Induction of Optimal Semi-distances for Individuals based on Feature Sets

Nicola Fanizzi, Claudia d'Amato, Floriana Esposito

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

Abstract. Many activities related to semantically annotated resources can be enabled by a notion of similarity among them. We propose a method for defining a family of semi-distances over the set of individuals in a knowledge base which can be used in these activities. In the line of works on distance-induction on clausal spaces, the family is parameterized on a committee of concepts. Hence, we also present a method based on the idea of simulated annealing to be used to optimize the choice of the best concept committee.

1 Introduction

Recently, a growing interest is being committed to alternative inductive procedures extending the scope of the methods that can be applied to concept representations. Some are based on a notion of similarity such as *case-based reasoning* [6], *retrieval* [5, 7], *inductive generalization* [10] and *conceptual clustering* [8] or *ontology matching* [16].

As pointed out in a seminal paper [2] concerning similarity in Description Logics (DL), most of the existing measures focus on the similarity of atomic concepts within simple hierarchies. Besides, alternative approaches are based on related notions of *feature* similarity or *information content* (see also [14]). All these approaches have been specifically aimed at assessing concept similarity. In the perspective of crafting similarity-based inductive methods for DL , the need for a definition of a semantic similarity measure for *individuals* arises, that is a problem that so far received less attention in the literature.

Recently, some dissimilarity measures for individuals in specific DL representations have been proposed which turned out to be practically effective for the targeted inductive tasks (e.g. the nearest-neighbor approach applied to retrieval [5]). Although these measures ultimately rely on the semantics of primitive concepts as elicited from the ABox, still they are partly based on structural criteria (a notion of normal form coupled with a *most specific concept* operator [1]) which determine also their main weakness: they are hardly scalable to deal with standard languages used in the current knowledge management frameworks. For example, in [5] the most specific concepts w.r.t. the ABox of individuals are first computed (or their approximations in a normal form) as expressed in $\mathcal{ALC}$, then a structural measure assesses the similarity of the resulting AND-OR trees, where, ultimately, the computation is based on the extensions of the primitive concepts in the leaves.

Therefore, we have devised a new family of dissimilarity measures for semantically annotated resources, which can overcome the aforementioned limitations. Our measures are mainly based on Minkowski's measures for Euclidean spaces induced by means of a proper method developed in the context of *multi-relational learning* [15]. Another source of inspiration was *rough sets* theory [13] which aims at the formal definition of vague sets (concepts) by means of their approximations determined by an *indiscernibility* relationship.

Namely, the measures are based on the degree of discernibility of the input individuals with respect to a committee of features, which are represented by concept descriptions expressed in DL. One of the advantages of these measures is that they do not rely on a particular language for semantic annotations. As such, these new measures are not absolute, since they depend on both the choice (and cardinality) of the features committee and the knowledge base they are applied to. Rather, they rely on statistics on individuals that are likely to be maintained by knowledge base management systems [9, 4], which can determine a potential speed-up in the measure computation during knowledge-intensive tasks. Furthermore, we also propose a way to extend the presented measures to the case of assessing concept similarity by means of the notion of *medoid* [11], i.e., in the DL context, the most centrally located individual in a concept extension w.r.t. a given metric.

Experimentally, it may be shown that the measures induced by large committees (e.g. including all primitive and defined concepts) can be sufficiently accurate (i.e. properly discriminating) when employed for classification tasks even though the committee of features employed were not the optimal one or if the concepts therein were partially redundant. Nevertheless, this has led us to investigate on a method to optimize the committee of features that serve as dimensions for the computation of the measure. To this purpose, the employment of genetic programming and randomized search procedures was considered. Finally we opted for an optimization procedure based on *simulated annealing* [3], a randomized approach that can overcome the problem of the local minima, i.e. finding a good solution w.r.t. the fitness function that is not globally optimal.

The remainder of the paper is organized as follows. The definition of the family of measures is proposed in Sect. 2, where we prove them to be semi-distances. In Sect. 3, we illustrate and discuss the method for optimizing the choice of concepts for the committee of features which induces the measures. Possible developments are finally examined in Sect. 4.

2 A Family of Semi-distances for Individuals

In the following, we assume that resources, concepts and their relationship may be defined in terms of a generic DL language endowed with the standard descriptive semantics (see the handbook [1] for a thorough reference).

For the measure definition, we simply consider a *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ containing a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$. The set of the individuals occurring in $\mathcal{A}$ will be denoted with $\mathsf{Ind}(\mathcal{A})$.

As regards the inference services, our measures require (non)membership queries performing ABox lookups or *instance-checking* [1]. The complexity depends on the

DL of choice, however much of the computational effort can be saved by means of pre-computation (see projection functions below).

2.1 A Family of Measures for Individuals

We focus on the problem of assessing the semantic similarity (or dissimilarity) of individuals in the context of a knowledge base expressed in DL. To the best of our knowledge, only few measures tackle this problem so far [5]. Following some ideas borrowed from machine learning [15], a family of totally semantic distance measures for individuals can be defined in the context of a knowledge base.

It can be observed that individuals lack a syntactic structure that may be exploited for a comparison. However, on a semantic level, similar individuals should *behave* similarly with respect to the same concepts, i.e. similar assertions should be shared by them. Therefore, we introduce novel dissimilarity measures for individuals, whose rationale is the comparison of their semantics w.r.t. a fixed number of dimensions represented by DL concept descriptions. Namely, individuals are compared on the grounds of their behavior w.r.t. a reduced (yet not necessarily disjoint) committee of features, represented by 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 semi-distance functions for individuals, inspired by Minkowski's metrics, can be defined as follows:

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

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

where $\forall i \in \{1, \ldots, m\}$ the i-th projection function *π_i is defined by:*

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

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

As an alternative, especially when a good number of assertions are available in the ABox, the measures can be approximated by defining the projection functions based on a simple ABox look-up:

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

2.2 Discussion

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

Proposition 1 (semi-distance). *For a fixed feature set and $p > 0$, function d_p is a semi-distance.*

Proof. In order to prove the thesis, given any three individuals $a, b, c \in \mathsf{Ind}(\mathcal{A})$ it must hold that:
1. $d_p(a, b) \geq 0$ (positivity)
2. $d_p(a, b) = d_p(b, a)$ (symmetry)
3. $d_p(a, c) \leq d_p(a, b) + d_p(b, c)$ (triangular inequality)
Now, we observe that:

1. *trivial, by definition*
2. *trivial, for the commutativity of the operators involved*
3. *the property follows for the properties of the power function:*

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

As such, these are only a semi-distances. Namely, it cannot be proved[1] that $d_p(a, b) = 0$ iff $a = b$. This is the case of *indiscernible* individuals with respect to the given set of features F.

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 may represent a sufficient number of (possibly redundant) features that are able to discriminate really different individuals.

It could be criticized that the subsumption hierarchy has not been explicitly involved. However, this may be actually yielded as a side-effect of the possible partial redundancy of the various concepts, which has an impact on their extensions and thus

[1] In case the *unique names assumption* were made, a further projection function can be introduced π_0, such that $|\pi_0(a) - \pi_0(b)| = 1$ iff $a \neq b$.

on the related projection function. A tradeoff is to be made between the number of features employed and the computational effort required for computing the related projection functions.

Compared to other distance (or dissimilarity) measures [2, 5], the presented functions do not depend on a specific language. Note that the computation of projection functions π_i ($i = 1, \ldots, m$) on the individuals can be performed in advance (with the support of KBMS [9, 4]) 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 case-based reasoning and all other instance-based methods including clustering algorithms.

Following the rationale of the average link criterion used in agglomerative clustering [11], the measures can be extended to the case of concepts, by recurring to the notion of medoids. The *medoid* of a group of individuals is the individual that has the highest similarity w.r.t. the others. Formally. given a group $G = \{a_1, a_2, \ldots, a_n\}$, the medoid is defined:

$$\mathrm{medoid}(G) = \operatorname*{argmin}_{a \in G} \sum_{j=1}^{n} d(a, a_j)$$

Now, given two concepts C_1, C_2, we can consider the two corresponding groups of individuals obtained by retrieval $R_i = \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{K} \models C_i(a)\}$, and their resp. medoids $m_i = \mathrm{medoid}(R_i)$ for $i = 1, 2$ w.r.t. a given measure d_p^{F} (for some $p > 0$ and committee F). Then we can define the function for concepts as follows:

$$d_p^{\mathsf{F}}(C_1, C_2) := d_p^{\mathsf{F}}(m_1, m_2)$$

3 Feature Set Optimization

Experimentally, we obtained satisfactory results[2] by testing the measure on distance-based classification. Nevertheless, various optimizations of the measures can be foreseen as concerns their parametric definition. Specifically, the choice of the concepts to be included in the committee – *feature selection* – will be examined. Among the possible committees, those that are able to better discriminate the individuals in the ABox ought to be preferred:

Definition 2 (good feature set). *Let* $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$ *be a set of concept descriptions. We call* F *a* good feature set *for the knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ *iff* $\forall a, b \in \mathsf{Ind}(\mathcal{A}), a \neq b : \exists i \in \{1, \ldots, m\} : \pi_i(a) \neq \pi_i(b)$.

Note that, when the function defined above adopts a good feature set, it has the properties of a metric on the related instance-space.

Since the function strongly depends on the choice of concepts included in the committee of features F, two immediate heuristics can be derived:

[2] Results omitted for lack of space. They are available in technical reports and papers to appear. See `http://lacam.di.uniba.it:8000/people/nicola.html`.

1. controlling the number of concepts of the committee (which has an impact also on efficiency), including especially those that are endowed with a real discriminating power;
2. finding optimal sets of discriminating features of a given cardinality, by allowing also their composition employing the specific constructors made available by the DL language of choice.

Both these heuristics can be enforced by means of suitable machine learning techniques especially when knowledge bases with large sets of individuals are available. Namely, part of the entire data can be drawn in order to induce optimal F sets, in advance with respect to the application of the measure for all purposes. The adoption of genetic programming has been considered for constructing optimal sets of features. Yet these algorithms are known to suffer from being possibly caught in local minima. An alternative may consist in employing a different probabilistic search procedure which aims at a global optimization. Thus a method based on simulated annealing [3] has been devised, whose algorithm is reported in Fig. 1.

Essentially the algorithm searches the space of all feature sets starting from an initial guess (determined by $\textsc{makeInitialFS}(\mathcal{K})$) based on the concepts (both primitive and defined) currently referenced in the knowledge base. The loop controlling the search is repeated for a number of times that depends on the temperature which gradually decays to 0, when the current feature committee can be returned. Meanwhile, this set is iteratively refined calling a suitable procedure $\textsc{randomSuccessor}()$. Then the fitness of the new feature set is compared to that of the current one determining the increment of energy ΔE. If this is positive then the candidate committee replaces the current one. Otherwise it will (less likely) be replaced with a probability that depends on ΔE.

As regards the heuristic $\textsc{fitnessValue}(\mathsf{F})$, it can be computed as the average *discernibility factor* [13] of the individuals w.r.t. the feature set. For example, given a set of individuals *IS* $= \{a_1, \ldots, a_n\} \subseteq \mathsf{Ind}(\mathcal{A})$ (the whole or just a sample of $\mathsf{Ind}(\mathcal{A})$ used to induce an optimal measure) the fitness function may be defined:

$$\textsc{fitnessValue}(\mathsf{F}) = k \cdot \sum_{1 \leq i < j \leq n} \sum_{k=1}^{|\mathsf{F}|} \mid \pi_k(a_i) - \pi_k(a_j) \mid$$

where k is a normalization factor which may be set to: $(1/m)\,(n \cdot (n-1)/4 - n)$, which depends on the number of couples of different individuals that really determine the fitness measure.

As concerns finding candidates to replace the current committee, the function $\textsc{randomSuccessor}()$ can be implemented by recurring to simple transformations of the feature set:

- adding (resp. removing) a concept C: $\mathsf{nextFS} \leftarrow \mathsf{currentFS} \cup \{C\}$
  (resp. $\mathsf{nextFS} \leftarrow \mathsf{currentFS} \setminus \{C\}$)
- randomly choosing one of the current concepts from $\mathsf{currentFS}$, say C;
  replacing it with one of its refinements $C' \in \textsc{ref}(C)$

Refining concept descriptions is language-dependent. For the case of $\mathcal{ALC}$ logic, refinement operators have been proposed in [12, 10]. Complete operators are to be preferred to ensure exploring the whole search-space

```
FeatureSet optimizeFeatureSet($\mathcal{K}$, $\Delta T$)
input   $\mathcal{K}$: Knowledge base
        $\Delta T$: function controlling the decrease of temperature
output  FeatureSet
static  currentFS: current Feature Set
        nextFS: next Feature Set
        Temperature: controlling the probability of downward steps
begin
currentFS $\leftarrow$ makeInitialFS($\mathcal{K}$)
for $t \leftarrow 1$ to $\infty$ do
        Temperature $\leftarrow$ Temperature $- \Delta T(t)$
        if (Temperature = 0)
            return currentFS
        nextFS $\leftarrow$ randomSuccessor(currentFS,$\mathcal{K}$)
        $\Delta E \leftarrow$ fitnessValue(nextFS) $-$ fitnessValue(currentFS)
        if ($\Delta E > 0$)
            currentFS $\leftarrow$ nextFS
        else   // replace FS with given probability
            replace(currentFS, nextFS, $e^{\Delta E}$)
end
```

Fig. 1. Feature Set optimization based on a Simulated Annealing procedure.

Given a suitable cooling schedule, the algorithm is known to find an optimal solution. To control the complexity of the process alternate schedules may be preferred that guaratee the construction of suboptimal solutions in polynomial time [3].

4 Conclusion and Extensions

We have proposed the definition of a family of semi-distances over the individuals in a DL knowledge base. The measures are not language-dependent yet they are parameterized on a committee of concepts. Therefore, we have also presented a randomized search method to find optimal committees. One of the advantages of the measures is that they are not language-dependent differently from previous proposals [5]. As previously mentioned, the subsumption relationships among concepts in the committee is not explicitly exploited in the measure for making the relative distances more accurate. The extension to the case of concept distance may also be ameliorated.

The measure may have a wide range of application of distance-based methods to knowledge bases. They have been integrated in an instance-based learning system implementing a nearest-neighbor learning algorithm: an experimentation on performing semantic-based retrieval proved the effectiveness of the new measures, compared to the outcomes obtained adopting other measures [5]. The next step concerns exploiting the measures in a conceptual clustering algorithm where clusters will be formed by grouping instances on the grounds of their similarity, possibly triggering the induction of new emerging concepts, as in [8].

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.J. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, *Working Notes of the 2005 International Description Logics Workshop, DL2005*, volume 147 of *CEUR Workshop Proceedings*. CEUR, 2005.

[3] E.K. Burke and G. Kendall. *Search Methodologies*. Springer, 2005.

[4] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, *Proceedings of the 2005 International Workshop on Description Logics, DL2005*, volume 147 of *CEUR Workshop Proceedings*. CEUR, 2005.

[5] 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.

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

[7] N. Fanizzi, C. d'Amato, and F. Esposito. Instance-based retrieval by analogy. In *Proceedings of the 22nd Annual ACM Symposium of Applied Computing, SAC2007*, volume 2, pages 1398–1402, Seoul, South Korea, 2007. ACM.

[8] N. Fanizzi, L. Iannone, I. Palmisano, and G. Semeraro. Concept formation in expressive description logics. In J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Proceedings of the 15th European Conference on Machine Learning, ECML2004*, volume 3201 of *LNAI*, pages 99–113. Springer, 2004.

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

[10] L. Iannone, I. Palmisano, and N. Fanizzi. An algorithm based on counterfactuals for concept learning in the semantic web. *Applied Intelligence*, 26(2):139–159, 2007.

[11] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[12] J. Lehmann. Concept learning in description logics. Master's thesis, Dresden University of Technology, 2006.

[13] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, 1991.

[14] A. Rodriguez. *Assessing semantic similarity between spatial entity classes*. PhD thesis, University of Maine, 1997.

[15] 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.

[16] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, IV:146–171, 2005.