Query Answering and Ontology Population: an Inductive Approach

Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito

Department of Computer Science, University of Bari
{claudia.damato|fanizzi|esposito}@di.uniba.it

Abstract. In order to overcome the limitations of deductive logic-based approaches to deriving operational knowledge from ontologies, especially when data come from distributed sources, inductive (instance-based) methods may be better suited, since they are usually efficient and noise-tolerant. In this paper we propose an inductive method for improving the instance retrieval and enriching the ontology population. By casting retrieval as a classification problem with the goal of assessing the individual class-memberships w.r.t. the query concepts, we propose an extension of the *k-Nearest Neighbor* algorithm for OWL ontologies based on an *entropic* distance measure. The procedure can classify the individuals w.r.t. the known concepts but it can also be used to retrieve individuals belonging to query concepts. Experimentally we show that the behavior of the classifier is comparable with the one of a standard reasoner. Moreover we show that new knowledge (not logically derivable) is induced. It can be suggested to the knowledge engineer for validation, during the ontology population task.

1 Introduction

Classification and query answering for retrieving resources in a knowledge base (KB) are important tasks. Generally these activities are performed by means of logical approaches that may fail when data comes from distributed sources, and is therefore exposed to inconsistency problems. This has given rise to alternative methods, such as *non-monotonic*, *paraconsistent* [8], *approximate* reasoning (see the discussion in [9]).

Inductive methods are known to be be quite efficient and more noise-tolerant, hence they seem suitable for contexts where knowledge is intended to be acquired from distributed sources. In this paper we propose an inductive *instance-based* method for *concept retrieval* [1] and *query answering* that may suggest new assertions which could not be logically derived, providing also a measure of their likelihood which may help dealing with the uncertainty caused by the inherent incompleteness of the KBs in the Semantic Web.

Namely, instance retrieval and query answering can be cast as classification problems, i.e. assessing the class-membership of the individuals in the KB w.r.t. some query concepts. Reasoning by analogy, similar individuals should likely belong to the extension of similar concepts. Moving from such an intuition,

an instance-based framework for retrieving resources contained in ontological KBs has been devised, to inductively infer (likely) consistent class-membership assertions that may be not logically derivable. As such, the resulting assertions may enrich the KBs since the method can also provide a likelihood measure for its outcomes. Then the time-consuming ontology population task can be facilitated since the knowledge engineer only has to validate such new knowledge, as also argued in [2].

Logic-based approaches to (approximate) instance retrieval have been proposed in the literature [13, 11]. We intend to apply inductive forms of reasoning borrowed from *machine learning*. Specifically, we propose an extension of the well-known *Nearest Neighbor* search (henceforth, *NN*) [14] to the standard representations of the SW (RDF through OWL). Our analogical approach is based on a dissimilarity measures for resources in these search space. The procedure retrieves individuals belonging to query concepts, by analogy with other training instances, namely on the grounds of the classification of the nearest ones (w.r.t. the dissimilarity measure). This approach may be quite efficient because it requires checking class-membership for a limited set of training instances yielding a decision on the classification of new instances.

From a technical viewpoint, extending the NN setting to the target representations founded in Description Logics (DL) [1], required suitable metrics whose definition could not be straightforward. In particular, a theoretical problem is posed by the *Open World Assumption* (OWA) that is generally made on the semantics of SW ontologies, differently from the typical standards of databases where the *Closed World Assumption* (CWA) is made. Moreover, the NN algorithms are devised for simple classifications where classes are assumed to be pairwise disjoint, which is quite unlikely in the Semantic Web context where an individual can be instance of more than one concept. Furthermore, dissimilarity measures that can cope with the semantics of expressive representations are necessary.

Most of the existing measures focus on concept (dis)similarity and particularly on the (dis)similarity of atomic concepts within hierarchies or simple ontologies (see the discussion in [3]). Conversely, for our purposes, a notion of dissimilarity between *individuals* is required. Recently, dissimilarity measures for specific description logics concept descriptions have been proposed [3, 4]. Although they turned out to be quite effective for the inductive tasks of interest, they are still partly based on structural criteria (a notion of normal form) which determine their main weakness: they are hardly scalable to deal with standard ontology languages.

In order to overcome these limitations, an extension of a semantic pseudo-metrics [7] is exploited. This language-independent measure assesses the dissimilarity of two individuals by comparing them on the grounds of their behavior w.r.t. a committee of features (concepts), namely those defined in the KB or that can be generated to this purpose[1]. In the former measures, all the features have

[1] The choice of optimal committees may be performed in advance through randomized search algorithms [7].

the same importance in determining the dissimilarity. However, it may well be that some features have a larger discriminating power w.r.t. the others. In this case, they should be more relevant in determining the dissimilarity value. Moving from this observation, we propose an extension of the measures presented in [7], where each feature of the committee is weighted on the grounds of the *quantity of information* that it conveys. This weight is then determined as an *entropy* measure, also used in attribute selection when building decision trees. The rationale is that the more general a feature (or its negation) is (i.e. low entropy) the less likely it may be usable for distinguishing the two individuals and vice versa.

The measure has been integrated in the NN procedure [4] and the classification of resources (individuals) w.r.t. a query concept has been performed through a voting procedure weighted by the neighbors' similarity. The resulting system allowed for an experimentation of the method on performing instance retrieval and query answering with a number ontologies drawn from public repositories. Its predictions were compared to assertions that were logically derived by a deductive reasoner. The experiments showed that the classification results are comparable (although slightly less complete) and also that the classifier is able to induce new knowledge that is not logically derivable. The experimentation also compared the outcomes obtained by the former measure, extended in this paper. Such a comparison showed that the measure presented in this paper may improve the classification results.

The paper is organized as follows. The basics of the instance-based approach applied to the standard representations are recalled in Sect. 2. The next Sect. 3 presents the semantic dissimilarity measures adopted in the retrieval procedure. Sect. 4 reports the outcomes of the experiments performed with the implementation of the procedure. Possible developments are finally examined in Sect. 5.

2 Resource Retrieval as Nearest Neighbor Search

2.1 Representation and Inference

In the following sections, we assume that concept descriptions are defined in terms of a generic sub-language based on OWL-DL that may be mapped to *Description Logics* with the standard model-theoretic semantics (see the handbook [1] for a thorough reference).

A *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$. $\mathcal{T}$ is a set of axioms that define concepts. $\mathcal{A}$ contains factual assertions concerning the resources, also known as individuals. Moreover, the *unique names assumption* may be made on the ABox individuals, that are represented by their URIs. The set of the individuals occurring in $\mathcal{A}$ will be denoted with $\mathsf{Ind}(\mathcal{A})$.

As regards the inference services, like all other instance-based methods, our procedure may require performing *instance-checking* [1], which roughly amounts to determining whether an individual, say a, belongs to a concept extension, i.e.

whether $C(a)$ holds for a certain concept C. Note that because of the OWA, a reasoner may be unable to give a positive or negative answer to a class-membership query. This service is provided proof-theoretically by a reasoner.

2.2 The Method

Query answering boils down to determining whether a resource belongs to a (query) concept extension. Here, an alternative inductive method is proposed for retrieving the resources that likely belong to a query concept. Such a method may also be able to provide an answer even when it may not be inferred by deduction, Moreover, it may also provide a measure of the likelihood of its answer.

In *similarity search* [14] the basic idea is to find the most similar object(s) to a query one (i.e. the one that is to be classified) with respect to a similarity (or dissimilarity) measure. We review the basics of the k-NN method applied to the Semantic Web context [4] context.

The objective is to induce an approximation for a discrete-valued target hypothesis function $h : IS \mapsto V$ from a space of instances IS to a set of values $V = \{v_1, \ldots, v_s\}$ standing for the classes (concepts) that have to be predicted. Note that normally $|IS| \ll |\mathsf{Ind}(\mathcal{A})|$ i.e. only a limited number of training instances is needed especially if they are prototypical for a region of the search space. Let x_q be the query instance whose class-membership is to be determined. Using a dissimilarity measure, the set of the k nearest (pre-classified) training instances w.r.t. x_q is selected: $NN(x_q) = \{x_i \mid i = 1, \ldots, k\}$.

In its simplest setting, the k-NN algorithm approximates h for classifying x_q on the grounds of the value that h is known to assume for the training instances in $NN(x_q)$, i.e. the k closest instances to x_q in terms of a dissimilarity measure. Precisely, the value is decided by means of a weighted majority voting procedure: it is simply the most *voted* value by the instances in $NN(x_q)$ weighted by the similarity of the neighbor individual.

The estimate of the hypothesis function for the query individual is:

$$\hat{h}(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} w_i \delta(v, h(x_i)) \tag{1}$$

where δ returns 1 in case of matching arguments and 0 otherwise, and, given a dissimilarity measure d, the weights are determined by $w_i = 1/d(x_i, x_q)$.

Note that the estimate function $\hat{h}$ is defined extensionally: the basic k-NN method does not return an intensional classification model (a function or a concept definition), it merely gives an answer for the instances to be classified.

It should be also observed that this setting assigns a value to the query instance which stands for one in a set of pairwise disjoint concepts (corresponding to the value set V). In a multi-relational setting this assumption cannot be made in general. An individual may be an instance of more than one concept.

The problem is also related to the CWA usually made in the knowledge discovery context. To deal with the OWA, the absence of information on whether

a training instance x belongs to the extension of the query concept Q should not be interpreted negatively, as in the standard settings which adopt the CWA. Rather, it should count as neutral (uncertain) information. Thus, assuming the alternate viewpoint, the multi-class problem is transformed into a ternary one. Hence another value set has to be adopted, namely $V = \{+1, -1, 0\}$, where the three values denote, respectively, membership, non-membership, and uncertainty, respectively.

The task can be cast as follows: given a query concept Q, determine the membership of an instance x_q through the NN procedure (see Eq. 1) where $V = \{-1, 0, +1\}$ and the hypothesis function values for the training instances are determined as follows:

$$h_Q(x) = \begin{cases} +1 & \mathcal{K} \models Q(x) \\ -1 & \mathcal{K} \models \neg Q(x) \\ 0 & \textit{otherwise} \end{cases}$$

i.e. the value of h_Q for the training instances is determined by the entailment[2] the corresponding assertion from the knowledge base.

Note that, being based on a majority vote of the individuals in the neighborhood, this procedure is less error-prone in case of noise in the data (e.g. incorrect assertions) w.r.t. a purely logic deductive procedure, therefore it may be able to give a correct classification even in case of (partially) inconsistent knowledge bases.

It should be noted that the inductive inference made by the procedure shown above is not guaranteed to be deductively valid. Indeed, inductive inference naturally yields a certain degree of uncertainty. In order to measure the likelihood of the decision made by the procedure (individual x_q belongs to the query concept denoted by value v maximizing the argmax argument in Eq. 1), given the nearest training individuals in $NN(x_q, k) = \{x_1, \ldots, x_k\}$, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$l(\textit{class}(x_q) = v | NN(x_q, k)) = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_Q(x_i))}{\sum_{v' \in V} \sum_{i=1}^{k} w_i \cdot \delta(v', h_Q(x_i))} \tag{2}$$

Hence the likelihood of the assertion $Q(x_q)$ corresponds to the case when $v = +1$.

3 A Semantic Pseudo-Metric for Individuals

As mentioned in the first section, various attempts to define semantic similarity (or dissimilarity) measures for concept languages have been made, yet they have still a limited applicability to simple languages [3] or they are not completely semantic depending also on the structure of the descriptions [4]. Moreover, for

[2] We use $\models$ to denote entailment, as computed through a reasoner.

our purposes, we need a function for measuring the similarity of individuals rather than concepts. It can be observed that individuals do not have a syntactic structure that can be compared. This has led to lifting them to the concept description level before comparing them (recurring to the notion of the *most specific concept* of an individual w.r.t. the ABox [1], yet this makes the measure language-dependent. Besides, it would add a further approximations as the most specific concepts can be defined only for simple DLs.

For the NN procedure, we intend to exploit a new measure that totally depends on semantic aspects of the individuals in the knowledge base.

3.1 The Family of Measures

The new dissimilarity measures are based on the idea of comparing the semantics of the input individuals along a number of dimensions represented by a committee of concept descriptions. Indeed, on a semantic level, similar individuals should behave similarly with respect to the same concepts. Following the ideas borrowed from [12], totally semantic distance measures for individuals can be defined in the context of a knowledge base.

More formally, the rationale is to compare individuals on the grounds of their semantics w.r.t. a collection of concept descriptions, say $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, which stands as a group of discriminating *features* expressed in the OWL-DL sub-language taken into account.

In its simple formulation, a family of distance functions for individuals inspired to Minkowski's norms L_p can be defined as follows [7]:

Definition 3.1 (family of measures). *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, a family of dissimilarity functions $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$ is defined as follows:*

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \quad d_p^{\mathsf{F}}(a, b) := \frac{1}{|\mathsf{F}|} \left[\sum_{i=1}^{|\mathsf{F}|} w_i \mid \delta_i(a, b) \mid^p \right]^{1/p}$$

where $p > 0$ and $\forall i \in \{1, \ldots, m\}$ the dissimilarity function δ_i is defined by:

$$\forall (a, b) \in (\mathsf{Ind}(\mathcal{A}))^2 \quad \delta_i(a, b) = \begin{cases} 0 & F_i(a) \in \mathcal{A} \wedge F_i(b) \in \mathcal{A} \\ 1 & F_i(a) \in \mathcal{A} \wedge \neg F_i(b) \in \mathcal{A} \text{ or} \\ & \neg F_i(a) \in \mathcal{A} \wedge F_i(b) \in \mathcal{A} \\ 1/2 & \text{otherwise} \end{cases}$$

or, model theoretically:

$$\forall (a, b) \in (\mathsf{Ind}(\mathcal{A}))^2 \quad \delta_i(a, b) = \begin{cases} 0 & \mathcal{K} \models F_i(a) \wedge \mathcal{K} \models F_i(b) \\ 1 & \mathcal{K} \models F_i(a) \wedge \mathcal{K} \models \neg F_i(b) \text{ or} \\ & \mathcal{K} \models \neg F_i(a) \wedge \mathcal{K} \models F_i(b) \\ 1/2 & \text{otherwise} \end{cases}$$

Note that the original measures [7] correspond to the case of uniform weights.

The alternative definition for the projections, requires the entailment of an assertion (instance-checking) rather than the simple ABox look-up; this can make the measure more accurate yet more complex to compute unless a KBMS is employed maintaining such information at least for the concepts in F.

In particular, we will consider the measures $d_1^{\mathsf{F}}(\cdot,\cdot)$ or $d_2^{\mathsf{F}}(\cdot,\cdot)$ in the experiments.

As regards the weights employed in the family of measures, they should reflect the impact of the single feature concept w.r.t. the overall dissimilarity. As mentioned, this can be determined by the quantity of information conveyed by a feature, which can be measured as its entropy. Namely, the extension of a feature F w.r.t. the whole domain of objects may be probabilistically quantified as $P_F = |F^{\mathcal{I}}|/|\Delta^{\mathcal{I}}|$ (w.r.t. the canonical interpretation $\mathcal{I}$). This can be roughly approximated with: $P_F = |\mathsf{retrieval}(F)|/|\mathsf{Ind}(\mathcal{A})|$. Hence, considering also the probability $P_{\neg F}$ related to its negation and that related to the unclassified individuals (w.r.t. F), denoted P_U, we may give an entropic measure for the feature:

$$H(F) = -\left(P_F \log(P_F) + P_{\neg F} \log(P_{\neg F}) + P_U \log(P_U)\right)$$

. These measures may be normalized for providing a good set of weights for the distance measures.

3.2 Discussion

It is easy to prove [7] that these functions have the standard properties for pseudo metrics (i.e. semi-distances [14]):

Proposition 3.1 (pseudo-metric). *For a given a feature set* F *and* $p > 0$, d_p^{F} *is a pseudo-metric.*

Proof. It is to be proved that:

1. $d_p(a,b) \geq 0$
2. $d_p(a,b) = d_p(b,a)$
3. $d_p(a,c) \leq d_p(a,b) + d_p(b,c)$

1. and 2. are trivial. As for 3., noted that

$$\begin{aligned}
(d_p(a,c))^p &= \frac{1}{m^p}\sum_{i=1}^{m} w_i \mid \delta_i(a,c) \mid^p \\
&\leq \frac{1}{m^p}\sum_{i=1}^{m} w_i \mid \delta_i(a,b) + \delta_i(b,c) \mid^p \\
&\leq \frac{1}{m^p}\sum_{i=1}^{m} w_i \mid \delta_i(a,b) \mid^p + \frac{1}{m^p}\sum_{i=1}^{m} w_i \mid \delta_i(b,c) \mid^p \\
&\leq (d_p(a,b))^p + (d_p(b,c))^p \leq (d_p(a,b) + d_p(b,c))^p
\end{aligned}$$

then the property follows for the monotonicity of the root function.

It cannot be proved that $d_p^{\mathsf{F}}(a,b) = 0$ iff $a = b$. This is the case of *indiscernible* individuals with respect to the given set of features F. To fulfill this property several methods have been proposed involving the consideration of equivalent classes of individuals or the adoption of a supplementary meta-feature F_0 determining the equality of the two individuals: $\delta_0(a,b) = 0$ if $a^{\mathcal{I}} = b^{\mathcal{I}}$ otherwise $\delta_0(a,b) = 1$.

Compared to other proposed dissimilarity measures [3, 4], the presented functions do not depend on the constructors of a specific language, rather they require only (retrieval or) instance-checking for computing the projections through class-membership queries to the knowledge base.

The complexity of measuring he dissimilarity of two individuals depends on the complexity of such inferences (see [1], Ch. 3). Note also that the projections that determine the measure can be computed (or derived from statistics maintained on the knowledge base) before the actual distance application, thus determining a speed-up in the computation of the measure. This is very important for algorithms that massively use this distance, such as all instance-based methods.

The measures strongly depend on F. Here, we make the assumption that the feature-set F represents a sufficient number of (possibly redundant) features that are able to discriminate really different individuals. The choice of the concepts to be included – *feature selection* – is beyond the scope of this work (see [7] for a randomized optimization procedure aimed at finding optimal committees). Experimentally, we could obtain good results by using the very set of both primitive and defined concepts found in the knowledge base.

Of course these approximate measures become more and more precise as the knowledge base is populated with an increasing number of individuals.

4 Experimentation

4.1 Experimental Setting

The NN procedure integrated with the pseudo-metric proposed in the previous section has been tested in a number of retrieval problems. To this purpose, we selected several ontologies from different domains represented in OWL, namely: Surface-Water-Model (SWM), NewTestamentNames (NTN) from the Protégé library[3], the Semantic Web Service Discovery dataset[4] (SWSD), the University0.0 ontology generated by the Lehigh University Benchmark[5] (LUBM), the BioPax glycolysis ontology[6] (BioPax) and the Financial ontology[7]. Tab. 1 summarizes details concerning these ontologies.

[3] `http://protege.stanford.edu/plugins/owl/owl-library`

[4] `https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Projects/xmedia/dl-tree.htm`

[5] `http://swat.cse.lehigh.edu/projects/lubm`

[6] `http://www.biopax.org/Downloads/Level1v1.4/biopax-example-ecocyc-glycolysis.owl`

[7] `http://www.cs.put.poznan.pl/alawrynowicz/financial.owl`

Table 1. Facts concerning the ontologies employed in the experiments.

Ontology	DL language	#concepts	#object prop.	#data prop.	#individuals
SWM	$\mathcal{ALCOF}(D)$	19	9	1	115
BioPAX	$\mathcal{ALCHF}(D)$	28	19	30	323
LUBM	$\mathcal{ALR}^{+}\mathcal{HI}(D)$	43	7	25	555
NTN	$\mathcal{SHIF}(D)$	47	27	8	676
SWSD	$\mathcal{ALCH}$	258	25	0	732
Financial	$\mathcal{ALCIF}$	60	17	0	1000

For each ontology, 20 queries were randomly generated by composition (conjunction and/or disjunction) of (2 through 8) primitive and defined concepts in each knowledge base. Query concepts were constructed so that each offered both positive and negative instances among the ABox individuals. The performance of the inductive method was evaluated by comparing its responses to those returned by a standard reasoner[8] as a baseline.

Experimentally, it was observed that large training sets make the distance measures (and consequently the NN procedure) very accurate. In order to make the problems more difficult, we selected limited training sets (TS) that amount to only 4% of the individuals occurring in each ontology. Then the parameter k was set to $\log(|TS|)$ depending on the number of individuals in the training set. Again, we found experimentally that much smaller values could be chosen, resulting in the same classification.

The simpler distances (d_1^{F}) were employed from the *original* family (uniform weights) and *entropic* family (weighted on the feature entropy), using all the concepts in the knowledge base for determining the set F with no further optimization.

4.2 Results

Standard measures. Initially the standard measures precision, recall, F_1-measure were employed to evaluate the system performance, especially when selecting the positive instances (individuals that should belong to the query concept). The outcomes are reported in Tab. 2. For each knowledge base, we report the average values obtained over the 20 random queries as well as their standard deviation and minimum-maximum ranges of values.

As an overall consideration we may observe that generally the outcomes obtained adopting the extended measure improve on those with the other one and appear also more stable (with some exceptions). Besides, it is possible to note that precision and recall values are generally quite good for all ontologies but SWSD, where especially recall is significantly lower. Namely, SWSD turned out to be more difficult (also in terms of precision) for two reasons: a very limited number of individuals per concept was available and the number of different concepts is larger w.r.t. the other knowledge bases. For the other ontologies

[8] We employed Pellet v. 1.5.1. See `http://pellet.owldl.com`

Table 2. Experimental results in terms of standard measures: averages ± standard deviations and [min,max] intervals.

Original Measure			
	precision	recall	F-measure
SWM	89.1 ± 27.3 [16.3;100.0]	84.4 ± 30.6 [11.1;100.0]	78.7 ± 30.6 [20.0;100.0]
BioPax	99.2 ± 1.9 [93.8;100.0]	97.3 ± 11.3 [50.0;100.0]	97,8 ± 7.4 [66.7;100.0]
LUBM	100.0 ± 0.0 [100.0;100.0]	71.7 ± 38.4 [9.1;100.0]	76.2 ± 34.4 [16.7;100.0]
NTN	98.8 ± 3.0 [86.9;100.0]	62.6 ± 42.8 [4.3;100.0]	66.9 ± 37.7 [8.2;100.0]
SWSD	74.7 ± 37.2 [8.0;100.0]	43.4 ± 35.5 [2.2;100.0]	54.9 ± 34.7 [4.3;100.0]
Financial	99.6 ± 1.3 [94.3;100.0]	94.8 ± 15.3 [50.0;100.0]	97.1 ± 10.2 [66.7;100.0]
Entropic Measure			
	precision	recall	F-measure
SWM	99.0 ± 4.3 [80.6;100.0]	75.8 ± 36.7 [11.1;100.0]	79.5 ± 30.8 [20.0;100.0]
BioPax	99.9 ± 0.4 [98.2;100.0]	97.3 ± 11.3 [50.0;100.0]	98,2 ± 7.4 [66.7;100.0]
LUBM	100.0 ± 0.0 [100.0;100.0]	81.6 ± 32.8 [11.1;100.0]	85.0 ± 28.4 [20.0;100.0]
NTN	97.0 ± 5.8 [76.4;100.0]	40.1 ± 41.3 [4.3;100.0]	45.1 ± 35.4 [8.2;97.2]
SWSD	94.1 ± 18.0 [40.0;100.0]	38.4 ± 37.9 [2.4;100.0]	46.5 ± 35.0 [4.5;100.0]
Financial	99.8 ± 0.3 [98.7;100.0]	95.0 ± 15.4 [50.0;100.0]	96.6 ± 10.2 [66.7;100.0]

values are much higher, as testified also by the F-measure values. The results in terms of precision are also more stable than those for recall as proved by the limited variance observed, whereas single queries happened to turn out quite difficult as regards the correctness of the answer.

The reason for precision being generally higher is probably due to the OWA. Indeed, in a many cases it was observed that the NN procedure deemed some individuals as relevant for the query issued while the DL reasoner was not able to assess this relevance and this was computed as a mistake while it may likely turn out to be a correct inference when judged by a human agent.

Because of these problems in the evaluation with the standard indices, especially due to the cases on unknown answers from the reference system (the reasoner) we thought to make this case more explicit by measuring both the rate of inductively classified individuals and the nature of the mistakes.

Alternative measures. Due to the OWA, cases were observed when, it could not be (deductively) ascertained whether a resource was relevant or not for a given query. Hence, we introduced the following indices for a further evaluation:

- *match rate*: number of individuals that got exactly the same classification ($v \in V$) by both the inductive and the deductive classifier with respect to the overall number of individuals (v vs. v);
- *omission error rate*: amount of individuals for which inductive method could not determine whether they were relevant to the query or not while they were actually relevant according to the reasoner (0 vs. ± 1);
- *commission error rate*: number of individuals (analogically) found to be relevant to the query concept, while they (logically) belong to its negation or vice-versa ($+1$ vs. -1 or -1 vs. $+1$);
- *induction rate*: amount of individuals found to be relevant to the query concept or to its negation, while either case is not logically derivable from the knowledge base (± 1 vs. 0);

Tab. 3 reports the outcomes in terms of these indices. Preliminarily, it is important to note that, in each experiment, the commission error was quite low or absent. This means that the inductive search procedure is quite accurate, namely it did not make critical mistakes attributing an individual to a concept that is disjoint with the right one. Also the omission error rate was generally quite low, yet more frequent than the previous type of error.

The usage of all concepts for the set F of d_1^{F} made the measure quite accurate, which is the reason why the procedure resulted quite conservative as regards inducing new assertions. In many cases, it matched rather faithfully the reasoner decisions. From the retrieval point of view, the cases of induction are interesting because they suggest new assertions which cannot be logically derived by using a deductive reasoner yet they might be used to complete a knowledge base [2], e.g. after being validated by an ontology engineer. For each candidate new assertion, Eq. 2 may be employed to assess the likelihood and hence decide on its inclusion (see next section).

If we compare these outcomes with those reported in other works on instance retrieval and inductive classification [4], where the highest average match rate observed was around 80%, we find a significant increase of the performance due to the accuracy of the new measure. Also the elapsed time (not reported here) was much less because of the different dissimilarity measure: once the values for the projection functions are pre-computed, the efficiency of the classification, which depends on the computation of the dissimilarity, was also improved.

As mentioned, we found also that a choice for smaller number of neighbors could have been made for the decision on the correct classification was often quite easy, even on account of fewer (the closest) neighbors. This yielded also that the likelihood of the inferences made (see Eq. 2) turned out quite high.

Likelihood and Top-k Answers A further investigation concerned the likelihood of the inductively answers provided by the NN procedure. In Tab. 4, we

Table 3. Results with alternative indices: averages ± standard deviations and [min,max] intervals.

Original Measure				
	match	commission	omission	induction
SWM	93.3 ± 10.3 [68.7;100.0]	0.0 ± 0.0 [0.0;0.0]	2.5 ± 4.4 [0.0;16.5]	4.2 ± 10.5 [0.0;31.3]
BioPax	99.9 ± 0.2 [99.4;100.0]	0.2 ± 0.2 [0.0;0.06]	0.0 ± 0.0 [0.0;0.0]	0.0 ± 0.0 [0.0;0.0]
LUBM	99.2 ± 0.8 [98.0;100.0]	0.0 ± 0.0 [0.0;0.0]	0.8 ± 0.8 [0.0;0.2]	0.0 ± 0.0 [0.0;0.0]
NTN	98.6 ± 1.5 [93.9;100.0]	0.0 ± 0.1 [0.0;0.4]	0.8 ± 1.1 [0.0;3.7]	0.6 ± 1.4 [0.0;6.1]
SWSD	97.5 ± 3.7 [84.6;100.0]	0.0 ± 0.0 [0.0;0.0]	1.8 ± 2.6 [0.0;9.7]	0.8 ± 1.5 [0.0;5.7]
Financial	99.5 ± 0.8 [97.3;100.0]	0.3 ± 0.7 [0.0;2.4]	0.0 ± 0.0 [0.0;0.0]	0.2 ± 0.2 [0.0;0.6]
Entropic Measure				
	match	commission	omission	induction
SWM	97.5 ± 3.2 [89.6;100.0]	0.0 ± 0.0 [0.0;0.0]	2.2 ± 3.1 [0.0;10.4]	0.3 ± 1.2 [0.0;5.2]
BioPax	99.9 ± 0.2 [99.4;100.0]	0.1 ± 0.2 [0.0;0.06]	0.0 ± 0.0 [0.0;0.0]	0.0 ± 0.0 [0.0;0.0]
LUBM	99.5 ± 0.7 [98.2;100.0]	0.0 ± 0.0 [0.0;0.0]	0.5 ± 0.7 [0.0;1.8]	0.0 ± 0.0 [0.0;0.0]
NTN	97.5 ± 1.9 [91.3;99.3]	0.6 ± 0.7 [0.0;1.6]	1.3 ± 1.4 [0.0;4.9]	0.6 ± 1.7 [0.0;7.1]
SWSD	98.0 ± 3.0 [88.3;100.0]	0.0 ± 0.0 [0.0;0.0]	1.9 ± 2.9 [0.0;11.3]	0.1 ± 0.2 [0.0;0.5]
Financial	99.7 ± 0.2 [99.4;100.0]	0.0 ± 0.0 [0.0;0.1]	0.0 ± 0.0 [0.0;0.0]	0.2 ± 0.2 [0.0;0.6]

report the average likelihoods computed (for all queries per ontology) during the previous experiments in case of induction of new consistent assertions (see Eq. 2), when the reasoner was not able to assess the membership. The first line reports the averages when answers were given based on the normalization of the likelihood over the 3 possible values. As expected, they are even higher when only the two cases $+1$ or -1 (membership, non-membership) are considered (see second line). As mentioned, since the distance measure accurately selected very similar neighbors, seldom tight cases occurred during the majority votes of the NN, hence the observed likelihood of the answers turned out quite high on average.

We also took into account the top-10 (positive) answers provided by the inductive procedure for the various queries, ranked according to the likelihood of the decision. Most of the values amounted to 100%. In order to assess the accuracy of such answers, we compared their related likelihood values to those of the

Table 4. Results (percentages) concerning the likelihood of the answers when the reasoner is not able to assess the class membership.

	SWM	BioPax	LUBM	NTN	SWSD	Financial
3-valued case	76.26	99.99	99.99	98.36	76.27	92.55
2-valued case	100.0	99.99	99.99	98.36	76.27	92.55

Table 5. Average differences of likelihood values (%) observed comparing the NN procedure to the reasoner on the top-10 (positive) answers.

	SWM	BioPax	LUBM	NTN	SWSD	Financial
likelihood diff.	0.0	0.2	0.0	0.3	2.5	0

deductive decisions made by the reasoner. Namely, we assigned a maximum likelihood of 100% to the decisions on membership (and 0% to the non-membership answer, if any) while 50% was assigned to cases when the reasoner was uncertain on the answer. The pairwise difference of likelihood are averaged over the top-10 answers of the various queries per each ontology. In Tab 5, we report such average difference. As expected such difference values are quite low, reflecting the fact that the top-ranked answers are also the most accurate ones.

5 Conclusions and Outlook

This paper explored the application of a distance-based procedure for semantic search applied knowledge bases represented in OWL. We extended a family of semantic dissimilarity measures based on feature committees [7] taking into account the amount of information conveyed by each feature based on an estimate of its entropy. The measure were integrated in an distance-based search procedure that can be exploited for the task of approximate instance retrieval which can be demonstrated to be effective even in the presence of incomplete (or noisy) information.

One of the advantages of the measures is that their computation can be very efficient in cases when statistics (on class-membership) are maintained by the KBMS [10]. As previously mentioned, the subsumption relationships among concepts in the committee is not explicitly exploited in the measure for making the relative distances more accurate. The extension to the case of concept distance may also be improved. Hence, scalability should be guaranteed as far as a good committee has been found and does not change also because of the locality properties observed for instances in several domains (e.g. social or biological networks).

The experiments made on various ontologies showed that the method is quite effective, and while its performance depends on the number (and distribution) of the available training instances, even working with quite limited training sets guarantees a good performance in terms of accuracy. Moreover, even if the measure accuracy embedded into the system depends on the chosen feature set,

the high accuracy registered for almost all considered data sets shows that the method can be applied to any domain and its performances are not connected to a particular domain. Besides, the procedure appears also robust to noise since it seldom made commission errors in the experiments carried out so far.

Various developments for the measure can be foreseen as concerns its definition. Namely, since it is very dependent on the features included in the committee, two immediate lines of research arise: 1) reducing the number of concepts saving those concepts which are endowed of a real discriminating power; 2) learning optimal sets of discriminating features, by allowing also their composition employing the specific constructors made available by the representation language of choice [7]. Both these objectives can be accomplished by means of machine learning techniques especially when ontologies with a large set of individuals are available. Namely, part of the entire data can be drawn in order to learn optimal feature sets, in advance with respect to the successive usage.

As mentioned, the distance measures are applicable to other instance-based tasks which can be approached through machine learning techniques. The next step has been plugging the measure in flat or hierarchical clustering algorithms where clusters would be formed grouping instances on the grounds of their similarity assessed through the measure [6, 5].

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] F. Baader, B. Ganter, B. Sertkaya, and U. Sattler. Completing description logic knowledge bases using formal concept analysis. In M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 230–235, Hyderabad, India, 2007.

[3] A. Borgida, T.J. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, *Working Notes of the International Description Logics Workshop*, volume 147 of *CEUR Workshop Proceedings*, Edinburgh, UK, 2005.

[4] C. d'Amato, N. Fanizzi, and F. Esposito. Reasoning by analogy in description logics through instance-based learning. In G. Tummarello, P. Bouquet, and O. Signore, editors, *Proceedings of Semantic Web Applications and Perspectives, 3rd Italian Semantic Web Workshop, SWAP2006*, volume 201 of *CEUR Workshop Proceedings*, Pisa, Italy, 2006.

[5] N. Fanizzi, C. d'Amato, and F. Esposito. Evolutionary conceptual clustering of semantically annotated resources. In *Proceedings of the IEEE International Conference on Semantic Computing, ICSC2007*, Irvine, CA, 2007. IEEE.

[6] N. Fanizzi, C. d'Amato, and F. Esposito. A hierarchical clustering procedure for semantically annotated resources. In R. Basili and M.T. Pazienza, editors, *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence, AI*IA2007*, volume 4733 of *LNAI*, pages 266–277. Springer, 2007.

[7] N. Fanizzi, C. d'Amato, and F. Esposito. Induction of optimal semi-distances for individuals based on feature sets. In D. Calvanese, E. Franconi, V. Haarslev, D. Lembo, B. Motik, A.-Y. Turhan, and S. Tessaris, editors, *Working Notes of the*

20th International Description Logics Workshop, DL2007, volume 250 of *CEUR Workshop Proceedings*, Bressanone, Italy, 2007.

[8] P. Haase, F. van Harmelen, Z. Huang, H. Stuckenschmidt, and Y. Sure. A framework for handling inconsistency in changing ontologies. In Y. Gil, V. Motta, E. Benjamins, and Mark A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference, ISWC2005*, number 3279 in LNCS, pages 353–367, Galway, Ireland, November 2005. Springer.

[9] P. Hitzler and D. Vrandečić. Resolution-based approximate reasoning for OWL DL. In Y. Gil, V. Motta, E. Benjamins, and Mark A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference, ISWC2005*, number 3279 in LNCS, pages 383–397, Galway, Ireland, November 2005. Springer.

[10] I. R. Horrocks, L. Li, D. Turi, and S. K. Bechhofer. The Instance Store: DL reasoning with large numbers of individuals. In V. Haarslev and R. Möller, editors, *Proceedings of the 2004 Description Logic Workshop, DL 2004*, volume 104 of *CEUR Workshop Proceedings*, pages 31–40. CEUR, 2004.

[11] R. Möller, V. Haarslev, and M. Wessel. On the scalability of description logic instance retrieval. In B. Parsia, U. Sattler, and D. Toman, editors, *Description Logics*, volume 189 of *CEUR Workshop Proceedings*. CEUR, 2006.

[12] M. Sebag. Distance induction in first order logic. In S. Džeroski and N. Lavrač, editors, *Proceedings of the 7th International Workshop on Inductive Logic Programming, ILP97*, volume 1297 of *LNAI*, pages 264–272. Springer, 1997.

[13] H. Wache, P. Groot, and H. Stuckenschmidt. Scalable instance retrieval for the semantic web by approximation. In M. Dean et al., editor, *WISE Workshops*, volume 3807 of *LNCS*, pages 245–254. Springer, 2005.

[14] P. Zezula, G. Amati, V. Dohnal, and M. Batko. *Similarity Search – The Metric Space Approach*. Advances in database Systems. Springer, 2007.