From Theoretical Framework to Generic Semantic Measures Library

Sébastien Harispe, Stefan Janaqi, Sylvie Ranwez, and Jacky Montmain

LGI2P, Ecole des mines d'Alès, Parc Scientifique G. Besse, F-30035 Nîmes Cedex 1
firstname.name@mines-ales.fr

Abstract. Semantic Measures (SMs) are of critical importance in multiple treatments relying on ontologies. However, the improvement and use of SMs are currently hampered by the lack of a dedicated theoretical framework and an extensive generic software solution. To meet these needs, this paper introduces a unified theoretical framework of graph-based SMs, from which we developed the open source Semantic Measures Library and toolkit, a solution that paves the way for design, computation and analysis of SMs. More information at dedicated website: http://www.semantic-measures-library.org.

Keywords: Semantic Similarity Measures, Unifying Framework, Semantic Measures Library, Software for Semantic Measure Computation.

1 Introduction

Numerous initiatives take advantage of ontologies to characterize entities, such as scientific publications, people or genes. To exploit such knowledge for information retrieval and knowledge discovery, inexact searches have to be performed, which require measures that are capable of assessing the degree of likeness of entities. Semantic measures (SMs) are designed for this purpose: to capture the relatedness of concepts (and by extension semantically characterized entities), by taking into account the semantic space in which they are defined. This paper explores graph-based SMs, here denoted SMs for convenience.

Study of SMs involves various communities which have designed numerous measures for specific treatments and ontologies [1, 2]. However, no unifying theoretical framework has been proposed to explicitly characterize core elements of SMs through the spectrum of similarity and graph theory. Indeed, given the lack of contributions focusing on theoretical aspects of SMs, studies are often restricted to a specific community and do not benefit to a broader target audience.

Likewise, most of software dedicated to SMs are developed for particular ontologies (e.g. WordNet, UMLS, Gene Ontology, MeSH, Disease Ontology [3–6]). This diversity of software solutions has drawbacks given that SM evaluations rely exclusively on empirical studies. Therefore, the lack of an extensive software solution slows down studies of SMs as well as the sharing of new findings related to the domain.

Y.T. Demey and H. Panetto (Eds.): OTM 2013 Workshops, LNCS 8186, pp. 739–742, 2013.

This paper proposes a unified theoretical framework and a generic software solution to mutualize efforts and advance the increasing role of SMs in numerous fields. The second section introduces the framework which defines a reduced set of abstract primitive functions that are commonly used for SM design. The third section presents the Semantic Measures Library, an extensive open source library dedicated to the computation, development and analysis of SMs.

2 A Unifying Theoretical Framework for Semantic Measures

Most SMs are expressed reducing an ontology to a graph $G = (C, E, R)$, with C being the set of classes represented as vertices and E the set of edges representing the oriented pairwise relationships defined between two classes. G is assumed to contain a taxonomical sub-graph defining a partial order $\preccurlyeq$ on C with $x \preccurlyeq y$ if x is a subclass of y. Moreover, classical knowledge bases are composed of a collection of entities which are conceptually characterized through classes, e.g. genes annotated by concepts [1] or PubMed articles indexed by MeSH descriptors. SMs used to compare entities are extension of those designed to compare pairs of classes. We therefore focus on SMs estimating the similarity of pairs of classes, that is to say a function $\mu: C \times C \rightarrow \mathbb{R}^+$. Formulas of existing measures proposed in the literature are not presented, we orient the reader to the numerous surveys for detailed presentations [1, 2].

This section defines a unifying theoretical framework aiming to: (i) better characterize SMs at a theoretical level through the definition of the primitive abstract functions on which SMs rely, (ii) distinguish common and relevant features of SMs, (iii) analyse measures' characteristics and properties, (iv) design and optimize SMs in accordance with usage contexts. Presented results expand previous works on the characterization of SMs through abstract formulations [3, 7–9]. Their main goal was to establish links between various models of semantic similarity rather than distinguishing the primitive functions governing measures design. Our contribution focuses mainly on the definition of those primitive functions. It is therefore not constrained to a specific model of similarity (e.g. feature, information theory, geometric and transformation models).

The set of primitive functions which can be used to express most abstract formulations of SMs is:

- $\rho(a)$: semantic representation of a class a (denoted $\tilde{a}$). With $\mathbb{K}$ a set composed of paths, classes or any subset of G, we define $\rho: C \rightarrow \mathbb{K}$.
- $\Theta(\tilde{a})$: salience of the representation of a class, i.e. the amount of information carried by a semantic representation $\tilde{a}$, $\Theta: \mathbb{K} \rightarrow \mathbb{R}^+$.
- $\Psi(\tilde{a}, \tilde{b})$: commonalities of the pair $(\tilde{a}, \tilde{b})$, $\Psi: \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$.
- $\Phi(\tilde{u}, \tilde{v})$: differences of the pair $(\tilde{a}, \tilde{b})$, $\Phi: \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$.
- $\zeta(\tilde{a}, \tilde{b})$: amount of information defined in the semantic space, which is not found in the couple of representations $(\tilde{a}, \tilde{b})$, $\zeta: \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}^+$.

Once the representation of a class and the corresponding operators used to estimate the commonalities and differences are defined, some abstract functions estimating similarity can be used. For instance, the similarity can be assessed based on an abstract expression of the *ratio model* (Sim_{RM}), initially proposed by Tversky to compare semantic elements represented as sets of features [9]:

$$Sim_{RM}(a,b) = \frac{\Psi(\tilde{a},\tilde{b})}{\alpha \cdot \Phi(\tilde{a},\tilde{b}) + \beta \cdot \Phi(\tilde{b},\tilde{a}) + \Psi(\tilde{a},\tilde{b})}$$

with $\tilde{a} = \rho(a)$, α and β two parameters defining the asymmetrical contribution of the operator Φ.

The unifying framework is not constrained to a specific measure but rather relies on primitive functions commonly used to define SMs. As an example, numerous similarity measure expressions rely on a function Θ estimating the amount of information carried by the representation of an element, e.g. the cardinality when compared elements are represented by sets. Such a function Θ enables the use of the abstracted form of two general expressions unifying binary measures: Gower & Legendre and Caillez & Kuntz parameterized measures [8]. Those abstract measures, among others, can be used to derive most of SMs through specific expressions of the abstract functions on which they rely $(\rho, \Theta, \Psi, \Phi)$ [7, 8].

To define a concrete measure, SM designers should specify the primitive functions distinguished by the framework. As an example: (i) representing a class by the set of its subclasses, $\rho(b) = A(b)$; (ii) defining the commonality classes as a function of the Information Content (IC) carried by their Most Informative Common Ancestor (MICA), $\Psi(\tilde{u},\tilde{v}) = IC(MICA_{u,v})$; (iii) evaluating the difference of two classes by $\Phi(\tilde{u},\tilde{v}) = IC(u) - IC(MICA_{u,v})$; (iv) setting Sim_{RM} with α and β equal to 1 corresponds to the measure proposed by Pirró & Euzenat [3].

Research on SMs is driven by a cycle consisting of three aspects: theoretical studies, software development and empirical studies. Another important aspect for SM studies is therefore to provide ways to take advantage of theoretical contributions through software solutions.

3 The Semantic Measures Library

The Semantic Measures Library (SML) is a generic software solution dedicated to SM computation, analysis and development. By generic, we mean that the SML can be used to take advantage of SMs on a wide range of semantic graphs. Following the abstraction layers defined in the previous section, a large collection of SMs has been implemented. Moreover, thanks to the code base available (e.g. graph algorithms), new measures can easily be designed and evaluated. The SML is written in JAVA and open-sourced under the GPL-compatible CeCILL licence. In addition, the SML-Toolkit enables non-developers to benefit from functionalities provided by the SML through command-line interfaces. Documentation, tutorials and downloads of both the SML and the toolkit are available at http://www.semantic-measures-library.org.

4 Conclusion

This paper defines a unifying theoretical framework for SMs. By underlining close relationships between measures, we distinguish the primitive functions required for their design. Besides facilitating the definition of countless new measures, this new insight allows us to study SMs which are grouped in accordance with the expressions of the abstract functions they rely on. This leads us to draw interesting perspectives regarding the theoretical studies of SMs.

To address the handicapping lack of a generic, extensive and efficient software solutions dedicated to SMs, we developed the open source Semantic Measures Library (SML) and associated toolkit. The SML, both highly tuneable and not limited to a particular semantic graph, is suited for diverse application contexts. Downloads and documentation are available at http://www.semantic-measures-library.org.

References

1. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS Computational Biology 5(7), e1000443 (2009)
2. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40(3), 288–299 (2007)
3. Pirró, G., Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 615–630. Springer, Heidelberg (2010)
4. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet:Similarity: measuring the relatedness of concepts. In: Proceedings of HLT-NAACL, Demonstration Papers, pp. 38–41 (2004)
5. Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., Wang, S.: GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics (Oxford, England) 26(7), 976–978 (2010)
6. Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., Li, X.: DOSim: An R package for similarity between diseases based on Disease Ontology. BMC Bioinformatics 12, 266 (2011)
7. Sánchez, D., Batet, M.: Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics 44(5), 749–759 (2011)
8. Blanchard, E., Harzallah, M., Kuntz, P.: A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Proceedings of 18th European Conference on Artificial Intelligence, pp. 20–24 (2008)
9. Cross, V., Yu, X., Hu, X.: Unifying ontological similarity measures: A theoretical and empirical investigation. International Journal of Approximate Reasoning 54(7), 861–875 (2013)