A New Ontology-Based Semantic Similarity Measure for Concepts Subsumed by Multiple Super Concepts

Ayesha Banu[1], Syeda Sameen Fatima[2], Khaleel Ur Rahman Khan[3]
[1]Department of Computer Science
Alluri Institute of Management Sciences
Kakatiya University, Warangal, A.P, India
[2]Department of Computer Science & Engineering
Univ. College of Engineering ,OU, Hyderabad, A.P, India
[3]Department of Computer Science & Engineering
ACE Engineering College, JNTU, Hyderabad, A.P, India
{ayeshaaims, sameenf}@gmail.com, khaleelrkhan@aceec.ac.in

ABSTRACT: *Semantic Similarity relates to computing the similarity between concepts of ontology. There exist four approaches to calculate the semantic similarity. The first approach is based on path length. Under this approach we studied and compared some of the measures on a bench mark dataset. Among the compared measures Wu & Palmer measure has the advantage of being simple to implement and has better performance compared to the other similarity measures. This measure considers only the depth of the LCS: (Least Common Subsumer) we call in our paper as Closest Common Parent for similarity computation. But there are complex and large taxonomies, covering thousands of interrelated concepts including several overlapping hierarchies, and extensive use of multiple inheritances (i.e. a concept is subsumed by several super concepts). For such taxonomies using only the LCS will ignore a great amount of explicit knowledge. To overcome this limitation we propose ontology based semantic similarity measure which extends Wu & Palmer measure by considering ASC :(All Subsumed Concepts). We compared both the measures on two benchmark datasets. The obtained results show that our measure gave improved similarity values compared to the Wu and Palmer measure.*

Keywords: Ontology, Semantic Similarity, Is-Taxonomy

Received: 10 October 2013, Revised 18 November 0213, Accepted 29 November 2013

© 2014 DLINE. All Rights Reserved

1. Introduction

Semantic Similarity is a measure which is used to compute the similarity between two concepts within ontology. This is regarded as a research subject mostly investigated in the fields of data processing, Artificial Intelligence, and linguistics. In particular, the field of the information retrieval is largely based on the similarity identification measures between documents [1]. We can distinguish four methods to determine the semantic similarity between concepts in ontology. The first approach computes similarity based on conceptual distance (also called edge based methods). The second approach use the information content of concepts for finding similarity (also called the node based method). The third approach is based on the features of the compared

concepts. The fourth approach is hybrid which combines the above approaches. Finding similar concepts in ontology is a core task in the area of ontology alignment/merging [2].

The remainder of this paper is organized as follows: Section 2 presents an overview of the four categories of semantic similarity measures. Section 3 explains seven different edge counting methods to compute semantic similarity. Section 4 presents comparison of four edge counting methods performed on Univ_Bench Ontology. Section 5 & 6 concentrates on the newly proposed measure. Section 7 gives the properties satisfied by the proposed measure and finally Section 8 concludes the paper. References are included in Section 9.

2. Semantic Similarity Measures

Several methods have been proposed for determining semantic similarity between concepts of ontology. They are divided in to four main categories [3]

2.1 Edge Counting Methods

These methods measure the similarity between two concepts c1, c2 by determining the path linking the terms in the taxonomy and the position of the terms in the taxonomy.

2.2 Information Content Methods

In this category, similarity measures are based on the Information content of each concept.

2.3 Feature based Methods

Measures that consider also the features of the concepts in order to compute similarity.

2.4 Hybrid Methods

Those methods combine ideas from the above three approaches in order to compute semantic similarity between c1 and c2.

We performed a survey [4] on the measures of all the four categories and Compared 8 different measures on two benchmark datasets. In this paper we studied 7 edge counting measures and compared four of them on univ_bench ontology. We identified the best performing measure and also its limitation. The measure proposed in this paper overcomes the limitation and which comes under the same category of measures.

3. Edge Counting methods to compute Semantic Similarity

This category of measures is based on how close the two concepts in the taxonomy are. Let O be ontology with set of concepts C. Let $c1, c2 \in C$.

3.1 Path Length Measure

Rada et al. [5] proposed this measure where $d(c1, c2)$ is the shortest path between the concepts $c1$, $c2$.

$$Dist_{Path}(c1, c2) = d(c1, c2) \tag{1}$$

3.2 Leacock & Chodorow

This measure [6] also uses the path length value along with the depth of the taxonomy given as

$$Sim_{LC}(c1, c2) = -\log\left(\frac{d(c1, c2)}{2D}\right) \tag{2}$$

Where $d(c1, c2)$ is the shortest path between the concepts $c1$, $c2$ and D is the depth of the taxonomy. This measure can also be written as

$$Sim_{LC} = \log\left(\frac{2D}{d(c1, c2)}\right) \tag{3}$$

Let us consider, for example, fragment of the Univ_Bench ontology [7] in OWL shown in Figure 1 to explain the semantic similarity calculation. We use Protégé 4.2 alpha [8] for visualizing the ontology in the form of Is-a taxonomy.

In this fragment path length between the concepts "*Employee*" and "*Student*" is 3 using node counting. The path length between concepts "*Lecturer*" and "*Professor*" is also 3. The similarity in these two cases is the same by Path length measure. However, intuitively speaking, the similarity between "*EMPLOYEE*" and "*STUDENT*" will be less than the similarity between "*LECTURER*" and "*PROFESSOR*".

Leacock and Chodorow also give same similarity value for both the pairs. For $D = 5$ SimLC(Employee, Student) = log (10/3) = 0.522 and Sim_{LC} (Lecturer, Professor) = log (10/3) = 0.522.

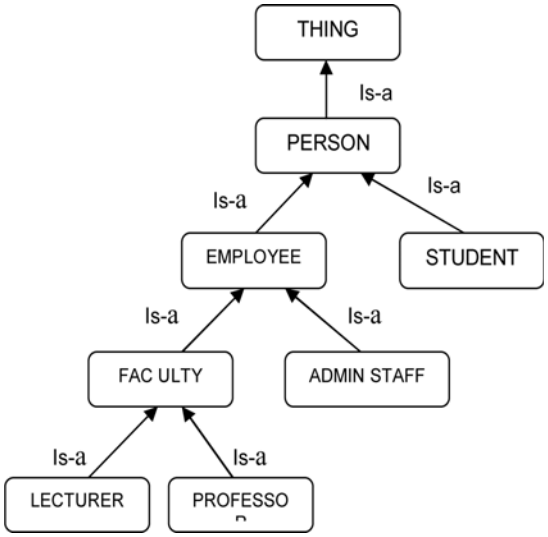

Figure 1. Part of Univ_Bench ontology

3.3 Wu and Palmer Measure

This measure [9] considers the depth of the Least Common Subsumer or the Closet Common Parent C_p for the concepts $c1$, $c2$. The measure is given as

$$Sim_{WP}(c1, c2) = \frac{2Np}{N1 + N2 + 2Np} \tag{4}$$

Np is the depth of Cp from root, N_1 is depth of $c1$ from Cp and N_2 is depth of $c2$ from Cp. $N1 + N2$ will result in shortest path between $c1$ and $c2$. Depth here is the number of is-a links.

$Sim_{\mathbf{WP}}$ (Employee, Student) = **2 * 1/1 + 1 + 2 * 1 = 2/4 = 0.5**

$Sim_{\mathbf{WP}}$ (Lecturer, Professor) = **2 * 3/1 + 1 + 2*3 = 6/8 = 0.75**

This measure shows the similarity between Lecturer and Professor is more than the similarity between Employee and Student.

To prove that this measure give accurate similarity values than the previous 2 measures we compared this measure against some other measures shown in Table 1 in section 4.

3.4 Li et al. Measure

This measure [10] combines the shortest path length (number of edges) between the concepts $c1$, $c2$ (L) and the depth of the closest common parent (Np). The measure is given as

$$Sim_{Li}(c1, c2) = e^{-\alpha L}\left[\frac{e^{\beta Np} - e^{-\beta Np}}{e^{\beta Np} + e^{-\beta Np}}\right] \tag{5}$$

$\alpha \geq 0$ and $\beta \geq 0$ are parameters scaling the contribution of shortest path length and depth respectively. As per Li et al. the optimal parameters are $\alpha = 0.2$ and $\beta = 0.6$. We use same values in our experiments also.

3.5 Mao et al. Measure

This Measure [11] defines a similarity measure using both shortest path information and number of descendents of compared concepts. The measure is given as

$$Sim_{Mao}(c1, c2) = \frac{\delta}{d(c1, c2)\log_2(1 + d(c1) + d(c2))} \quad (6)$$

Where $d(c1, c2)$ is the number of edges between c1 and $c2$, $d(c1)$, $d(c2)$ the number of descendants of $c1$, $c2$.

δ is a constant whose value is set to 0.9. If both the concepts are leaf concepts with no descendants then the measure results in zero denominator value.

3.6 Concept Specificity Measure

Al-Mubaid & Nguyen[12] propose a similarity measure where they assume every branch of the ontology at root node as one cluster. A common specificity value is calculated as

$$\text{CSpec}(c1, c2) = D - depth(\text{LCS}(c1, c2)) \quad (7)$$

D is the depth of the cluster to which the concepts $c1$, $c2$ belong and depth (LCS ($c1$, $c2$)) is the depth of the Least Common Subsumer i.e. the closest common parent of $c1$, $c2$. We can write formula (7) as

$$\text{CSpec}(c1, c2) = D - Np \quad (8)$$

Using this value the similarity value is computed by

$$\text{Sim}_{\text{CS}}(c1, c2) = \log[(path - 1)^{\alpha}(\text{CSpec})^{\beta} + k] \quad (9)$$

Path is the shortest path between $c1$, $c2$. $\alpha > 0$, $\beta > 0$ are contribution factors of two features; k is a constant.

The values of α, β, k are set to 1 experimentally.

3.7 Super Concept based Similarity Measure

This measure proposed by *M*. Batet et al. [13] for $c1, c2 \in C$. define a set

$T(Ci) = \{Ci\} \cup \{Cj \in C \mid Cj \text{ is the super concept of } Ci\}$.The similarity between 2 concepts $c1$, $c2$ is given as

$$Sim_{log} = -\log_2\left[\frac{|T(c1) \cup T(c2)| - |T(c1) \cap T(c2)|}{|T(c1) \cup T(c2)|}\right] \quad (10)$$

4. Comparison of Similarity Measures

In this section we perform an experiment on a benchmark dataset for comparing the similarity measures in order to find the best performing measure. We consider Univ_Bench ontology [7] in OWL which describe data pertaining to universities and their departments. The choice of this ontology is justified by the fact that it presents a field about which users are familiar. This ontology is developed for benchmarking reasons and it contains 45 concepts around 4 major subjects like Work, Organization, People and Publication. We randomly selected 10 concept pairs from all subjects and compared four measures given in equations (2) (4) (5) (9) of section 3. The measures used are represented by LCH- Leacock & Chodorow, Li - Li et al. Measure, WP- Wu and Palmer, CSpec- Concept Specificity Measure. This comparison is only to test which measure among the mostly used semantic similarity measures under this category perform well. The comparison is shown below.

We computed the average value of all the ten concept- pairs over which we ranked the measures. WP:Wu & Palmer measure ranked first in our comparison. All the measures in this category use the concept of Closest Common Parent suggested by Wu & Palmer as it is considered to provide accurate similarity values.

In next section we show the limitation of Wu & Palmer measure and propose a new measure. We will also show its comparative results on two bench mark datasets.

CONCEPT PAIR	Li	LCH	WP	Cspec
1)AsstProf- AssoProf	0.65	0.52	0.8	0.301
2)Lecturer - Professor	0.62	0.52	0.75	0.477
3)University- Institute	0.36	0.52	0.5	0.301
4)TeachingCourse-ResearchWork	0.36	0.52	0.5	0.301
5)UGStudent – ReseachAsst	0.55	0.52	0.66	0.6
6)Clerk-Systemworker	0.62	0.52	0.75	0.477
7)Dean - Chair	0.65	0.52	0.8	0.301
8)AssoProf- VisitProf	0.65	0.52	0.8	0.301
9)GradLevelCourse – ReschLevelCourse	0.55	0.52	0.66	0.301
10)Chair – AsstProf	0.65	0.52	0.8	0.301
Average(Rank)	**0.56(2)**	**0.52(3)**	**0.7(1)**	**0.36(4)**

Table 1. Comparison of 4 different measures on Univ_Bench Ontology

5. New Semantic Similarity Measure

Analyzing the path-based methods, we notice that these measures consider the shortest path between a pair of concepts, and it is the sum of Is-a links between each of the concepts and their Least Common Subsumer LCS. If one or both concepts subsume

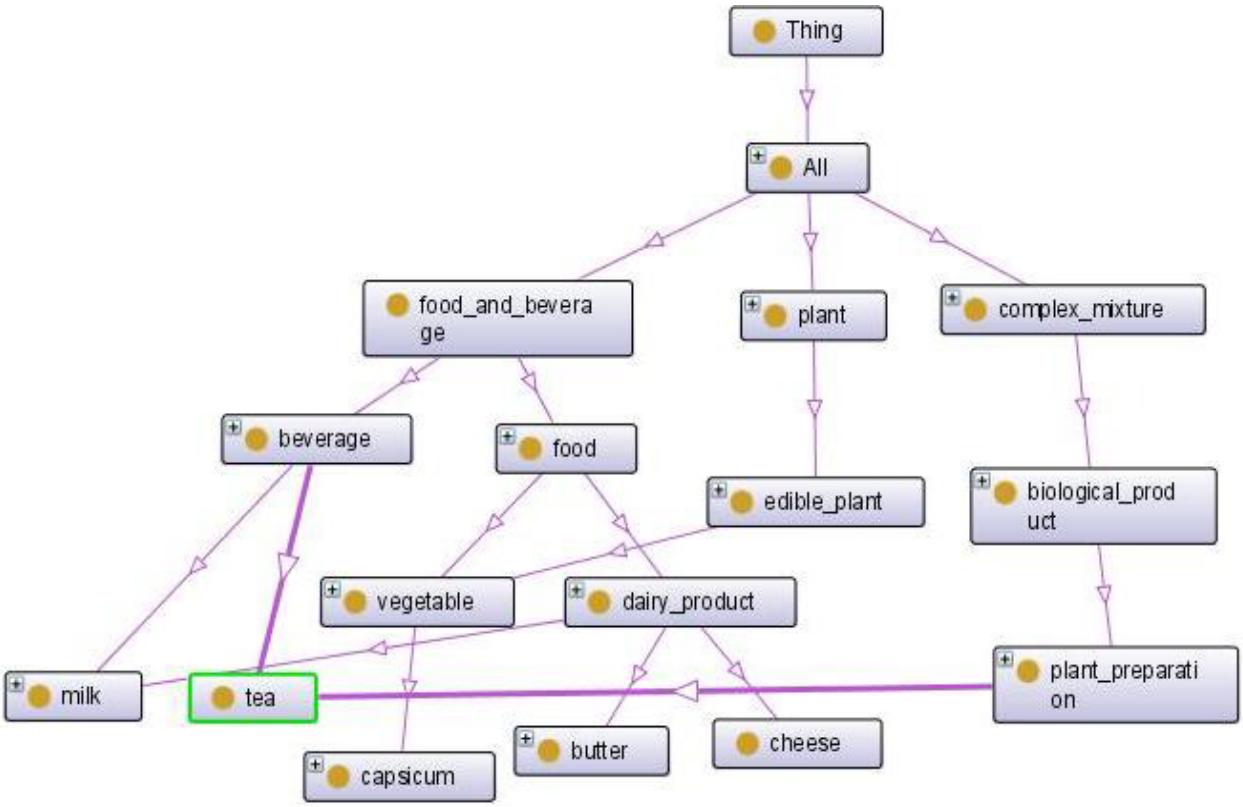

Figure 2. A Snippet of MeSH Ontology

more than one super concept i.e. inherit from several is-a hierarchies, all possible paths between the two concepts are calculated, but only the shortest one is kept for consideration.

Wu & Palmer measure includes depth of the Closest Common Parent from the root which improves the measure comparatively.

But for complex and large taxonomies, covering thousands of interrelated concepts including several overlapping hierarchies, and an extensive use of multiple inheritance (i.e. a concept is subsumed by several super concepts) considering only the LCS waste a great amount of explicit knowledge. This can be a limitation of Wu & Palmer measure.

Examples for such a large ontology with several concepts showing multiple inheritances are MeSH.owl Ontology [14] and Human.owl Ontology [15]. MeSH ontology is an OWL ontology which is used as a benchmark for several experimental analysis which as per [16] is a collection of 229698 classes around 108 major subjects. Human.owl is OAEI bench mark dataset.

We show a small snippet of MeSH ontology below with concepts subsumed by multiple super concepts.

We propose a measure considering the depth of all the subsumed concepts instead of only the depth of LCS while computing similarity. This can give more improved results. The proposed measure abbreviated by

ASC-All Subsumed Concepts is given as $Sim_{ASC}(c1, c1) = \dfrac{\sum_{k=1}^{n} 2N_{kp}}{N1 + N2 + \sum_{k=1}^{n} 2N_{kp}}$ (11)

K is number of common parents along all paths for the concepts whose similarity is computed. We keep the value of $N_1 + N_2$ same as Wu & Palmer measure as it gives the shortest path from $c1$ to $c2$. Our major concentration in the proposed measure is on the depth of the super concepts.

The following algorithm shows the working of the measure proposed in the above equation (11) for any two concepts of an input ontology.

Algorithm SemSim_ASC (C1, C2)

Input: O (Input Ontology), **C1, C2**(any two concepts of **O**)

Output: Semantic Similarity between C1 and C2

1. Let **C** be set of all concepts of O & **rt** be the root of O

2. For C1, C2 ∈ C

a. Extract SP_{1i} : set all parent nodes for $C1$ along path i
SP_{2i}: set all parent nodes for $C2$ along path i

b. Compute depth ($C1_{ji}$) :depth of C1 from parent j along path i
depth ($C2_{ji}$) : depth of $C2$ from parent j along path i

c. ∀ paths i find CP ($C1$, $C2$): Common Parent

3. ∀ paths i Compute

a. Np = depth (CP, rt)

b. depth ($C1$, CP)

c. depth ($C2$, CP)

4. $\text{SemSim\_ASC}(C1, C2) = \dfrac{\sum_{k=1}^{n} 2 * Nkp}{\min_{\forall i}(depth(c1, cp)) + \min_{\forall i}(depth(c2, cp)) + \sum_{k=1}^{n} 2 * Nkp}$

K: number of common parents

6. Experimental Results

We compare the proposed measure with Wu & Palmer over 29 concept pairs of the two benchmark datasets shown in table 2

below.

Dataset 1: MeSH Ontology: MeSH(Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors(concepts) in a hierarchical structure that permits searching at various levels of specificity. There are 26,853 descriptors in 2013 MeSH [17].

Dataset 2: Human Ontology: This dataset is taken from the Anatomy track of OAEI Ontology Alignment Evaluation Initiative 2013 Campaign . Since 2004, OAEI organizes evaluation campaigns aiming at evaluating ontology matching technologies [18]

Dataset	Concept	Concept Pair	SIM_{WP}	SIM_{ASC}
DATA SET – 1 : MeSH Ontology (Mesh. OWL)	FOOD & BEVERAGE	(Milk , Mineral Water)	0.75	0.66
		(Milk , Tea)	0.75	0.8
		(Milk , Cheese)	0.66	0.85
		(Milk , Coffee)	0.75	0.8
		(Food , Beverage)	0.66	0.66
		(Beer , Wine)	0.8	0.8
		(Cheese , Ice cream)	0.8	0.8
		(Cheese , Butter)	0.8	0.875
		(Fruit , Vegetable)	0.75	0.857
		(Fruit , Meat)	0.75	0.8
	BODY REGION	(Forehead , Eye)	0.8	0.83
		(Nose , Eye)	0.8	0.87
		(Scalp , Face)	0.75	0.75
		(Face , Ear)	0.75	0.8
	EYE DISEASE	(Low Vision , Color Vision Defect)	0.75	0.75
		(Low Vision , Amblyopia)	0.75	0.8
		(Amblyopia , Diplopia)	0.75	0.8
	PERSON	(Working Women , Preg Women)	0.75	0.75
		(Women Physician , Women Dentist)	0.75	0.9
		(Mother , Father)	0.75	0.75
		(Mother , Single Parent)	0.75	0.83
DATASET – 2: Human Anatomy: Human. owl	ORGAN	(Lung , Liver)	0.66	0.66
		(Ovary , Fallopian Tubes)	0.66	0.875
		(Skin , Kidney)	0.66	0.66
		(Brain , Spinal Cord)	0.66	0.85
	MICRO ANATOMY (TISSUE)	(Lung Tissue , Gastric Tissue)	0.75	0.75
		(Lung Tissue , Tonsillar Tissue)	0.75	0.8
		(Tonsillar Tissue , Splenic Tissue)	0.75	0.85
		(Lymph Node Tissue , Tonsillar Tissue)	0.75	0.85

Table 2. Comparison of WP & ASC on part of MeSH & Human Ontology

The table above shows for the concepts like Food, Beverage and Beer, Wine the similarity value is same for both the measures because they are subsumed by only one super concept. Thus our measure gives the same best result as of Wu & Palmer for such concepts.

For other concepts which are subsumed by many super concepts our proposed measure shows good improvement in the similarity value.

For example in first block we observe that WP shows same similarity value for Milk, Mineralwater and Milk, Tea. We can observe

that the similarity can never be same. It should be more for Milk,Tea than for Milk,Mineralwater. This happens because WP measure considers only the Closest Common Parent. Our measure clearly shows that SimASC (Milk, Tea) = 0.8 is more than SimASC (Milk, Mineralwater) = 0.66.

The second block shows SimWP (Milk, Cheese) = 0.66 is less than SimWP(Milk,Coffee) = 0.0.75. But it is a known fact that Milk and Cheese are more similar than Milk and Coffee. Our measure shows that SimASC (Milk, Cheese) > SimASC (Milk,Coffee).

Table 2 shows many such noticeable differences in similarity values resulted by Wu & Palmer measure and the proposed measure. It brings a considerable improvement by considering all subsumed concepts. Figure 3 below shows the comparison of our proposed measure with the Wu & Palmer measure for few conceptpairs. The concept-pairs are taken over X-axis and similarity values on Y-axis.

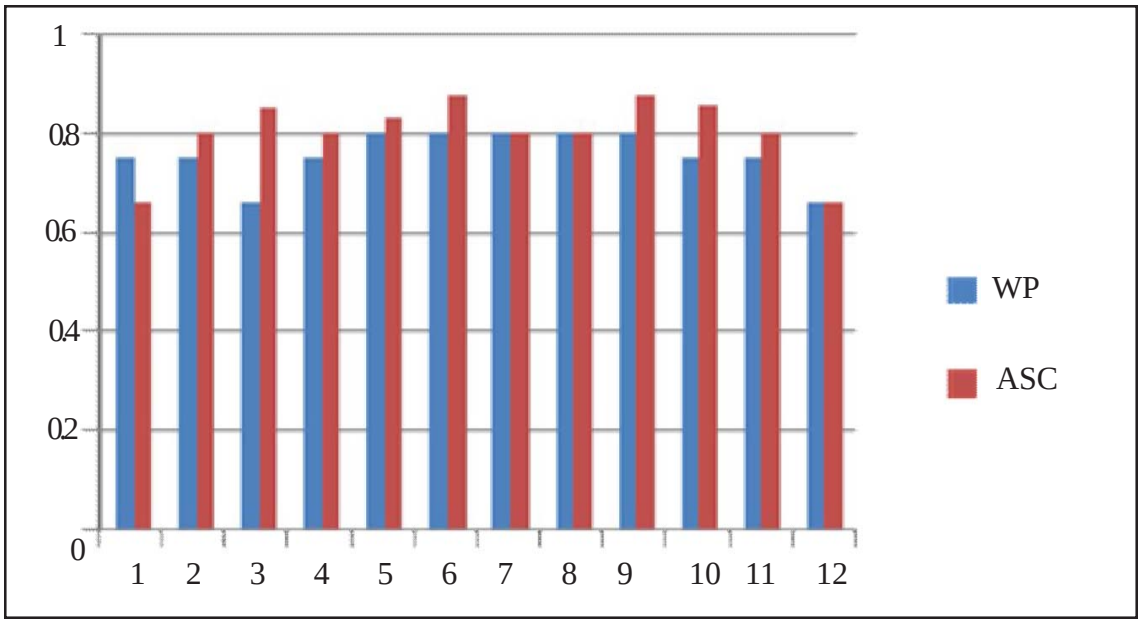

Figure 3. Comparison of the effectiveness of our measure compared to the Wu and Palmer measure

Semantic similarity measures play a key role in ontology alignment/mapping, Information Retrieval, Information Integration and other applications involving comparison between concepts. Ontology mapping aims to find the concepts which are similar between any two ontology's of some specific domain. These measures help in computing the similarity. They also provide necessary Semantic context information for information retrieval applications.

7. Property of Proposed Similarity Measure

In this section we enumerate some properties of similarity measure [19]. These properties depend on a particular application; sometimes a property will be useful, sometimes it will be undesirable. The function of similarity which we propose ensures the following properties: For any three concepts $c1$, $c2$ and $c2$ of ontology O:

1) **Nonnegativity:** $\mathrm{Sim}_{ACS}(c1, c2) \geq 0$

2) **Symmetry:** $\mathrm{Sim}_{ACS}(c1, c2) = \mathrm{Sim}_{ACS}(c2, c1)$

3) **Triangle inequality:** $\mathrm{Sim}_{ACS}(c1, c2) + \mathrm{Sim}_{ACS}(c2, c3) \geq \mathrm{Sim}_{ACS}(c1, c3)$

4) **Strong triangle inequality:** $\mathrm{Sim}_{ACS}(c1, c2) + \mathrm{Sim}_{ACS}(c1, c3) \geq \mathrm{Sim}_{ACS}(c2, c3)$

8. Conclusions

In this work we have presented seven semantic similarity measures in an Is-a taxonomy based on the notion of edge counting. We performed the first experiment on the Univ_Bench ontology and compared four measures to find which measure performs

better. The results show Wu & Palmer measure as the rank 1 measure.

Yet identifying the limitation I this measure we proposed a semantic similarity measure (Sim_{ASC}) which is an extension of the Wu & Palmer measure(Sim_{WP}) which works on concepts in the Is-a taxonomy and which is subsumed by multiple super concepts. The comparison between WP & ASC on concept - pairs of the Is-a taxonomy of the MeSH and Human ontology show that our measure gives improved results over WP. The proposed measure takes knowledge of overlapping concepts during similarity computation. This also basically satisfy the properties of similarity measures.

References

[1] Baeza-Yates, R., Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass.

[2] Noy, N. F., Musen, M. (2000). PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. *In*: Proceedings of AAAI-2000, Austin, Texas. MIT Press/AAAI Press.

[3] Angelos Hliaoutakis, Giannis Varelas, Epimeneidis Voutsakis, Euripides G. M. Petrakis, Evangelos Milios. (2006). Information Retrieval by Semantic Similarity, *Intern. Journal on Semantic Web and Information Systems* (IJSWIS), 3 (3) 55–73, July/Sept. Special Issue of Multimedia Semantics.

[4] Ayesha Banu, Syeda Sameen Fatima, Khaleel Ur Raham Khan. (2013). A Survey and Comparison of WordNet Based Semantic Similarity Measures, *International Journal of Computer Science And Technology* (IJCST). 4 (2) 456-461, April - June.

[5] Rada, R., Mili, H., Bichnell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics. p 17-30.

[6] Claudia Leacock, Martin Chodorow. (1998). Combining local context and WordNet similarity for word sense identification. In [20].

[7] http://www.lehigh.edu/~zp2/2004/0401/univ-bench.owl

[8] http://protege.stanford.edu/download/protege/4.3/ installanywhere/Web_Installers/

[9] Wu, Z., Palmer, M. (1994). Verb semantics and lexical selection. *In*: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, p. 133-138.

[10] Yuhua Li, Zuhair A. Bandar, David McLean. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering, 15 (4) 871-882, July/August.

[11] Mao, W., Chu, W. W. (2002). Free text medical document retrieval via phrased-based vector Space model, *In*: Proc. of AMIA'02, San Antonio,TX.

[12] Al-Mubaid, H., Nguyen, H. A. (2006). A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain, *In*: Proc. The 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society EMBS, New York, USA, September.

[13] Batet, M., Sánchez, D., Valls, A. (2011). An ontology- based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatic*s. In press.

[14] http://bike.snu.ac.kr/sites/default/files/meshonto.owl

[15] http://oaei.ontologymatching.org/2013/ anatomy/index.html

[16] http://bioportal.bioontology.org/ontologies/3019

[17] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

[18] http://oaei.ontologymatching.org/2013/

[19] Veltkamp, R. C., Latecki, L. J. (2006). Properties and Performances of Shape Similarity Measures.

[20] Christianne Fellbaum. (1998). WordNet: An Electronic Lexical Database. The MIT press.