Rare diseases knowledge management: the contribution of proximity measurements in OntoOrpha and OMIM.

X. AIMÉ, J. CHARLET, F. FURST, P. KUNTZ, F. TRICHE and F. DHOMBRES

a ORPHANET, INSERM US14, France
b INSERM U872, Paris, France

c AP-HP, Paris, France

d Université Pierre et Marie Curie, Paris, France

e Service de diagnostic prénatal, Pôle de péринatalité, Hôpital Armand Trousseau, Paris, France

f MIS, Université Picardie Jules Verne, Amiens, France
g LINA CNRS UMR6241, Université de Nantes, Nantes, France

Abstract. In this paper, we introduce an application of Proxima and define a new measure of proximity between two concepts present in an ontology. The approach is based on the three dimensions of a conceptualization: intension with relations between concepts, expression with terms denoting concepts, and extension with instances of concepts. This preliminary work, in the field of rare diseases, involved the Orphanet Ontology of Rare Diseases (OntoOrpha) and corpus of texts extracted from Online Inheritance in Man (OMIM). The proximity measurements are consistent with an appropriate representation of groups of diseases in the ontology, which are derived from the Orphanet classifications of rare diseases. Other semantic relations are explored and new perspectives in medical knowledge curation are proposed.

Keywords. Proximity measure, Semiotic ontology, Rare diseases, Knowledge discovery.

Introduction

The notion of proximity can cover at least two types of association including (i) the conceptual similarity, which is an analogy between two concepts that share a number of features, and (ii) the conceptual proximity in the strict sense, which is a cognitive link established between two concepts in a given context. For example, the concepts Bone and Fracture are close, they often occur together. However, they are not similar. The concepts Femoral neck and Humeral neck are very similar in nature and in structure. However, they are usually not close. This distinction, between proximity and similarity, was introduced in cognitive psychology by the Gestalt theory [3]. The law of proximity states that we group items that appear closely in the same perceptive zone.

1 Corresponding Author, 96 rue Didot, 75014, Paris, France.
The law of similarity states that details are grouped by functional or descriptive features. In [1], we propose to distinguish these two notions with two distinct measures. In this paper, we present a use case of our measure of proximity: Proxima. The evaluation of proximity is closely linked to the co-occurrence measures. Currently, this notion has been highlighted in many activities related to lexicography and similarity [7, 8, 9, 10]. We use a semiotic approach of Knowledge Engineering, which takes the three dimensions of knowledge (introduced by Morris and Peirce [4]) into account: (i) the intension of concepts, in other words, the semantics attached to them, (ii) the extension of concepts with their instances, and (iii) the expression of concepts with the associated terms denoting them. In this paper, we explore the Orphanet Ontology of Rare Diseases (OntoOrpha) [5] and a corpus of texts on human genes and genetic disorders (OMIM) by the scope of a new proximity measurement (Proxima [1]). This work suggests new approaches for medical ontology curation and validation and for information retrieval.

1. Material and methods

Proxima evaluates the proximity between two concepts according to the three semiotic dimensions. We use corpora of texts from OMIM, supposedly representative of this cognitive universe to calculate the extensional and expressional components of Proxima. In our work, two concepts are considered close if (1) in an intensional point of view, they are connected by many properties, (2) in an extensional point of view, terms denoting instances of these concepts are often present together in the textual corpus, and (3) in an expressional point of view, terms denoting concepts are often present together in the text corpus. The proximity between two concepts \(c_1\) and \(c_2\) is formally defined by the function (1) \(\text{Proxima} : C \times C \rightarrow [0,1]\)

\[
\text{Proxima} (c_1,c_2) = \alpha. \text{int}_p(c_1,c_2) + \beta. \text{exp}_p (c_1,c_2) + \gamma. \text{ext}_p (c_1,c_2) \tag{1}
\]

The importance of each component into the conceptualization of a domain can vary according to individuals or domain. Three coefficients respectively named \(\alpha\), \(\beta\) and \(\gamma\) are introduced to take into account the relative importance of these three components. We impose that these coefficients are not negative. Their values are in the interval \([0,1]\), and \(\alpha + \beta + \gamma = 1\). Although the coefficient values can be arbitrarily set, or calibrated by experiments, we proposed a method to automatically compute approximations in [2]. We considered that the triplet \((\alpha, \beta, \gamma)\) characterizes the coordinates of the cognitive user in the semiotic triangle. From an intensional level, the proximity between two concepts can be evaluated in an ontology (OntoOrpha) by the ratio between the number of properties between them and the total number of properties having as domain one of the two concepts. If there is at least one property between the two concepts, the function \(\text{int}_p : C \times C \rightarrow [0,1]\) is formally defined by (2).

\[
\text{int}_p(c_1,c_2) = \begin{cases} 
0, & \text{if } |p_1| + |p_2| = 0 \\
\left[1 - \log \left( \frac{|p_1| + |p_2|}{|p_1| + |p_2|} \right) \right]^{-1} & \text{otherwise}
\end{cases} \tag{2}
\]
Where (i) $p_1$ is the set of properties having $c_1$ as domain, (ii) $p_2$ the set of properties having $c_2$ as domain, (iii) $p_{12}$ the set of properties having $c_1$ as domain and $c_2$ as range, and (iv) $p_{21}$ the set of properties having $c_2$ as domain and $c_1$ as range.

The expressional component is calculated from a corpus of texts representative of the domain considered for the endgroup. From an expressional level, the more two concepts are denoted by terms present together in the same documents, the closer they are. Specifically, the expressional component of Proxima is even higher than the number of documents including together the terms of both concepts compared to the number of documents where at least one of two concepts appears. The function $exp_p : C \times C \rightarrow [0,1]$ is formally defined by (3).

$$
exp_p(c_1, c_2) = \left( \frac{nbBoth_c(c_1, c_2)}{nbOne_c(c_1, c_2)} \right)^{1/2}
$$

Where $nbBoth_c (c_1, c_2)$ is the number of documents in which at least one term denoting the concept $c_1$ and at least one term denoting the concept $c_2$ are present, and $nbOne_c (c_1, c_2)$ is the number of documents in which at least one term denoting $c_1$ or $c_2$ is present. As there are no instances in our ontology, the extensional component of Proxima was not used in this experiment.

OntoOrpha is a rare diseases ontology with a terminology (with labels and synonyms in 6 languages) derived from human curated Orphanet\(^2\) knowledge bases [5]. The current version of the ontology contains 11,169 concepts, 20 relations, no instance, and is available on BioPortal website\(^3\).

![Figure 1. Partial view of the core of the Orphanet Ontology of Rare Diseases (OntoOrpha).](image)

Online Mendelian Inheritance in Man (OMIM) is a database of human genes and genetic disorders\(^4\). The text of each and every 26,151 entries has been extracted and we retained only the $TI$ (title) and $TX$ (text) fields in order to build the text corpus. Using this ontology and this corpus, we have determined the values of $\alpha$ and $\beta$ ($\gamma=0$, because there are no instances in the ontology). The ratio $\alpha/\beta$ was derived from the ratio between the number of concepts designated indirectly in the corpus\(^5\) and the number of concepts designated directly by terms in the corpus. The pairs of concepts identified by the Proxima value were analyzed and reviewed by a medical expert who assessed the

---

\(^2\) Orphanet, the portal for rare diseases and orphan drugs : [http://www.orpha.net](http://www.orpha.net)

\(^3\) Available on the website: [http://bioportal.bioontology.org/ontologies/1586](http://bioportal.bioontology.org/ontologies/1586)


\(^5\) The term tetralogy of Fallot refers directly to the concept Tetralogy of Fallot, and indirectly to the concepts Conotruncal heart malformation and Congenital heart malformation.
semantic of the relation between these concepts. Comparison between intensional and extensional components of Proxima value was performed by a paired t-test\(^6\).

![Figure 2. Distribution of semantic relations according to Proxima values.](image)

**Table 1.** Semantic relations in 356 pairs of reviewed concepts: proportions according to the Proxima values and mean of differences between intensional and expressional proximities.

<table>
<thead>
<tr>
<th>Semantic Relation</th>
<th>HIGH PROXIMA VALUES (&gt;0.4)</th>
<th>INTERMEDIATE PROXIMA VALUES (0.2-0.4)</th>
<th>LOW PROXIMA VALUES (&lt;0.2)</th>
<th>Mean of the differences (\text{int}_p-\text{exp}_p)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>partOf</td>
<td>42.3 [34.6-50.1]</td>
<td>25.0 [16.5-33.5]</td>
<td>1.0 [0-3.0]</td>
<td>.102 &lt; 10(^{-5})</td>
<td></td>
</tr>
<tr>
<td>isA</td>
<td>21.2 [14.7-27.6]</td>
<td>22.0 [13.9-30.1]</td>
<td>32.0 [22.9-41.1]</td>
<td>.182 &lt; 10(^{-5})</td>
<td></td>
</tr>
<tr>
<td>geneOf</td>
<td>25.6 [18.8-32.5]</td>
<td>39.0 [29.4-48.6]</td>
<td>17.0 [9.6-24.4]</td>
<td>- .121 &lt; 10(^{-5})</td>
<td></td>
</tr>
<tr>
<td>hasSubtype</td>
<td>7.7 [3.5-11.9]</td>
<td>6.0 [1.3-10.7]</td>
<td>0.0</td>
<td>-.122 &lt; 10(^{-5})</td>
<td></td>
</tr>
<tr>
<td>signOf</td>
<td>6 [0-6]</td>
<td>6.0 [1.3-10.7]</td>
<td>43.0 [33.3-52.7]</td>
<td>.039 .017</td>
<td></td>
</tr>
<tr>
<td>inheritanceOf</td>
<td>0 [0-5]</td>
<td>0 [0-4.7]</td>
<td>5.0 [7-9.3]</td>
<td>.057 .172</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>2.6 [-1.5-0]</td>
<td>2.0 [0-4.7]</td>
<td>2.0 [0-4.7]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Results

There were 62,378,865 pairs of concepts in the ontology, and 36,915 pairs of concepts had at least one relation between them. Specifically in this subset, 15,367 pairs of concepts were present at least once in the corpus of text. Thus, we restricted our experiments to this subset. The Proxima values range from .05 to .73 (mean = .18, sd = .12), with \(\alpha=.73, \beta=.27\) and \(\gamma=0\). The expert has reviewed 356 pairs of concepts and has identified the semantic relation involved in these pairs. 97.8\% (95\% CI= [96.2-99.3]) of these semantic relations were asserted in the ontology. The distribution of the semantic relations varied significantly according to the Proxima values (cf. Figure 2).

The partOf relation was observed for pairs of concepts with high Proxima values. The geneOf relation linked approximately 27\% of the reviewed pairs of concepts, regardless of the Proxima values. The signOf relation linked pairs of concepts with lower Proxima values. Table 1 shows the reviewed concept pairs and the proportion of involved semantic relations. The intensional and expressional components of Proxima

values differed significantly, depending on the semantic relation: partOf, isA and signOf were more intensionally weighted, whereas geneOf and hasSubtype were more expresionally weighted (Table 1).

3. Discussion

Our work on the differences between the intension (OntoOrpha) and expression (OMIM) proximities has shown that, although two concepts can be close in terms of relation, they are not necessarily found together in texts, and vice-versa. Practically, the type of semantic relationship involved affected strongly the occurrence. The partOf relation had high Proxima values with the highest intensional proximity. This observation was consistent with an appropriate representation of groups of disorders in OntoOrpha. In fact, this part of the ontology was derived from Orphanet knowledge bases where a Group of disorders constitutes one level in the Orphanet classifications of rare diseases. The geneOf relation had a wide range of Proxima values with a high expresionial component. Consequently, an adjustment of the intensional component of Proxima was necessary to give an appropriate weight to this relation in the context of rare diseases. The signOf relation had very low Proxima values. Even though the intensional component was higher than the expressional component, this is a line of argument for an evolution of the clinical signs organization in the ontology (which is actually a work in progress in Orphanet) and also for using a dedicated ontology of signs such as the Human Phenotype Ontology (HPO) [6]. This preliminary work opens new perspectives: (1) for medical ontology curation, which can be partly driven by knowledge discovery approaches: using corpora of texts enables the discovery of new relations (with the lowest intensional component), (2) for medical ontology validation: the validation of the ontology to represent a subject domain is best when the coverage quality of the corpora of texts reference is high, and (3) for information retrieval: request extension as well as result filtering and sorting can also be based on proximity calculations.

References