Instance-based Query Answering with Semantic Knowledge Bases

Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito

Dipartimento di Informatica, Università degli Studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{fanizzi|claudia.damato|esposito}@di.uniba.it

Abstract. A procedure founded in *instance-based learning* is presented, for performing a form of analogical reasoning on knowledge bases expressed in a wide range of ontology languages. The procedure exploits a novel semi-distance measure for individuals, that is based on their semantics w.r.t. a number of dimensions corresponding to a committee of features represented by concept descriptions. The procedure can answer by analogy to class'membership queries on the grounds of the classification of a number of training instances (the nearest ones w.r.t. the semi-distance measure). Particularly, it may also predict assertions that are not logically entailed by the knowledge base. In the experimentation, where we compare the procedure to a logical reasoner, we show that it can be quite accurate and augment the scope of its applicability, outperforming previous prototypes that adopted other semantic measures.

1 Introduction

In the perspective of knowledge sharing and reuse of the *social* vision of the Semantic Web, new services are required aiming at noise-tolerant and efficient forms of reasoning. From this perspective, instance-based inductive methods applied to multi-relational domains appear particularly well suited. Indeed, they are known to be both very efficient and noise-tolerant and noise is always harmful in contexts where knowledge is to be acquired from distributed sources.

A relational instance-based framework for the Semantic Web context has been devised (based on a similarity measure) to derive (by analogy) both consistent consequences from the knowledge base and, possibly, also new assertions which were not previously logically derivable. The main idea is that similar individuals, by analogy, should likely belong to similar concepts. Specifically, we derive a classification procedure that constitutes a relational form of the *Nearest Neighbor* algorithm (*NN*, henceforth) [5], a well-known approach to *lazy learning*.

These algorithms are efficient because no explicit hypothesis has to be learned. Rather, the workload is shifted towards the classification phase when knowledge concerning the training instances is used to classify new ones. Particularly, this only requires checking the assertions for a limited set of instances training on such concepts and making a decision (classification) for new query instances.

From a technical viewpoint, upgrading NN algorithms to work on multi-relational representations [5], like the concept languages used in the Semantic Web [4], required novel similarity measures that are suitable for such representations. This adaptation could not be straightforward. In particular, a theoretical problem has been posed by the *Open World Assumption* (OWA) that is generally made in the target context, differently from typical the machine learning settings where the *Closed World Assumption* (CWA) is the standard. Besides, in the standard NN multi-class setting, different concept are assumed to be disjoint, which typically cannot hold in a Semantic Web context. As pointed out in [2], most of the existing measures focus on the similarity of atomic concepts within hierarchies or simple ontologies. Moreover they have been conceived for assessing *concept* similarity. On the other hand, for our purposes, a notion of similarity between *individuals* is required.

Recently, dissimilarity measures for specific description logics concept descriptions have been proposed [3]. Although they turned out to be quite effective for the inductive tasks of interest [4], they were still partly based on structural criteria (a notion of normal form) which determine their main weakness: they are hardly scalable to deal with standard languages, such as OWL-DL, commonly used for knowledge bases.

In this paper we introduce a new semantic dissimilarity measure which can overcome these limitations. Following some ideas introduced in [6], we present a new family of measures that is suitable a wide range of ontology languages (RDF through OWL) since it is merely based on the discernibility of the input individuals with respect to a fixed set of features represented by concept definitions (hypotheses). As such the new measures are not absolute, yet they depend on the knowledge base they are applied to.

The measure and the NN procedure have been integrated in a system that allowed for an extended experimentation the method on performing instance retrieval with real ontologies drawn from public repositories comparing its predictions to the assertions that were logically derived by a standard reasoner. These experiments show that the novel measure considerably increases the effectiveness of the method with respect to the past experiments where the same procedure was integrated with other structural/semantic dissimilarity measures [4]. Moreover, as expected, an increase of accuracy was observed with the increase of the dimensions employed for the measure, which proposes further lines of development for the presented measure.

The paper is organized as follows. The basics of the instance-based approach applied to ontology representations are recalled in Sect. 2. The next Sect. 3 presents the novel semantic similarity measure adopted with the inductive procedure. Successively, Sect. 4 reports the outcomes of experiments performed with its implementation. Possible developments are finally examined in Sect. 5.

2 A Nearest Neighbor Approach to Instance-checking

In the following, we assume that concept descriptions are defined in terms of a generic ontology language that may be mapped to some description logic 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 concept definitions. $\mathcal{A}$ contains assertions (facts, data) concerning the world state. Moreover, normally the *unique names assumption* is made on the ABox individuals. 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*, which amounts to determining whether an individual, say a, belongs to a concept extension, i.e. whether $C(a)$ holds for a certain concept C.

Given an ontology, a classification method can employed for predicting the concepts to which a new individual it is likely to belong. These individuals are supposed to be described by assertions in the knowledge base. Such an classification procedure may also suggest new assertions about such an individual which cannot not be inferred by deduction, in analogy with the class-membership of other similar instances.

We review the basics of the k-Nearest Neighbor method in the semantic web context [4] and propose how to exploit this classification procedure for inductive instance checking and query answering. It is ascribed to the category of lazy learning, since the learning phase is reduced to memorizing instances of the target concepts pre-classified by an expert. Then, during the classification phase, a notion of similarity over the instance space is employed to classify a new instance in analogy with its neighbors.

The objective is to induce an approximation for a discrete-valued target 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.

Let x_q be the query instance whose class-membership is to be checked. 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 majority voting procedure: it is simply the most *voted* value by the instances in $NN(x_q)$.

A problem with this formulation is that it takes into account similarity only when selecting those instance to be included in the neighborhood. Therefore a modified setting is generally adopted, that is based on weighting the vote according to the distance of the query instance from the training instances:

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

where δ is a function that returns 1 in case of matching arguments and 0 otherwise, and, given a distance measure d, the weights are determined by $w_i = 1/d(x_i, x_q)$ or $w_i = 1/d(x_i, x_q)^2$.

Note that the hypothesis function $\hat{h}$ is defined only extensionally, since 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 query instances to be classified.

It should be also observed that a strong assumption of this setting is that it can be employed to assign the query instance to the concept from a set of pairwise disjoint concepts (those corresponding to the value set V). This is an assumption that cannot be always made. In our setting, indeed, an individual might be an instance of more than one concept. In a more general setting that does not assume the pairwise disjointness of the concepts, given a query concept Q, the membership of an instance x_q may be checked through a NN classification procedure, transforming the multi-class problem into a binary one. Therefore, a simple binary value set $V = \{-1, +1\}$ may be employed. Then, a hypothesis h_Q is computed for performing inductive query answering:

$$\hat{h}_Q(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} \frac{\delta(v, h_Q(x_i))}{d(x_q, x_i)^2} \tag{2}$$

where the value of h_Q for the training instances x_i is simply determined to the occurrence $(+1)$ or absence (-1) of the corresponding assertion $Q(x_i)$ in the ABox. Alternately, Q may return $+1$ when $Q(x_i)$ can be inferred[1] from the knowledge base ($\mathcal{K} \models Q(x_i)$), and -1 otherwise.

The problem with non-explicitly disjoint concepts 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 certain training instance x is likely to belong to the extension of the query concept Q should not be interpreted negatively, as in the standard machine learning settings which adopt the CWA. Rather, it should count as neutral information. Thus, another value set has to be adopted for the classification of the neighboring training instances, namely $V = \{-1, +1, 0\}$, where the three values denote, respectively, occurrence of the assertion, occurrence of the opposite assertion and absence of both:

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

Occurrence can be easily computed with a look-up in the ABox, therefore the overall complexity of the procedure depends on the number $k \ll |\mathsf{Ind}(\mathcal{A})|$, that is the number of times the distance measure is needed.

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 (i.e. incorrect

[1] In the following $\models$ will denote entailment, as computed through a reasoner.

assertions in the ABox), therefore it may be able to give a correct classification even in case of (partially) inconsistent knowledge bases.

Again, a more complex procedure may be devised by simply substituting the notion of occurrence (absence) of assertions in (from) the ABox with the one of derivability (non-derivability) from the whole knowledge base, i.e. $\mathcal{K} \models Q(x)$, $\mathcal{K} \models \neg Q(x)$ and neither of the previous relations, respectively. Although this may exploit more information and turn out to be more accurate, it is also much more computationally expensive, since the simple look-up in the ABox must be replaced with a logical inference (instance checking). However much of the computation can be performed in advance w.r.t. to the classification phase and the number of inferences needed is bounded by k.

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 concept denoted by value v maximizing the argmax argument in Eq. (2)), given the nearest training individuals in $NN(x_q)$, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$l(class(x_q) = v | NN(x_q)) = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_Q(x_i))}{\sum_{v' \in V} \sum_{i=1}^{k} w_i \cdot \delta(v', h_Q(x_i))} \tag{3}$$

3 A Semantic Semi-Distance 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 [2] or they are not completely semantic depending also on the structure of the descriptions [3]. Moreover, for 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].

For the nearest-neighbor classification procedure recalled in Sect. 2, we have developed a new measure with a definition that totally depends on semantic aspects of the individuals in the knowledge base.

3.1 The Measure

On a semantic level, similar individuals should behave similarly with respect to the same concepts. We introduce a novel measure for assessing the similarity of individuals in a knowledge base, which is based on the idea of comparing their semantics along a number of dimensions represented by a committee of

concept descriptions. Following the ideas borrowed from [6], we can define totally semantic distance measures for individuals in the context of a knowledge base.

The rationale of the new measure is to compare them on the grounds of their behavior w.r.t. a given set of hypotheses, that is 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 language taken into account.

In its simple formulation, a family of distance functions for individuals inspired to Minkowski's distances can be defined as follows:

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

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

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

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 1 & F_i(x) \in \mathcal{A} \\ 0 & \neg F_i(x) \in \mathcal{A} \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

The superscript F will be omitted when the set of hypotheses is fixed.

As an alternative, like in the definition of the hypothesis function for the NN procedure, the definition of the measures can be made more accurate by considering entailment rather than the simple ABox look-up, when determining the values of the projection functions:

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 1 & \mathcal{K} \models F_i(x) \\ 0 & \mathcal{K} \models \neg F_i(x) \\ \frac{1}{2} & \textit{otherwise} \end{cases}$$

In particular, we will consider the following measures:

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \quad d_1(a, b) := \frac{1}{m} \sum_{i=1}^{m} \mid \pi_i(a) - \pi_i(b) \mid$$

or:

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \quad d_2(a, b) := \frac{1}{m} \sqrt{\sum_{i=1}^{m} (\pi_i(a) - \pi_i(b))^2}$$

3.2 Discussion

It is easy to prove that these functions have the standard properties for semi-distances:

Proposition 3.1 (semi-distance). *For a fixed hypothesis set and $p > 0$, given any three instances $a, b, c \in \mathsf{Ind}(\mathcal{A})$. it holds that:*

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

Proof.

1. *trivial*
2. *trivial*
3. *Noted that*

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

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

It cannot be proved that $d_p(a, b) = 0$ iff $a = b$. This is the case of *indiscernible* individuals with respect to the given set of hypotheses F.

Compared to other proposed distance (or dissimilarity) measures [2], the presented function does not depend on the constructors of a specific language, rather it requires only retrieval or instance-checking service used for deciding whether an individual is asserted in the knowledge base to belong to a concept extension (or, alternatively, if this could be derived as a logical consequence).

Note that the π_i functions ($\forall i = 1, \ldots, m$) for the training instances, that contribute to determine the measure with respect to new ones, can be computed in advance thus determining a speed-up in the actual computation of the measure. This is very important for the measure integration in algorithms which massively use this distance, such as all instance-based methods.

The underlying idea for the measure is that similar individuals should exhibit the same behavior w.r.t. the concepts in 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. Experimentally, we could obtain good results by using the very set of both primitive and defined concepts found in the ontology.

Table 1. Ontologies employed in the experiments.

ontology	*DL*	*#concepts*	*#obj. prop*	*#data prop*	*#individuals*
FSM	$\mathcal{SOF}(D)$	20	10	7	37
S.-W.-M.	$\mathcal{ALCOF}(D)$	19	9	1	115
SCIENCE	$\mathcal{ALCIF}(D)$	74	70	40	331
FINANCIAL	$\mathcal{ALCIF}$	60	17	0	652
NTN	$\mathcal{SHIF}(D)$	47	27	8	676

4 Experiments

4.1 Experimental Setting

In order to test the inductive instance-checking NN procedure proposed in Sect. 2, integrated with the new dissimilarity measure, we have applied it to a number of retrieval problems. To these purposes, we selected a number of different ontologies represented in OWL, namely: FSM, SURFACE-WATER-MODEL, SCIENCE and NEWTESTAMENTNAMES from the Protégé library[2], the FINANCIAL ontology[3] employed as a testbed for the PELLET reasoner. Table 1 summarizes important details concerning the ontologies employed in the experimentation.

The FSM ontology describes the domain of *finite state machines* using the $\mathcal{SOF}(D)$ language. It is made up of 20 (primitive and defined) concepts (some of them are explicitly declared to be disjoint), 10 object properties, 7 datatype properties, 37 distinct individual names. About half of the individuals are asserted as instances of a single concept and are not involved in any role (object property) assertion. SURFACE-WATER-MODEL is an $\mathcal{ALCOF}(D)$ ontology describing the domain of the surface water and the water quality models. It is made up of 19 concepts (both primitive and defined) with no specification about their disjointness, 9 object properties, 115 distinct individual names; each of them is an instance of a single class and only some of them are involved in object properties. The SCIENCE ontology describes scientific facts in $\mathcal{ALCIF}(D)$. It is made up of 74 concepts, 70 object properties, 331 individual names. FINANCIAL is an $\mathcal{ALCIF}$ ontology that describes the domain of eBanking. It is made up of 60 (primitive and defined) concepts (some of them are declared to be disjoint), 17 object properties, and no datatype property. It contains 17941 distinct individual names. From the original ABox, we randomly extracted assertions for 652 individuals. NEWTESTAMENTNAMES (developed for the *Semantic Bible* Project) describes facts related to the New Testament. It contains of 47 concepts, 27 object properties, 676 individual names.

The experiment was quite intensive involving the classification of all the individuals in each ontology; namely, the individuals were checked through the

[2] `http://protege.stanford.edu/plugins/owl/owl-library`
[3] `http://www.cs.put.poznan.pl/alawrynowicz/financial.owl`

inductive procedure to assess wether they were to be retrieved as instances of a query concept. Therefore, 15 queries were randomly generated by conjunction/disjunction of primitive or defined concepts of each ontology. The performance was evaluated comparing its responses to those returned by a standard reasoner[4] as a baseline.

The experiment has been repeated twice adopting different procedures according to the size of the corresponding ABox (measured by $|\mathsf{Ind}(\mathcal{A})|$): a leave-one-out cross validation for the smaller ontologies (FSM and S.-W.-M.) and a ten-fold cross validation one for the larger ones Applying the k-NN method, we chose $\sqrt{|\mathsf{Ind}(\mathcal{A})|}$, as a value for k, as advised in the instance-based learning literature. Yet we found experimentally that much smaller values could be chosen, resulting in the same classification. We employed the simpler version of the distance (d_1) utilizing all the concepts in the ontology for determining the set F.

For each concept in the ontology, we measured the following parameters for the evaluation:

- *match rate*: number of cases of individuals that got exactly the same classification by both classifiers with respect to the overall number of individuals;
- *omission error rate*: amount of unlabeled individuals (our method could not determine whether it was an instance or not) while it was to be classified as an instance of that concept;
- *commission error rate*: amount of individuals (analogically) labeled as instances of a concept, while they (logically) belong to that concept or viceversa
- *induction rate*: amount of individuals that were found to belong to a concept or its negation, while this information is not logically derivable from the knowledge base

We report the average rates obtained over all the concepts in each ontology and also their standard deviation.

4.2 Retrieval Employing the New Measure in the NN Procedure

By looking at Tab. 2 reporting the experimental outcomes (mean values and standard deviations), preliminarily it is important to note that, for every ontology, the commission error was low. This means that the procedure is quite accurate: it did not make critical mistakes i.e. cases when an individual is deemed as an instance of a concept while it really is an instance of another disjoint concept.

If we compare these outcomes with those reported in previous papers [4], where the average accuracy on the same was slightly higher than 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 lowered because, once the values for the π's functions are pre-computed, the efficiency of the classification, which depends a lot on the computation of the dissimilarity, gains a lot of speed-up.

[4] We employed PELLET: `http://pellet.owldl.com`

Table 2. Results (average±std-dev.) of the experiments with the method employing the new semantic measure.

	match rate	*commission rate*	*omission rate*	*induction rate*
FSM	97.7 ± 3.00	2.30 ± 3.00	0.00 ± 0.00	0.00 ± 0.00
S.-W.-M.	99.9 ± 0.20	0.00 ± 0.00	0.10 ± 0.20	0.00 ± 0.00
SCIENCE	99.8 ± 0.50	0.00 ± 0.00	0.20 ± 0.10	0.00 ± 0.00
FINANCIAL	90.4 ± 24.6	9.40 ± 24.5	0.10 ± 0.10	0.10 ± 0.20
NTN	99.9 ± 0.10	0.00 ± 7.60	0.10 ± 0.00	0.00 ± 0.10

The usage of all concepts for the set F made the measure quite accurate, which is the reason why the procedure resulted quite conservative as regards inducing new assertions. It rather matched faithfully the reasoner decisions. A noteworthy difference was observed for the case of the FINANCIAL ontology for which we find the lowest match rate and the highest variability in the results over the various concepts. On a careful examination of the experimentation with this ontology, we found that the average results were lowered by a concept whose assertions, having been poorly sampled from the initial ontology, could not constitute enough evidence to our inductive method for determining the correct classification. The same problem, to a lesser extent, were found also with the FSM ontology which was the one with the least number of assertions. This shows that the weaker side of any instance-based procedure is really when data are too sparse or non evenly distributed.

As mentioned, we found also that a lower value for k could have been chosen, for in many cases the decision on the correct classification was easy to make even on account of a few (the closest) neighbor instances. This yields also the likelihood of the inference made (see Eq. (3)) turned out quite high.

4.3 Varying the Hypothesis Set

In the previous experiments all concepts involved in an ontology were used for inclusion in the hypothesis set F. We sensed that the inherent redundancy helped a lot the measure accuracy. Yet larger sets yield more effort to be made for computing the measures. Nevertheless, it is well known that the NN approach suffers when lots of further irrelevant attributes for describing the instances are added. Thus, we have tested also how the variation of hypotheses (concept descriptions) belonging to the set F could affect the performance of the measure. We expected that with an increasing number of hypotheses considered in F, the accuracy of the measure would increase accordingly.

To test this claim experimentally, one of the ontologies considered the previous experiment was considered. We performed repeatedly (three times) a leave-one-out cross validation with an increasing percentage of concepts randomly selected for F w.r.t. the overall number of primitive or defined concept names in the ontology. The average results returned by the system are depicted in Fig. 1. Numerical details of such outcomes are given in Table 3.

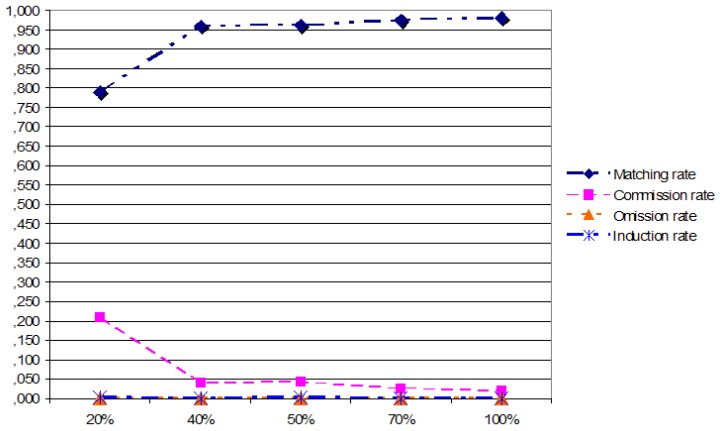

Fig. 1. Average results varying the number of hypotheses in the set F.

As expected, it is possible to see that the accuracy of the decisions (*match rate*) is positively correlated with the number of concepts included in F. The same outcomes were obtained by repeating similar experiments with other ontologies. It should be observed that in some cases the concepts randomly selected for inclusion in F actually turned out to be a little redundant (by subsumption or because of a simple overlap between their extension) This suggest a line of further investigation that will concern finding minimal subsets of concepts to be used for the measure.

5 Conclusions and Future Work

This paper explored the application of an instance-based learning procedure for analogical reasoning applied to concept representations adopted in the Semantic Web context. We defined a novel semantic similarity measure that has a wide scope of application to methods which require the assessment of the semantic (dis)ssimilarity of individuals. Particularly, in this paper we employed it integrated in an instance-based instance-checking procedure in the task of instance

Table 3. Average results varying the number of hypotheses in the set F.

	match	*commission*	*omission*	*Induction*
% of concepts	*rate*	*rate*	*rate*	*rate*
20%	79.1	20.7	0.00	0.20
40%	96.1	03.9	0.00	0.00
50%	97.2	02.8	0.00	0.00
70%	97.4	02.6	0.00	0.00
100%	98.0	02.0	0.00	0.00

retrieval (predicting class-membership) which can be effective even in the presence of missing (or noisy) information in the knowledge bases.

The experiments made on various ontologies showed that the method is quite effective, and, as expected, its performance depends on the number (and distribution) of the available training instances. Besides, the procedure is robust to noise since it seldom made commission errors in the experiments that have been carried out so far.

Various developments for the measure can be foreseen as concerns its definition. Namely, since it is very dependant on the concepts included in the committee of features, 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. 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 measure is applicable to other instance-based tasks which can be approached through machine learning techniques. The next step will be plugging the measure in a hierarchical clustering algorithm where clusters would be formed grouping instances on the grounds of their similarity assessed through the measure.

References

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

[2] A. Borgida, T. 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.

[3] C. d'Amato, N. Fanizzi, and F. Esposito. A dissimilarity measure for $\mathcal{ALC}$ concept descriptions. In *Proceedings of the 21st Annual ACM Symposium of Applied Computing, SAC2006*, volume 2, pages 1695–1699, Dijon, France, 2006. ACM.

[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] W. Emde and D. Wettschereck. Relational instance-based learning. In L. Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning, ICML96*, pages 122–130. Morgan Kaufmann, 1996.

[6] 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.