Ontology-based semantic similarity: Formalizing the Superconcept-Based Distance

Montserrat Batet[1]

Universitat Rovira i Virgili
Department of Computer Science and Mathematics
Intelligent Technologies for Advanced Knowledge Acquisition Research Group
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain.
{montserrat.batet}@urv.cat

Abstract. Classical path length based semantic similarities compute the similarity between a pair of concepts of an ontology as the minimum inter-link distance between those concepts, omitting the rest of taxonomical knowledge. In this report, an ontology-based semantic similarity is introduced, which considers the amount of shared and non shared taxonomical knowledge involving a pair of concepts to asses the semantic similarity. In this report it is presented the triangle inequality demonstration of the proposed measure.

1 Introduction

The computation of the semantic similarity/distance between concepts is an active trend. The semantic similarity/distance between a pair of concepts quantifies how words extracted from documents or textual descriptions are alike, and it is based on the estimation of semantic evidence observed in some knowledge source. In that sense, taxonomies and, more generally ontologies are considered as a graph model in which semantic relations are modeled as links between concepts.

In the literature several similarity measures based on the exploitation of the geometrical model of ontologies have been proposed [1–3]. However, those measures are only based on the exploitation of the minimum path length between a pair of concepts losing a lot of taxonomical information.

In general, all those measures are similarities or dissimilarities, and not distances, so very often they will violate triangular inequality, due to the natural implicit uncertainty of linguistic data. In most cases, this question is interesting only from a formal point of view; however, for practical applications in knowledge engineering, the developed approaches do not generally require this constraining property [4]. However, in some specific applications it is required that metrical properties are guaranteed.

Thus, in order to take as much taxonomical information provided by the ontology as possible and to provide a semantic distance that guarantees the metric properties, a new way to compute the semantic distance between concepts in an ontology was proposed [5].

In this report it is presented the demonstration of the triangle inequality property.

1.1 Ontologies

Ontologies have emerged in the last years as a fundamental tool for formalizing and representing domain knowledge. An ontology can be defined as a formal, explicit specification of a shared conceptualization [6]. Ontologies are composed at least by classes (concepts of the domain), relations (different types of binary associations between concepts or data values). Over this knowledge structure, logical axioms (restrictions) can be defined to further describe classes and instances (real world individuals) can be created from class definitions.

Formally, an ontology O is presented as an object model composed by a set of concepts or classes C, which are *taxonomically* related by the transitive is-a relation $H^c \in C \times C$, called concept hierarchy or taxonomy, and *non-taxonomically* related by named object relations $R^* \in C \times C \times$ String.

2 Superconcept Based Distance

Path length-based measures only consider the minimum path between a pair of concepts, omitting the rest of the taxonomical knowledge available in the ontology, wasting a great amount of relevant knowledge.

So, we defined a proposal to compute the distance between a pair of concepts considering overlapping and non-overlapping taxonomical knowledge between concepts. The Superconcept-based Distance is defined as:

Definition: Superconcept-based Distance (SCD)

$$d_{\mathcal{SCD}}(c_i, c_j) = \sqrt{\frac{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)| - |\mathcal{A}(c_i) \cap \mathcal{A}(c_j)|}{|\mathcal{A}(c_i) \cup \mathcal{A}(c_j)|}} \tag{1}$$

where $\mathcal{A}(c_i) = \{c_j \in C | c_j$ is superconcept of $c_i \ \in H^C \vee \ c_i = c_j\}$.

In previous works [5], it was observed that this measure outperforms classical path length based measures.

3 Demonstration of the triangle inequality property

Proof: Demonstrations related to properties identity and symmetry are simple. We can demonstrate the triangle inequality:

Let:

$$A = \{c_j \in C | H^C(c_1, c_j) \vee c_1 = c_j\}$$
$$B = \{c_j \in C | H^C(c_2, c_j) \vee c_2 = c_j\}$$
$$C = \{c_j \in C | H^C(c_3, c_j) \vee c_3 = c_j\}$$

where, $H^C(c_i, c_j)$ means that c_j is a superconcept of c_i.

So the distance between c_1 and c_2 is:

$$d_{SCD}(c_1, c_2) = \sqrt{\frac{|A \cup B| - |A \cap B|}{|A \cup B|}}$$

We demonstrate the triangle inequality by:

$$\sqrt{\frac{|A \cup B| - |A \cap B|}{|A \cup B|}} + \sqrt{\frac{|B \cup C| - |B \cap C|}{|B \cup C|}} \geq \sqrt{\frac{|A \cup C| - |A \cap C|}{|A \cup C|}}$$

Let

$$\begin{aligned} x = |A \cup B| \text{ and } p = |A \cap B| \text{ with } 1 \leq p \leq x \\ y = |B \cup C| \text{ and } q = |B \cap C| \text{ with } 1 \leq q \leq y \\ z = |A \cup C| \text{ and } r = |A \cap C| \text{ with } 1 \leq r \leq z \end{aligned}$$

$p = 1$ when $A \cap B$ =root node of the ontology, $q = 1$ when $B \cap C$ =root node of the ontology, and $r = 1$ when $A \cap C$ =root node of the ontology, and $p = x$ when $c_1 = c_2$, $q = x$ when $c_2 = c_3$, and $r = x$ when $c_1 = c_3$.

So the triangular inequality can be expressed as:

$$\sqrt{\frac{x-p}{x}} + \sqrt{\frac{y-q}{y}} \geq \sqrt{\frac{z-r}{z}}$$

, that is:

$$\frac{\sqrt{x-p}}{\sqrt{x}} + \frac{\sqrt{y-q}}{\sqrt{y}} \geq \frac{\sqrt{z-r}}{\sqrt{z}}$$

$$\frac{\sqrt{(x-p)yz}}{\sqrt{xyz}} + \frac{\sqrt{(y-q)xz}}{\sqrt{xyz}} \geq \frac{\sqrt{(z-r)xy}}{\sqrt{xyz}}$$

This is equivalent to demostrate:

$$\sqrt{(x-p)yz} + \sqrt{(y-q)xz} \geq \sqrt{(z-r)xy}$$

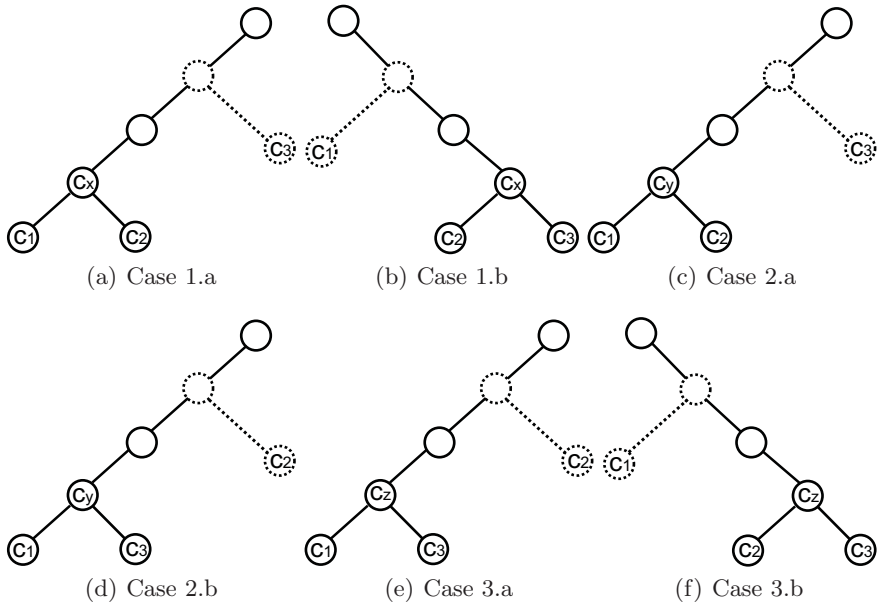
$$\sqrt{xyz - pyz} + \sqrt{xyz - xqz} \geq \sqrt{xyz - xyr}$$

The concept hierarchical structure of an ontology fulfills that:

$$\begin{aligned} A \cap B = B \cap C \text{ or} \\ A \cap B = A \cap C \text{ or} \\ B \cap C = A \cap C \end{aligned} \tag{2}$$

Let to see that. Supose:

1. $A \cap B \neq B \cap C \Rightarrow$ (1.a) $\exists c_x \in A \cap B, c_x \notin B \cap C$ or (1.b) $\exists c_x \notin A \cap B, c_x \in B \cap C$
2. $A \cap B \neq A \cap C \Rightarrow$ (2.a) $\exists c_x \in A \cap B, c_x \notin A \cap C$ or (2.b) $\exists c_x \notin A \cap B, c_x \in A \cap C$
3. $B \cap C \neq A \cap C \Rightarrow$ (3.a) $\exists c_y \in A \cap C, c_y \notin B \cap C$ or (3.b) $\exists c_y \notin A \cap C, c_y \in B \cap C$



(a) Case 1.a (b) Case 1.b (c) Case 2.a

(d) Case 2.b (e) Case 3.a (f) Case 3.b

Fig. 1. Intersection cases

These six possible cases are graphically represented in figure 1. From these figures is easy to see that:

$$\begin{aligned}
\text{If expression 1.a} = true \text{ (Fig. 1(a))} &\Rightarrow \text{expression 3} = false.\\
\text{If expression 1.b} = true \text{ (Fig. 1(b))} &\Rightarrow \text{expression 2} = false.\\
\text{If expression 2.a} = true \text{ (Fig. 1(c))} &\Rightarrow \text{expression 3} = false.\\
\text{If expression 2.b} = true \text{ (Fig. 1(d))} &\Rightarrow \text{expression 1} = false.\\
\text{If expression 3.a} = true \text{ (Fig. 1(e))} &\Rightarrow \text{expression 1} = false.\\
\text{If expression 3.b} = true \text{ (Fig. 1(f))} &\Rightarrow \text{expression 2} = false.
\end{aligned}$$

So, expression 2 is true.

So, let to demonstrate the triangle inequality in those three cases.

1. **Case** $A \cup B = B \cup C$**:**
   If $A \cap B = B \cap C \Rightarrow p = q$ and $p \leq r$

$$\sqrt{xyz - pyz} + \sqrt{xyz - xpz} \geq \sqrt{xyz - xyr}$$

$$(\sqrt{xyz - pyz} + \sqrt{xyz - xpz})^2 \geq (\sqrt{xyz - xyr})^2$$

$$xyz - pyz + xyz - xpz + 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq xyz - xyr$$

$$xyz + xyr - pz(x + y) + 2\sqrt{xyz - pyz}\sqrt{xyz - xpz} \geq 0$$

In this case we can rewrite x, y and z as:

$$\begin{aligned} x &= a+b+p \\ y &= b+c+p \\ z &= a-(r-p)+c-(r-p)+r = a+2p+c-r \end{aligned}$$

,where $a = |A|-p$, $b = |B|-p$, and $c = |C|-p$. Then,

$$\begin{aligned} &a^2b+4abp+2abc-abr+a^2c+4acp+ac^2-acr+a^2p+3ap^2-apr+b^2a+2b^2p+b^2c \\ &\quad -b^2r+4bcp+bc^2-bcr+4bp^2-2bpr+3cp^2+c^2p+-cpr+2p^3-p^2r+abr+acr \\ &\quad +apr+b^2r+bcr+2bpr+cpr+p^2r-a^2p-2abp-4ap^2-2acp-4bp^2-4p^3-4cp^2 \\ &\quad -2bcp-c^2p+apr+2bpr+2p^2r+cpr+2\sqrt{xyz-pyz}\sqrt{xyz-xpz} \geq 0 \end{aligned}$$

$$\begin{aligned} &a^2b+2abp+2abc+a^2c+2acp+ac^2+b^2a+2b^2p+b^2c+2bcp+bc^2+cpr+2p^2r-ap^2 \\ &\quad -2p^3-cp^2+apr+2bpr+2\sqrt{xyz-pyz}\sqrt{xyz-xpz} \geq 0 \end{aligned}$$

Taking in account that $p \leq r$:

$$\begin{aligned} &a^2b+2abp+2abc+a^2c+2acp+ac^2+b^2a+2b^2p+b^2c+2bcp+bc^2+\underbrace{2p^2r-2p^3}_{\geq 0} \\ &\quad +\underbrace{apr-ap^2}_{\geq 0}+\underbrace{cpr-cp^2}_{\geq 0}+2bpr+2\sqrt{\underbrace{xyz-pyz}_{\geq 0}}\sqrt{\underbrace{xyz-xpz}_{\geq 0}} \geq 0 \end{aligned}$$

So, the triangle inequality is demonstrated when $A \cap B = B \cap C$.

2. Case $A \cap B = A \cap C$: If $A \cup B = A \cup C \Rightarrow p = r$ and $p \leq q$.

$$\sqrt{xyz-pyz}+\sqrt{xyz-xqz} \geq \sqrt{xyz-xyp}$$

$$xyz-pyz+xyz-xqz+2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq xyz-xyp$$

$$xyz+yp(x-z)-xqz+2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq 0$$

In this case we can rewrite x, y and z as:

$$\begin{aligned} x &= a+b+p \\ y &= b-(q-p)+c-(q-p)+q = b+2p+c-q \\ z &= a+c+p \end{aligned}$$

,where $a = |A|-p$, $b = |B|-p$, and $c = |C|-p$. Then,

$$\begin{aligned} &a^2b+2a^2p+a^2c-a^2q+2abc+4acp+ac^2-acq+4abp+4ap^2-2apq+ab^2-abq+b^2c \\ &\quad +4bcp+bc^2-bcq+b^2p+3bp^2-bpq+3cp^2+c^2p-cpq+2p^3-p^2q+b^2p+2bp^2 \\ &\quad -2cp^2-c^2p-bpq+cpq-a^2q-acq-2apq-abq-bcq-bpq-cpq-p^2q \\ &\quad +2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq 0 \end{aligned}$$

Taking in account that $q \leq c+p$ and $q \leq b+p$:

$$\underbrace{a^2p+a^2b-a^2q}_{\geq 0}+\underbrace{a^2p+a^2c-a^2q}_{\geq 0}+\underbrace{acp+ac^2-acq}_{\geq 0}+\underbrace{acp+abc-acq}_{\geq 0}$$
$$+\underbrace{2ap^2+2acp-2apq}_{\geq 0}+\underbrace{2ap^2+2abp-2apq}_{\geq 0}+\underbrace{abp+ab^2-abq}_{\geq 0}+\underbrace{abp+abc-abq}_{\geq 0}$$
$$+\underbrace{bcp+b^2c-bcq}_{\geq 0}+\underbrace{bcp+bc^2-bcq}_{\geq 0}+\underbrace{2bp^2+2b^2p-2bpq}_{\geq 0}+\underbrace{bp^2+bcp-bpq}_{\geq 0}$$
$$+\underbrace{2p^3+2bp^2-2p^2q}_{\geq 0}+\underbrace{cp^2+bcp^2-cpq}_{\geq 0}+2\sqrt{\underbrace{xyz-pyz}_{\geq 0}}\sqrt{\underbrace{xyz-xqz}_{\geq 0}}\geq 0$$

So, the triangle inequality is demonstrated when $|A \cup B| = |A \cup C|$.

3. Case $B \cap C = A \cap C$:

f $B \cap C = A \cap C \Rightarrow q = r$ and $q \leq p$.

$$\sqrt{xyz-pyz}+\sqrt{xyz-xqz} \geq \sqrt{xyz-xyq}$$
$$xyz-pyz+xyz-xqz+2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq xyz-xyq$$
$$xyz+xq(y-z)-pyz+2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq 0$$

In this case we can rewrite x, y and z as:

$$x = a-(p-q)+b-(p-q)+p = a+b+2q-p$$
$$y = b+c+q$$
$$z = a+c+q$$

,where $a = |A|-p$, $b = |B|-p$, and $c = |C|-p$ and $q \leq p$. Then,

$$a^2b+ab^2+2abq-abp+2abc+b^2c+4bcq-bcp+2abq+b^2q+3bq^2-bqp+a^2c+4acq$$
$$-acp+ac^2+bc^2+2c^2q-c^2p+4cq^2-2cqp+a^2q+3aq^2-aqp$$
$$+2q^3-q^2p-a^2q+b^2q+2bq^2-2aq^2$$
$$-bqp+aqp-abp-bcp-bqp-acp-c^2p-2cqp-aqp-q^2p$$
$$+2\sqrt{xyz-pyz}\sqrt{xyz-xqz} \geq 0$$

Taking in account that $p \leq a+q$ and $p \leq b+q$:

$$\underbrace{a^2b+abq-abp}_{\geq 0}+\underbrace{ab^2+abq-abp}_{\geq 0}+\underbrace{abc+bcq-bcq}_{\geq 0}+\underbrace{b^2c+bcq-bcp}_{\geq 0}+\underbrace{2b^2q+2bq^2-2bqp}_{\geq 0}$$
$$+\underbrace{abq+bq^2-bqp}_{\geq 0}+\underbrace{a^2c+acq-acp}_{\geq 0}+\underbrace{abc+acq-acp}_{\geq 0}+\underbrace{ac^2+c^2q-c^2p}_{\geq 0}+\underbrace{bc^2+c^2q-c^2p}_{\geq 0}$$
$$+\underbrace{2acq+2cq^2-2cqp}_{\geq 0}+\underbrace{2bcq^2+2cq^2-2cqp}_{\geq 0}+\underbrace{abq+aq^2-aqp}_{\geq 0}+\underbrace{2bq^2+2q^3-2q^2p}_{\geq 0}$$
$$+2\sqrt{\underbrace{xyz-pyz}_{\geq 0}}\sqrt{\underbrace{xyz-xqz}_{\geq 0}}\geq 0$$

So, the triangle inequality is demonstrated when $|B \cap C| = |A \cap C|$.

Acknowledgments Montserrat Batet would acknowledge the help of Dr. Juan A. Rodríguez.

References

1. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd annual Meeting of the Association for Computational Linguistics, New Mexico, USA, Association for Computational Linguistics (1994) 133 –138
2. Rada, R., Mili, H., Bichnell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics **9(1)** (1989) 17–30
3. Leacock, C., Chodorow, M.: WordNet: An electronic lexical database. In: Combining local context and WordNet similarity for word sense identification. MIT Press (1998) 265–283
4. Blanchard, E., Harzallah, M., Kuntz, P.: A generic framework for comparing semantic similarities on a subsumption hierarchy. In Ghallab, M., Spyropoulos, C.D., Fakotakis, N., Avouris, N.M., eds.: Procedeings of 18th European Conference on Artificial Intelligence (ECAI). Volume 178., Patras, Greece, IOS Press (July 21-25 2008) 20–24
5. Batet, M., Sánchez, D., Valls, A., Gibert, K.: Ontology-based semantic similarity in the biomedical domain. In: 12th Conference on Artificial Intelligence in Medicine (AIME'09). WS. Intelligent Data Analysis in Biomedicine and Pharmacology, Italy (2009)
6. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. IEEE Transactions on Data and Knowledge Engineering **25(1-2)**(1-2) (1998) 161–197