



## Similarity Matching of Ontology in Semantic Web

### *Semantik Web'de Ontoloji Benzerliklerinin Eşleşmesi*

Ayşe Salman\* 

Maltepe University, Department of Computer Engineering, Istanbul, Turkey

#### Abstract

Matching Ontologies becomes an important task for many applications in Semantic Web. This paper investigates effective similarity match between ontologies by considering similarity on two levels. We first consider the similarity of the linguistic properties of the ontology entities which takes in consideration both morphological and semantics of the entities. This is then combined with measuring the similarity of the ontology structure as represented by RDF graph. This similarity is derived by constructing a graph from the matched nodes and use it to calculate the measure of structure similarity.

**Keywords:** Linguistic similarity, Ontology matching, RDF (Resource Description Framework) graph, Similarity graph, Similarity measure, Structure similarity

#### Öz

Ontolojileri eşleştirmek birçok Semantik Web uygulaması için önemli bir görev haline gelmiştir. Bu makale, iki benzerlik seviyesi kullanarak ontolojiler arasındaki etkili benzerlik eşleşmesini araştırmaktadır. Bu çalışmada, ilk olarak varlıkların hem morfolojisini hem de semantiğini hesaba katarak, ontoloji varlıklarının dilsel özelliklerinin benzerliklerinin incelenmesi ile başlamaktadır. Bu daha sonra bir RDF grafiği ile temsil edilen ontoloji yapısının bir karşılaştırması ile birleştirilmektedir. Bu benzerlik, eşleşen düğümlerden bir grafik oluşturularak ve yapı benzerliğinin ölçüsünü hesaplamak için kullanılarak hesaplanmaktadır.

**Anahtar Kelimeler:** Dilsel benzerlik, Ontoloji eşleştirme, RDF (Kaynak Tanımlama Çerçevesi) grafiği, Benzerlik ölçüsü, Eşleşen grafik, Yapı benzerliği

### 1. Introduction


Ontology matching is a key challenge in Semantic Web (Berners-Lee et al. 2001). It is the problem of finding semantic mappings among ontologies for data integration and reuse. To operate effectively, the Semantic Web must be able to make explicit the semantics of Web resources via ontologies, which software agents use to automatically process these resources. Hence, large number of ontologies are constructed covering many domains. Given the nature of the web, being decentralised and lack common criterion for building ontology, many of these ontologies will have overlapping domains, while many others may describe similar domains but using different terminologies. This

heterogeneity presents problems however, particularly in terms of redundancy and ambiguity.

To achieve some degree of interoperability to enable integration and sharing of data across different applications and organizations, virtually any application that involves multiple ontologies must establish correspondences among them (Ushold 2003). Hence, the development of tools to measure the degree of similarity among ontologies has received much attention (Liu et al. 2021).

To be fully meaningful the similarity must consider the use of both linguistic and structural matchings hence adopts two phases to compute similarity of the ontologies and return a measure of the degree of similarity. In this work the measure is considered a real number in the unit interval  $[0,1]$ , with 0 means no similarity and 1 means complete similarity.

\*Corresponding author: aysesalmantr@gmail.com

Ayşe Salman  [orcid.org/0000-0003-2649-3061](https://orcid.org/0000-0003-2649-3061)



## 2. Determining Linguistic Similarity Among Entities

The first phase of the analysis is to compute linguistic similarity among ontologies entities. Generally, linguistic similarity between two entities relies on both morphological and semantic of the entities. Ontologies are usually represented as attributed directed graphs and since the emergence of the Semantic Web, such graph has been standardized by the World Wide Web Consortium (Berners-Lee et al. 2001) as set of RDF triples of the form <subject, predicate, object>. Attribute name or label of a node or an arc in RDF graph can be a URI or a literal, whether string, word or variable, and nodes can also be blanks. Morphological similarity of the ontology entities can be done by comparing their character strings, while semantic similarity is done by comparing their meanings.

To measure morphological similarity, we can use *Levenshtein string edit distance* (Rice et al. 1997). In this method, the similarity of two strings is taken as the minimum number of single characters edit operations: deletions, insertions, or substitutions, which are required in order to transform one string into the other. Thus, the morphological similarity measure  $Sim_{morph}$  is given by:

$$Sim_{morph}(n_i, n_j) = 1 - \frac{ED(n_i, n_j)}{\max(|n_i|, |n_j|)} \quad (1)$$

Where  $ED(n_i, n_j)$  is the Levenshtein's edit distance for label strings of the two entities  $n_i$  and  $n_j$  as character strings irrespective of the types of the labels.

To compare the semantics of ontology entities, their character strings should be *tokenized* using a *tokenization* algorithm such as NLTK (Ramasubramanian and Ramya 2013, Bird et al. 2009). Tokenization is an NLP basic process that partitions a character string into a sequence of units of letters called *tokens* by using delimiters (such as punctuations, upper case, digits etc.) and discarding the non-letter ones. If any such tokens are words, we can measure their semantic similarity by using WordNet (Miller et al. 2006). WordNet is a lexical database that organizes nouns and verbs into a taxonomy of *is-a* relations. Several methods are devised in the bibliography for measuring the correlation or similarity between two terms from the WordNet ontology (Qin et al. 2009, Blanchard et al. 2005, Budanitsky and Hirst 2006, Lin 1998, Ding et al. 2005, Resnik 1995). However, the results on the evaluation of the performance of these similarity methods done by Verlas et al. (Varelas et al. 2005) have shown fairly general agreements among them. Here we

select similarity measures based on counting edges between concepts i.e., their depth in the WordNet is-a taxonomy e.g., the works in (Wu and Palmer 1994, Algergawy et al. 2010), Hence to measure the semantic similarity between two tokens  $t_1$  and  $t_2$  we use Wu and Palmer's formula (2) below that is illustrated in Figure 1.

$$Sim_{Tok}(t_1, t_2) = \frac{2 \times depth(LCS)}{depth(t_1) + depth(t_2)} \quad (2)$$

$$= \frac{2 \times D_0}{D_1 + D_2 + 2 \times D_0}$$

where  $D_1$  and  $D_2$  are the numbers of is-a edges from nodes representing tokens  $t_1$  and  $t_2$  to node  $t_0$ , their first mutual parent (called the Least Common Subsumer LCS) that subsumes them both.  $D_0$  is the number of is-a edges from  $t_0$  to the root of the hierarchy. When tokens are not proper words which means cannot be found in WordNet, then only morphological similarity is considered.

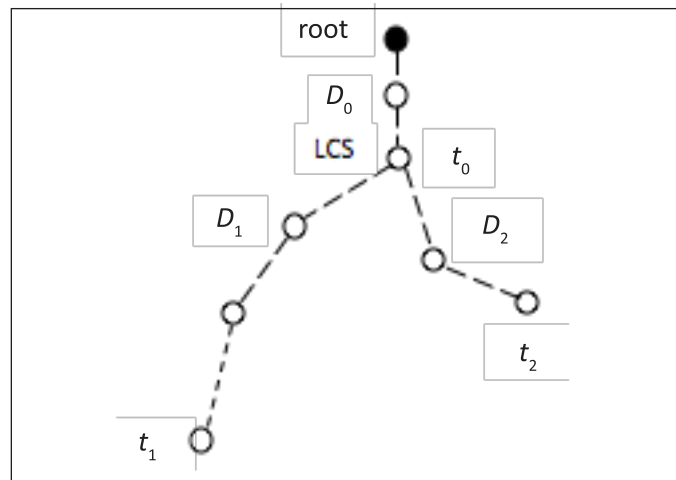


Figure 1. Illustration of the semantic similarity measure in formula (2).

The semantic similarity  $Sim_{Sem}$  of the labels of two entities  $n_i$  and  $n_j$  can be taken as the semantic similarity of their corresponding token sets  $T_1$  and  $T_2$  (Nayak et al. 2007, Salman 2020, Wang et al. 2021). This can be calculated by adding up the best (maximum) similarity of every token in one set to the tokens in the other. The summation is then averaged by dividing over the sum of the two sets cardinalities as shown in formula (3) below:

$$Sim_{Sem}(n_i, n_j) = \frac{\sum_{t_1 \in T_1} [\max_{t_2 \in T_2} Sim_{Tok}(t_1, t_2)] + \sum_{t_2 \in T_2} [\max_{t_1 \in T_1} Sim_{Tok}(t_1, t_2)]}{|T_1| + |T_2|} \quad (3)$$

The linguistic similarity  $Sim_{Ling}(n_i, n_j)$  of two entities  $n_i$  and  $n_j$  can then be calculated by weighted sum combining morphological and semantic similarities (Zhang et al. 2008, Salman 2020, Liu et al. 2021, Lv et al. 2020) and can be defined as:

$$Sim_{Ling}(n_i, n_j) = p \times Sim_{morph}(n_i, n_j) + q \times Sim_{Sem}(n_i, n_j) \quad (4)$$

where,  $p, q \in [0, 1]$  and  $p + q = 1$

### 3. Structure Similarity Matching

Linguistic similarity as shown in section 2 considers the similarity of the ontology entities without considering the similarity of ontology structure. As the ontology is represented by RDF graph, similarity of ontology structure is turned into similarity matching of RDF graphs. Exact graph matchings have been shown in the literature to be NP-complete, e.g., in (Klyne 2004). However, we rely in RDF graph matching on characteristics of RDF graph (Zhu et al. 2002). An RDF graph representing ontology is simply a set of triples of the form <subject, property, object> (Klyne 2004) and we can define it as follows.

**Definition:** An RDF graph is a directed labelled graph  $G = (N, A)$  where (1)  $N$  is a set of nodes whether represent *subjects* or *objects*, (2)  $A$  is a set of arcs represent *properties* (called also predicates) and (3) each arc  $a \in A$  connects subject node  $s \in N$  to object  $o \in N$  in that direction, representing triple  $\langle s, p, o \rangle$ .

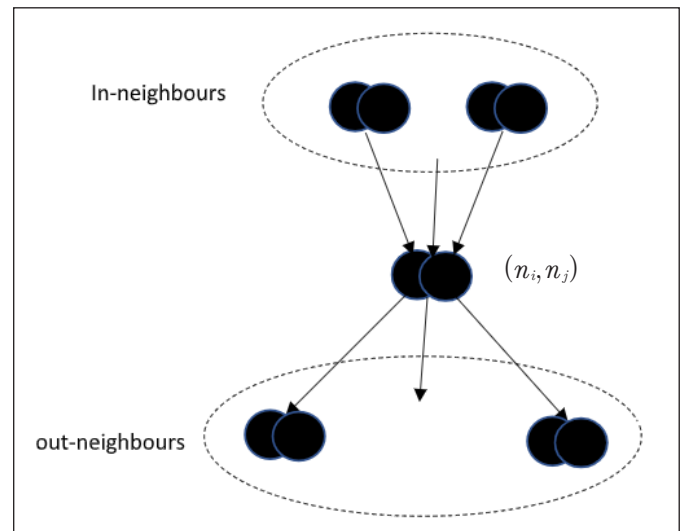
For computing the similarity between two RDF graphs, matching is based on the intuition that if two subject nodes are matched then their objects are also matched if their arcs properties were matched. This idea was proposed in Similarity Flooding work by Melnik et al [8] and also by G. Jeh et al in SimRank (Jeh and Widom 2002). The matching between two graphs can be put in the form of a graph, we call this *similarity graph*, by pairing matched nodes according to the that intuition. Similar procedure was also used in the work in (Zhang et al. 2008) to create matching tree. Hence to compute the similarity between two RDF graphs  $g$  and  $g'$  we first construct a *similarity graph*  $G(g, g')$  according to the following rule. Each node in the similarity graph is an element from  $g \times g'$  such that  $((n_1, n_1'), p, (n_2, n_2')) \in G(g, g') \Leftrightarrow (n_1, p, n_2) \in g$  and  $(n_1', p, n_2') \in g'$  and  $n_1$  is matched to  $n_1'$ . Using graph  $G(g, g')$ , the structure similarity between  $n_i$  in  $g$  and  $n_j$  in  $g'$  is the average of the similarity  $Sim_{Ling}(n_i, n_j)$  and the similarity of the nodes neighbours to node  $(n_i, n_j)$

, in-neighbours and out-neighbours, as illustrated in Figure 2. Similar method was devised in the work in (Zhang et al. 2012) but it uses statements representing triples rather than individual nodes and also in the work in (Zhang et al. 2008) that uses similarity tree rather than similarity graph. The structural similarity  $Sim_{struc}$  of two nodes  $n_i$  and  $n_j$  is then calculated by the following formula:

$$Sim_{struc}(n_i, n_j) = \alpha \times Sim_{Ling}(n_i, n_j) + \beta \times \frac{\sum_{k=1}^{|I(n_i, n_j)|} Sim_{Ling} I_k(n_i, n_j)}{|I(n_i, n_j)|} + \gamma \times \frac{\sum_{k=1}^{|O(n_i, n_j)|} Sim_{Ling} O_k(n_i, n_j)}{|O(n_i, n_j)|} \quad (5)$$

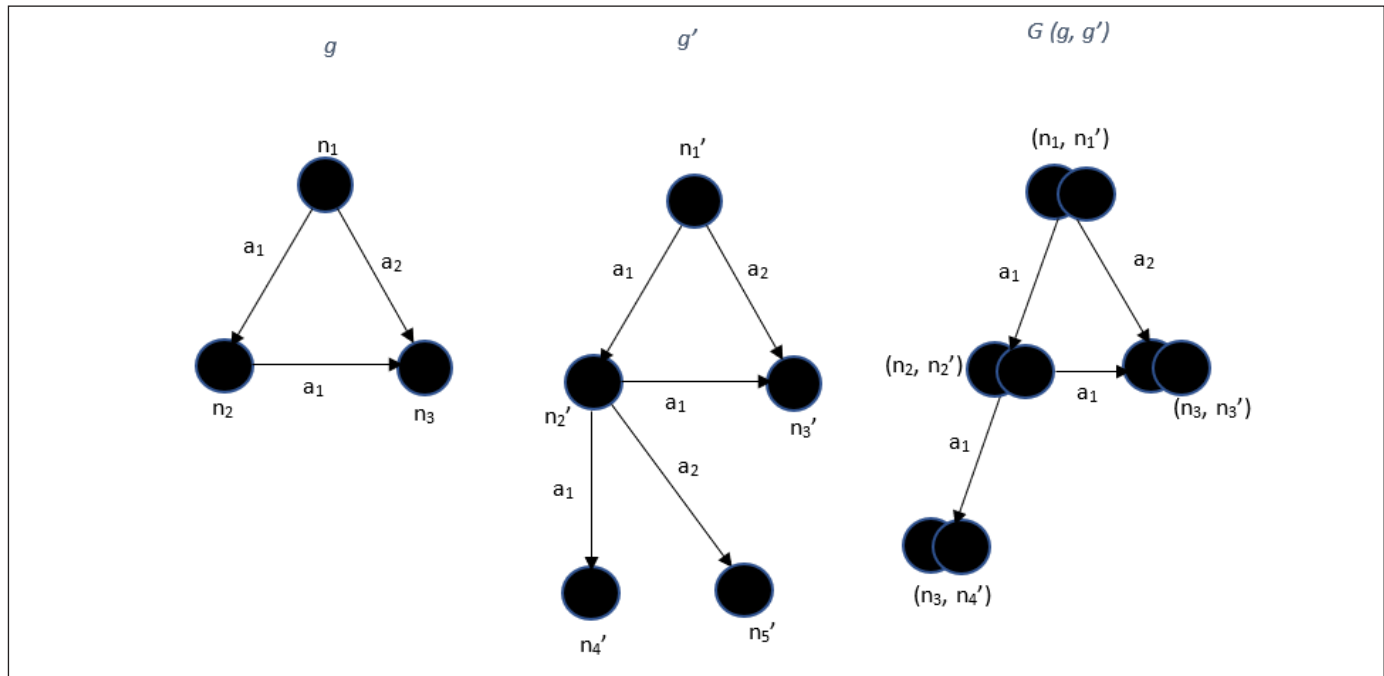
$\alpha, \beta, \gamma \in [0, 1]$  and  $\alpha + \beta + \gamma = 1$

Where in-degree  $|I(n_i, n_j)|$  (number of arcs leaving  $(n_i, n_j)$  as subject node), out-degree  $|O(n_i, n_j)|$  (number of arcs entering  $(n_i, n_j)$  as object node). Individual in-neighbour is denoted by  $I_k(n_i, n_j)$ , for  $1 \leq k \leq |I(n_i, n_j)|$  and individual out-neighbour is denoted by  $O_k(n_i, n_j)$ , for  $1 \leq k \leq |O(n_i, n_j)|$ .



**Figure 2.** Neighbouring nodes of the similarity graph node  $(n_i, n_j)$ .

To illustrate how the similarity graph is computed we take the simple example displayed in Figure 3 below. Two ontology RDF graphs  $g$  and  $g'$  are displayed in Figure 3-a and Figure 3-b. To match these two RDF graphs, we create the Similarity Graph  $G(g, g')$  in Figure 3-c in which each node is represented by a pair of matched nodes of  $g$  and  $g'$  starting from the first matched pairs  $(n_1$  and  $n_1'$  in the figure) then follow the corresponding matched arcs.



**Figure 3.** Illustration of creating the similarity graph.

To calculate the similarity between two RDF graphs  $g$  and  $g'$  we can then use the following formula:

$$Sim(g, g') = \frac{\sum_1^n Sim_{struct}(n_i, n_j) \times D(n_i, n_j)}{\sum_1^n D(n_i, n_j)} \quad (6)$$

Where  $n$  is the number of nodes in the similarity graph and  $D(n_i, n_j)$  is the node  $(n_i, n_j)$  degree = indegree  $|I(n_i, n_j)|$  + outdegree  $|O(n_i, n_j)|$ .

#### 4. Conclusion

The work presented in this paper has taken an approach for effective measure of ontology similarity. This is done by combining two measures of similarity. One is the measure of linguistic similarity of ontology entities which considers both morphological and semantics of the entities. The second measure is the similarity of the ontology structure using its representation as RDF graph. This measure avoids the complexity of graph isomorphism by relying on the triples characteristics of RDF graphs to create a similarity graph. While this is a step in the development of the similarity problem, in future work more analysis still required particularly in handling the semantics of the different types of node labelling, in addition to testing and evaluating the method on real data.

#### 5. References

- Algergawy, A., Nayak, R., & Saake, G. (2010). Element similarity measures in XML schema matching. *Information Sciences*, 180(24), 4975-4998.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, 284(5), 34-43.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- Blanchard, E., Harzallah, M., Briand, H., & Kuntz, P. (2005). A typology of ontology-based semantic measures. *EMOI-INTEROP*, 160, 3-11.
- Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational linguistics*, 32(1), 13-47.
- Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J. J., & McBride, B. (2014). RDF 1.1 concepts and abstract syntax. *W3C recommendation*, 25(02), 1-22.
- Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., & Kolari, P. (2005). Finding and ranking knowledge on the semantic web. In *International Semantic Web Conference* (pp. 156-170). Springer, Berlin, Heidelberg.
- Graves, A., Adali, S., & Hendler, J. (2008). A Method to Rank Nodes in an RDF Graph. In *International Semantic Web Conference (Posters & Demos)* (Vol. 401).

- Jeh, G., & Widom, J. (2002, July).** Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 538-543).
- Klyne, G. (2004).** RDF Concepts and Abstract Syntax W3C Recommendation. <http://www.w3.org/TR/rdf-concepts/>.
- Lin, D. (1998).** An information-theoretic definition of similarity. In *Icml* (Vol. 98, No. 1998, pp. 296-304).
- Liu, X., Tong, Q., Liu, X., & Qin, Z. (2021).** Ontology matching: state of the art, future challenges and thinking based on utilized information. *IEEE Access*.
- Ly, Q., Jiang, C., & Li, H. (2020).** Solving ontology meta-matching problem through an evolutionary algorithm with approximate evaluation indicators and adaptive selection pressure. *IEEE Access*, 9, 3046-3064.
- Melnik, S., Garcia-Molina, H., & Rahm, E. (2002).** Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proceedings 18th international conference on data engineering* (pp. 117-128). IEEE.
- Miller, G. A., Fellbaum, C., Teng, R., Wolff, S., Wakefield, P., Langone, H., & Haskell, B. (2006).** WordNet: A lexical database for the English language. *Cognitive Science Lab, Princeton University*, <http://www.cogsci.princeton.edu/wn>.
- Motik, B., & Patel-Schneider, P. (2012).** OWL 2 Web Ontology Language Mapping to RDF Graphs.
- Nayak, R., & Tran, T. (2007).** A progressive clustering algorithm to group the XML data by structural and semantic similarity. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(04), 723-743.
- Qin, P., Lu, Z., Yan, Y., & Wu, F. (2009).** A new measure of word semantic similarity based on wordnet hierarchy and dag theory. In *2009 International Conference on Web Information Systems and Mining* (pp. 181-185). IEEE.
- Ramasubramanian, C., & Ramya, R. (2013).** Effective pre-processing activities in text mining using improved porter's stemming algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(12), 4536-4538.
- Resnik, P. (1995).** Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
- Rice, S. V., Bunke, H., & Nartker, T. A. (1997).** Classes of cost functions for string edit distance. *Algorithmica*, 18(2), 271-280.
- Salman, A. (2020).** Similarity matching of XML schema. *Karaelmas Fen ve Mühendislik Dergisi*, 10(1), 121-129.
- Schneider, P., Hayes, P., & Horrocks, I. (2004).** OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation. *World Wide Web Consortium (W3C)*.
- Uschold, M. (2003).** Where are the semantics in the semantic web?. *Ai Magazine*, 24(3), 25-25.
- Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., & Milios, E. E. (2005).** Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management* (pp. 10-16).
- Wang, Y., Li, Y., Fan, J., Ye, C., & Chai, M. (2021).** A survey of typical attributed graph queries. *World Wide Web*, 24(1), 297-346.
- Wu, Z., & Palmer, M. (1994).** Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.
- Zhang, D., Song, T., He, J., Shi, X., & Dong, Y. (2012).** A similarity-oriented RDF graph matching algorithm for ranking linked data. In *2012 IEEE 12th International Conference on Computer and Information Technology* (pp. 427-434). IEEE.
- Zhang, R., Wang, Y., & Wang, J. (2008).** Research on ontology matching approach in semantic web. In *2008 International Conference on Internet Computing in Science and Engineering* (pp. 254-257). IEEE.
- Zhu, H., Zhong, J., Li, J., & Yu, Y. (2002).** An approach for semantic search by matching RDF graphs. In *FLAIRS Conference* (pp. 450-454).