Reasoning by Analogy in Description Logics through Instance-based Learning

Claudia d'Amato, Nicola Fanizzi, Floriana Esposito
Dipartimento di Informatica, Università degli Studi di Bari
Campus Universitario, Via Orabona 4, 70125 Bari, Italy
{claudia.damato | fanizzi | esposito}@di.uniba.it

***Abstract*— This work presents a method founded in instance-based learning for inductive (memory-based) reasoning on ABoxes. The method, which exploits a semantic dissimilarity measure between concepts and instances, can be employed both to answer class membership queries and to predict new assertions that may be not logically entailed by the knowledge base. In a preliminary experimentation, we show that the method is sound and it is actually able to induce new assertions that might be acquired in the knowledge base.**

I. Introduction

Most of the research on ontology reasoning has been focusing on methods based on deductive reasoning. However, important tasks that are likely to be provided by new generation knowledge-based systems, such as construction, revision, population and evolution are likely to be supported also by inductive methods. This has brought to an increasing interest in *machine learning* and *knowledge discovery* methods applied to ontological representations (see [1], [2] and, more recently, [3], [4], [5], [6]).

We propose an algorithm which is based on a notion of concept similarity for performing a form of lazy learning on typical ontological representations. Namely, by combining an instance-based (analogical) approach with a notion of semantic dissimilarity, this paper intends to demonstrate the applicability of inductive reasoning in this field which can be considered another form of approximate reasoning (see discussion in [7]). In particular, we have adapted the general instance-based learning approach like the k-*Nearest Neighbor* [8] to the specific multi-relational setting for ontology languages. A couple of technical problems had to be solved for this adaptation to ontology representations: 1) the *Open World Assumption* (*OWA*) that is made in this context; 2) in this multi-class problem setting disjunction cannot be assumed by default.

The standard ontology languages are founded in Description Logics (henceforth DLs) as they borrow the language constructors for expressing complex concept definitions. Instance-based methods depend on a similarity measure defined on the instance space. In this perspective, a variety of measures for concept representations have been proposed (see [9] for a survey). As pointed out in [10], most of these measures focus on the similarity of atomic concepts within hierarchies or simple ontologies, based on a few relations. Thus, it becomes necessary to investigate similarity in more complex languages.

It has been observed that, adopting richer representations, the structural properties have less and less impact in assessing semantic similarity. Hence, the vision of similarity based only on a structural (graph-based) approach, such as in [11], [12] may fall short. We have proposed some dissimilarity measures for non trivial DL languages, based on the semantics conveyed by the ABox assertions, which are suitable for being used in instance-based methods [13], [14]. These measures elicit the underlying semantics by querying the knowledge base for assessing the concept extensions, estimated through their *retrieval* [15], as also hinted in [16]. Besides, the overall similarity is also (partially) influenced by the concepts which are related through role restrictions. Moreover, in many other typical tasks (e.g. conceptual clustering or definition), it is necessary to assess the similarity between concepts (resp. individuals). By recurring to the notion of *most specific concept* of an individual with respect to an ABox [15], as representatives of the individuals at the concept level, the measures for concepts can be extended to such cases.

This analogical reasoning procedure like this may be employed to answering class membership queries through analogical rather than deductive reasoning which may be more robust with respect to noise and is likely to suggest new knowledge (which was not logically derivable). Specifically we developed the method also for an application of semantic web service discovery where services are annotated in DLs.

Another application might regard supporting various tasks for the knowledge engineer, such as the acquisition of candidate assertions for enriching ontologies with partially populated ABoxes: the outcomes given by the procedure can be utilized as recommendations. Indeed, as we show in the experimentation, the newly induced assertions are quite accurate (commission errors, i.e. predicting a concept erroneously, were rarely observed). In turn, the outcomes of the procedure may also trigger other related tasks such as induction (revision) of (faulty) knowledge bases.

The paper is organized as follows. In the next section, the representation language is briefly presented. Two concept dissimilarity measures are recalled and exploited in a modified version of the k-NN classification procedure. The results of a preliminary experimental evaluation of the method using these two measures are shown and, finally, possible developments of the method are examined.

II. $\mathcal{ALC}$ Knowledge Bases

We recall the basics of $\mathcal{ALC}$ [17], a logic which adopts constructors supported by the standard ontology languages (see the DL handbook [15] for a thorough reference).

In DLs, descriptions are inductively defined starting with a set N_C of *primitive concept* names and a set N_R of *primitive roles*. Complex descriptions are built using primitive concepts and roles and the constructors showed in the following. The semantics of the descriptions is defined by an *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where $\Delta^{\mathcal{I}}$ is a non-empty set, the *domain* of the interpretation, and $\cdot^{\mathcal{I}}$ is the *interpretation function* that maps each $A \in N_C$ to a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and each $R \in N_R$ to $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$.

The *top* concept $\top$ is interpreted as the whole domain of objects $\Delta^{\mathcal{I}}$, while the *bottom* concept $\bot$ corresponds to $\emptyset$. Complex descriptions can be built in $\mathcal{ALC}$ using the following constructors. The language supports *full negation*: given any description C, denoted $\neg C$, it amounts to $\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$. Concept *conjunction*, denoted $C_1 \sqcap C_2$, yields the extension $C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$; dually, concept *disjunction*, denoted $C_1 \sqcup C_2$, yields the union $C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$. Finally, the *existential restriction*, denoted $\exists R.C$, is interpreted as $\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}})\}$ and the *value restriction* $\forall R.C$ has as its extension the set $\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}}((x,y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}})\}$.

The main inference is *subsumption* between concepts based on their semantics: given two descriptions C and D, C *subsumes* D, denoted by $C \sqsupseteq D$, iff for every interpretation $\mathcal{I}$ it holds that $C^{\mathcal{I}} \supseteq D^{\mathcal{I}}$. When $C \sqsupseteq D$ and $D \sqsupseteq C$ then they are equivalent, denoted with $C \equiv D$.

A *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$:

- $\mathcal{T}$ is the set of definitions $C \equiv D$, meaning that, for every interpretation $\mathcal{I}$, $C^{\mathcal{I}} = D^{\mathcal{I}}$, where C is the concept name and D is its description constructed as above;
- $\mathcal{A}$ contains concept and role assertions about the world-state, e.g. $C(a)$ and $R(a,b)$, meaning that, for every $\mathcal{I}$, $a^{\mathcal{I}} \in C^{\mathcal{I}}$ and $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$.

A related inference used in the following is *instance checking*, that is deciding whether an individual is an instance of a concept [18], [15]. Conversely, it may be necessary to find the concepts which an individual belongs to (*realization problem*), especially the most specific one (see Def. 5).

Semantically equivalent (yet syntactically different) descriptions can be given for the same concept. Nevertheless, equivalent concepts can be reduced to a normal form by means of rewriting rules that preserve their equivalence [19]:

A description D is in $\mathcal{ALC}$ *normal form* iff $D \equiv \bot$ (then $D := \bot$) or if $D \equiv \top$ (then $D := \top$) or if $D = D_1 \sqcup \cdots \sqcup D_n$ ($\forall i = 1, \ldots, n$, $D_i \not\equiv \bot$) with

$$D_i = \bigsqcap_{A \in \mathsf{prim}(D_i)} A \sqcap \bigsqcap_{R \in N_R} \left[\forall R.\mathsf{val}_R(D_i) \sqcap \bigsqcap_{E \in \mathsf{ex}_R(D_i)} \exists R.E \right]$$

where:

- $\mathsf{prim}(D_i)$ is the set of all (negated) primitive concepts occurring at the top level of D_i;
- $\mathsf{val}_R(D_i)$ is the conjunction $C_1^i \sqcap \cdots \sqcap C_n^i$ in the value restriction of role R, if any (otherwise $\mathsf{val}_R(D_i) = \top$);
- $\mathsf{ex}_R(D_i)$ is the set of concepts in the existential restrictions of the role R.

For any R, every sub-description in $\mathsf{ex}_R(D_i)$ and $\mathsf{val}_R(D_i)$ is in normal form.

In the following, let $\mathcal{L} = \mathcal{ALC}/_{\equiv}$ be the quotient set of $\mathcal{ALC}$ descriptions in normal form.

III. Dissimilarity Measures in Description Logics

We recall two definition of dissimilarity measures for $\mathcal{ALC}$ descriptions expressed in normal form [13], [14]. These measures are based on both the structure and the semantics of the concept descriptions.

A. Overlap Dissimilarity Measure

The first measure, is derived from a measure of the overlap between concepts, that can be defined as follows:

Definition 1 (overlap function): Let $\mathcal{I}$ be the canonical interpretation of the ABox $\mathcal{A}$. The *overlap function* f is defined[1]: $f : \mathcal{L} \times \mathcal{L} \mapsto R^+$ defined for descriptions $C, D \in \mathcal{L}$, with $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$

disjunctive level:

$$f(C,D) := f_{\sqcup}(C,D) = \begin{cases} \infty & C \equiv D \\ 0 & C \sqcap D \equiv \bot \\ \max_{i,j} f_{\sqcap}(C_i, D_j) & \text{otherwise} \end{cases}$$

conjunctive level:

$$f_{\sqcap}(C_i, D_j) := f_P(\mathsf{prim}(C_i), \mathsf{prim}(D_j)) + \\ + \lambda(f_{\forall}(C_i, D_j) + f_{\exists}(C_i, D_j))$$

with $\lambda \in [0,1]$.

primitive concepts:

$$f_P(P_1, P_2) := \frac{|R(P_1) \cup R(P_2)|}{|(R(P_1) \cup R(P_2)) \setminus (R(P_1) \cap R(P_2))|}$$

where $R(P) = \bigcap_{A \in P} A^{\mathcal{I}}$ and $f_P(P_1, P_2) = \infty$ when $R(P_1) = R(P_2)$.

value restrictions:

$$f_{\forall}(C_i, D_j) := \sum_{R \in N_R} f_{\sqcup}(\mathsf{val}_R(C_i), \mathsf{val}_R(D_j))$$

existential restrictions:

$$f_{\exists}(C_i, D_j) := \sum_{R \in N_R} \sum_{k=1}^{N} \max_{p=1,\ldots,M} f_{\sqcup}(C_i^k, D_j^p)$$

where $C_i^k \in \mathsf{ex}_R(C_i)$ and $D_j^p \in \mathsf{ex}_R(D_j)$ and we suppose w.l.o.g. that $N = |\mathsf{ex}_R(C_i)| \geq |\mathsf{ex}_R(D_j)| = M$, otherwise the indices N and M as well as C_i and D_j are to be exchanged in the formula above.

[1]The name $\mathcal{A}$ of the ABox is omitted for simplicity.

The function f represents a measure of the overlap between two descriptions expressed in $\mathcal{ALC}$ normal form. It measures the similarity of the input concepts based on the similarity between their extensions (approximated with the retrieval) and also on the similarity of the concepts reached by the role restrictions. Namely, It is defined recursively beginning from the top level of the descriptions (a disjunctive level) up to the bottom level represented by (conjunctions of) primitive concepts.

Overlap at the disjunctive level is treated as the maximum overlap between the disjunctive forms of the input concepts. At conjunctive levels, instead of simply considering the similarity like in Tversky's measure [20] (as we do for prim's), that could turn out to be null in case of disjoint retrieval sets for the input concepts, we distinguish the overlaps between the prims and those between the concepts in the scope of the various role restrictions[2]; the contribution of the overlap of concepts reached through role restrictions can be penalized by tweaking the parameter λ. The measure for primitive concepts resembles Tversky's measure and it represents a semantic baseline since it depends on the semantics of the knowledge base, as conveyed by the ABox assertions. This is in line with to the ideas in [16], [10], where semantics is elicited as a probability distribution over the domain of the interpretation.

As for the role restrictions overlap, for the universal restrictions we simply add the overlap measures varying the role, while for the existential restrictions the measure is trickier: borrowing the idea of the existential mappings we consider, per role, all possible matches between the concepts in the scope of existential restrictions, then we consider the maximal sum of overlaps resulting from all the matches.

Now, it is possible to derive a dissimilarity measure based on f as follows:

Definition 2 (overlap dissimilarity measure): The *overlap dissimilarity measure* is a function $d : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$ such that, given two concept descriptions $C, D \in \mathcal{L}$, $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$:

$$d(C, D) := \begin{cases} 1 & \text{if } f(C, D) = 0 \\ 0 & \text{if } f(C, D) = \infty \\ \frac{1}{f(C,D)} & \text{otherwise} \end{cases}$$

Function d simply measures the level of dissimilarity between two concepts as the inverse of the overlap function f. Particularly, if $f(C, D) = 0$, i.e. there is no overlap between the considered concepts, then d must indicate that the two concepts are totally different, indeed $d(C, D) = 1$, the maximum value of its range. If $f(C, D) = \infty$ this means that the two concepts are totally overlapped and consequently $d(C, D) = 0$ that means that the two concept are indistinguishable, indeed d assumes the minimum value of its range. If the considered concepts have a partial overlap then their dissimilarity is inversely proportional to their overlap, since in this case $f(C, D) > 1$ and consequently $0 < d(C, D) < 1$.

[2]We tried also other solutions, such as considering minima or products of the three overlap measures, yet experimentally this did not yield a better performance whereas the computation time was increased. For practical reasons, the maximal measure ∞ has been replaced with large numbers depending on the cardinality of the set of individuals in the ABox: $|\mathsf{Ind}(\mathcal{A})|$.

B. A Dissimilarity Measure Based on Information Content

As discussed in [12], a measure of concept (dis)similarity can be derived from the notion of *Information Content* (IC) that, in turn, depends on the probability of an individual to belong to a certain concept. Now, differently from other works, which assume that a probability distribution for the concepts in an ontology is known, we derive it from the knowledge base, from the distribution that can be estimated therein [16], [10].

In order to approximate this probability for a certain concept C, we recur to its extension w.r.t. the considered ABox in a fixed interpretation. Namely, we chose the *canonical interpretation* $\mathcal{I}_\mathcal{A}$, which is the one adopting the set of individuals mentioned in the ABox as its domain and the identity as its interpretation function [15]. Now, given a concept C its probability is estimated by:

$$pr(C) = |C^{\mathcal{I}_\mathcal{A}}|/|\Delta^{\mathcal{I}_\mathcal{A}}|$$

Finally, we can compute the information content of a concept, employing this probability:

$$IC(C) = -\log pr(C)$$

A measure of the concept dissimilarity is now formally defined [14]:

Definition 3 (IC dissimilarity): Let $\mathcal{A}$ be an ABox with canonical interpretation $\mathcal{I}$. The *information content dissimilarity measure* is a function $g : \mathcal{L} \times \mathcal{L} \mapsto R^+$ defined recursively for any two normal form concept descriptions $C, D \in \mathcal{L}$, with $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$

disjunctive level:

$$g(C, D) := g_\sqcup(C, D) = \begin{cases} 0 & \text{if } C \equiv D \\ \infty & \text{if } C \sqcap D = \bot \\ \min\limits_{\substack{1 \le i \le n \\ 1 \le j \le m}} g_\sqcap(C_i, D_j) & \text{otherwise} \end{cases}$$

conjunctive level:

$$g_\sqcap(C_i, D_j) := g_P(\mathsf{prim}(C_i), \mathsf{prim}(D_j)) + \\ +\lambda(g_\forall(C_i, D_j) + g_\exists(C_i, D_j))$$

with $\lambda \in [0, 1]$.

primitive concepts:

$$g_P(P_i, P_j) := \begin{cases} \infty & \text{if } P_i \sqcap P_j \equiv \bot \\ \frac{IC(P_i \sqcap P_j)+1}{IC(\mathsf{LCS}(P_i,P_j))+1} & \text{otherwise} \end{cases}$$

value restrictions:

$$g_\forall(C_i, D_j) := \sum_{R \in N_R} g_\sqcup(\mathsf{val}_R(C_i), \mathsf{val}_R(D_j))$$

existential restrictions:

$$g_\exists(C_i, D_j) := \sum_{R \in N_R} \sum_{k=1}^{N} \min_{p=1,\ldots,M} g_\sqcup(C_i^k, D_j^p)$$

where $C_i^k \in \mathsf{ex}_R(C_i)$ and $D_j^p \in \mathsf{ex}_R(D_j)$ and we suppose w.l.o.g. that $N = |\mathsf{ex}_R(C_i)| \geq |\mathsf{ex}_R(D_j)| = M$, otherwise the indices N and M are to be exchanged in the formula above.

The function g represents a measure of the dissimilarity between two descriptions expressed in $\mathcal{ALC}$ normal form. It is defined recursively beginning from the top level of the descriptions (a disjunctive level) up to the bottom level represented by (conjunctions of) primitive concepts.

Now g has values in $[0, \infty]$. It may be useful to derive a normalized dissimilarity measure as shown in the following.

Definition 4 (normalized IC dissimilarity): Let $\mathcal{A}$ be an ABox with canonical interpretation $\mathcal{I}$. The *normalized information content dissimilarity measure* is the function $d : \mathcal{L} \times \mathcal{L} \mapsto [0, 1]$, such that given the concept descriptions in $\mathcal{ALC}$ normal form $C = \bigsqcup_{i=1}^{n} C_i$ and $D = \bigsqcup_{j=1}^{m} D_j$, let

$$d(C, D) := \begin{cases} 0 & \text{if } g(C, D) = 0 \\ 1 & \text{if } g(C, D) = \infty \\ 1 - \frac{1}{g(C,D)} & \text{otherwise} \end{cases}$$

C. Measuring the Dissimilarity between Individuals

The notion of *Most Specific Concept* is commonly exploited for lifting individuals to the concept level.

Definition 5 (most specific concept): Given an ABox $\mathcal{A}$ and an individual a, the *most specific concept* of a w.r.t. $\mathcal{A}$ is the concept C, denoted $\mathsf{MSC}_{\mathcal{A}}(a)$, such that $\mathcal{A} \models C(a)$ and for any other concept D such that $\mathcal{A} \models D(a)$, it holds that $C \sqsubseteq D$.

In case of cyclic ABoxes expressed in a DL with existential restrictions the MSC may not be expressed by a finite description [15], yet it may be often approximated.

On performing experiments related to another similarity measure exclusively based on concept extensions [21], we noticed that, recurring to the MSC for lifting individuals to the concept level, just falls short: indeed the MSCs may be too specific and unable to include other (similar) individuals in their extensions. By comparing descriptions reduced to the normal form we have given a more structural definition of dissimilarity. However, since MSCs are computed from the same ABox assertions, reflecting the current knowledge state, this guarantees that structurally similar representations will be obtained for semantically similar concepts.

Let us recall that, given the ABox, it is possible to calculate the most specific concept of an individual a w.r.t. the ABox, $\mathsf{MSC}(a)$ or at least its approximation $\mathsf{MSC}^k(a)$ up to a certain description depth k. In some cases these are equivalent concepts but in general we have that $\mathsf{MSC}^k(a) \sqsupseteq \mathsf{MSC}(a)$.

Given two individuals a and b in the ABox, we consider $\mathsf{MSC}^k(a)$ and $\mathsf{MSC}^k(b)$ (supposed in normal form). Now, in order to assess the dissimilarity between the individuals, d measure can be applied to these descriptions:

$$d(a, b) := d(\mathsf{MSC}^k(a), \mathsf{MSC}^k(b))$$

This may turn out to be handy in several tasks, namely both in inductive reasoning (construction, repairing of knowledge bases) and in information retrieval.

D. Discussion

We proved in [13], [14] that these measures are really dissimilarity measures according to the formal definition [22], considering that their input (what is actually compared) is equivalence classes in $\mathcal{L}$, i.e. $\mathcal{ALC}$ concept descriptions with the same normal form.

As previously mentioned, we have also attempted slightly different measure definitions (e.g. considering minima or products at the conjunctive level that sound more intuitive) which experimentally did not prove more effective than the simple (additive) definition given above.

The computational complexity of the measures depends on the complexity of the required reasoning services for computing the concepts retrieval. Namely both subsumption and instance-checking are P-space for the $\mathcal{ALC}$ logic[18], yet such inference can be computed once and preliminarily, before the measures are computed for the method we will present in the following.

Obviously when computing the dissimilarity measures for cases involving individuals, the extra cost of computing MSCs (or their approximations) has to be added.

Nevertheless, in practical applications, these computations may be efficiently carried out exploiting the statistics that are maintained by the DBMSs query optimizers. Besides, the counts that are necessary for computing the concept extensions could be estimated by means of the probability distribution over the domain.

IV. A Nearest Neighbor Classification Procedure in Description Logics

We briefly review the basics of the k-Nearest Neighbor method (*k-NN*) and propose how to exploit the classification procedure for inductive reasoning. In this lazy approach to learning, a notion of distance (or dissimilarity) measure for the instance space is employed to classify a new instance.

Let x_q be the instance that requires a classification. Using a dissimilarity measure, the set of k nearest pre-classified instances is selected. The objective is to learn a discrete-valued target function $h : IS \mapsto V$ from a space of instances IS to a set of values $V = \{v_1, \ldots, v_s\}$. In its simplest setting, the k-NN algorithm approximates h for new instances x_q on the ground of the value that h assumes in the neighborhood of x_q, i.e. the k closest instances to the new instance in terms of a dissimilarity measure. Precisely, it is assigned according to the value which is *voted* by the majority of instances in the neighborhood. The the classification function $\hat{h}$ can be formally defined as follows:

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

where δ is a function that returns 1 in case of matching arguments and 0 otherwise. Note that the hypothesized function $\hat{h}$ is defined only extensionally, that is the k-NN method does not return an intensional classification model (e.g. a function or a

concept definition), it merely gives an answer for new query instances to be classified, employing the procedure above.

This simple formulation does not take into account the similarity among instances, except when selecting the instances to be included in the neighborhood. Therefore a modified setting is generally adopted, weighting the vote on the grounds of the similarity of the instance:

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

where, usually, $w_i = 1/d(x_i, x_q)$ or $w_i = 1/d(x_i, x_q)^2$.

Usually this method is employed to classify vectors of features in some n-dimensional instance space (e.g. often $IS = \mathbf{R}^n$). Let us now turn to adapt the method to the more complex context of DLs descriptions. Preliminarily, it should be observed that a strong assumption of this setting is that it can be employed to assign a value (e.g. a class) to a query instance among a set of values which can be regarded as a set of pairwise disjoint concepts/classes. This is an assumption that cannot be always valid. In this case, indeed, an individual could be an instance of more than one concept.

Let us consider a new value set $V = \{C_1, \ldots, C_s\}$ of concepts C_j that may be assigned to a query instance x_q. If they were to be considered as disjoint (like in the standard machine learning setting), the decision procedure would adopt the hypothesis function defined in Eq. (1), with the query instance assigned the *single* concept voted by the weighted majority of instances in the neighborhood.

In the general case considered in this paper, when the disjointness of the classes cannot be assumed (unless explicitly stated in the TBox), one can adopt another answering procedure, decomposing the multi-class problem into smaller binary classification problems (one per concept). Therefore, a simple binary value set ($V = \{-1, +1\}$) is to be employed. Then, for each single concept (say C_j), a hypothesis $\hat{h}_j$ is computed on the fly during the classification phase:

$$\hat{h}_j(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} \frac{\delta(v, h_j(x_i))}{d(x_q, x_i)^2} \quad \forall j \in \{1, \ldots, s\} \quad (2)$$

where each function h_j, simply indicates the occurrence $(+1)$ or absence (-1) of the corresponding assertion in the ABox for the k training instances x_i: $C_j(x_i) \in \mathcal{A}$. Alternately[3], h_j may return $+1$ when $C_j(x_i)$ is logically entailed by the knowledge base $\mathcal{K}$, and -1 otherwise.

The problem with non-explicitly disjoint concepts is also related to the *Closed World Assumption* usually made in the context of Information Retrieval and Machine Learning. That is the reason for adapting the standard setting to cope both with the case of non-disjoint concepts and with the OWA which is commonly made in the Semantic Web context. To deal with the OWA, the absence of information on whether a certain instance x belongs to the extension of concept C_j should not be interpreted negatively; rather, it should count as neutral information. Thus, one can still adopt the decision procedure in Eq. (2), however another value set has to be adopted for the h_j's, namely $V = \{-1, 0, +1\}$, where the three values denote, respectively, positive occurrence, absence and negative occurrence (positive for the concept negation). Formally:

$$h_j(x) = \begin{cases} +1 & C_j(x) \in \mathcal{A} \\ 0 & C_j(x) \notin \mathcal{A} \wedge \neg C_j(x) \notin \mathcal{A} \\ -1 & \neg C_j(x) \in \mathcal{A} \end{cases}$$

Again, a more complex procedure may be devised by simply substituting the notion of occurrence (absence) of assertions in (from) the ABox with one based on logic entailment (denoted with $\vdash$) from the knowledge base, i.e. $\mathcal{K} \vdash C_j(x)$, $\mathcal{K} \nvdash C_j(x)$ nor $\mathcal{K} \nvdash \neg C_j(x)$ and $\mathcal{K} \vdash \neg C_j(x)$, respectively. Although this may help reaching the precision of deductive reasoning, it is also much more computationally expensive, since the simple lookup in the ABox must be replaced with instance checking.

From a computational viewpoint this procedure could be implemented to provide an answer even more efficiently than with a standard deductive reasoner. Indeed, once the retrieval of the primitive concepts is computed, the dissimilarity measures can be easily computed by means of a dynamic programming algorithm. Besides, the various measures could be maintained in an ad hoc data structure which would allow for an efficient retrieval of the nearest neighbors, such as the kD-trees or ball trees [23]

V. Experiments

A. Experimental Setting

In order to assess the validity of the presented method with the dissimilarity measures proposed in Sect. III, we have applied it to the instance classification problem, with four different ontologies represented in OWL: FSM, SURFACE-WATER-MODEL from the Protégé library[4], the FINANCIAL ontology[5] employed as a testbed for the PELLET reasoner and a small FAMILY ontology written in our lab. Although they are represented in languages that are different from $\mathcal{ALC}$, we simply discarded these details when constructing the MSC approximations to be able to apply the presented measures.

FAMILY is an $\mathcal{ALCF}$ ontology describing *kinship* relationships. It is made up of 14 concepts (both primitive and defined), some of them are declared to be disjoint, 5 object properties, 39 distinct individual names. Most of the individuals are asserted to be instances of more than one concept, and are involved in more than one role assertions. This ontology has been written to have a small yet more complex case with respect to the following ones. Indeed, while the other ontologies are more regular, i.e. only some concepts are employed in the assertions (the others are defined only intensionally), in the FAMILY ontology every concept

[3]For the sake of simplicity and efficiency, this case will not be considered in the following.

[4]See the webpage: `http://protege.stanford.edu/plugins/owl/owl-library`

[5]See the webpage: `http://www.cs.put.poznan.pl/alawrynowicz/financial.owl`

has at least one instance asserted. The same happens for the assertions on roles; particularly, there are some cases where role assertions constitute a chain from an individual to another one, by means of other intermediate assertions.

The FSM ontology describes the domain of *finite state machines* using the $\mathcal{SOF}(D)$ language. It is made up of 20 (both 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).

SURFACE-WATER-MODEL is an $\mathcal{ALCOF}(D)$ ontology describing the domain of the surface water and the water quality models. It is based on the *Surface-water Models Information Clearinghouse* (SMIC) of the USGS. Namely, it is an ontology of numerical models for surface water flow and water quality simulation. The application domain of these models comprises lakes, oceans, estuaries etc.. These models are classified based on their availability, application domain, dimensions, partial differential equation solver, and characteristics types. It is made up of 19 concepts (both primitive and defined) without any specification about 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.

FINANCIAL is an $\mathcal{ALCIF}$ ontology that describes the domain of eBanking. It is made up of 60 (both 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.

The classification method was applied to all the individuals in each ontology; namely, the individuals were checked to assess if they were instances of the concepts in the ontology through the analogical method. The performance was evaluated comparing its responses to those returned by a standard reasoner[6] as a baseline.

Specifically, for each individual in the ontology the MSC is computed and enlisted in the set of training (or test) examples. Each example is classified applying the adapted k-NN method presented in the previous section. As a value of k we chose $\sqrt{|\mathsf{Ind}(\mathcal{A})|}$, as advised in the instance-based learning literature.

The experiment has been repeated twice adopting a leave-one-out cross validation procedure with both the dissimilarity measures defined in Section III.

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 vice-versa
- *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.

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

TABLE I

AVERAGE RESULTS OF THE EXPERIMENTS WITH THE METHOD EMPLOYING THE MEASURE BASED ON OVERLAP.

	Match Rate	Commission Rate	Omission Rate	Induction Rate
FAMILY	.654±.174	.000±.000	.231±.173	.115±.107
FSM	.974±.044	.026±.044	.000±.000	.000±.000
S.-W.-M.	.820±.241	.000±.000	.064±.111	.116±.246
FINANCIAL	.807±.091	.024±.076	.000±.001	.169±.076

B. Experiments Employing the Overlap Measure

By looking at Tab. I reporting the experimental outcomes with the dissimilarity measure based on the overlap (see Def. 2), preliminarily it is important to note that, for every ontology, the commission error was quite low. This means that the classifier 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.

In particular, by looking at the outcomes related to the FAMILY ontology, it can be observed that the match rate is the lowest while the highest rate of omission errors was reported. This may be due to two facts: 1) very few individuals were available w.r.t. the number of concepts[7]; 2) sparse data situation: instances are irregularly *spread* over the concepts, that is they might be instances of different concepts, which are sometimes disjoint. Hence the MSC approximations that were computed also resulted very different one from another, which reduces the possibility of significantly matching similar MSCs. This is a known drawback of the Nearest-Neighbor methods. Nevertheless, as mentioned above, it is important to note that the algorithm did not make any commission error and it is able to infer new knowledge (11%).

As regards the FSM ontology, we have observed the maximum match rate with respect to the classification given by the logic reasoner. Moreover, differently from the other ontologies, both the omission error rate and induction rate were null. A very limited percentage of incorrect classification cases was observed. These outcomes were probably due to the fact that individuals in this ontology are quite regularly divided by the assertions on concepts and roles, so computing their MSCs, these are all very similar to each other and so the amount

[7]Instance-based methods make an intensive use of the information about the individuals and improve their performance with the increase of the number of instances considered.

TABLE II

AVERAGE RESULTS OF THE EXPERIMENTS WITH THE METHOD EMPLOYING THE MEASURE BASED ON INFORMATION CONTENT.

	Match Rate	Commission Rate	Omission Rate	Induction Rate
FAMILY	.608±.230	.000±.000	.330±.216	.062±.217
FSM	.899±.178	.096±.179	.000±.000	.005±.024
S.-W.-M.	.820±.241	.000±.000	.064±.111	.116±.246
FINANCIAL	.807±.091	.024±.076	.000±.001	.169±.046

of information they convey is very low. A choice of a lower number k of neighbors could probably help committing those residual errors.

For the same reasons, also for the SURFACE-WATER-MODEL ontology quite a high rate of matching classifications was reported (yet less then with the previous ontology); moreover, some cases of omission error (6%) were observed. The induction rate was about 12% which means that for this ontology our classifier always assigned individuals to the correct concepts but, in some cases, it could also induce new assertions. Since this rate represents assertions that were not logically deducible from the ontology and yet they were inferred inductively by the analogical classifier, these figures would be a positive outcome (provided this knowledge were deemed as correct by an expert). Particularly, in this case the increase of the induction rate has been due to the presence of assertions of mutual disjointness for some of the concepts.

Results are no different also for the case of the experiments with FINANCIAL ontology that largely exceeds the others in terms of number of concepts and individuals. Namely, the observed match rate is again above the 80% and the rest of the cases are comprised in the induction rate (17%), leaving a limited margin to residual errors. This corroborates a fact about the NN learners, that is their reaching better and better performance in the limit, as long as new training instances become available. Actually, performing a 10-fold cross validation we obtained almost the same results.

C. Experiments with the Information Content Measure

The average results obtained by adopting the procedure with the measure based on information content (see Def. 4) are reported in Table II.

By analyzing this table it is possible to note that no sensible variation was observed in the classifications performed using the first dissimilarity measure. Particularly, with both measures the method correctly classified all the individuals, without commission errors. The reason is that, in most of the cases, the individuals of these ontologies are instances of one concept only and they are involved in a few roles (object properties). Some of the figures are slightly lower than those observed in the other experiment: this is confirmed by a higher variability.

Surprisingly, the results on the larger ontologies (S.-W.-M. and FINANCIAL) perfectly match those obtained with the other measure which is probably due to the fact that we used a leave-one-out cross-validation mode which yielded a high value for the number k of training instances for the neighborhood and it is well known that the NN procedure becomes more and more precise as more instances can be considered. The price to be paid was a longer computation time.

VI. CONCLUSIONS AND FUTURE WORK

In this work we have coupled with a method for inductive inference on ABoxes with two composite concept similarity measures. The method is based on the classification in analogy with the majority of neighbor training individuals. It can be naturally exploited for predicting/suggesting missing information about individuals thus enabling a sort of inductive retrieval. For example, it may be employed a query to the KB may be issued. Specifically we are targeting also the task of service discovery when for semantically annotated web service. Even more so the outcomes of our method may be decisive for enabling a series of other inductive tasks such as clustering, case-based reasoning, , etc..

The proposed method is able to induce new assertions, in addition to the knowledge already derivable by means of a reasoner. Then it seems to be able to enhance the standard instance-based classification. Particularly, an increase in accuracy was observed when the instances increase in number and are homogeneously spread.

The presented measure can be refined tweaking the weighting factor λ for decreasing the impact of the dissimilarity between nested sub-concepts in the descriptions on the determination of the overall value.

Lately, we have defined another measure for $\mathcal{ALN}$ [24]. Hence a natural extension may concern the definition of dissimilarity measures in more expressive DLs languages. For example, a normal form for $\mathcal{ALCN}$ can be obtained based on those for $\mathcal{ALN}$ and $\mathcal{ALCN}$. Then, by exploiting the notion of existential mappings [25], the measure presented in this paper may be extended to more expressive DLs. We are currently developing other kind of semantic similarities based on the idea of the hypothesis-driven distance.

Besides, as mentioned, this method could be extended with different (yet still computationally tractable) answering procedures grounded on statistical inference (non-parametric tests based on ranked distances), in order to accept answers as correct with a high degree of confidence. Furthermore, the k-NN method in its classical form is particularly suitable for the automated induction of missing values for (scalar or numeric) datatype properties of an individual as an estimate derived from the values of the datatypes for the surrounding individuals.

Kernels are another means to express a notion of similarity is some unknown feature space. We are working at the definition of kernel functions on DLs representations [26], thus allowing the exploitation of kernel methods efficiency (e.g. the support vector machines) in a multi-relational setting.

ACKNOWLEDGMENTS

This work was partially supported by the regional interest projects DIPIS (*DIstributed Production as Innovative System*) and DDTA (*Distretto Digitale Tessile Abbigliamento*) in the

context of the tasks related to the semantic web services discovery.

REFERENCES

[1] W. Cohen and H. Hirsh, "Learning the CLASSIC description logic," in *Proceedings of the 4th International Conference on the Principles of Knowledge Representation and Reasoning*, P. Torasso, J. Doyle, and E. Sandewall, Eds. Morgan Kaufmann, 1994, pp. 121–133.

[2] J.-U. Kietz and K. Morik, "A polynomial approach to the constructive induction of structural knowledge," *Machine Learning*, vol. 14, no. 2, pp. 193–218, 1994.

[3] S. Staab, A. Maedche, C. Nedellec, and P. Wiemer-Hastings, Eds., *ECAI2000 Workshop on Ontology Learning*, ser. CEUR WS Proceedings, 2000, vol. 31.

[4] S. Staab, A. Maedche, C. Nedellec, and E. Hovy, Eds., *IJCAI2001 Workshop on Ontology Learning*, 2001.

[5] F. Esposito, N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro., "Knowledge-intensive induction of terminologies from metadata," in *ISWC2004, Proceedings of the 3rd International Semantic Web Conference*, ser. LNCS, F. van Harmelen, S. McIlraith, and D. Plexousakis, Eds., vol. 3298. Springer, 2004, pp. 441–455.

[6] M. d'Aquin, J. Lieber, and A. Napoli, "Decentralized case-based reasoning for the Semantic Web," in *Proceedings of the 4th International Semantic Web Conference, ISWC2005*, ser. LNCS, Y. Gil, V. Motta, E. Benjamins, and M. A. Musen, Eds., no. 3279. Springer, 2005, pp. 142–155.

[7] P. Hitzler and D. Vrandečić, "Resolution-based approximate reasoning for OWL DL," in *Proceedings of the 4th International Semantic Web Conference, ISWC2005*, ser. LNCS, Y. Gil, V. Motta, E. Benjamins, and M. A. Musen, Eds., no. 3279. Springer, 2005, pp. 383–397.

[8] T. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[9] M. Rodríguez, "Assessing semantic similarity between spatial entity classes," Ph.D. dissertation, University of Maine, 1997.

[10] A. Borgida, T. Walsh, and H. Hirsh, "Towards measuring similarity in description logics," in *Working Notes of the International Description Logics Workshop*, ser. CEUR Workshop Proceedings, Edinburgh, UK, 2005.

[11] M. W. Bright, A. R. Hurson, and S. H. Pakzad, "Automated resolution of semantic heterogeneity in multidatabases." *ACM Transaction on Database Systems*, vol. 19, no. 2, pp. 212–253, 1994.

[12] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999.

[13] C. d'Amato, N. Fanizzi, and F. Esposito, "A dissimilarity measure for the $\mathcal{ALC}$ description logic," in *Semantic Web Applications and Perspectives, 2nd Italian Semantic Web Workshop SWAP2005*, P. Bouquet and G. Tummarello, Eds., vol. 166. Trento, Italy: CEUR, 2005.

[14] ——, "A dissimilarity measure for $\mathcal{ALC}$ concept descriptions," in *Proceedings of the 21st Annual ACM Symposium of Applied Computing, SAC2006*, H. Haddad, Ed. Dijon, France: ACM, 2006, pp. 1695–1699.

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

[16] F. Bacchus, "Lp, a logic for representing and reasoning with statistical knowledge," *Computational Intelligence*, vol. 6, pp. 209–231, 1990.

[17] M. Schmidt-Schauß and G. Smolka, "Attributive concept descriptions with complements," *Artificial Intelligence*, vol. 48, no. 1, pp. 1–26, 1991.

[18] F. M. Donini, M. Lenzerini, D. Nardi, and A. Schaerf, "Deduction in concept languages: From subsumption to instance checking," *Journal of Logic and Computation*, vol. 4, no. 4, pp. 423–452, 1994. [Online]. Available: citeseer.ist.psu.edu/donini94deduction.html

[19] S. Brandt, R. Küsters, and A.-Y. Turhan, "Approximation and difference in description logics," in *Proceedings of the International Conference on Knowledge Representation*, D. Fensel, F. Giunchiglia, D. McGuinness, and M.-A. Williams, Eds. Morgan Kaufmann, 2002, pp. 203–214.

[20] A. Tversky, "Features od similarity," *Psycological Review*, vol. 84, no. 4, pp. 327–352, 1997.

[21] C. d'Amato, N. Fanizzi, and F. Esposito, "A semantic similarity measure for expressive description logics," in *Proceedings of Convegno Italiano di Logica Computazionale, CILC05*, A. Pettorossi, Ed., Rome, Italy, 2005.

[22] H. Bock., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, 1999.

[23] I. H. Witten and E. Frank, *Data Mining*, 2nd ed. Morgan Kaufmann, 2005.

[24] N. Fanizzi and C. d'Amato, "A similarity measure for the $\mathcal{ALN}$ description logic," in *Proceedings of Convegno Italiano di Logica Computazionale, CILC06*, Bari, Italy, 2006, http://cilc2006.di.uniba.it/download/camera/15_Fanizzi_CILC06.pdf.

[25] R. Kusters and R. Molitor, "Computing least common subsumers in $\mathcal{ALEN}$," in *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI2001*, B. Nebel, Ed., 2001, pp. 219–224.

[26] N. Fanizzi and C. d'Amato, "A declarative kernel for $\mathcal{ALC}$ concept descriptions." in *Proceedings of the 16th International Symposium on Methodologies for Intelligent Systems, ISMIS 2006*, ser. LNAI, F. Esposito, Z. Ras, D. Malerba, and G. Semeraro, Eds., vol. 4203. Springer, 2006, pp. 322–331.