A COMBINED FUZZY SEMANTIC SIMILARITY MEASURE IN OWL ONTOLOGIES

Vincenzo Cannella, Giuseppe Russo, Pierluca Sangiorgi and Roberto Pirrone
Universita' degli Studi Palermo, Dipartimento Ingegneria Informatica - DINFO
Viale delle Scienze ed. 6 p.3, 90128 Palermo, Italy

Keywords: Semantic Similarity Measure, OWL Ontologies, Semantic Web, Fuzzy Measure.

Abstract: An algorithm is presented in this paper to calculate a semantic similarity measure inside an OWL ontology. The formulation is based on a combined measure taking into account the two most important aspects involved in the similarity computation. These are the structural properties of a concept, and the information content inside the ontology. We define a fuzzy system to blend these information sources with a training process over some ontologies. Finding a similarity measure between concepts of an ontology is a fundamental topic to accomplish information exchange on the Web. Through this measure it is possible to perform sophisticated queries over the web where the user is able to request concepts with a predefined similarity (or even dissimilarity) degree.

1 INTRODUCTION

The definition of a measure for semantic similarity is a very important task to accomplish in many processes such as clustering and data mining, database schema mapping, word sense disambiguation, information indexing and information filtering. Estimation of semantic similarity measure in the same ontology (Gruber, 1993) or between distinct ontologies is a central need for the processes involving the information exchange over the Web as defined in the cornerstone article from Tim Berneers-Lee about the Semantic Web (Berners-Lee and Lassila, 2001) (Berners-Lee et al., 2006). This work deals with a possible definition of a semantic similarity measure inside an ontology described through a OWL-DL file (Helfin, 2004). OWL is the W3C standard for ontology representation and is widely used in most applications over the web.

Defining a semantic similarity measure requires a first agreement about the meaning of the term "similarity". In many works such as (Resnik, 1995) a terminological clarification has been proposed. The more used terms in the literature are *similarity*, *relatedness* and *distance*. The distinction between *similarity* and *relatedness* regards the use of a functional relation between concepts. Two concepts are *related* if there is a functional relation connecting them in some way. As an example the subsumption relation (class-instance) is a functional relation, and so is the meronymy relation or whatever relation a user can define inside a particular domain of interest. The *distance* term is intended as the opposite of *similarity*. Semantic distance, however, could be used with respect to distance between related concepts and distance between similar concepts. The most used term in literature is *similarity* but in this paper we refer to this term in a broader way that is close to *relatedness*. The proposed approach for this work is the definition and the implementation of a combined semantic similarity measure to be computed on OWL-Dl ontologies. The proposed measure is based on ontology structure and on the information provided by the attributes and the relations that are defined inside the ontology. For our purposes similarity is defined as follows: The more two concepts are close in terms of their structural properties and are correlated, the more they are similar. The measure is a parametric one. Parameters tuning is achieved by a fuzzy algorithm that mixes the components of the measure definition.

The rest of the paper is arranged as follows: in the next paragraph some related works are presented. Third paragraph presents the proposed solution and the algorithm. Next, some experimental results are reported. Finally, some conclusions and future work are presented.

Cannella V., Russo G., Sangiorgi P. and Pirrone R. (2008).
A COMBINED FUZZY SEMANTIC SIMILARITY MEASURE IN OWL ONTOLOGIES.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 181-186
DOI: 10.5220/0001522801810186

2 RELATED WORKS

Many approaches have been proposed in literature to define a measure for semantic similarity between concepts. There are two fundamental categories of measures. The first one is based on the topological distance between concepts in the ontology graph, while the second one makes use of their information content. Some other approaches try to combine these two aspects.

2.1 Graph Distance Models

A natural approach to define a measure in an ontology is to use the distance between concepts because an ontology is a graph where the edges represent relations between concepts and the nodes are the concepts themselves. An earlier example of this measure was proposed by (Rada et al., 1989) where the distance between two concepts c_i and c_j is defined as

$$dist(c_i,c_j) = \min_{p \in path(c_i,c_j)} len_e(p). \quad (1)$$

where $path(c_i,c_j)$ defines a possible path connecting c_i and c_j. Starting from the Rada's definition a new measure is defined in (Leacock and Chodorow, 1998) where some adjustments are adopted. The major problem of this approach is related to the consideration that all the edges in the graph represent uniform distances. This assumption is trivially wrong when dealing with an ontology where the edges along the path between two concepts can represent different levels of generalization. So they cannot be assumed to have the same weight in computing the distance. Concepts with different depth with respect to the root node are differently related. The more they are general the more they are distant. A natural solution is to weight the path with some function that takes into account also the depth level of the concept. In (Wu and Palmer, 1995) a measure is presented as the sum of two different values.

$$sim(c_i,c_j) = \frac{2 * \min len_e(p_{comm})}{\min len_e(p) + 2 \min len_e(p_{comm})}. \quad (2)$$

where $len_e(p_{comm})$ is an intermediate distance defined starting from the most common subsumer of the two nodes. Most recent approaches like (Castano et al., 2004) concentrate their focus on the determination of the weights related to the relation with neighbor nodes.

2.2 Information Theory Models

The models inspired to information theory require some additional information to define the similarity measure. Usually the information is given by a corpus of documents related to the ontology. In (Resnik, 1995) a measure is defined on the following hypothesis: the more two concepts share common information, the more they are similar. The formulation for the similarity is the following:

$$sim(c_i,c_j) = \max_{c \in S(c_i,c_j)} [-\log p_c]. \quad (3)$$

Where S is the set of concepts that subsumes both c_i and c_j. The frequency p_c of the concept inside the corpus, gives the added information value. In (Jiang and Conrath, 1997) the distance between two concepts is computed as the difference between the sum of the information content of the two concepts and the information content of their most specific common subsumer. In (Lin, 1999) a similarity that takes into account the information shared by two concepts, like Resnik, but also the difference between them is proposed.

2.3 Combined Models

Another possible approach is to combine the previously presented techniques, to gain the benefits provided by both the models. The structural approach for ontologies expressed in OWL is restrictive because an OWL file contains many information about classes, attributes and relations that can be used. Furthermore a combined approach can overcome some problems due to the facts that in an ontology both structural and information based contents are relevant. This approach seems to be very promising and is the one used in our work. In (Nguyen and Al-Mubaid, 2006) a new measure for similarity is defined, which uses a new feature called *common specificity* (CSpec) besides the path length feature. The *CSpec* feature is derived from the information content obtained from single concepts and the overall ontology, given a related document corpus. The formulation is:

$$CSpec(c_i,c_j) = IC_{max} - IC(LCS(c_i,c_j)). \quad (4)$$

where IC_{max} is the maximum IC (ontology information content) of concept nodes in the ontology and $LCS(c_i,c_j)$ is the least common subsumer of c_i and c_j.
In (Wang et al., 2006) the HIC-AIC algorithm to compute similarity is presented, which is based on hierarchal information content (HIC) and attributes information content (AIC). This approach defines the similarity measure as the sum of the two values starting from an useful ontology model definition that we have extended.

3 PROPOSED SOLUTION

3.1 Ontology Definition

As previously stated, the structure-based and information-based methods for similarity calculation depend on the ontology structure and neglect many information such as relations and attributes. In the following an approach based on hierarchical content and behavioral information content is presented. The used ontology model definition that is suitable for our purposes is a 6-tuple defined as: $O = \{C, R, A, H^c, att, rel\}$ where:

- C: concepts set.
- R: relations set.
- A: attributes set.
- H^c: hierarchy structure, is a particular relation called concept taxonomy. $H^c(c_1, c_2)$ means that c_2 is a sub-concept of c_1.
- *att*: the concept-attribute function. It relates the concepts of set C and the attributes of set A. $att(c_1, a_1)$ means that c_1 has an a_1 attribute.
- *rel*: the concept-relation function. It relates the concepts of set C and the relations of set R. $rel(c_1, r_1)$ means that c_1 has a r_1 relation that connect it with other concepts.

In this model only the functional part of ontologies is explicitly stated, while the descriptive one is discarded. This is formed by the lexicon of concepts, the relations and the mapping functions of the functional and descriptive parts. The purpose is to have a clear description and to focus the attention to really significant terms to be used in practice.

3.2 Assumptions and Definitions

The proposed solution is based on some reasonable assumptions about the domain that are summarized in the following. The definitions are reported to help the explanation of some crucial aspects of the solution.
Assumption 1: Given two concepts c_1 and c_2, if a_1 is an attribute of c_1 and c_2 is a sub-concept of c_1, then a_1 is an attribute of c_2. In other words, a sub-concept inherits all the attributes of his father node.
Assumption 2: Given two concepts c_1 and c_2, if c_1 has a r_1 relation and c_2 is a sub-concept of c_1, then c_2 has a r_1 relation. In other words, a sub-concept inherits all the direct relations of his father node.
Definition 1: Ancestor, denoted as $Ancestor(c_1, c_2)$. Given three concepts c_1, c_2 and c_3, if c_2 is a sub-concept of c_1 or c_3 is a sub-concept of c_1 and c_2 is a sub-concept of c_3 or c_3 is a sub-concept of c_1 and c_3 is an ancestor of c_2 then c_1 is an ancestor of c_2.
Definition 2: Most Recent Common Ancestor, denoted as $MRCA(c_1, c_2)$. Given two concepts c_1 and c_2, the most recent common ancestor of c_1 and c_2 is the concept c which is ancestor of both c_1 and c_2 and has the minimum number of ancestors of c_1 and c_2 in its descendants nodes.
Definition 3: Common Attribute, denoted as $CA(a_1, c_1, c_2)$. Given two concepts c_1 e c_2, if a_1 is an attribute of c_1 and c_2, then a_1 is a common attribute of c_1 and c_2.
Definition 4: Common Relation, denoted as $CR(r_1, c_1, c_2)$. Given two concepts c_1 and c_2, if r_1 is a relation of c_1 and c_2, then r_1 is a common relation of c_1 and c_2.
Definition 5: Not-Inherited Common Attribute, denoted as $NICA(a_1, c_1, c_2)$. Given two concepts c_1 and c_2, if a_1 is a common attribute of c_1 and c_2, and a_1 isn't an attribute of the most recent common ancestor of c_1 and c_2, then a_1 is a not-inherited common attribute of c_1 and c_2.
Definition 6: First Child of Most Recent Common Ancestor, denoted as $FCM(c_x, c_1, c_2)$. Denotes the first concept c_x to cross starting from the Most Recent Common Ancestor(c_1, c_2) to reach the concept c_1 through the shortest path between the concepts c_1 and c_2. $FCM(c_y, c_1, c_2)$ denotes the other concept child c_y to cross reaching the concept c_2.

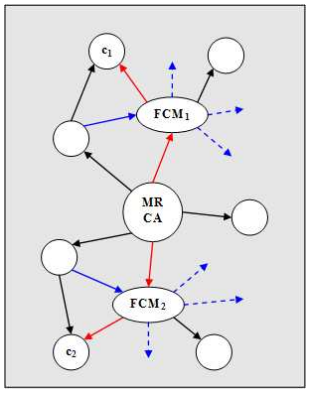

Figure 1: A typical ontology fragment.

Figure 1 shows a fragment of an ontology where it is possible to see the two concepts to compare, the most recent common ancestor, his first child, the shortest path between c_1 and c_2 through is-a relations (highlighted in red) and the relations of the two children to be compared (highlighted in blue). Starting from these definitions we are able to formalize our algorithm for the semantic similarity measure.

3.3 Proposed Algorithm

The proposed method is based on the assumption that "two concepts are much more similar how less distant and more correlated among them they are". This simple definition makes intuitive the use of a fuzzy system (Zadeh, 1992) to produce a similarity measure of two concepts as a function of distance and correlation. The measure will range in $[0,1]$.
The distance component is defined by the following:

$$dist(c_1,c_2) = ShortestWeightedPath(c_1,c_2). \quad (5)$$

This term is based on a structural approach (Graph Distance Model). It denotes the distance between two ontology concepts. It's expressed as the minimum weighted sum of relations crossed to reach c_2 from c_1.
The correlation component is given by:

$$corr(c_1,c_2) = \left(1+\frac{N_i}{T_a}\right) * \left(1+\frac{N_r}{T_r.}\right) \quad (6)$$

where N_i is the number of NICAs of c_1 and c_2, T_a is the total number of attributes of c_1 and c_2, N_r is the number of the CRs of FCMs of c_1 and c_2, T_r is the total number of relations of the FCMs of c_1 and c_2. This term is based on a behavioral approach and denotes the correlation between two ontology concepts in terms of their specialization considering the functional aspect given by attributes and relations. In our scenario, we consider that two concepts are correlated, in a behavioral sense, "if they are able to express something similar". The expression in (6) is made by two gain terms. It means that correlation increases if c_1 and c_2 share many attributes not inherited from their MRCAs and if their ancestors share many relations. This is a global measure of how c_1 and c_2 carry a similar information content. The relations to consider in this formula are only those giving an informative contribution, while structural information (like is-a, subclass-of, etc) is neglected. To proceed in measuring similarity, we have to estimate at first the relevant terms like ancestors, common and not common attributes, common and not common relations. In this way, we are able to measure Distance and Correlation. These two values are used as crisp inputs in a fuzzy system (see Fig 2). We defined the fuzzy sets for the inputs and the similarity output using the following fuzzy rules:

- IF closed AND correlated THEN very similar
- IF closed AND average correlated THEN similar
- IF closed AND not correlated THEN average similar
- IF far AND not correlated THEN very dissimilar
- IF far AND average correlated THEN dissimilar
- IF far AND correlated THEN average dissimilar
- IF average closed AND correlated THEN similar
- IF average closed AND average correlated THEN average similar
- IF average closed AND not correlated THEN average dissimilar

We used Gaussian membership functions for the inputs, while the output has triangular one. Crisp similarity is obtained from the output fuzzy set using the centroid rule.

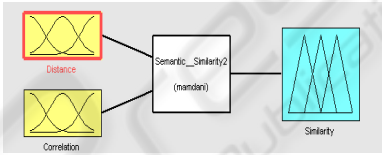

Figure 2: The Fuzzy System.

The proposed algorithm is based on measure combining the well known structural approach, the Shortest Weighted Path, and an information-based one, which relies on behavioral considerations about the concepts in the ontology. In particular, the second term of the measure estimates the behavioral specialization of each node inside the ontology. This result considering the analysis of attributes and relations. Differently from other related works, relations inside the ontology have particular relevance for the algorithm. Relations are able to provide information regarding not only the ontology structure, but also the contents the ontology deals with.

3.4 Experimental Results

Before performing experiments the fuzzy system has to be trained to tune the parameters of the membership functions belonging to the fuzzy sets. For the training phase we used some ontologies taken from a selection of OWL ontologies (Protege', 2004). Such ontologies have been specifically developed for the Semantic Web and provided by the Protege' team of the Stanford Medical Informatics at the Stanford University School of Medicine. We have trained our fuzzy system using the Matlab Fis Editor module to check our fuzzy rules and to find the appropriate membership functions for the two input variables Distance and Correlation and the output variable Similarity. Figures 3, 4, 5 show the membership functions adopted for these three values.

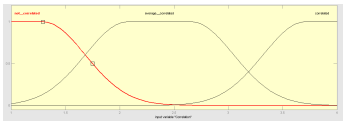

Figure 3: Correlation.

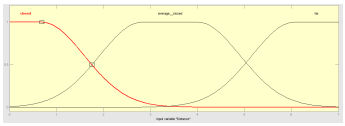

Figure 4: Distance.

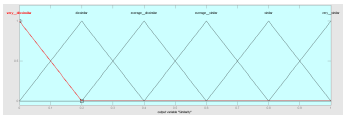

Figure 5: Similarity.

We did not manipulate the output membership functions for Similarity. We used standard triangular functions equally distributed over the whole range. This prevents the system from over-fitting particular ontological structures. In this way we let the system free to adapt to those cases when the user wants to enhance the contribution of a single input variable (Correlation or Distance) according to the arrangement of the ontology under investigation. Figure 6 shows the surface describing the Similarity distribution, which reflects the adoption of equally spaced membership functions. We tested the method on an

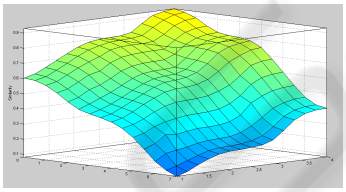

Figure 6: Surface Similarity Distribution.

ontology about the tourism domain contributed by Holger Knublauch. This ontology is composed by 35 concepts, 21 attributes of 6 different types, 20 relations and its graphical representation is shown in figure 7.

Finally, we have compared the results with Resnik's algorithm and Lin's algorithm using respectively the equation (3) and the equation (7) below.

$$sim(c_1,c_2) = \frac{2 * \log(P(C))}{\log(P(C_1)) + \log(P(C_2))}. \qquad (7)$$

In the literature these measures usually includes a concept when computing its offspring. We excluded the parent from the computation of its offspring due to the use of OWL ontologies where the super-concept `owl:Thing` is defined by default. The results are shown in the following table:

Table 1: Experimental Results.

Concept 1 - Concept 2	R. Sim	L. Sim	F. C. Sim
Capital - National Park	0.640	0.640	0.404
Bunjee Jumping - Surf	0.502	0.502	0.454
Destination - Acc.	0.012	0.015	0.490
Museum - Town	0.012	0.012	0.195
Beach - City	0.640	0.503	0.511

The obtained values need to be explained in detail because of the different interpretation with respect to the other methods. This approach is able to carry out the semantic aspect considering the domain of the ontology and making a measure related to the context.

Let's consider the result obtained in semantic similarity between Accommodation and Destination concepts. Generally, we can say that this two terms are dissimilar and traditional methods confirm that, but if we consider this terms in an appropriate context, they could be semantically similar (related). In our experimental domain about tourism they share the relation hasAccommodation - hasDestination and result more similar respect to traditional methods that neglect the relations. We propose this as a better approach to Semantic Web technologies in System Integration, where using domain aspects is a crucial topic for a good optimization of the systems. The evaluation of the distribution of semantic similarity values in the surface (see Fig. 6) gives the possibility to set many different thresholds to discriminate the similar concepts to be presented to users. The measure is not linear and the attention is focused on [0.4 0.7] range. As an example the value 0.404 between Capital and National Park in the domain is just over the minimum range acceptable (0.4) if we want to present more additional information. The value 0.454 between Bunjee-Jumping and Surf like the value 0.511 for Beach-City are acceptable and are the result of the improvement given by the combined measure. The first couple is closer so the structural contribution gives an higher similarity value while the second couple shares a relation. The obtained surface makes clear that for our purposes it is possible to add additional and pertinent results, related to context as decreasing rating to end users that make queries in these systems.

3.5 Conclusions and Future Work

In this paper an algorithm to calculate a semantic similarity measure inside an OWL ontology was presented. This similarity method is based on a fuzzy

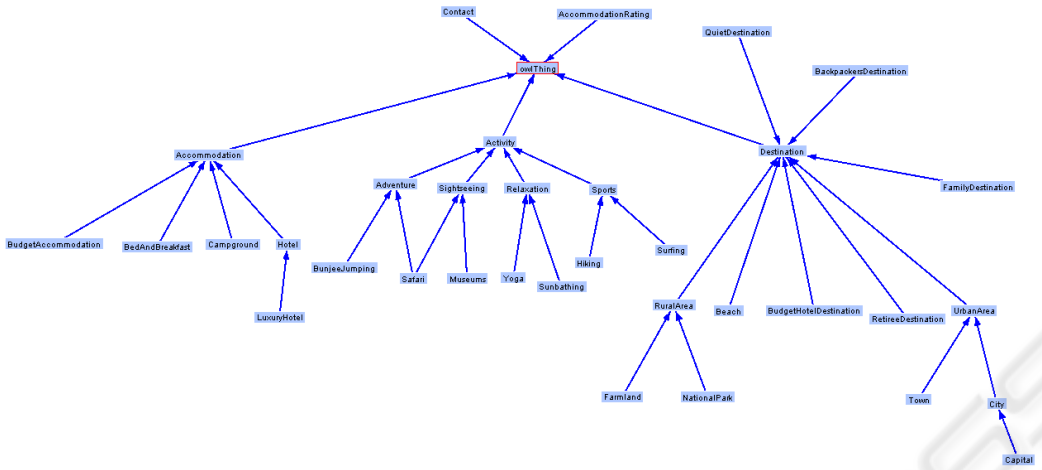

Figure 7: A view of the ontology.

mix of a behavioral information approach and a structural one. The proposed measure results efficient when it is used in ontologies with many relations apart of the structural ones (is_a, part_of, etc.). Thanks to relations, is possible to retrieve significant results for the used domain, otherwise neglected by other methods. Our future work will provide an extension in the similarity calculation using all the properties of the OWL ontologies like the definition of union, intersection, restriction and all the other properties provided from the OWL standard. We are also trying tho integrate the proposed algorithm in a complete system at the presentation level to obtain replies that are richer than in the typical SPARQL environments.

REFERENCES

Berners-Lee, T. and Lassila, O. (2001). The semantic web. In *Scientific American*.

Berners-Lee, T., Shadbolt, N., and Hall, W. (2006). The semantic web revisited. In *IEEE Intelligent Systems*.

Castano, S., Ferrara, A., Montanelli, S., and Racca, G. (2004). Semantic information interoperability in open networked systems. In *Proc. of the Int. Conference on Semantics of a Networked World (ICSNW), in cooperation with ACM SIGMOD 2004*.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. In *Knowledge Acquisition*.

Helfin, J. (2004). Web ontology language (owl): Use cases and requirements. In *W3C Recommendation*.

Jiang, J. J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*.

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In *WordNet: An electronic lexical database*. MIT Press.

Lin, D. (1999). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann.

Nguyen, H. and Al-Mubaid, H. (2006). A combination-based semantic similarity measure using multiple information sources. In *IEEE International Conference on Information Reuse and Integration*.

Protege' (2004). Protege' ontology library. http://protegewiki.stanford.edu/index.php/. [Online; accessed October 2007].

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. In *IEEE Transactions on Systems, Man, and Cybernetics*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of fhe fhe International Joint Conference on ArtificiaI Intelligence (IJCAI)*.

Wang, G.-H., Wang, Y.-D., and Guo, M.-Z. (2006). An ontology-based method for similarity calculation of concepts in the semantic web. In *IEEE International Conference on Machine Learning and Cybernetics*, pages 100–105.

Wu, Z. and Palmer, M. (1995). Verb semantics and lexical selection. In *Proceedings of32nd Annual Meeting of the Associations for Computational Linguistics*.

Zadeh, L. A. (1992). Knowledge representation in fuzzy logic. In Yager, R. R. and Zadeh, L. A., editors, *An Introduction to Fuzzy Logic Applications in Intelligent Systems*, pages 1–25, Boston. Kluwer.