ON ALIGNING INTERESTING PARTS OF ONTOLOGIES

Christos Tatsiopoulos, Basilis Boutsinas
Department of Business Administration, University of Patras, Artificial Intelligence Research Center, 26500, Rio, Greece

Konstantinos Sidiropoulos
School of Engineering and Design, Brunel University West London, Uxbridge, Middlesex, UB8 3PH, London, U.K.

Keywords: Ontology Merging, Ontology Alignment.

Abstract: Ontology merging/alignment is one of the most important tasks in ontology engineering. It is imposed by the decentralized nature of both the WWW and the Semantic Web, where heterogeneous and incompatible ontologies can be developed by different communities. Usually, ontology merging/alignment is based on an ontology mapping that has been established in a previous phase. In this paper, we define a new problem within the alignment process: the problem of detecting and then updating only interesting parts of an ontology, based on the knowledge included in another one. To this end, we define and evaluate a number of different measures of interestingness of parts of ontologies. We also present experimental results for their evaluation on test ontologies.

1 INTRODUCTION

Ontologies are used to handle complex situations, due to the continuous growth of the Semantic Web, where they are used to describe the semantics of the data. Due to the decentralized nature of both the WWW and the Semantic Web, it is inevitable that different communities within the so-called information society represent and treat the same basic concepts in different ways. For example, the basic concept ``Person" is treated entirely differently in a medical ontology than in a business one. The need for merging/alignment arises when such ontologies have to be integrated.

In this paper, we consider the task of aligning source ontologies, which may use different vocabularies and may have overlapping content. More specifically, we define a new problem within the alignment process: the problem of updating each one of the source ontologies using the knowledge included in the other but the update is performed only to the parts of each source ontology which are considered interesting by its designer. Of course, updating is taking place if the knowledge included in the interesting parts of the one ontology is a superset of knowledge included in the interesting parts of the ontology to be updated.

All merging and alignment techniques presented in the literature (such as Chimaera, FCA-Merge, PROMPT, Ontomorph, OntoDNA) consider merging or alignment of the entire input ontologies. The key ideas of any of them could be applied to the alignment of interesting parts after the latter have been located.

We reduce the problem to automatic detection of the interesting parts based only on the structure of the ontology and not to any user input. This problem is important for agent oriented applications of ontologies, whenever an agent need to update its knowledge from other agents. In this paper, we define and evaluate a number of different measures of interestingness of parts of ontologies. The term "interestingness" was first used in data mining as a measure of how much interesting an extracted data mining rule is, with respect to a user judgement. "Entropy" and "support" are such measures of interestingness.

There is not any generally accepted definition of interestingness. In fact, each of the proposed measures concerns a different aspect of what "interestingness" in ontologies could mean. Note that the proposed measures exploit only the structure of the source ontologies. Thus, the proposed measures are application independent.

Tatsiopoulos C., Boutsinas B. and Sidiropoulos K. (2009).
ON ALIGNING INTERESTING PARTS OF ONTOLOGIES.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 363-367

Interestingness of concepts within an ontology has already been explored. The DIaMOND plug-in for Protégé (d'Entremont and Storey, 2006) defines interestingness based on user's browsing activities. It continuously calculates the degree of interest for each concept by tracking user's navigation activities on an ontology. In (Tu, et al., 2005) the importance of concepts is used in filtering large scale ontologies in order to obtain efficient visualizations of them. More specifically, the importance of a concept is calculated as a weighted combination of the depth of the concept in the ontology and the sum of the importance of its direct child nodes. In (Wu, et al., 2008) interestingness is defined for both concepts and relations based on the structure of the ontology: 1) a concept is more important if there are more relations starting from the concept, 2) a concept is more important if there is a relation starting from the concept to a more important concept, 3) a concept is more important if it has a higher relation weight to any other concept and 4) a relation weight is higher if it starts from a more important concept.

In the rest of the paper we first present the proposed measures of interestingness (Section 2) along with their evaluation on test ontologies (Section 3). Finally, we conclude (Section 4).

2 THE PROPOSED INTERESTINGNESS MEASURES

We consider that the alignment of interesting parts of two source ontologies follows a mapping process, where correspondences between elements of the source ontologies are established (see for instance Doan et al., 2002, Giunchiglia et al., 2004, Hovy, 1998, Kalfoglou and Schorlemmer, 2002, McGuinness et al., 2000, Noy and Musen, 2000, Prasad et al., 2002, Stumme and Maedche, 2001, Tang et al., 2006, Tatsiopoulos and Boutsinas, 2009).

Given such a mapping and a set of interesting parts for each source ontology, an alignment algorithm could be used in order to make the interesting parts of ontology O_1 consistent and coherent with ontology O_2. Of course, such alignment is applied if O_2 represents more knowledge than O_1, as far the interesting parts of the latter is concerned.

The contribution of this paper concerns the step of the identification of interesting parts. Any mapping and alignment algorithms could be used. In what follows we define some measures of interestingness. They are all based on the structure of the ontology. Note that we have investigated several others. However the proposed ones exhibit significant results during empirical tests.

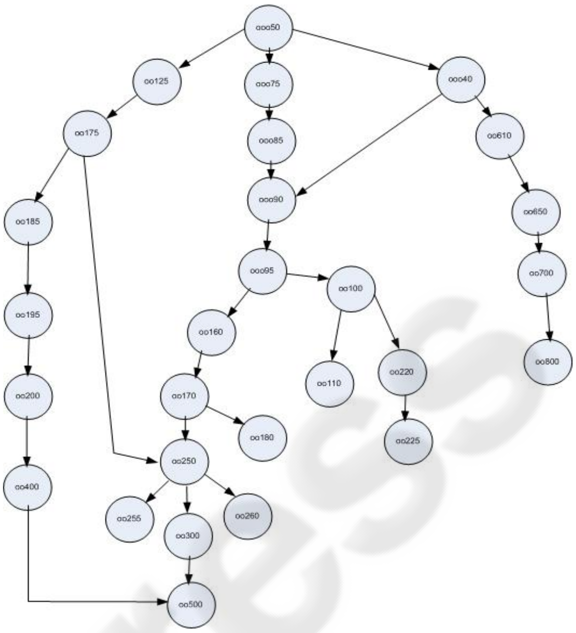


Figure 1: The test ontology.

The proposed measures can be applied to ontology structures forming a directed acyclic graph. Thus, it supports multiple inheritance. The required formal definition of input ontologies contains two core items shared by most formal definitions of an ontology in the literature: concepts and a hierarchical IS-A relation. Thus, we define an ontology as a pair G=(C,r), where C is a set of concepts and r is a partial order on C, i.e. a binary relation $r \in C \times C$ which is reflexive, transitive, and antisymmetric.

The proposed measures are all based on detecting interesting concepts within an ontology. Then, we consider interesting parts of the ontology the subgraphs rooted at these interesting concepts. Thus, the following measures assign a value to each concept representing its interestingness:

1) Percentage Direct Child Nodes (rel_cD%): the number of nodes which are directly connected to a specific node, as percentage of the number of nodes in the ontology. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, in ontology shown in Fig. 1, node oo250 (value=10.71%) is more interesting than oo180 (value=0%) or even oo40 (value=7.14%), since it has more direct child nodes.

2) Percentage Indirect Child Nodes (rel_cI%): the number of nodes in the subgraph rooted at a specific node, as percentage of the number of nodes in the ontology. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, node ooo95 (value=42.86%) is more interesting than oo185 (value=14.29%), since there are more nodes in the subgraph rooted at ooo95. Also, now node ooo40 (value=64.29%) is more interesting than oo250 (value=14.29%) for the same reason. Note that in an ontology with a deep concept structure, very general concepts (owl:Thing in the extreme case) would get higher value than very specific ones. One could claim that a specific concept is the interesting one, not the fact that something is an owl:Thing. However, we consider interesting parts those rooted at interesting concepts. Thus, according to rel_cI%, large subgraphs are more interesting than smaller ones, either the latter are disjoin or subsubgraphs. Note that in implementation level interesting concepts are searched within concepts with level greater than a threshold t, i.e., $l(i)>t$.

3) Percentage Brother Nodes (rel_b%): the number of Direct Child Nodes of the father node(s), i.e., of immediate ancestor(s), of a specific node, as percentage of the number of nodes in the ontology. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, node oo250 (value=7.14%) is more interesting than node oo180 (value=3.57%) and node ooo90 (value=3.57%), since it has more brother nodes having both oo175 and oo170 as father nodes.

4) Mean Distance of Brother Nodes (mdisbr): the mean distance of a specific node from its Brother Nodes. The distance of two nodes d(x,y) is calculated using the dissimilarity measure presented in (Boutsinas and Papastergiou, 2008). The dissimilarity between any two attribute values is repesented by the distance between the corresponding nodes of the tree structure as defined by the following formula: $d(X,Y) = 1/fl(X,Y) * Average((l(X)-fl(X,Y))/max(p(X), (l(Y)-fl(X,Y))/ max(p(Y))) * (p(X,Y)/ (max(p(X))+max(p(Y)))$, where X and Y represent any two nodes, fl(X,Y) is the level of the nearest common father node of X and Y nodes, i.e. the level of the nearest common predecessor, l(X) is the level of node X, i.e. the depth of the node, max(p(X)) is the length of the maximum path starting from the root to a leaf and containing node X, p(X,Y) is the length of the directed path (number of edges) connecting X and Y ($p(X,X)=0$). If there is not a path connecting X and Y then $p(X,Y)=p(X,fl(X,Y))+p(Y,fl(X,Y))$.

Mean Distance of Brother Nodes is calculated by the following algorithm:

```
for each Brother Node i of node j
    calculate d(i,j)
    set count+=1, dsum+=d(i,j)
return dsum/count
```

Note that the lower is the value of d(X,Y) the greater is the interestingness. For example, node oo110 (value=0.003%) is more interesting than node oo610 (value=0.0115%), since its father node is located deeper in the ontology. Finally, note that similarity (1-mdisbr) could be used instead of dissimilarity, for compatibility with the rest measures.

5) Network Density of range k (nden(k)): Network Density of range k of a specific node i is the number of nodes that are connected to or can be reached from i, via a path of length at most k, which does not include direction changes. Note that we have implemented the calculation of nden(k) dynamically. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, for node oo610 nden(2)=4, since there are 2 ancestor nodes (ooo40, ooo50) and two successor ones (oo650,oo700). Thus, it is more interesting than node oo110 (nden(2)=2).

6) Percentage Incoming Paths (in%): the indegree ($d_{in}(i)$) of a node i, i.e., the number of edges which have i as their end-node, as percentage of the total incoming and outcoming paths, i.e., of indegree plus outdegree of i. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, node ooo90 (value=66.67%) is more interesting than node oo250 (value=40%).

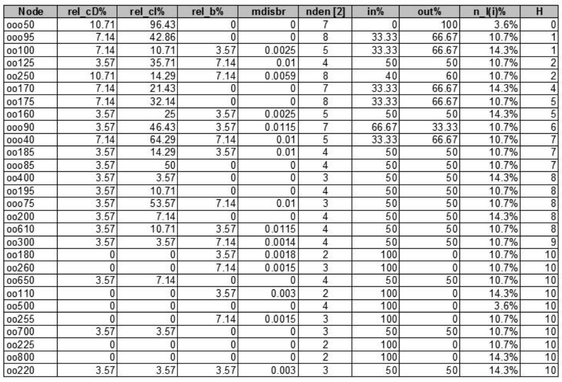
7) Percentage Outcoming Paths (out%): the outdegree ($d_{out}(i)$) of a node i, i.e., the number of edges which have i as their start-node, as percentage of the total incoming and outcoming paths, i.e., of indegree plus outdegree of i. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, node oo250 (value=60%) is more interesting than node oo90 (value=33.33%).

8) Percentage Level distribution (n_l(i)%): Level distribution of a specific node i is the number of nodes belonging to the level of node i, i.e., l(i), (i.e., the length of the maximum path -number of edges- from the root to node i), as percentage of the number of nodes in the ontology. Note that the higher is the value of the measure for a node, the greater is its interestingness. For example, there are 4 nodes on the same level with node oo100 (value=14.3%) which is more interesting than node oo250 (value=10.7%), since there are 3 nodes on its level.

3 EMPIRICAL RESULTS

Empirical tests aim at evaluating the defined interestingness measures on test ontology G_1 shown in Fig. 1. It is constructed with respect to Gene Ontology (http://www.geneontology.org/). Node naming is compatible with Gene Ontology.

To evaluate the defined measures we used a node ranking with respect to their interestingness, defined by a human expert (a researcher of the Institute for Language & Speech Processing http://www.ilsp.gr). The values for all the measures, along with expert's ranking (column H) are presented in Fig. 2. Note that expert's ranking assigns "1" to the most interesting ("0" is assigned only to root node).



Node	rel_cD%	rel_cI%	rel_b%	mdisbr	nden [2]	in%	out%	n_l(i)%	H
ooo50	10.71	96.43	0	0	7	0	100	3.6%	0
ooo95	7.14	42.86	0	0	8	33.33	66.67	10.7%	1
oo100	7.14	10.71	3.57	0.0025	5	33.33	66.67	14.3%	1
oo125	3.57	35.71	7.14	0.01	4	50	50	10.7%	2
oo250	10.71	14.29	7.14	0.0059	8	40	60	10.7%	2
oo170	7.14	21.43	0	0	7	33.33	66.67	14.3%	4
oo175	7.14	32.14	0	0	8	33.33	66.67	10.7%	5
oo160	3.57	25	3.57	0.0025	5	50	50	14.3%	5
ooo90	3.57	46.43	3.57	0.0115	7	66.67	33.33	10.7%	6
ooo40	7.14	64.29	7.14	0.01	5	33.33	66.67	10.7%	7
oo185	3.57	14.29	3.57	0.01	4	50	50	10.7%	7
ooo85	3.57	50	0	0	4	50	50	10.7%	7
oo400	3.57	3.57	0	0	3	50	50	14.3%	8
oo195	3.57	10.71	0	0	4	50	50	10.7%	8
ooo75	3.57	53.57	7.14	0.01	3	50	50	10.7%	8
oo200	3.57	7.14	0	0	4	50	50	14.3%	8
oo610	3.57	10.71	3.57	0.0115	4	50	50	10.7%	8
oo300	3.57	3.57	7.14	0.0014	4	50	50	10.7%	9
oo180	0	0	3.57	0.0018	2	100	0	10.7%	10
oo260	0	0	7.14	0.0015	3	100	0	10.7%	10
oo650	3.57	7.14	0	0	4	50	50	10.7%	10
oo110	0	0	3.57	0.003	2	100	0	14.3%	10
oo500	0	0	0	0	4	100	0	3.6%	10
oo255	0	0	7.14	0.0015	3	100	0	10.7%	10
oo700	3.57	3.57	0	0	3	50	50	10.7%	10
oo225	0	0	0	0	2	100	0	10.7%	10
oo800	0	0	0	0	2	100	0	14.3%	10
oo220	3.57	3.57	3.57	0.003	3	50	50	14.3%	10

Figure 2: Summarized results.

For instance, nodes oo100, ooo95, and oo250 are of great interestingness to the human expert, with oo100 to be the most interesting and then the ooo95, and the oo250.

After testing several test ontologies, like the one in Fig. 1, we can conclude that many measures assign interestingness in a way that reflects expert's first choices.

Moreover, some measures discover additional interesting nodes. For instance, both the n_l(i)% and the out% measure identify both oo100 and oo170 as interesting, assigning the value 14.3% & 66.67% respectively. This result was returned to the expert for additional comments. Then, this result was accepted as valid according to expert's criteria. However, some other measures have not provided successful results in a consistent manner.

4 CONCLUSIONS

We defined a new problem within the alignment process: the problem of aligning only interesting parts of ontologies. To tackle the problem we defined and evaluated a number of different measures of interestingness of parts of ontologies, each one representing different semantics of interestingness.

Despite the support or the controversy of the statement that ontology mapping/alignment is similar to database schema matching/integration (Kalfoglou and Schorlemmer, 2003, Noy and Klein, 2004), the proposed measures could be applied to both of them.

We are now working on an integration of the different measures, for instance by introducing a unified model as a function of them: f(w_1×rel_cD, ..., w_8×n_l(i)), where w_i is a weight. Preliminary results show that a different unified model is needed for each different type of structure. For instance, a unified model of the form w1×rel_cD% + w2×rel_cI% + w3×rel_b% + w4×mdistr + w5×nden[2] + w6×in% + w7×out% + w8×n_l(i)% results in almost 100% accuracy w.r.t. expert, for shallow ontologies, where w1 ranges between 0-60, w2 0-55, w3 0-80, w4 0-95, w5 0-90, w6 0-100, w7 0-45 and w8 0-80.

Moreover, we are now working on taking into consideration the linguistic analysis of concepts represented by the nodes with respect to a corpus of documents or the WWW. More specifically, we are investigating the use of term weighting techniques adopted in text mining (such as Document Frequency, mean TFIDF, Term Frequency Variance, etc.).

Finally, we are working on applying the proposed measures for detection of interesting parts to a system for knowledge transferring between mobile phones storing ontologies and holding by tourists.

REFERENCES

Boutsinas, B., Papastergiou, T., 2008. On clustering tree structured data with categorical nature, Pattern Recognition, Elsevier Science, 41, pp. 3613-3623.

d'Entremont, T., Storey, M.-A., 2006. Using a degree-of-interest model for adaptive visualizations in protégé, 9th International Protégé Conference.

Doan, A., Madhavan, J., Domingos, P., Halevy, A., 2002. Learning to map between ontologies on the semantic web, Proceedings of the 11th International World Wide Web Conference (WWW 2002), pp. 303-319.

Giunchiglia, F., Shvaiko, P., Yatskevich, M., 2004. S-Match: an Algorithm and an Implementation of Semantic Matching, Proceedings of 1st European Semantic Web Symposium, pp. 61-75.

Hovy, E., 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses, Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), Granada, Spain, pp. 535-542.

Kalfoglou, Y., Schorlemmer, M., 2003. Ontology Mapping: the State of the Art, The Knowledge Engineering Review, 18(1), 2003, pp. 1-37.

Kalfoglou, Y., Schorlemmer, M., 2002. Information-flow-based ontology mapping, LNCS 2519, pp. 1132–1151.

McGuinness, D.L., Fikes, R., Rice, J., Wilder, S., 2000. An Environment for Merging and Testing Large Ontologies, Proceedings of the Seventh International Conference on Principles of Knowledge, Representation and Reasoning (KR2000), Breckenridge, Colorado, pp. 483-493.

Noy, N.F., Klein, M., 2004. Ontology evolution: not the same as schema evolution, Knowledge and Information Systems, 6(4), pp. 428-440.

Noy, N.F., Musen, M., 2000. PROMPT: algorithm and tool for automated ontology merging and alignment, Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, pp. 450-455.

Prasad, S., Peng, Y., Finin, T., 2002. Using Explicit Information To Map Between Two Ontologies, Proceedings of the Workshop on Ontologies in Agent Systems, Bologna, Italy, Vol. 66, pp. 52-56.

Stumme, G., Maedche, A., 2001. Ontology Merging for Federated Ontologies on the Semantic Web, Proceedings of the International Workshop for Foundations of Models for Information Integration, Viterbo, Italy.

Tang, J., Li, J., Liang, B., Huang, X., Li, Y., Wang, K., 2006. Using Bayesian decision for ontology mapping, Journal of Web Semantics, 4(4), pp. 243–262.

Tatsiopoulos, C., Boutsinas, B., 2009. Ontology Mapping based on association rule mining, 11th International Conference on Enterprise Information Systems, Milan Italy.

Tu, K., Xiong, M., Zhang, L., Zhu, H., Zhang, J., Yu, Y., 2005. Towards imaging largescale ontologies for quick understanding and analysis, LNCS 3729.

Wu, G., Li, J., Feng, L., Wang, K., 2008. Identifying Potentially Important Concepts and Relations in an Ontology , Intern. Semantic Web Conf., pp.33-49.