Query Answering and Ontology Population: an Inductive Approach

Claudia d'Amato Nicola Fanizzi Floriana Esposito

Department of Computer Science
University of Bari

ESWC 2008 ⋄ Tenerife, June 4, 2008

Contents

Introduction & Motivations

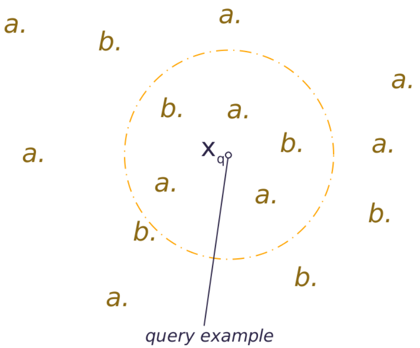
- In the SW context, reasoning is performed through deductive-based inference
- Purely logic methods may fail when data sources are distributed and potentially incoherent
  - This has given rise to *alternative methods* such as approximate and inductive reasoning
- **Focus** on *Query Answering* task i.e. finding the extension of a query concept
  - *It can can be cast* as a problem of establishing the class membership of the individuals in a KB.
  - It can be solved by the use of *instance-based methods* that are known to be both *very efficient* and *fault-tolerant* compared to the classic logic-based methods.
  - The *Nearest Neighbor approach* is adopted

Knowledge Base Representation

- OWL representation founded in Description Logics (DL):
- Knowledge base: $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$
  - TBox $\mathcal{T}$: a set of DL concept definitions
  - ABox $\mathcal{A}$: assertions (facts) about the world state
  - Ind($\mathcal{A}$): set of Individuals (resources) in the ABox
- Inference service of interest from the KBMS:
  - *instance-checking*: decision procedure that assess if an individual is instance of a certain concept or not
    - Sometimes a simple lookup may be sufficient

Nearest Neighbor Classification

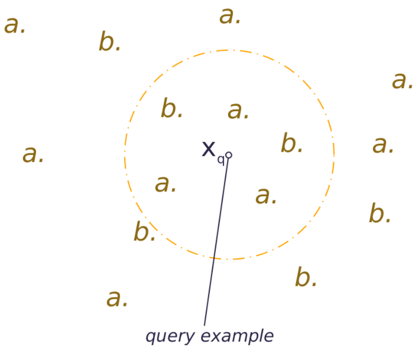
classes: a, b $k = 5$



$class(x_q) \leftarrow ?$

Nearest Neighbor Classification

classes: a, b $k = 5$



$class(x_q) \leftarrow$ **a**

Technical Problems

1. Generally applied to *feature vector* representation
   → *upgrade k-NN to more expressive representations*
2. Classification: classes considered as *disjoint*
   → *cannot assume disjointness of all concepts*
3. An implicit *Closed World Assumption* is made in ML
   → *cope with the Open World Assumption made in SeWeb*

Customization to DLs

1. Definition of a dissimilarity measure applicable to ontological knowledge
2. Alternative classification procedure adopted:
   - multi-class problem *decomposed* into smaller *binary classification problems* (one per target concept)
   - For each query concept Q:
     binary classification $\{-1, +1\}$
3. Extend the possible results with a *third value* 0 representing unknown classification: $\{-1, 0, +1\}$

Weighted majority voting criterion is applied

Realized k-NN algorithm

- **Training Phase:** All training examples (individuals in the KB) are memorized jointly with the classes to which they belong to
- **Testing Phase:**
  - For each test example x_q, given a dissimilarity measure d, the k training elements less dissimilar from x_q are determined, hence

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

where $V = \{-1, 0, +1\}$; $\delta(a, b) = 1$ if $a = b$; $\delta(a, b) = 0$ if $a \neq b$; $\omega_i = 1/d(x_q, x_i)$ and

$$h_j(x) = \begin{cases} +1 & C_j(x) \in \mathcal{A} \quad (\mathcal{K} \models C_j(x)) \\ -1 & \neg C_j(x) \in \mathcal{A} \quad (\mathcal{K} \models \neg C_j(x)) \\ 0 & \textit{otherwise} \end{cases}$$

Semi-Distance Measure: Rationale

- **IDEA**: *on a semantic level, similar individuals should behave similarly w.r.t. the same concepts*
- Following HDD **[Sebag 1997]**: individuals can be compared on the grounds of their behavior w.r.t. a given set of hypotheses $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, that is a collection of (primitive or defined) concepts **[Fanizzi et al. @ DL 2007]**
  - F stands as a group of *discriminating features* expressed in the considered language
- **Proposed Extention:** Features are weighted w.r.t. their *discriminating power* in determining the dissimilarity value.
  - Weights determined on the ground of *information conveyed* that is measured with the notion of *entropy*
- As such, the new measure *totally depends on semantic* aspects of the individuals in the KB

Semantic Semi-Dinstance Measure: Definition

Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a KB and let $\mathsf{Ind}(\mathcal{A})$ be the set of the individuals in $\mathcal{A}$. Given sets of concept descriptions $\mathsf{F} = \{F_1, F_2, \dots, F_m\}$ in $\mathcal{T}$, a *family of semi-distance functions* $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto \mathbb{R}^+$ is defined as follows:

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

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

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

Defining Feature Weight

- Features are weighted w.r.t. their *discriminating power* in determining the dissimilarity value.
  - Weights determined on the ground of *the quantity information conveyed* ⇒ measured as the *entropy* of the feature
- **Rationale: the more general a feature (or its negation) is (low entropy) the less usable it is for distinguishing the two individuals and vice versa**
- The probability of a feature F is approximated as $P_F = |\text{retrieval}(F)|/|\text{Ind}(\mathcal{A})|$
- Considering also $P_{\neg F}$ related to its negation and that related to the unclassified individuals (w.r.t. F), denoted P_U, the entropic measure of F is given by:

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

Distance Measure: Example

$\mathcal{T} = \{$ Female $\equiv \neg$Male, Parent $\equiv \forall$child.Being $\sqcap \exists$child.Being,
Father $\equiv$ Male $\sqcap$ Parent,
FatherWithoutSons $\equiv$ Father $\sqcap \forall$child.Female$\}$

$\mathcal{A} = \{$ Being(ZEUS), Being(APOLLO), Being(HERCULES), Being(HERA),
Male(ZEUS), Male(APOLLO), Male(HERCULES),
Parent(ZEUS), Parent(APOLLO), $\neg$Father(HERA),
God(ZEUS), God(APOLLO), God(HERA), $\neg$God(HERCULES),
hasChild(ZEUS, APOLLO), hasChild(HERA, APOLLO),
hasChild(ZEUS, HERCULES), $\}$

Suppose $\mathsf{F} = \{F_1, F_2, F_3, F_4\} = \{$Male, God, Parent, FatherWithoutSons$\}$.

Let us compute the distances (with $p = 1$):

$d_1^{\mathsf{F}}(\text{HERCULES}, \text{ZEUS}) =$

$(\overline{\omega}_{\text{Male}} \cdot |1-1| + \overline{\omega}_{\text{God}} \cdot |0-1| + \overline{\omega}_{\text{Parent}} \cdot |1/2-1| + \overline{\omega}_{\text{FatherWithoutSons}} \cdot |1/2-0|)/4$

Computation $\overline{\omega}_i$ Trivial $\Rightarrow$ Omitted

Experimental Setting

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

- 20 query concept (randomly generated) considered for each ontology
- All the individuals in each ontology have been classified; $k = log|TrainingSet|$ where $TrainingSet = |Ind(\mathcal{A})| \cdot 4\%$
- d_1^F employed considering both *uniform feature weights* and *entropic feature weights*; $F =$ all concepts in the ontology
- *10-fold* cross validation
- Performance compared with a standard reasoner (Pellet).

Evaluation in terms of standard IR measures

Average ± standard deviation and [min.;max.] intervals.

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

Outcomes: Discussion

- Precision and Recall quite high
  - except for SWSD where precision was significantly lower since *a very limited number of individuals per concept was available*
  - the *entropic measure* improve results w.r.t. the one using uniform weights
- Recall less than precision $\Rightarrow$ due to the OWA
  - *Many cases in which the reasoner does not return any result differently from the classifier*
  - **Behavior registered as mistake while it may likely turn out to be a correct inference when judged by a human agent.**

$$\Downarrow$$

  - *In order to distinguish between inductively classified individuals and real mistakes additional indices have been considered.*

Additional Evaluation Parameters

- *match rate*: cases of match of the classification returns by both procedures.
- *omission error rate*: cases when our procedure cannot decide (0) while the reasoner gave a classification (± 1)
- *commission error rate*: cases when our procedure returned ± 1 while the reasoner gave the opposite outcome ∓ 1
- *induction rate*: cases when the reasoner cannot decide (0) while our procedure gave a classification (± 1)

Additional Outcomes

Average ± standard deviation and [min.;max.] intervals.

Uniform Weight Measure				
	match	commission	omission	induction
SWM	93.3 ± 10.3 [68.7;100.0]	0.0 ± 0.0 [0.0;0.0]	2.5 ± 4.4 [0.0;16.5]	4.2 ± 10.5 [0.0;31.3]
BioPax	99.9 ± 0.2 [99.4;100.0]	0.2 ± 0.2 [0.0;0.06]	0.0 ± 0.0 [0.0;0.0]	0.0 ± 0.0 [0.0;0.0]
LUBM	99.2 ± 0.8 [98.0;100.0]	0.0 ± 0.0 [0.0;0.0]	0.8 ± 0.8 [0.0;0.2]	0.0 ± 0.0 [0.0;0.0]
NTN	98.6 ± 1.5 [93.9;100.0]	0.0 ± 0.1 [0.0;0.4]	0.8 ± 1.1 [0.0;3.7]	0.6 ± 1.4 [0.0;6.1]
SWSD	97.5 ± 3.7 [84.6;100.0]	0.0 ± 0.0 [0.0;0.0]	1.8 ± 2.6 [0.0;9.7]	0.8 ± 1.5 [0.0;5.7]
Financial	99.5 ± 0.8 [97.3;100.0]	0.3 ± 0.7 [0.0;2.4]	0.0 ± 0.0 [0.0;0.0]	0.2 ± 0.2 [0.0;0.6]

Entropic Measure				
	match	commission	omission	induction
SWM	97.5 ± 3.2 [89.6;100.0]	0.0 ± 0.0 [0.0;0.0]	2.2 ± 3.1 [0.0;10.4]	0.3 ± 1.2 [0.0;5.2]
BioPax	99.9 ± 0.2 [99.4;100.0]	0.1 ± 0.2 [0.0;0.06]	0.0 ± 0.0 [0.0;0.0]	0.0 ± 0.0 [0.0;0.0]
LUBM	99.5 ± 0.7 [98.2;100.0]	0.0 ± 0.0 [0.0;0.0]	0.5 ± 0.7 [0.0;1.8]	0.0 ± 0.0 [0.0;0.0]
NTN	97.5 ± 1.9 [91.3;99.3]	0.6 ± 0.7 [0.0;1.6]	1.3 ± 1.4 [0.0;4.9]	0.6 ± 1.7 [0.0;7.1]
SWSD	98.0 ± 3.0 [88.3;100.0]	0.0 ± 0.0 [0.0;0.0]	1.9 ± 2.9 [0.0;11.3]	0.1 ± 0.2 [0.0;0.5]
Financial	99.7 ± 0.2 [99.4;100.0]	0.0 ± 0.0 [0.0;0.1]	0.0 ± 0.0 [0.0;0.0]	0.2 ± 0.2 [0.0;0.6]

Additional outcomes: Discussion

- *Commission error* almost null on average
- *Omission error rate* almost null
- *Induction Rate* not null
  - **new knowledge (not logically derivable) is induced** ⇒ it can be used for making the *ontology population task semi-automatic*
  - exception for LUBM and BioPax ontologies, where individuals are instances of the same concepts (most of the time a single concept) and this does not allow to induce new knowledge.
  - For the other ontologies, induced knowledge can be found ⇒ *individuals are instances of many concepts* and they are *homogeneously spread* w.r.t. the several concepts.

Likelihood of the inductive assertions

Since inductive results are not certain, the likelihood of the decision made by the procedure could be also measured:

- given the nearest training individuals in $NN(x_q, k) = \{x_1, \ldots, x_k\}$, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

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

Likelihood of the inductive assertions: Results

	SWM	NTN	SWSD	FINANCIAL
3-valued case	76.26	98.36	76.27	92.55
2-valued case	100.0	98.36	76.27	92.55

- *First row* $\Rightarrow$ likelihood based on the normalization over the 3 possible values $(0, +1, -1)$.
- *Second row* $\Rightarrow$ likelihood based on the normalization over the 2 possible values $(+1, -1)$.
  - *Likelihood increases only for* SWM $\Rightarrow$ this in the only case in which example labeled with 0 are selected as neighbors.
- *High likelihood values* $\Rightarrow$ the distance function selects very similar examples w.r.t. the query instance

Conclusions & Future Work

Conclusions: Proposed and inductive method for performing concept retrieval that is:

- comparable with a deductive reasoner (even working with quite limited training sets)
- able to induce new knowledge not logically derivable

Future works:

- Investigate feature building/selection for reducing the effort in computing individual distance

The End

That's all!

Questions ?