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

1 Introduction & Motivation

2 Concept Retrieval by Semantic Nearest Neighbor Search

3 A Semantic Semi-Distance Measure for DLs

4 Experimentation
   • Setting
   • Evaluation Parameters
   • Experimental Results

5 Conclusions and Future Works
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 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: \( K = \langle T, A \rangle \)
    - TBox \( T \): a set of DL concept definitions
    - ABox \( A \): assertions (facts) about the world state
    - \( \text{Ind}(A) \): 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: $a, b$  \hspace{1cm} k = 5

$\text{class}(x_q) \leftarrow ?$
Nearest Neighbor Classification

classes:  $a, b$  \quad k = 5

$\text{class}(x_q) \leftarrow a$
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

\[
\hat{h}_j(x_q) := \arg\max_{v \in V} \sum_{i=1}^{k} \omega_i \cdot \delta(v, h_j(x_i)) \quad \forall j \in \{1, \ldots, s\} \quad (1)
\]

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

\[
h_j(x) = \begin{cases} 
+1 & C_j(x) \in \mathcal{A} \quad (\mathcal{K} \models C_j(x)) \\
-1 & \neg C_j(x) \in \mathcal{A} \quad (\mathcal{K} \models \neg C_j(x)) \\
0 & \text{otherwise}
\end{cases}
\]
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 $F = \{F_1, F_2, \ldots, F_m\}$, 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-Distance Measure: Definition

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

$$
\forall a, b \in \text{Ind}(\mathcal{A}) \quad d^F_p(a, b) := \frac{1}{m} \left[ \sum_{i=1}^{m} \omega_i \cdot | \pi_i(a) - \pi_i(b) |^p \right]^{1/p}
$$

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

$$
\forall a \in \text{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 
1 & F_i(a) \in \mathcal{A} \quad (\mathcal{K} \models F_i(a)) \\
0 & \neg F_i(a) \in \mathcal{A} \quad (\mathcal{K} \models \neg F_i(a)) \\
1/2 & \text{otherwise}
\end{cases}
$$
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
  $$P_F = |\text{retrieval}(F)|/|\text{Ind}(A)|$$

- Considering also $P_{¬F}$ 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:
  $$H(F) = - (P_F \log(P_F) + P_{¬F} \log(P_{¬F}) + P_U \log(P_U))$$
Distance Measure: Example

\[ T = \{ \text{Female} \equiv \neg \text{Male}, \quad \text{Parent} \equiv \forall \text{child}. \text{Being} \sqcap \exists \text{child}. \text{Being}, \]
\[ \quad \text{Father} \equiv \text{Male} \sqcap \text{Parent}, \]
\[ \quad \text{FatherWithoutSons} \equiv \text{Father} \sqcap \forall \text{child}. \text{Female} \} \]

\[ A = \{ \text{Being(ZEUS)}, \text{Being(APOLLO)}, \text{Being(HERCULES)}, \text{Being(HERA)}, \]
\[ \quad \text{Male(ZEUS)}, \text{Male(APOLLO)}, \text{Male(HERCULES)}, \]
\[ \quad \text{Parent(ZEUS)}, \text{Parent(APOLLO)}, \neg \text{Father(HERA)}, \]
\[ \quad \text{God(ZEUS)}, \text{God(APOLLO)}, \text{God(HERA)}, \neg \text{God(HERCULES)}, \]
\[ \quad \text{hasChild(ZEUS, APOLLO)}, \text{hasChild(HERA, APOLLO)}, \]
\[ \quad \text{hasChild(ZEUS, HERCULES)}, \} \]

Suppose \( F = \{ F_1, F_2, F_3, F_4 \} = \{ \text{Male, God, Parent, FatherWithoutSons} \} \).

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

\[ d_1^F(\text{HERCULES}, \text{ZEUS}) = \]
\[ (\omega_{\text{Male}} \cdot |1 - 1| + \omega_{\text{God}} \cdot |0 - 1| + \omega_{\text{Parent}} \cdot |1/2 - 1| + \omega_{\text{FatherWithoutSons}} \cdot |1/2 - 0|)/4 \]

Computation \( \omega_i \); Trivial \( \Rightarrow \) Omitted
### Experimental Setting

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

- **20** query concept (randomly generated) considered for each ontology.
- All the individuals in each ontology have been classified; 
  \[ k = \log|\text{TrainingSet}| \]  
  where \( \text{TrainingSet} = |\text{Ind}(A)| \cdot 4\% \)
- \( d_{I}^{F} \) 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.

<table>
<thead>
<tr>
<th></th>
<th>Uniform Weight Measure</th>
<th>Entropic Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>SWM</td>
<td>89.1</td>
<td>84.4</td>
</tr>
<tr>
<td></td>
<td>[16.3;100.0]</td>
<td>[11.1;100.0]</td>
</tr>
<tr>
<td>BioPax</td>
<td>99.2</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td>[93.8;100.0]</td>
<td>[50.0;100.0]</td>
</tr>
<tr>
<td>LUBM</td>
<td>100.0</td>
<td>71.7</td>
</tr>
<tr>
<td></td>
<td>[100.0;100.0]</td>
<td>[9.1;100.0]</td>
</tr>
<tr>
<td>NTN</td>
<td>98.8</td>
<td>62.6</td>
</tr>
<tr>
<td></td>
<td>[86.9;100.0]</td>
<td>[4.3;100.0]</td>
</tr>
<tr>
<td>SWSD</td>
<td>74.7</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>[8.0;100.0]</td>
<td>[2.2;100.0]</td>
</tr>
<tr>
<td>Financial</td>
<td>99.6</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>[94.3;100.0]</td>
<td>[50.0;100.0]</td>
</tr>
</tbody>
</table>

C. d’Amato  Inductive Query Answering and Ontology Population
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 ⇒ 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.

- 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 $\pm$ standard deviation and [min.;max.] intervals.

<table>
<thead>
<tr>
<th></th>
<th>Uniform Weight Measure</th>
<th></th>
<th>Entropic Measure</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>match</td>
<td>commission</td>
<td>omission</td>
<td>induction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWM</td>
<td>93.3 ± 10.3</td>
<td>0.0 ± 0.0</td>
<td>2.5 ± 4.4</td>
<td>4.2 ± 10.5</td>
</tr>
<tr>
<td></td>
<td>[68.7;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;16.5]</td>
<td>[0.0;31.3]</td>
</tr>
<tr>
<td>BioPax</td>
<td>99.9 ± 0.2</td>
<td>0.2 ± 0.2</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>[99.4;100.0]</td>
<td>[0.0;0.06]</td>
<td>[0.0;0.0]</td>
<td>[0.0;0.0]</td>
</tr>
<tr>
<td>LUBM</td>
<td>99.2 ± 0.8</td>
<td>0.0 ± 0.0</td>
<td>0.8 ± 0.8</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>[98.0;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;2.0]</td>
<td>[0.0;0.0]</td>
</tr>
<tr>
<td>NTN</td>
<td>98.6 ± 1.5</td>
<td>0.0 ± 0.1</td>
<td>0.8 ± 1.1</td>
<td>0.6 ± 1.4</td>
</tr>
<tr>
<td></td>
<td>[93.9;100.0]</td>
<td>[0.0;0.4]</td>
<td>[0.0;3.7]</td>
<td>[0.0;6.1]</td>
</tr>
<tr>
<td>SWSD</td>
<td>97.5 ± 3.7</td>
<td>0.0 ± 0.0</td>
<td>1.8 ± 2.6</td>
<td>0.8 ± 1.5</td>
</tr>
<tr>
<td></td>
<td>[84.6;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;9.7]</td>
<td>[0.0;5.7]</td>
</tr>
<tr>
<td>Financial</td>
<td>99.5 ± 0.8</td>
<td>0.3 ± 0.7</td>
<td>0.0 ± 0.0</td>
<td>0.2 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>[97.3;100.0]</td>
<td>[0.0;2.4]</td>
<td>[0.0;0.0]</td>
<td>[0.0;6.0]</td>
</tr>
<tr>
<td></td>
<td>SWM</td>
<td>97.5 ± 3.2</td>
<td>0.0 ± 0.0</td>
<td>2.2 ± 3.1</td>
</tr>
<tr>
<td></td>
<td>[89.6;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;10.4]</td>
<td>[0.0;5.2]</td>
</tr>
<tr>
<td></td>
<td>BioPax</td>
<td>99.9 ± 0.2</td>
<td>0.1 ± 0.2</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>[99.4;100.0]</td>
<td>[0.0;0.06]</td>
<td>[0.0;0.0]</td>
<td>[0.0;0.0]</td>
</tr>
<tr>
<td></td>
<td>LUBM</td>
<td>99.5 ± 0.7</td>
<td>0.0 ± 0.0</td>
<td>0.5 ± 0.7</td>
</tr>
<tr>
<td></td>
<td>[98.2;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;1.8]</td>
<td>[0.0;0.0]</td>
</tr>
<tr>
<td></td>
<td>NTN</td>
<td>97.5 ± 1.9</td>
<td>0.6 ± 0.7</td>
<td>1.3 ± 1.4</td>
</tr>
<tr>
<td></td>
<td>[91.3;99.3]</td>
<td>[0.0;1.6]</td>
<td>[0.0;4.9]</td>
<td>[0.0;7.1]</td>
</tr>
<tr>
<td></td>
<td>SWSD</td>
<td>98.0 ± 3.0</td>
<td>0.0 ± 0.0</td>
<td>1.9 ± 2.9</td>
</tr>
<tr>
<td></td>
<td>[88.3;100.0]</td>
<td>[0.0;0.0]</td>
<td>[0.0;11.3]</td>
<td>[0.0;0.5]</td>
</tr>
<tr>
<td></td>
<td>Financial</td>
<td>99.7 ± 0.2</td>
<td>0.0 ± 0.0</td>
<td>0.0 ± 0.0</td>
</tr>
<tr>
<td></td>
<td>[99.4;100.0]</td>
<td>[0.0;0.1]</td>
<td>[0.0;0.0]</td>
<td>[0.0;0.6]</td>
</tr>
</tbody>
</table>
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 $\text{NN}(x_q, k) = \{x_1, \ldots, x_k\}$, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$l(\text{class}(x_q) = v | \text{NN}(x_q, k)) = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_Q(x_i))}{\sum_{v' \in V} \sum_{i=1}^{k} w_i \cdot \delta(v', h_Q(x_i))}$$  \hspace{1cm} (2)
Likelihood of the inductive assertions: Results

<table>
<thead>
<tr>
<th></th>
<th>SWM</th>
<th>NTN</th>
<th>SWSD</th>
<th>Financial</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-valued case</td>
<td>76.26</td>
<td>98.36</td>
<td>76.27</td>
<td>92.55</td>
</tr>
<tr>
<td>2-valued case</td>
<td>100.0</td>
<td>98.36</td>
<td>76.27</td>
<td>92.55</td>
</tr>
</tbody>
</table>

- **First row** ⇒ likelihood based on the normalization over the 3 possible values \((0, +1, -1)\).
- **Second row** ⇒ likelihood based on the normalization over the 2 possible values \((+1, -1)\).
  - **Likelihood increases only for SWM** ⇒ this in the only case in which example labeled with 0 are selected as neighbors.
  - **High likelihood values** ⇒ 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
That’s all!

Questions?