texttt{?x ?p ?o . FILTER NOT EXISTS}\ Q(C, \texttt{?x})\ \}, \tag{11}$$

has the problem of treating negation as failure, like in databases, where the closed-world assumption is made. Since we want to get as close as possible to an open-world semantics, $Q(\neg C, x)$ should be defined differently, as the union of the concepts that are disjoint from C. One might try to express this as the set of individuals x that are instances of a concept C' such that no individual $z \in C^{\mathcal{I}}$ is an instance of C', yielding the query

$$\begin{aligned} Q(\neg C, \texttt{?x}) = \{\ & \texttt{?x a ?dc .} \\ & \texttt{FILTER NOT EXISTS}\ \{\ \texttt{?z a ?dc .}\ \ Q(C, \texttt{?z})\ \}\ \}, \end{aligned} \tag{12}$$

where `?z` is a variable that does not occur anywhere else in the query. This translation is conceptually more satisfactory than the one in Equation 11, but it just pushes the problem one step further, because this way of testing whether two concepts are disjoint is based on negation as failure too. The only way to be certain that two classes are disjoint would be to find an axiom to this effect in the ontology:

$$Q(\neg C, \texttt{?x}) = \{\ \texttt{?x a ?dc . ?dc owl:disjointWith}\ C\ \}, \tag{13}$$

otherwise, either we find an individual which is an instance of both classes, and thus we know the two classes are not disjoint, or we don't, in which case the two classes may or may not be disjoint. The fact is, very few DisjointClasses axioms are currently found in existing ontologies. For example, in the DBpedia ontology, the query `SELECT ?x ?y { ?x owl:disjointWith ?y }` executed on November 22, 2013 returned 17 solutions only.

To compare these three alternative definitions of $Q(\neg C, \texttt{?x})$, we may refer to the diagram in Figure 1. We wish to estimate the actual extent of $(\neg C)^{\mathcal{I}}$. Clearly, $Q(C, \texttt{?x})$ (in dark grey) underestimates the real extent of $C^{\mathcal{I}}$ (in light gray). Therefore, we may say that Equation 11 overestimates the real extent of $(\neg C)^{\mathcal{I}}$, in the sense that it will regard as instances of $\neg C$ all individuals a for which "a `a` C" is not found in the RDF repository.

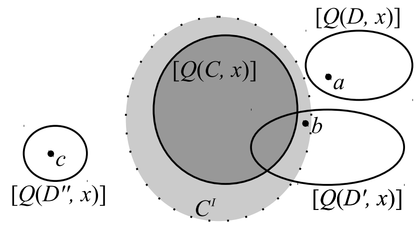


Fig. 1. A schematic illustration of the heuristics used to capture negation under the open world assumption. D'' is a concept which is declared to be disjoint with C in the RDF repository.

Now, if b is such that "b `a` C" is not known, but "b `a` D'" is known for some class D' and some instances of D are known to be also instances of C, then it might well be that b is an instance of C as well. If, however, a is such that "a `a` C" is not known and no instance of D is known that is also an instance of C, then we are more likely to believe that a is not an instance of C. Therefore Equation 12 regards as instances of $\neg C$ fewer individuals, those for which it is highly likely that they do not belong in C.

On the other hand, Equation 13 certainly underestimates $(\neg C)^{\mathcal{I}}$, to the point of considering it empty if "D'' `owl:disjointWith` C" is not declared in the RDF repository. Furthermore, an individual may be an instance of $\neg C$ even though it is not an instance of a class disjoint with C!

To sum up, Equation 11 is too optimistic, Equation 13 too pessimistic, and Equation 12 somewhere in the middle. Following the old adage "virtue stands in the middle", adopting Equation 12 looks like a sensible choice.

We will end this section by arguing that a suitable definition of confirmation to adopt in this framework is Scheffler and Goodman's *selective confirmation* [12], which characterizes a confirmation as a fact not simply satisfying an axiom, but, further, favoring the axiom rather than its contrary. For instance, the occurence of a black raven *selectively confirms* the axiom $\mathsf{Raven} \sqsubseteq \mathsf{Black}$ because it both confirms it and fails to confirm its negation, namely that there exist ravens that are not black. On the contrary, the observation of a green apple does not contradict $\mathsf{Raven} \sqsubseteq \mathsf{Black}$, but it does not disconfirm $\mathsf{Raven} \not\sqsubseteq \mathsf{Black}$ either; therefore, it does not selectively confirm $\mathsf{Raven} \sqsubseteq \mathsf{Black}$.

4 Evaluation on Subsumption Axiom Testing

The semantics of subsumption axioms of the form $\mathsf{SubClassOf}(C\ D)$ ($C \sqsubseteq D$ in description logic syntax) is $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$, which may also be written $x \in C^{\mathcal{I}} \Rightarrow x \in D^{\mathcal{I}}$. Therefore, $\text{content}(C \sqsubseteq D) = \{D(a) : C(a) \text{ in the RDF store}\}$, because, if $C(a)$ holds, $C(a) \Rightarrow D(a) = \neg C(a) \vee D(a) = \top \vee D(a) = D(a)$. The support

$u_{C \sqsubseteq D}$ of the axiom can thus be computed with the following SPARQL query:

$$\texttt{SELECT (count(DISTINCT ?x) AS ?u) WHERE } \{Q(C, \texttt{?x})\}. \tag{14}$$

In order to compute $ARI(C \sqsubseteq D)$, we must provide a computational definition of $u^+_{C \sqsubseteq D}$ and $u^-_{C \sqsubseteq D}$. We start with the following statements:

– confirmations are individuals i such that $i \in [Q(C, x)]$ and $i \in [Q(D, x)]$;
– counterexamples are individuals i such that $i \in [Q(C, x)]$ and $i \in [Q(\neg D, x)]$.

This may be translated into SPARQL queries to compute $u^+_{C \sqsubseteq D}$ and $u^-_{C \sqsubseteq D}$:

$$\begin{array}{l}\texttt{SELECT (count(DISTINCT ?x) AS ?numConfirmations)} \\ \texttt{WHERE } \{\ Q(C, \texttt{?x})\ Q(D, \texttt{?x})\ \}\end{array} \tag{15}$$

and

$$\begin{array}{l}\texttt{SELECT (count(DISTINCT ?x) AS ?numCounterexamples)} \\ \texttt{WHERE } \{\ Q(C, \texttt{?x})\ Q(\neg D, \texttt{?x})\ \}\end{array} \tag{16}$$

respectively. Notice that an $i \in [Q(C, x)]$ such that $i \notin [Q(D, x)]$ does not contradict $C \sqsubseteq D$, because it might well be the case that the assertion "i `a` D" is just missing. Likewise, an $i \in [Q(\neg D, x)]$ such that $i \in [Q(\neg C, x)]$ will not be treated as a confirmation, based on our choice to regard as evidence in favor of a hypothesis only selective confirmations.

We evaluated the proposed scoring heuristics by performing tests of subsumption axioms using DBpedia 3.9 in English as the reference RDF fact repository. In particular, on April 27, 2014, we downloaded the DBpedia dumps of English version 3.9, generated in late March/early April 2013, along with the DBpedia ontology, version 3.9. This local dump of DBpedia, consisting of 812,546,748 RDF triples, has been bulk-loaded into Jena TDB and a prototype for performing axiom tests using the proposed method has been coded in Java, using Jena ARQ and TDB to access the RDF repository.

All experiments have been performed on a Fujitsu CELSIUS workstation equipped with twelve six-core Intel Xeon CPU E5-2630 v2 processors at 2.60GHz clock speed, with 15,360 KB cache each, 128 GB RAM, 4 TB of disk space with a 128 GB SSD cache, under the Ubuntu 12.04.4 LTS 64-bit operating system.

We performed two experiments of different type: an explorative test of systematically generated subsumption axioms and an exhaustive test of all subsumption axioms in the DBpedia ontology.[4]

For the former experiment, we systematically generated and tested subsumption axioms involving atomic classes only according the following protocol: for each of the 442 classes C referred to in the RDF repository, we construct all axioms of the form $C \sqsubseteq D$ such that C and D share at least one instance. Classes D are obtained with query `SELECT DISTINCT ?D WHERE` $\{Q(C, \texttt{?x})$`. ?x a ?D`$\}$. Due to the sheer number of axioms thus generated, and to the long time it takes to test them,[5] this experiment could not complete at the time of writing and

[4] Results available at URL `http://www.i3s.unice.fr/~tettaman/RDFMining/`.
[5] Up to 27 hours for axiom SubClassOf(dbo:Eukaryote dbo:Artist)!

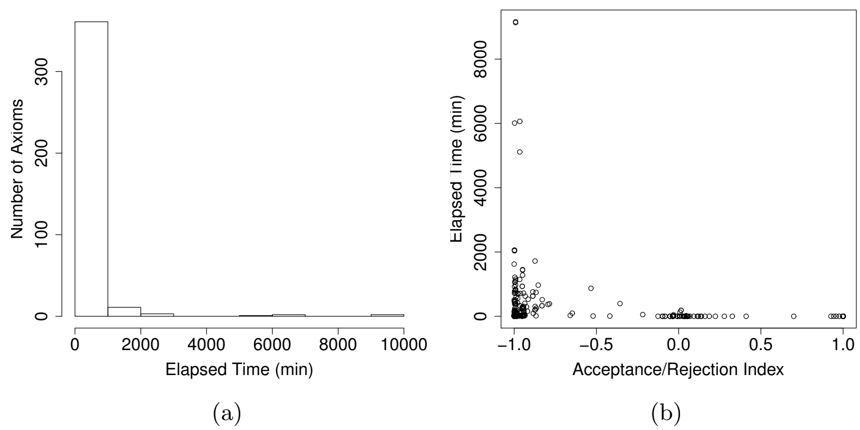


Fig. 2. A histogram showing the distribution of test time of systematically generated SubClassOf axioms (a), and a plot of the time taken for testing as a function of ARI.

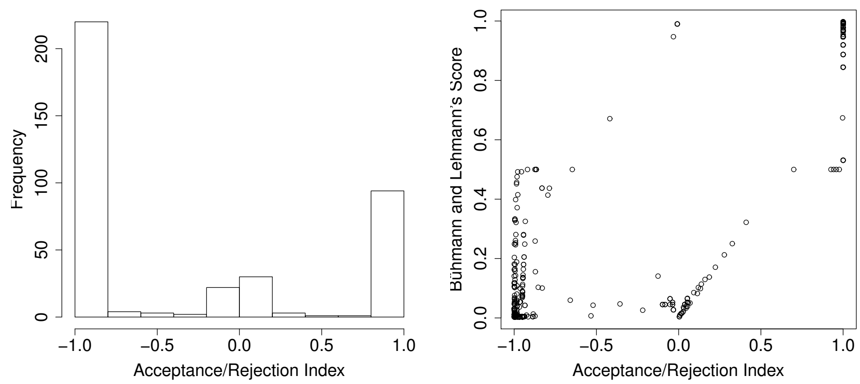
is still underway; however, a sufficient number of axioms was tested to allow gathering some statistics. Figure 2a shows the distribution of test time, while the plot of time vs. ARI of axioms in Figure 2b suggests that the time it takes to test an axiom is inversely proportional to its ARI. This is good news, because it suggests that putting a time-out on the test would be an acceptable heuristics to decide whether to accept or reject a candidate axiom, for an axiom which takes too long to test will likely end up having a very negative ARI.

By construction, all axioms generated in this experiment have at least one confirmation and, as a consequence, non-zero possibility (thus ARI > -1).

The ARI values of systematically generated axioms tend to cluster around the three values -1, 0, and 1 (see Figure 3a).

To assess the discriminatory ability of the proposed scoring heuristics, we have evaluated these results by sorting the 380 tested axioms by their ARI and by manually tagging each of them as either *true* or *false* based on common sense. Three out of the 78 axioms with an ARI of 1 are clearly false: SubClassOf(dbo:Eukaryote dbo:Species), SubClassOf(dbo:OrganisationMember dbo:SportsTeamMember), and SubClassOf(dbo:SportCompetitionResult dbo:OlympicResult). However, no scoring heuristics based on counting would be able to tell these 3 false positives from the true ones. Proceeding by decreasing ARI, the first false axiom encountered was SubClassOf(dbo:TennisLeague skos:Concept), with an ARI of 0.699854,[6] followed by the true axiom SubClassOf(dbo:Road gml:_Feature), with an ARI of 0.410399, followed by a large number of false axioms, starting

[6] The confirmations of this and several other axioms of the form SubclassOf(C skos:Concept) with a much lower ARI are obvious mistakes: no individual should be a skos:Concept. This seems now to have been fixed in the live version of DBpedia.



(a) (b)

Fig. 3. A histogram showing the distribution of the acceptance/rejection index of systematically generated SubClassOf axioms (a), and the relationship between the acceptance/rejection index and the probability-based score used in [2] (b).

with SubClassOf(schema:Product gml:_Feature) with an ARI of 0.326187. This positive ARI means no counterexamples were found; the 2,806 confirmations are all instances of classes dbo:Aircraft and dbo:Ship having, strangely enough, geographical coordinates.

A few seemingly true axioms are found with an ARI around zero. They are SubClassOf(schema:School schema:Place), SubClassOf(schema:School dbo:Place), with an ARI of 0.00830433, and SubClassOf(schema:School dbo:EducationalInstitution), SubClassOf(schema:School dbo:Organisation), SubClassOf(schema:School schema:EducationalOrganization), and SubClassOf(schema:School schema:Organization), with an ARI of -0.00830433.

A dubious case is SubClassOf(schema:Museum dbo:Building), with an ARI of -0.031754, 3,957 confirmations and two counterexamples :Saint_Peter's_Basilica and :US_90. We tagged it as false, because one could imagine a museum hosted on a ship or an open-air museum, but we admit to our choice being debatable.

This distribution of true and false axioms suggests $\mathrm{ARI}(\phi) > 1/3$ as the optimal acceptance criterion for a candidate axiom ϕ. This would yield 4 false positives and 6 false negatives (corresponding to a 97.37% accuracy). However, it appears that the misclassification of the above axioms is to blame on mistakes in DBpedia. This highlights the potential for the proposed heuristics as a tool for RDF data validation: confirmations and counterexamples of axioms with ARI around zero is where the search for bugs should focus.

It is interesting to compare these results with those one would obtain by using a probability-based score. Figure 3b provides a comparison by plotting each axiom according to its ARI (X-axis) and its score computed as in [2] (Yaxis). First of all it is clear that both scores tend to agree in the extremes, with some notable exceptions, but behave quite differently in all other cases. The probabilistic score with the 0.7 threshold suggested by [2] gives 13 false negatives (7 more than the ARI) and 4 false positives (the three false axioms with ARI of 1, plus SubClassOf(schema:Museum dbo:Building), which we tagged as false). Among the false negatives, there are five axioms with an ARI of 1 which are rejected by the probabilistic score: they are all of the form SubClassOf(dbo:TennisLeague D). Then we find SubClassOf(schema:School schema:Place) and SubClassOf(schema:School dbo:Place), which were also rejected by the ARI, and six axioms of the form SubClassOf(C gml:_Feature), which have ARIs comprised between 0.41 and 0.997. Finally, most false axiom candidates get an ARI close to -1, whilst their probabilistic scores are almost evenly distributed between 0 and 0.5. We might say that, besides being more accurate, ARI gives clearer indications than the probabilistic score.

For the second, more validation-oriented experiment, we extracted an exhaustive list of SubClassOf axioms in the DBpedia ontology, in functional syntax, with the query

```
SELECT DISTINCT concat("SubClassOf(<",str(?x),"> <",str(?y),">)")
WHERE { ?x a owl:Class . ?x rdfs:subClassOf ?y }
```

thus obtaining 541 axioms. Testing them took "only" 1 h 23 min 31 s, due to the fact that most of these axioms have a positive ARI and can thus be tested relatively rapidly.

A large number of these axioms (143) turned out to have $u_\phi = 0$ (empty content) thus their ARI is 0. For 28 axioms, a negative ARI signals the presence of seemingly erroneous facts: the following table gives a few examples of axioms $C \sqsubseteq D$ with their conterexamples (instances of the subclass that also belong to a class disjoint with the superclass, i.e., a such that $C(a)$ and $E(a)$ with $E^{\mathcal{I}} \cap D^{\mathcal{I}} = \emptyset$: in this case, either $C(a)$ is wrong or $E(a)$ is).

Axiom	**Counterexamples**
SubClassOf(dbo:LaunchPad dbo:Infrastructure)	:USA
SubClassOf(dbo:Brain dbo:AnatomicalStructure)	:Brain [*sic*]
SubClassOf(dbo:Train dbo:MeanOfTransportation)	:New_Jersey_Transit_rail_operations, :ALWEG
SubClassOf(dbo:ProgrammingLanguage dbo:Software)	:Ajax
SubClassOf(dbo:PoliticalParty dbo:Organisation)	:Guelphs_and_Ghibellines, :-,[7] :New_People's_Army, :Syrian

5 Conclusion

We have proposed a candidate axiom scoring heuristics, based on possibility theory, to be used for automatic axiom induction from RDF data and, ultimately, to provide a solid basis for ontology learning. In addition, we have developed a framework based on the proposed heuristics, which uses the model-theoretic semantics of OWL 2 and SPARQL queries to test candidate axioms.

[7] That is `<http://dbpedia.org/resource/->`. This IRI is dereferenced to the "hyphen-minus" resource.

The results of experimental evaluation on the DBpedia dataset clearly indicate that the proposed heuristics is suitable for tasks such axiom induction and ontology learning and, furthermore, may be beneficial as a tool for ontology and knowledge-base validation.

One may object that, being based on possibility theory, our scoring heuristics is less objective than a probability-based scoring method. However, we have argued in Section 2 that scoring heuristics based on probability are doomed to be arbitrary and subjective or, in other words, *qualitative* and, therefore, hardly more rigorous or objective than the proposed approach. The experimental results corroborate this claim.

Future work includes extending the experimental evaluation to more general sets of candidate axioms and enlarging the test base by including additional RDF datasets from the LOD. We also plan on improving the implementation of the framework by setting a time-out on query evaluation to reduce the computational overhead of axiom testing.

References

1. Agresti, A., Coull, B.A.: Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52(2) (May 1998)
2. Bühmann, L., Lehmann, J.: Universal OWL axiom enrichment for large knowledge bases. In: EKAW 2012. pp. 57–71
3. Cuenca Grau, B., Motik, B., Patel-Schneider, P.: OWL 2 web ontology language direct semantics (second edition). W3C recommendation, W3C (December 2012)
4. Dubois, D., Prade, H.: Fuzzy sets and probability: Misunderstandings, bridges and gaps. Fuzzy Sets and Systems 40(1), 143–202 (1991)
5. Fanizzi, N., d'Amato, C., Esposito, F.: DL-FOIL concept learning in description logics. In: ILP 2008. pp. 107–121
6. Fleischhacker, D., Völker, J., Stuckenschmidt, H.: Mining RDF data for property axioms. In: OTM 2012. pp. 718–735
7. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: A theoretical framework for ontology evaluation and validation. In: SWAP 2005
8. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: ESWC 2006. pp. 140–154
9. Hellmann, S., Lehmann, J., Auer, S.: Learning of OWL class descriptions on very large knowledge bases. Int. J. Semantic Web Inf. Syst. 5(2), 25–48 (2009)
10. Muggleton, S., De Raedt, L., Poole, D., Bratko, I., Flach, P., Inoue, K., Srinivasan, A.: ILP turns 20: Biography and future challenges. Machine Learning 86, 3–23 (2012)
11. Poveda-Villalón, M., del Carmen Suárez-Figueroa, M., Gómez-Pérez, A.: Validating ontologies with OOPS! In: EKAW 2012. pp. 267–281
12. Scheffler, I., Goodman, N.: Selective confirmation and the ravens: A reply to Foster. The Journal of Philosophy 69(3), 78–83 (Feb 10 1972)
13. Tartir, S., Budak Arpinar, I., Sheth, A.P.: Ontological evaluation and validation. In: Theory and Applications of Ontologies: Computer Applications. Springer (2010)
14. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–28 (1978)