International Journal of Engineering & Technology, 7 (4) (2018) 2161-2165

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.14153
Research paper

An ontology-based semantic similarity metric to empower semantic search

Suraiya Parveen [1] *, Ranjit Biswas [1]

[1] Department of Computer Science & Engineering, School of Engineering Sciences and Technology Jamia Hamdard University, New Delhi, 110065
**Corresponding author E-mail: husainsuraiya@gmail.com*

Abstract

Heterogeneity in documents is a challenge for information Retrieval. The keyword search focuses on matching the keywords with web repositories. It does not consider the synonyms or semantically similar words. The heterogeneity of the content makes retrieval inadequate. Semantic search helps to capture more appropriate results using domain ontology. Keyword search is extended with the help of similar concepts of ontology. Similarity between the ontological concepts is recognized to get appropriate search results. Once the semantic similarity among the concepts is known, more relevant documents can be retrieved. In this paper, we propose a metric based on traditional methods, combined with computational techniques to measure the similarity between concepts. The paper gives the concept of DOT (Domain Ontology Tree). It uses conventional definitions of the Tree (Data Structure) for ontology and proposes a method of partitioning to calculate the similarity. The method is based on IS-A hierarchical relationship. We have implemented a prototype system for the support of the proposed method, and also compared it with existing methods, the results are encouraging.

Keywords: *Heterogeneity; Ontology; Metric; Semantic Similarity; Domain Ontology Tree (Dot).*

1. Introduction

Information is processed Data; and the Data on the web is represented in the form of Text, Image, Audio, Video etc. Heterogeneity in web documents means similar contents are expressed in different ways in web documents. This makes the information retrieval difficult. The existing systems focus on matching the keywords and overlook synonyms and related words. This causes inadequacy in retrieving relevant documents. To achieve efficient retrieval, the system must know the meaning and relationship of the concepts. Domain ontology gives this capability to the system. The ontology is expected to impart semantics to the data so that its heterogeneity can be managed effectively. Metadata in the form of RDFS, XML and ontology enables the semantic web to compute the data and provide relevant information. The search techniques embedded with semantics empower Information Retrieval and get back more relevant information from web.

Domain ontology includes concepts; which are usually the keywords used for search. The keyword search looks for matching the particular word or key phrases in digital documents, and not for the document semantics. For example, the word *advisor* is semantically similar to the word *consultant*. Domain ontology contains concepts of the domain and their relationship. The Domain Ontology can expand the search for the concept *advisor* to the concept *consultan*t and *guide*. This makes the semantic search possible and retrieval more efficient. Computational techniques need to be developed to measure similarity in the concepts. The proposed semantic similarity measure makes use of

IS-A relationship. The measure quantifies the similarity of concepts and the Similarity Score calculated by the method makes the system recognize the semantically similar concepts. When the semantic similarity between the concepts is known, the search involves semantics too. This is how semantic similarity empowers search techniques. We have developed a prototype system to demonstrate the methodology. The system is compared with existing systems [3] [5] for evaluation.

Semantic search can be very effective to address current challenges of Information Retrieval. Many methods have been proposed till today for similarity measurement, they have their own merits and demerits. The literature survey classifies the different solutions in three major categories: Node based, Path based and a hybrid of both. The present method is a path based method for calculating the similarity. The rest of the paper is organized as below; Section 2 introduces the semantic similarity and its relevance in semantic search. Section 3 presents the terminology and methodology used in this paper. Section 4 gives the design of proposed metric. In Section 5 is devoted to experimental evaluation of our method, results and its comparison with other methods. The conclusion is presented in section 6.

2. Semantic similarity

"Semantic means study of meaning." Present web has enormous possibilities, yet to explore. The semantic web is embedded with ontology and metadata so that it can be converted in machine readable format. Ontologies are represented in the form of RDF graphs. RDF graph is a set of nodes and links. The concepts of ontology are contained in nodes of RDF graph and the links between nodes establish the relationship among concepts. Graphical representation of ontology contain new refinement of concepts of the domain using IS-A relationship. This means concepts within the ontology are semantically linked to each other. To quantify the similarity between concepts, it is logical to incorporate path length or path weight. Many established methods [1] - [4] use path length or path weight, in the form of semantic distance, to measure similarity. Literature survey [5] - [10] explores the various proposed solutions to calculate similarity in ontology. These Semantic similarity measures give different solutions, to quantify the similarity, between the concepts of ontology.

3. Proposed method for measuring concept similarity

In this section, we present our methodology to measure similarity in concepts. The concepts of ontology follow hierarchal arrangement; we have named it as a Domain Ontology Tree. To find the relationship between ontology structures, we have used IS-A relationship. The edges of the DOT are assigned weights and semantic distance between the concepts is calculated. Semantic similarity is inversely proportional to semantic distance. The proposed method for calculating similarity is elaborated in section 4. We have modified the terminology [3] as below:

3.1. Terminology

1) Root Node:

Root is the most basic concept C_0. Where C is base concept and numeric value signifies the level of the root.

2) Level:

Level is a rank in the hierarchy, the root is the super class in IS-A hierarchy at level zero.

3) Depth of Concept

Depth of concept is calculated as

Depth of C_i = (level of C_i+1) Where C_i is any of the concepts within the ontology.

The depth of the root is one.

4) Depth of Domain ontology tree

The longest path in the tree, starting from the root node to the deepest leaf node,

Represented as Depth $_{max =}$ (level $_{max}$ + 1)

5) Parent sequence nodes:

The set of nodes, obtained by back tracking the concept node in question, including parent and parent of the parent till root node.

6) Common Parent

The Common parent of two concept nodes is obtained by back tracking the concepts (C_i, C_j) nodes till a common node is found in sub-tree.

7) Sibling nodes

Nodes have same parent.

4. The design of semantic similarity metric

Ontology can be represented using IS-A relationship among concepts within a domain. Graph and tree structure is appropriate selection for similarity measurement. In DOT (Domain Ontology Tree) nodes are the concepts and relationship between concepts is organized using links as shown in figure 1. Ontology is represented mathematically in this study. Ontology consists of a finite set of concepts. Each concept has attributes. Our method of similarity calculation considers only concepts.

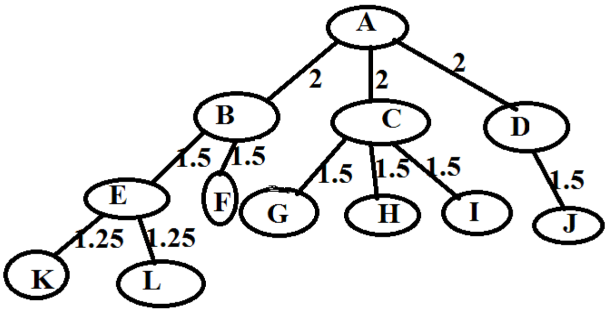

Fig. 1: Domain Ontology Tree with Edge Weights.

a) Definition of Domain Ontology Tree

- The root node contains the base concept of ontology. At $level_1$, the remaining nodes of ontology are partitioned into ‘n’ sub-trees. C_1, $C_{2,}$ ------, C $_{n,}$ which are succeeding concepts of base concept; and can be considered as sub hierarchy of ontology.
- Each sub-tree in DOT at $level_1$ is a partition among the concepts of the domain. The weight of the links between the root node and the nodes at $level_1$ is kept highest to consider the partition of base concept [4] [11].
- At every next level, the nodes contain new refinement of concepts. Concepts become more specific, so weights gradually decrease at lower levels of DOT.
- Sub-Tree of any concept at any level in DOT is also a Tree structure and follows the same nomenclature.

Figure 1 illustrates above definition of a domain ontology tree containing 12 Nodes. Let the above DOT is an ontology extract of some Domain D. Here A is the base concept of ontology. The remaining nodes are divided into 3 sub-trees (sub-hierarchies) at level $_{1;}$ each one contains similar concepts of the domain.

b) Quantification of proposed measure

To measure the similarity between concepts, we require weight allocation to the edges of DOT. The technique is studied and applied in some previous work [4], [11], [12]. We borrow their original perception as it is strengthening our thought of partitioning of concepts. In ontology hierarchy, concepts become more specific at lower levels, hence weight of edges decreases. The method used for weight calculation (A) gives maximum weight to the edges at level1; minimum weight to the edges of concepts at leaf nodes. The formula for weight allocation is

Weight of edge w (ci, cj) = 1 + $\frac{1}{k^{(\text{depth}\, c_j)}}$ (A)

k is a predefined factor greater than1, k is a constant and control the rate of decrement of weight of edges along with the depth of ontology hierarchy. The value of k is set as 2.

c) Similarity measurement

The generalized concepts at level $_1$ gradually turn to be more specific at the next levels. Hence weights assigned to the edges proportionally decreasing with each next level of DOT as shown in figure1. Path distance between two concepts is inversely proportional to the similarity [4].

Sim(s,t)=1/distance(s,t) [13]

In our method, for the distance between two concepts, we are using semantic path length which is the sum of weights of edges from C_i to C_j. Both the concepts are back tracked, taking the shortest path, including the immediate common parent of both concepts.

Example:

Semantic Distance between L and F (fig.1)

Edges Path L-E-B-F

Where B is the immediate common parent of L and F

Semantic distance {1.25+1.5+1.5} = 4.25

Semantic Similarity Sim ($_{L,\ F}$) = 1/distance ($_{L,\ F}$) [13] = $\frac{1}{4.25}$ = 0.23

This method has been applied for calculating similarities in concepts of existing methods, picked as case studies.

4.1. The metric

The Metric for similarity measure is developed by the Domain ontology tree, the weight assigned to edges and semantic path length. The Metric can be expressed in a string notation as shown below and is easily convertible in algorithm for computation.

DOT=A[B{E(K,L),F}][C{(G,H,I)}][D{(J)}] (B)

The above metric establish:
The level one of the ontology hierarchy; divide the base concept into sub-concepts, these are actually sub- concepts with different characteristics in ontology. The next levels carry specific concepts; which become more specific as level increases. The immediate parent, child concept pair is most similar, at lower levels as compared to the upper levels; siblings follow the same pattern ie sibling pair is more similar at lower levels. The concept pairs, which belong to different sub hierarchy, are least similar.

4.2. Similarity objectives

The proposed algorithm meets the four criteria of similarity measures [14]. The range of semantic similarity is from 0 to 1.

1) Non-Negativity: Similarity value cannot be less than zero.
2) Identity: Sim (A, A) = Sim (B, B) =1
3) Symmetry:Sim (A, B) = Sim (B, A)
4) Uniqueness: Sim (A, B) =1 → A=B,

5. Experimental evaluation of proposed metric

To support our metric, designed using DOT and string notation, we have experimented with a Piece of Univ-bench ontology [15], as shown in Figure 3. We have demonstrated the proposed methodology; we have picked some concept pairs and calculated the similarity of each pair. The similarity of the same concept pairs is also calculated using Wu and Palmer method [5]. The results obtained by our method are arranged side by side with Wu and Palmer in table 1. But the purpose is not to compare the results with Wu-Palmer. We are exploring the hierarchical nature of ontology. The Results and Analysis prove the hypothesis very clearly.
The Wu Palmer measuring technique is shown in figure 2 of section 5.1.

5.1. Wu and palmer similarity measure

Wu and Palmer method is based on Rada et al [10]. Both the methods are used to measure similarity using "IS-A" hierarchical relations in the ontology.

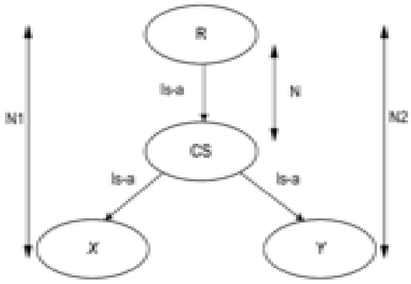

Fig. 2: WU and Palmer Measure.

Wu and Palmer method is based on counting edges from Ci to root (N1), Cj to root (N2) and common parent of both concepts to root (N).

Sim WP = $\frac{2N}{(N1+N2)}$

Now we apply the proposed method and Wu - Palmer method to the piece of Univ. Bench ontology, shown in Figure 3. The results and analysis is presented in Table1. The purpose of our method is to make use of hieratical nature of ontology in similarity calculation. The results obtained by this method also strengthen our hypothesis of using hierarchal nature of ontology in semantic similarity measurement.

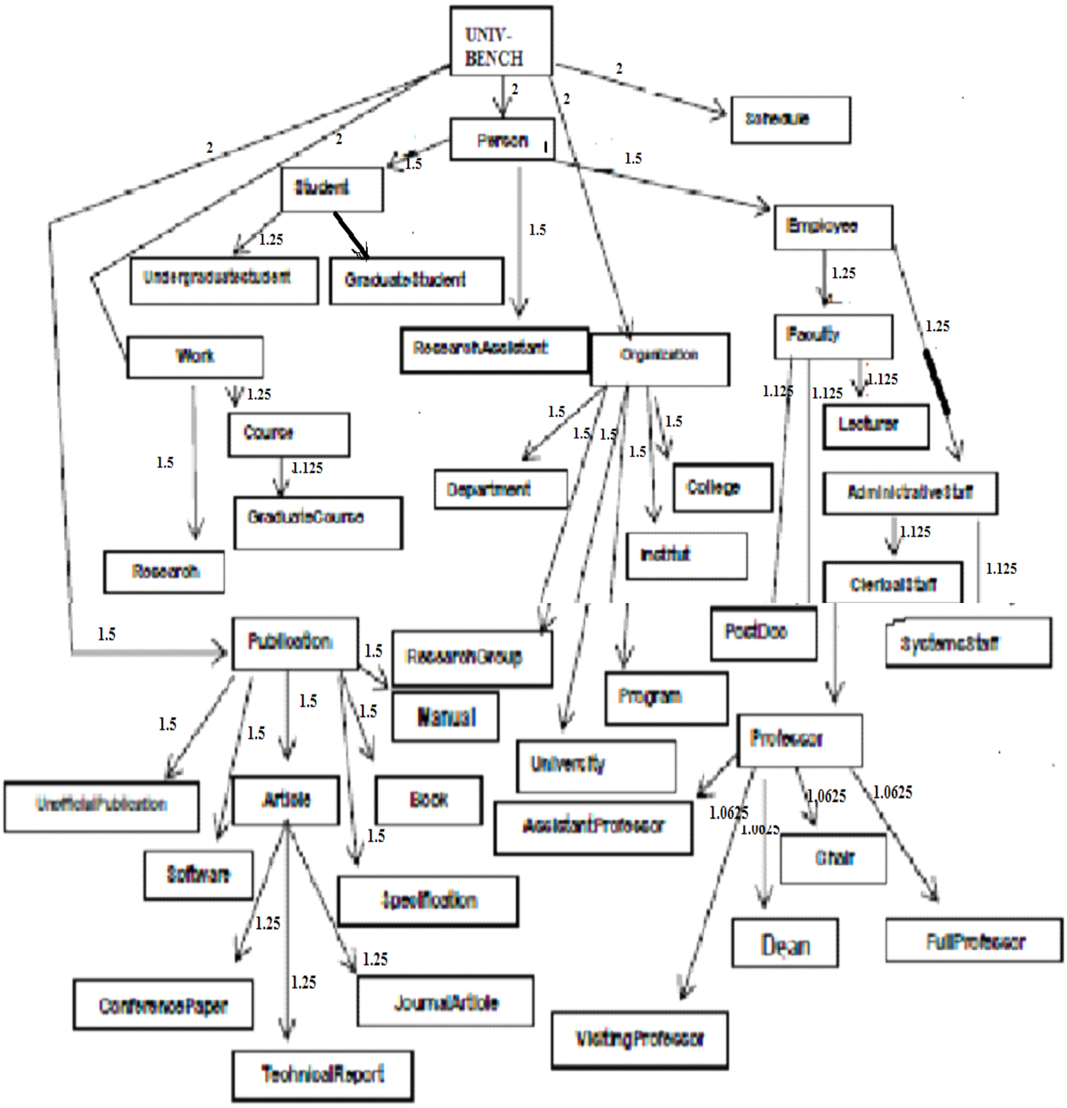

Fig. 3: UNIV-Bench Ontology.

C_i	C_j	Sim wp	Our method	Analysis
Organization	College	0.66	0.66	Parent-child
Person	ResearchAssistant	0.66	0.66	Parent-child
Publication	Book	0.66	0.66	Parent-child
Employee	Faculty	0.4	0.8	Parent-child
Article	Technical Report	0.4	0.8	Parent-child
Professor	Full Professor	0.88	0.94	Parent-child
Book	Article	0.5	0.33	Sibling
Faculty	Admin. Staff	0.73	0.33	Sibling
College	Department	0.5	0.33	Sibling
Course	Research	0.5	0.33	Sibling
Clerical staff	System staff	0.40	0.44	Sibling
Sys staff	Professor	0.66	0.44	Sibling
Visiting Professor	Full Professor	0.8	0.47	Sibling
Dean	Chair	0.8	0.47	Sibling
Visiting Professor	System Staff	0.44	0.17	Different sub -hierarchy
System staff	Dean	0.44	0.17	Different sub –hierarchy
System Staff	Professor	0.5	0.21	Different sub -hierarchy
ResearchAssistant	Faculty	0.4	0.23	Different sub –hierarchy
Research	GraduateCourse	0.4	0.25	Different sub -hierarchy
Person	Schedule	0	0.25	Different sub –hierarchy
Person	Work	0	0.25	Different sub –hierarchy

Table 1: Results and Analysis

Table 1 shows results and analysis of results. Here we are not comparing our results with Wu- Palmers results for same concept pairs. It is only to convey that similarity values of our method are mapping with the hierarchal nature of ontology. Analysis shows similarity values of concepts are clearly divided into three categories. The concept pair with immediate parent-child relationship has high similarity values. Concept pair having a sibling relationship has lesser similarity values. The concept pair which belongs to different a sub hierarchy is least similar.

To validate our method, we have compared it with one more existing system [3]. This work also calculates semantic similarity in ontological concepts. However, they consider some important factors which influence similarity and developed a formula to calculate the edge weight; whereas we focus on the inheritance relationship in ontological concepts to calculate semantic similarity.

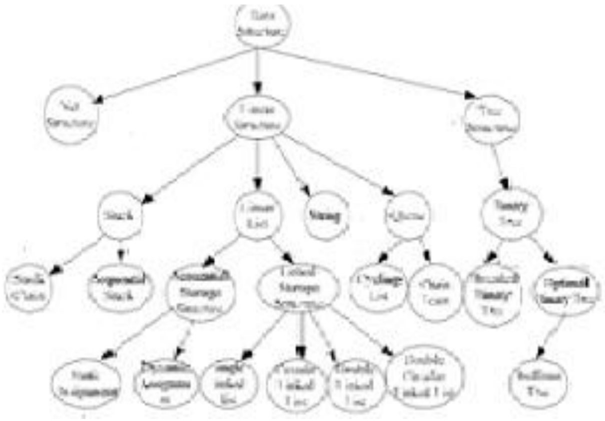

Fig. 4: Data Structure Ontology.

Table 2: Results with Data Structure ontology

Similarity of Concept Pair	Method [3]	Expert Value [3]	Our Method	Remark
Linear Structure, Linear List	88.7	90	80	Parent-child
Huffman Tree, Static Assignment	33.6	20	8.5	Different sub Hierarchy
Optimal Binary tree, Huffman Tree	99.8	100	88	Parent-Child

The work employs a piece of ontology “Data Structure” as shown in figure 3. The aim of selecting this research for comparison is to evaluate and validate our proposal. The results obtained by our method with the results in [3] are shown in table 2. The results of the experimental evaluation confirm our proposal. Using this metric, semantic score of the concepts is measured; higher score means highly similar and less score means less similarity in concepts. When a user tries to search for a concept; system expands the search using semantic similarity. The search resolves heterogeneity of the concept with the help of domain ontology. The search results provide more efficient Information retrieval.

5.2. Outcome of experimental evaluation

Attempting regular trials and testing of the method with experts, we have come across with following inferences:

- Levels of Domain Ontology Tree play a very important role in similarity measurement.
- Immediate Parent and child are semantically closest and the similarity between them is directly proportional to depth of concepts.ie deeper the concepts, semantically closer they are.
- Lower level siblings are semantically more similar than upper level siblings.
- Parent and child similarity is more as compared to siblings of the same sub-tree.
- Concept pair within the same sub hierarchy is semantically closer as compared to the concepts contained in two different sub hierarchies.

6. Conclusion

This paper presents a measure of semantic similarity between concepts in domain ontology. The contribution of this research work is to use inheritance relationship of ontological concepts to design, a measure, to get semantic similarity values among concept pairs of domain ontology. The method employs semantic similarity to empower semantic search by resolving heterogeneity issues in web documents. Semantic similarity facilitates in intensifying keyword search. It includes semantics to the search, in order to make information retrieval more efficient.

The proposed method calculates the similarity between two concepts and gives a semantic score. The method is based on IS – A hierarchy and demonstrates how to calculate the similarity score in any two concepts of domain ontology. The experiments conducted for validation of proposed metric and comparative analysis with existing methods, strengthen our approach. In future, the proposed methodology can be easily extended for semantic relatedness of concepts for deeper search. We plan to work for, ranking the retrieved documents according to the semantic similarity of the keyword and content.

References

[1] Djamel Guessoum, Moeiz Miraoui, Chakib Tadj "modification of Wu and Palmer Semantic Similarity Measure "UBICOMM 2016: The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp 42-46. 2016. ISBN: 978-1-61208-505-0.

[2] Zhao Guozeng "Research on concept similarity calculation method based on the semantic grid" Journal of Chemical and Pharmaceutical Research, 2015, 7 (3): 476-481ISSN: 0975-7384CODEN (USA).

[3] Wenjie Li, Qiuxiang Xia," A Method of Concept Similarity Computation Based on Semantic Distance "Advanced in Control Engineering and Information Science. SciVerseScienceDirect procedia Engineering 15 (2011) 3854 – 3859, 877-7058 © 2011 Published by Elsevier Ltd.

[4] Jike Ge, Yuhui Qiu,"Concept Similarity Matching Based on Semantic Distance" Fourth International Conference on Semantics, Knowledge and Grid 0-7695-3401-5/08 2008 IEEE.

[5] Z. Wu and M. Palmer. “Verb semantics and lexical selection”. In Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics, pp 133-138. 1994. https://doi.org/10.3115/981732.981751.

[6] D. Lin. “An Information-Theoretic Definition of similarity”. In Proceedings of the fifteenth International Conference on MachineLearning (ICML'98).MorganKaufmann: MadisonWI, pp.296-304.1998.

[7] P. Resnik (1995). “Using information content to evaluate semantic similarity in taxonomy”. In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.

[8] N. Ho and F. Cédrick. “Lexical Similarity based on Quantity of Information Exchanged-Synonym Extraction”. In the Proceeding of Conf. RIVF’04, February 2-5, 2004. Hanoi, Vietnam.

[9] J.H. Lee, M.H. Kim and Y.J. Lee. “Information Retrieval Based on Conceptual Distance in IS-A Hierarchy”. Journal of Documentation 49, pp 188-207, 1993. https://doi.org/10.1108/eb026913.

[10] Rada, H. Mili, E. Bichnell, and M. Blettner, “Development and application of a metric ton semantic net”. IEEE Transaction on Systems, Man, and Cybernetics. pp 17-30. 1989. https://doi.org/10.1109/21.24528.

[11] J. Zhong, H. Zhu, J. Li and Y. Yu, “Conceptual graph matching for semantic search,” The 2002 International Conference on Computational Science (ICCS2002), Amsterdam, pp. 92-106, 2002.

[12] Y. Ganjisaffar, H. Abolhassani, M. Neshati and M. Jamali, “A Similarity Measure for OWL-S Annotated Web Services,” Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 621-624, 2006. https://doi.org/10.1109/WI.2006.26.

[13] Troy Simpson, Thanh Dao, “WordNet-based semantic similarity measurment”2010.

[14] R. C. Veltkamp and L.J. Latecki. “Properties and Performances of Shape Similarity Measures”. 2006.

[15] M. Dean, G. Schreiber, S. Bechhofer, F. van Harmelen, J.Hendler, I. Horrocks, and L. A. Stein, “OWL web ontology language reference,” W3C Recommendation February 10, 2004.