# Instance-based Retrieval by Analogy

Nicola Fanizzi
Dipartimento di Informatica
Università degli Studi di Bari
Bari - Italy
fanizzi@di.uniba.it

Claudia d'Amato
Dipartimento di Informatica
Università degli Studi di Bari
Bari - Italy
claudia.damato@di.uniba.it

Floriana Esposito
Dipartimento di Informatica
Università degli Studi di Bari
Bari - Italy
esposito@di.uniba.it

## ABSTRACT

This work presents a method for retrieval in knowledge bases expressed in Description Logics, founded in the *instance-based learning.* The procedure implements the *disjunctive version space* approach exploiting a notion of semantic difference. The method can be employed both to answer to class-membership queries, even though the answers are not logically entailed by the knowledge base, e.g. there are some inconsistent assertions due to heterogeneous sources. In addition, it may also predict/suggest new assertions The method has been implemented and tested in an experimentation, where we show that it is sound and effective.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

## Keywords

Nearest Neighbor, Disjunctive Version Space, Description Logics, Analogy

## 1. INTRODUCTION AND MOTIVATION

Many important tasks that are likely to be provided by new generation knowledge-based systems, such as classification, construction, revision, population are likely supported by inductive methods. In order to support these tasks and overcome the inherent complexity of classic logic-based inference other forms of reasoning are being investigated, both deductive, such as *non-monotonic*, *paraconsistent* [8], *approximate* reasoning (see the discussion in [9]), *case-based reasoning* [5] and inductive-analogical forms such as inductive *generalization* [3] and *specialization* [7].

All these approaches aim at noise-tolerant and efficient forms of reasoning. From this viewpoint, instance-based inductive methods [6] are particularly well suited. Indeed, they are known to be both very efficient and fault-tolerant compared to the classic logic-based methods, being noise always a danger in contexts where knowledge is distributed and acquired from heterogeneous sources. Two kinds of noise may be identified. The first kind may be introduced by inconsistency in the knowledge base: some of the methods aforementioned methods are able deal with this problem by means of approximate reasoning or by spotting and repairing specific parts of the knowledge bases. A second kind of noise is due to incorrect knowledge that does not strictly cause inconsistency, nevertheless it may yield incomplete/inconsistent conclusions with respect to the intended meaning of the concepts in considered domain.

Instance-based algorithms, which can be suitable for these cases, have been mainly applied to attribute-value representations. Upgrading the algorithms to work on multi-relational representations [6], namely on the concept languages used in the Semantic Web, founded in Description Logics (DLs) [1] (see Sect. 2), requires specific adjustments.

An instance-based framework for DLs was devised (see Sect. 3) for exploiting a dissimilarity measure to derive inductively (by analogy) both consistent consequences from the knowledge base and also new assertions which were not previously logically derivable. In turn, this enables also other related reasoning services such as classification, retrieval and clustering. Particularly, classification can be performed even in absence of a definition for the target concept in the knowledge base by analogy with a set of training assertions on such a concept (provided by an expert).

Specifically, we elaborate on classification procedures based on *lazy learning*, namely a relational form of the well-known *Nearest Neighbor* (NN) approach [10]. The baseline idea is that similar individuals, by analogy, should likely belong to similar concepts. The adaptation to the context of DLs concept languages 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 data mining settings where the *Closed World Assumption* (CWA) is the standard. Besides, in the standard NN multi-class setting, different classes are often assumed to be disjoint, which is not typical in the context of the Semantic Web.

These ideas have been further pushed forward (see Sect. 4) by considering another form of instance-based learning: the disjunctive version space approach, adapted to a DL framework. In this setting, the neighborhood of an individual w.r.t. a target concept is determined on the grounds of its similarity to other training individuals in the knowledge base which are known to belong to that concept. Instead of

*SAC'07* March 11-15, 2007, Seoul, Korea

using a similarity measure, like in the basic NN approach, the notion of neighborhood is based on class-membership queries performed on a training set of individuals: An individual is said to belong to the neighborhood of a positive example, when it belongs to the conjunction of the concepts that differentiate that instance from each negative example. In turn each such a concept can be regarded as a disjunction of features that separate a positive from a negative example. Thus belonging to the neighborhood of positive instances of a target concept gives a criterion to decide on the membership of an individual. The procedure is not crisp, since a number of mistakes can be tolerated, blaming them to the noise in the data. The method has been implemented so that some preliminary experimental results with real ontologies can be presented (Sect. 5).

## 2. REPRESENTATION AND INFERENCE SERVICES

The basics of $\mathcal{ALC}$ and inference in DL are briefly recalled. This logic adopts constructors supported by the standard Web ontology languages (see the DL handbook [1] for a thorough reference). Actually, the methods presented in the next sections may be made less language dependent through suitable approximations.

In DLs, concept descriptions are defined in terms of a set $N_C$ of *primitive concept* names and a set $N_R$ of *primitive roles*. The semantics of the concept 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 $\Delta^{\mathcal{I}}$, 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 concept description $C$, denoted $\neg C$, it amounts to $\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$. The *conjunction* of concepts, denoted with $C_1 \sqcap C_2$, yields an extension $C_1^{\mathcal{I}} \cap C_2^{\mathcal{I}}$ and, dually, concept *disjunction*, denoted with $C_1 \sqcup C_2$, yields $C_1^{\mathcal{I}} \cup C_2^{\mathcal{I}}$. Finally, there are two restrictions on roles: the *existential restriction*, denoted with $\exists R.C$, and interpreted as the set $\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}} : (x, y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$ and the *value restriction*, denoted with $\forall R.C$, whose extension is $\{x \in \Delta^{\mathcal{I}} \mid \forall y \in \Delta^{\mathcal{I}} : (x, y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}$.

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[1] $C \equiv D$, meaning $C^{\mathcal{I}} = D^{\mathcal{I}}$, where $C$ is atomic (the concept name) and $D$ is an arbitrarily complex description defined as above. $\mathcal{A}$ contains assertions on the world state, e.g. $C(a)$ and $R(a, b)$, meaning that $a^{\mathcal{I}} \in C^{\mathcal{I}}$ and $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$. Moreover, normally the *unique names assumption* is made on the ABox individuals. These are denoted with $\mathsf{Ind}(\mathcal{A})$.

In this context the most common inference is the semantic notion of *subsumption* between concepts:

Definition 2.1. *Given two concept descriptions $C$ and $D$, $D$* subsumes *$C$, denoted by $C \sqsubseteq D$, iff for every interpretation $\mathcal{I}$ it holds that $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. When $C \sqsubseteq D$ and $D \sqsubseteq C$, they are* equivalent*, denoted with $C \equiv D$.*

Another important inference for reasoning with individuals requires finding the concepts which an individual belongs to, namely, the most specific one:

Definition 2.2. *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 $C \sqsubseteq D$.*

Unfortunately, for many non-trivial DL languages, such as $\mathcal{ALC}$, the exact $\mathsf{MSC}$ may not be always expressed with a finite acyclic description [1] interpreted with the descriptive semantics presented earlier, yet it may be approximated [3, 2], which is satisfactory for inductive approaches.

## 3. A NEAREST NEIGHBOR PROCEDURE IN DL

The basics of the Nearest Neighbor approach [10] are recalled showing how to exploit a classification procedure for inductive reasoning and retrieval.

In this lazy-learning approach the learning phase is reduced to memorizing training instances of the target concepts pre-classified by an expert. Then, during the classification phase, a notion of similarity for the instance space is employed to classify a new instance in analogy with its neighbor. Given an ontology, a classification method can be employed for assigning a individual with the concepts it is likely to belong.

Let $x_q$ be the instance that must be classified. Using a similarity measure (or any other distance function), the set of the $k$ nearest pre-classified instances w.r.t. $x_q$ is selected. The objective is to learn an estimate of a hypothesis function for the target concept membership $h : \mathsf{TI} \mapsto V$ from a space of training instances $\mathsf{TI}$ to a set of values $V = \{v_1, \ldots, v_s\}$ standing for the classes to be assigned. In its simplest setting, the algorithm approximates $h$ for $x_q$ on the ground of the value that $h$ assumes for the training instances in the neighborhood of $x_q$, i.e. the $k$ closest instances to the new instance in terms of a dissimilarity measure. Precisely, this instance is assigned a class according to the value which is *voted* by the majority of instances in the neighborhood. This setting takes into account similarity only when selecting the instances to be included in a neighborhood.

A more general setting is based on weighting the vote according to the distance of the query instance from the training instances:

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

where $\hat{h}$ is the estimated hypothesis function, $\delta$ is the *Kronecker symbol*, a function that returns 1 in case of matching arguments and 0 otherwise, and, given a distance measure $d$, $w_i = 1/d(x_i, x_q)$ or $w_i = 1/d(x_i, x_q)^2$. In the case of the $\mathcal{ALC}$ DL, the similarity measure in [4] could be employed.

Note that the hypothesis function $\hat{h}$ is defined only extensionally, therefore the $k$-NN method does not return an intensional classification model (a function or a concept definition), it merely gives an answer for new query instances to be classified, employing the procedure mentioned above.

It should be observed that a strong assumption made in this setting is that it can be employed to assign the query instance to the class from a set of values which can be regarded as a set of pairwise disjoint concepts. This is a simplifying assumption that cannot be always valid. In our setting, indeed, an individual could be an instance of more than one

[1] The cases of general axioms or cyclic definitions will not considered here.

concept. Let us consider a value set $V = \{C_1, \ldots, C_s\}$, of possibly overlapping concepts $C_j$ $(1 \leq j \leq s)$ that may be assigned to a query instance $x_q$. If the classes were disjoint as in the standard setting, the decision procedure defining the hypothesis function is the same as in Eq. (1), with the query instance assigned the *single* class of the majority of instances in the neighborhood. In the general case, when the pairwise disjointness of the concepts cannot be assumed, one can adopt another classification procedure, decomposing the multi-class problem into smaller binary classification problems (one per target concept).

The problem with non-explicitly disjoint concepts is also related to the CWA usually made in the knowledge discovery context. That is the reason for adapting the standard setting to cope both with the case of generally non-disjoint classes 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 training instance $x$ belongs to the extension of concept $C_j$ should not be interpreted negatively, as shown before. Rather, it should count as neutral information. Thus, a ternary value set has to be adopted for the $h_j$'s, namely $V = \{-1, 0, +1\}$, where the values denote, respectively, membership[2], non-membership and absence of information:

$$h_j(x) = \begin{cases} +1 & \mathcal{K} \vdash C_j(x) \\ -1 & \mathcal{K} \vdash \neg C_j(x) \\ \ \ 0 & o.w. \end{cases}$$

The checks could have been pre-computed for the KB, 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 assertions in the ABox), therefore it may be able to give an answer, requiring only the correctness of the training instances classification.

## 4. APPLYING A DISJUNCTIVE VERSION SPACE APPROACH TO DL

Another analogy-based method for retrieval can be derived from Sebag's notion of *Disjunctive Version Space* [11]. Differently from the NN approach based on distances presented in the previous section, the population of an individual's neighborhood is performed inducing definitions for the query concept on the grounds of its examples and counterexamples available in the knowledge base.

Given a query concept $C_j$ $(1 \leq j \leq s)$, for each training instance $x \in \mathsf{TI}$ such that $C_j(x)$ holds (positive example for $C_j$), a hypothesis $H_j^x$ may be generated (*on-the-fly*) by considering the subset of counterexamples for $C_j$, denoted $\overline{\mathsf{E}}_j = \{\overline{x} \in \mathsf{TI} \mid \mathcal{K} \vdash \neg C_j(\overline{x})\} \subseteq \mathsf{TI}$ and finding a maximally discriminating description $D(x, \overline{x}) \in \mathcal{L}$ for each $\overline{x} \in \overline{\mathsf{E}}_j$. Namely, the hypothesis $H_j^x$ may be regarded as the conjunction of such $D(x, \overline{x})$'s, to be induced varying the counterexamples $\overline{x}$ in $\overline{\mathsf{E}}_j$:

$$H_j^x = \bigsqcap_{\overline{x} \in \overline{\mathsf{E}}_j} D(x, \overline{x})$$

In order to produce a(n approximation of) description $D(x, \overline{x})$, one possibility is considering the difference

$$D(x, \overline{x}) := \mathsf{MSC}^p(x) - \mathsf{MSC}^p(\overline{x})$$

where $p$ is a fixed depth that may depend on the ABox depth (see the final remark in this section). and the symbol $-$ denotes Teege's *difference operator* for DL descriptions [12]. In the case of $\mathcal{ALC}$, we have:

$$D(x, \overline{x}) := D_x \sqcup \neg D_{\overline{x}}$$

where $D_x = \mathsf{MSC}^p(x)$ and $D_{\overline{x}} = \mathsf{MSC}^p(\overline{x})$.

Now, for each training individual $x$, the individual under classification $x_q$ will belong to $x$'s neighborhood w.r.t. $C_j$ iff it belongs to the related hypothesis $H_j^x$. The *neighbor instance set* of $x_q$ w.r.t. $C_j$ is defined as follows

$$N_j(x_q) := \{x \in \mathsf{TI} \mid \mathcal{K} \vdash H_j^x(x_q)\}$$

The classification procedure can be defined again as a majority vote for the classes $V = \{-1, 0, +1\}$:

$$\hat{h}_j(x_q) := \operatorname*{argmax}_{v \in V} \sum_{x \in N_j(x_q)} \delta(v, H_j^x)$$

Yet, in this case the procedure may be biased by the different numbers of training instances in $N_j(x_q)$ voting for the negative or neutral cases. Hence, we rather consider the proportions of votes over the total number of training individuals classified with the three values of $V$:

$$\hat{h}_j(x_q) := \operatorname*{argmax}_{v \in V} \sum_{x \in N_j(x_q)} w_j^v \cdot \delta(v, H_j^x)$$

where the weighting factor $w_j^v = \#(v, N_j(x_q))/\#(v, \mathsf{TI})$ denotes the count of neighbor instances w.r.t. $x_q$ voting for value $v$ for concept $C_j$ over the total number of training individuals belonging to the same class.

***Discussion.*** As suggested in [11], the procedure can be parametrized on precision and recall. Indeed it may become more noise-tolerant by admitting an amount of consistency errors (say $\varepsilon$) in deciding whether a training instance belongs to the neighborhood: there may be up to $\varepsilon \cdot |\overline{\mathsf{E}}_j|$ cases, i.e. a number of counterexamples $\overline{x} \in \overline{\mathsf{E}}_j$, for which $x_q \notin D(x, \overline{x})$ and yet the membership to a neighborhood will be assumed as acceptable $(x \in N_j(x_q))$.

Besides, the method can be tuned also w.r.t. the completeness, by adjusting the specificity of $D(\cdot, \cdot)$ according to a number of features to be considered (say $M$) as a separation between positive and negative instances.

Better and language-independent definitions of $D(\cdot, \cdot)$ can be considered, that may be based only on the available assertions, as the algorithm only requires to know whether a new individual belongs to the neighborhood or not and this can be specified also with no involvement of the concept level.

Namely, in order to extend the applicability to more expressive languages than $\mathcal{ALC}$, an alternate way for building the discriminating definitions $D(\cdot, \cdot)$. The method adopted here is suitable for logics endowed with a notion of difference and a further approximation had to be made on the construction of the $\mathsf{MSC}$'s. Nevertheless, the algorithm is not extremely language-dependent: Any other method that can induce concept descriptions which are able to explain a positive instance and rule out a single negative one would be acceptable.

[2] Here $\vdash$ indicates the instance checking service to be provided by a reasoner. This proof-theoretic interpretation could be replaced by weaker and/or more efficiently computable procedures, such as a paraconsistent derivation

## 5. EXPERIMENTS

We present the outcomes of experiments carried out for testing the feasibility of the method illustrated in the previous section. Its implementation was tested on answering queries w.r.t. four ontologies drawn from the Protégé library[3], endowed with an different numbers of individuals, namely: the FSM, SURFACE-WATER-MODEL, SCIENCE, and NEWTESTAMENTNAMES. Some are expressed in larger DLs than $\mathcal{ALC}$. This affected the construction of the MSC's approximations, which turned out to be more general than those that could be produced in the original DLs.

FSM is an $\mathcal{SF}(D)$ KB describing finite state machines. It is made up of 20 concepts, 10 object properties, 7 datatype properties, 37 individual names. SURFACE-WATER-MODEL is an $\mathcal{ALO}(D)$ ontology describing water quality models. It is based on the *Surface-water Models Information Clearinghouse* of the US Geological Survey. It deals with numerical models for surface water flow and water quality simulation. These models are classified according to their availability, domain, dimensions, and characteristic types. It is made up of 19 concepts, 9 object properties, 115 individual names. SCIENCE is an $\mathcal{ALCIF}(D)$ KB and describes scientific facts. It is made up of 74 concepts, 70 object properties, 331 individual names. The NEWTESTAMENTNAMES is an $\mathcal{SHIN}(D)$ KB which describes facts related to the New Testament (*Semantic Bible* Project). It is made up of 47 concepts, 27 object properties, 676 individual names.

The classification method presented in the previous section was applied to each test ontology, by generating 15 random queries based on the concepts and roles therein; each query is a complex concept made up of a variable random number (from 2 up to 11) of (primitive and defined) concepts found in the knowledge base.

A naïve retrieval procedure required, for each test query, every individual is considered to determine if it belongs to the answerer set (+1) or not (-1), or it is neutral 0, i.e. unknown answer) w.r.t. the ontology. Specifically, for each training individual an MSC approximation was precomputed and assigned to the set of examples or counterexamples w.r.t. the query concept. Each test individual is then classified applying the method presented in the previous section. For the smaller knowledge bases leave-one-out cross validation procedure was used, while for the larger ones a 10-fold cross validation was performed. We intended to assess whether our method is able to retrieve instances correctly, i.e. its performance was compared to the relevance determined by an expert, whose role was made by a reasoner[4]. Additionally, it should also be able to induce by analogy new (previously unknown) class-membership assertions that cannot be logically inferred.

Particularly, for each ontology and for each concept, four rates have been computed: *match rate*, *omission error rate*, *commission error rate*, *induction rate*. The match rate is the proportion of instances retrieved exactly as a reasoner would do. The omission error is related to completeness. It measures the amount of relevant individuals w.r.t. a certain query (i.e. the answer is ±1) that were not retrieved (answer 0). The commission error is related to soundness. It measures the amount of individuals whose relevance was mismatched, i.e. they were retrieved when they belonged to the negation of the query concept or vice-versa. The induction rate measures the amount of individuals found as relevant (answer ±1) even though the expert cannot give an answer (i.e. the reasoner returns `unknown`). Thus, commission error may be more harmful than omission error. A high induction rate means that the procedure was actually able to suggest new assertions that are likely to be valid and can be validated by a knowledge engineer.

Tab. 1 reports the experimental results. Per each ontology, we report the average rates for the measures discussed above and also the interval of values assumed for the 15 random queries during the cross validation experiments.

Primarily, by looking at table, it is important to note that, for every ontology, the commission error was null on average and the variance is also quite low. This means that the classifier has never made critical mistakes because no individual has been deemed as an instance of a concept while really it is an instance an disjoint class. Also the omission error rate is almost null. The highest value was observed on the FSM ontology, likely caused by the very few individuals in it. The performance is comparable to the reasoner, as the high match rate values show. Yet this yields low induction rates.

As such, the method appears to be sound. As regards the completeness, we observed during the first experiment that the amount of individuals that vote for the `unknown` class w.r.t. the query was generally high for these ontologies, thus biasing the final decision. Although this may be satisfactory compared to the reasoner's performance, the final goal is to overcome the inherent incompleteness due to the OWA and try to induce the real classification of an individual (w.r.t. the intended meaning of the domain modeled by an ontology). Therefore, we tweaked the procedure, by decreasing the impact of the individuals classified as unknown during the voting phase. Specifically, it is modified by answering `unknown` only when the number of neighboring positives and negatives is balanced; in the rest of the cases, the new procedure gives an answer to a binary classification problem depending on the their majority.

Again, the method proved sound (null commission error rate). As regards completeness, we observed in two cases a shift from the match rate towards the induction rate: i.e. the system actually suggested a classification, even in presence of a high rate of individuals classified as `unknown`. Indeed, for the SURFACE-WATER-MODEL and the NEWTESTAMENTNAMES ontologies, the incompleteness is caused by the lack of information about concept disjointness, thus yielding no counterexamples which were needed to assess the membership to the target query. For the other cases, the outcomes are almost similar to those observed in the previous experiment. Namely, or the FSM ontology this should be caused by the number of disjointness axioms (46) which greatly helps the procedure. Moreover, most of the its individuals are instances of a single concept. This latter situation applies also the SCIENCE ontology.

Concluding, we have observed that the proposed method is often able to induce new assertions in addition those that were already logically derivable from the knowledge base. Particularly, an increase in prediction accuracy was observed when the instances are homogeneously spread and information about concept disjointness is available. Besides, the method confirmed its tolerance to noise as a very low com-

[3] Located at the webpage: `http://protege.stanford.edu/download/ontologies.html`

[4] We employed PELLET: `http://www.mindswap.org`

**Table 1: Results of the experiments.**

| | Ontology | *measure* | *match rate* | *induction rate* | *omission err. rate* | *commission err. rate* |
|---|---|---|---|---|---|---|
| 1st Experiment | fsm | *avg.* | 0.92 | 0.00 | 0.08 | 0.00 |
| | | *range* | 0.84 - 1.00 | 0.00 - 0.00 | 0.00 - 0.16 | 0.00 - 0.00 |
| | Surface-Water-Model | *avg.* | 0.92 | 0.07 | 0.01 | 0.00 |
| | | *range* | 0.57 - 1.00 | 0.00 - 0.43 | 0.00 - 0.03 | 0.00 - 0.00 |
| | Science | *avg.* | 0.94 | 0.06 | 0.00 | 0.00 |
| | | *range* | 0.04 - 1.00 | 0.00 - 0.96 | 0.00 - 0.00 | 0.00 - 0.00 |
| | NewTestamentNames | *avg.* | 0.98 | 0.00 | 0.02 | 0.00 |
| | | *range* | 0.78 - 1.00 | 0.00 - 0.00 | 0.00 - 0.22 | 0.00 - 0.00 |
| 2nd Experiment | fsm | *avg.* | 0.81 | 0.00 | 0.19 | 0.00 |
| | | *range* | 0.30 - 1.00 | 0.00 - 0.00 | 0.00 - 0.68 | 0.00 - 0.03 |
| | Surface-Water-Model | *avg.* | 0.54 | 0.46 | 0.00 | 0.00 |
| | | *range* | 0.02 - 1.00 | 0.00 - 0.99 | 0.00 - 0.01 | 0.00 - 0.00 |
| | Science | *avg.* | 0.98 | 0.02 | 0.00 | 0.00 |
| | | *range* | 0.71 - 1.00 | 0.00 - 0.29 | 0.00 - 0.00 | 0.00 - 0.00 |
| | NewTestamentNames | *avg.* | 0.45 | 0.55 | 0.00 | 0.00 |
| | | *range* | 0.01 - 1.00 | 0.00 - 0.99 | 0.00 - 0.00 | 0.00 - 0.00 |

mission error was observed. In order to assess the compliance with the real relevance of an individual (*intended meaning*) w.r.t. the test queries probably a human expert would be more suitable than a reasoner.

## 6. CONCLUSIONS AND FUTURE WORK

A merely deductive approach to retrieval may fall short in the real cases of heterogeneous knowledge bases integrating distributed knowledge sources. That was the reason for investigating other forms of retrieval based on different membership decision procedures. Specifically, we have proposed a way to adapt instance-based learning and the disjunctive version space approach, to the retrieval task.

An instance-based learning method applied to DL representations that may serve to predict/suggest missing information about individuals in a knowledge base. Besides, the procedure is robust to noise and never made commission errors in the experiments that have been carried out so far.

Future work will concern a further investigation on ways to make the method more language-independent, so to apply it to more expressive DL languages, as those implemented in OWL. Moreover, we are studying the possibility of providing, together with each individual classification, also an estimate of its probability. Besides, the application of the method to the problem of Semantic Web Service discovery and retrieval is also foreseen.

## 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 semantic web services discovery tasks.

## 7. REFERENCES

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

[2] Baader, F., and Küsters, R. Non-standard inferences in description logics: The story so far. In *Mathematical Problems from Applied Logic. New Logics for the XXIst Century*, D. Gabbay, S. S. Goncharov, and M. Zakharyaschev, Eds., vol. 4 of *International Mathematical Series*. Kluwer/Plenum Publishers, 2005.

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

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

[5] d'Aquin, M., Lieber, J., and Napoli, A. Decentralized case-based reasoning for the Semantic Web. In *Proceedings of the 4th International Semantic Web Conference, ISWC2005* (2005), Y. Gil, V. Motta, E. Benjamins, and M. A. Musen, Eds., vol. 3279 of *LNCS*, Springer, pp. 142–155.

[6] Emde, W., and Wettschereck, D. Relational instance-based learning. In *Proceedings of the Thirteenth International Conference on Machine Learning, ICML96* (1996), L. Saitta, Ed., Morgan Kaufmann, pp. 122–130.

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

[8] Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., and Sure, Y. A framework for handling inconsistency in changing ontologies. In *Proceedings of the 4th International Semantic Web Conference, ISWC2005* (2005), Y. Gil, V. Motta, E. Benjamins, and M. A. Musen, Eds., vol. 3279 of *LNCS*, pp. 353–367.

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

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

[11] Sebag, M. Delaying the choice of bias: A disjunctive version space approach. In *Proceedings of the Thirteenth International Conference on Machine Learning, ICML96* (1996), L. Saitta, Ed., Morgan Kaufmann, pp. 444–452.

[12] Teege, G. A subtraction operation for description logics. In *Proceedings of the 4th International Conference on Principles of Knowledge Representation and Reasoning* (1994), P. Torasso, J. Doyle, and E. Sandewall, Eds., Morgan Kaufmann, pp. 540–550.