A Comparison of Semantic Similarity Measures

Ayesha Banu Mohd
Vaagdevi College of Engineering
India
ayesha_b@vaagdevi.edu.in

ABSTRACT: *In natural language processing, identifying semantic similarity is a major challenge for which several measures have been introduced and tested. There are four types of semantic measures which includes the following. First are the Ontology Based measures that depend on the closeness of the concepts in the taxonomy. Ontology-based measure detects the similarity in terms of the path linking the concepts and position of the concept in the hierarchy. The second one is related to the Information Content of concepts is considered to find the semantic similarity. The third and fourth includes the Feature-based similarity and Hybrid similarity measures. Semantic similarity measurements have wider impact in other areas such as data mining, computational intelligence, linguistics, information retrieval systems and so on. Whatever the measure is used, arriving at the quality is a prominent question. For identifying the effectiveness, the systems like the Psycholinguistic evaluation is used to justify the similarity measure quality. The comparison of the proposed semantic similarity measure values is being carried out with the expert opinion. The values are statistically tested suing Pearson Correlation Coefficient is used to test the quality of the similarity measure. When the correlation between the computational method value and the human assessment values. This work described the proposed ASC semantic similarity measure and its Psycholinguistic evaluation versus the opinion of the experts.*

Keywords: Semantic Similarity, Ontology, Information Retrieval, Psycholinguistic evaluation, Correlation Coefficient

Received: 18 August 2021, Revised 9 December 2021, Accepted 17 December 2021

DOI: 10.6025/jcl/2022/13/1/1-8

1. Introduction

Computing semantic similarity is mostly applied in the areas of Information Retrieval, Information Integration and other areas of application where the concepts are compared with each other [1,2] to discover ontology mapping [3], to validate ontology mappings[4] and word-sense disambiguation [5]. Ontology-based approaches depend on the closeness of the concepts in the taxonomy. If O represents ontology, C represents a set of all the concepts of ontology O and C1, C2, C. These measures find the similarity in terms of the path linking the concepts and position of the concept in the hierarchy. These measures are easy to implement but always require to work on rich and consistent ontologies. Information Content based approaches [6, 7, 8] exploit the notion of Information Content (IC) value of the concepts. The IC value is computed from the taxonomy or large corpus like

WordNet. According to IC based measures the semantic relatedness between the concepts is related to the information they share in common. The more the information is shared, the more is the similarity. This is definitely better compared to ontology-based measures. The features of a term contain knowledge and valuable information about the term. Feature Based Semantic Similarity Measures use this feature as a basic source for computing the similarity. These measures use more semantic knowledge by considering both commonalities and differences of compared concepts. Hybrid Semantic Similarity Measures are combination these approaches. This paper give a brief of some important semantic similarity measures under the first two approaches and explains the problem in similarity computation when the concepts exhibit multiple inheritances. The paper also briefs the proposed measures which overcome the problem of multiple subsumptions. In order to evaluate the proposed semantic similarity measures, a similarity experiment has been conducted to collect ratings of similarity provided by human subjects. This paper explains the evaluation experiment done to prove the quality and accuracy of the proposed measures.

2. Ontology Based Semantic Similarity Measures

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 $c_1, c_2 \in C$.

2.1. Path Length Measure

Rada et al. [9] proposed this measure

$$\boldsymbol{Dist\ Path(c1, c2) = d(c1, c2)} \qquad (1)$$

where $d(c1,c2)$ is the shortest path between the concepts $c1$, $c2$.

2.2. Leacock & Chodorow Measure

This measure [10] also uses the path length value along with the depth of the taxonomy given as

$$\textbf{Sim}_{\text{LC}}\,\textbf{(c1, c2)=}\ \log\left(\frac{2\text{D}}{\text{d(c1, c2)}}\right) \qquad (2)$$

Where d(c1, c2) is the shortest path between the concepts c1, c2 and D is the depth of the taxonomy.

2.3. Wu and Palmer: The previous two measures mainly depend on the shortest distance between t he concepts for which similarity is computed.

$$\textbf{Sim}_{\text{WP}}\,\textbf{(c1, c2)} = \frac{2Np}{N_1 + N_2 + 2Np} \qquad (3)$$

The similarity is defined as the closeness of the concepts in the hierarchy. In Wu & Palmer measure [11] CP is the Closest Common Parent (CCP) of C1 and C2. N1 is the number of edges from C1 to CP. N2 is the number of edges from C2 to CP. NP is the number of edges from CP to the root. N1+N2 is the shortest path between C1 and C2. Depth is the number of links or edges. The wu & palmer measure shows a considerable improvement in similarity value compared to other two measures. This measure has laid a foundation to many other measures in this category. Slimani et.al.[12] and Ganeshan et.al. [13] also propose measures extending wu & palmer. Many more measures under this category are explained in [19].

The advantage of Edge-Based measures is their simplicity. They always depend on the "is-a" hierarchy of input ontology. It also takes the low computational cost to evaluate these measures. The limitations that affect the performance of these measures are they always rely on the shortest path between the concepts. But, these measures when applied on large ontologies like Word Net & MeSH, which support multiple inheritances, they ignore most of the taxonomical knowledge which is modeled in the ontology explicitly.

3. Information Content Based Semantic Similarity Measures

3.1. Resnik's Measure

In view of the limitations of edge-counting approaches, Resnik proposed to complement the taxonomical knowledge provided by ontology with a measure of the information content of concepts computed from corpora like WordNet. The idea behind this semantic similarity is that the similarity of two concepts is related to information they share in common.

3.1. Resnik's Measure

As per resnik [14] for any concept C the information content is given as **IC (C) = -log P(C),**

Where $P(C)$ is the probability of the concept C in the corpora. The probability is computed as **P(C) = freq(C) / N**

$$Sim_{RES}(c1, c2) = \mathrm{IC(CCP(C1, C2))} \tag{4}$$

3.2. Jiang and Conrath measure: This measure [15] is proposed using resniks measure

$$\mathrm{SimJC(c1, c2) = (IC\ (c1) + IC(c2) – 2^*\ Sim_{RES}\ (c1, c2)} \tag{5}$$

This is based on quantifying the length of the taxonomical links as the difference between the IC of a concept and its subsumer. When comparing term pairs, they compute their distance by subtracting the sum of the IC of each term alone from the IC of their CCP(Closest Common Parent).

3.3. Lin's Measure

As per Lin[16] the similarity between two terms should be measured as the ratio between the amount of information needed to state their commonality and the information needed to fully describe them. His measure considers commonality in the same manner as Resnik's approach on one hand and the IC of each concept alone on the other hand.

$$\mathrm{Sim}_{LIN}\mathrm{(c1, c2) = 2\ ^*\ Sim_{RES}\ (c1, c2) / IC\ (c1) + IC\ (c2)} \tag{6}$$

3.4. Lord et al. Measure: This measure [17] is given as

$$\mathrm{Sim_{Lord}(c1, c2)=1\text{-}\ Sim_{RES}\ (c1, c2)} \tag{7}$$

3.5. Seco et al. Measure

This measure [18] considers the hyponyms of the WordNet to calculate the Information Content value.

The similarity function is obtained by normalizing and applying a linear transformation to the Jiang and Conrath formula. They argue that the more hyponyms a concept has the less information it expresses and concepts that are leaf nodes are the most specified in the taxonomy so the information they express is maximal. The information content value is computed as

$$\mathrm{ICWN(c) = 1\text{-}[log(hypo(c)+1) / log(maxwn)]} \tag{8}$$

where the function hypo returns the number of hyponyms of a given concept and maxwn is a constant that is set to the maximum number of concepts that exist in the taxonomy. The similarity value using can be given as

$$\mathrm{Sim(c1,c2)=}\ 1-\frac{\mathrm{ic_{wn}(c1)+ic_{wn}(c2)-2^*sim_{res}(c1,c2)}}{2} \tag{9}$$

Simres corresponds to Resnik's similarity function but now accommodating ICWN values.

In a is-a taxonomy the hyponyms of any concept are same as the descendants of the concept in the hierarchy. Using this statement revised measure of Seco is given as

$$IC_{ONT}(C) = 1 - \left[\frac{\log(num\_desc(C)+1)}{\log(Max_{ONT})}\right] \tag{10}$$

The number of successors or descendants for a concept C is represented by num_desc(C) and a total number of concepts in the ontology is given MaxONT. The normalized values for this measure range between [0...1]. the value of ICONT is 1 for the leaf concept is 0 for the root concept. We used equation (10) for computing IC value in the proposed measures.

The semantic similarity measures based on Ontology and Information Content always depend on the structure i.e., position of terms in the taxonomy and information content value respectively. The structure and information content are not directly comparable when concepts are taken from two different ontologies. This limitation can be addressed by Feature Based and Hybrid semantic similarity measures. A detailed survey and comparison of all the semantic similarity measures is presented in [19].

4. The Problem of Multiple Subsumption

Several measures have been proposed to compute the semantic similarity between any two concepts within a given ontology. Edge-Based Measures consider the depth of the closest common parent (CCP) and Information-Content Based Measures take the information content value of the CCP for computing the similarity between any two concepts. At present, there are many complex and large taxonomies which cover thousands of interrelated concepts and use several multiple inheritances. In such cases considering only the depth or information content of CCP ignores a large amount of explicit knowledge.

According to Cross, hu [20] According to Cross & Hu when one or both the concepts in the taxonomy are subsumed by multiple super concepts then the depth of all the concepts is computed and Np is assumed the smallest one. According to Resnik [14] compute the IC value for all subsuming concepts and retain the highest value as it is considered to be the most informative concept. But these measures may not produce accurate results by considering only least depth value or highest IC value. Most of the domain knowledge will be ignored that affects the resulting Semantic Similarity value.

5. ASC Based Semantic Similarity Measures

To address the problem of multiple inheritances while computing the semantic similarity between any two concepts of a taxonomy, this work proposes a measure called ASC: All Subsumed Concepts, which considers the depth of all the parent concepts instead of considering only the depth of the Closest Common Parent(CCP). Wu and Palmer measure is taken as base for the newly proposed measure. The survey [19] also reveals that remaining measures proposed later also adopt the principle of Wu and Palmer. [21] The proposed measure considers all the super-concepts which belong to all the possible taxonomical paths for the concepts evaluated. This captures as much semantic evidence as possible when the concepts represent multiple inheritances.

The proposed ASC: All Subsumed Concepts measure, Algorithm, its working principle and implementation with results is explained in [22]. To be brief enough, the algorithm takes any two concepts of an ontology as input and returns similarity between concepts as a numerical value. For both the concepts, set of all subsumers (parent concepts) are taken along all the paths of the taxonomy. For every parent j in the path i the concept depth is computed and for all paths i the common parent CP (c1, c2) is found. For all common parents CP we compute depth (CP, rt), depth (c1, CP) and depth (c2, CP).

The similarity value is computed using the measure (11). K represents all the common parents.

$$\text{SemSim\_ASC(C1,C2)=} \frac{\sum_{k=1}^{n} 2*N_{kp}}{\min_{\forall i}\left(depth(c1,cp)\right)+\min_{\forall i}\left(depth(c2,cp)\right)+\sum_{k=1}^{n} 2*N_{kp}} \qquad (11)$$

ASC based measure performs better than the Wu & Palmer measure especially for those concepts which exhibit multiple subsumptions [22].

Resnik's , Jiang & Conrath , Lord's and Lin's measures are considered as four basic IC based semantic similarity measures as they lay foundation for many other measures in this category. Many comparisons have been done on these measures[23][24][25] over different datasets to test which measure perform well in computing the semantic similarity value. As observed from the related work done, and also the survey results [19] the two measures Lin and Jiang & Conrath show higher level of correlation ranking 1 or 2. Hence the proposed ASC based measure consider only these two measures as a basis to form the new semantic similarity measure.

This measure computes semantic similarity considering the Information Content(IC) value of all the super concepts along all the possible paths connecting the two concepts. The Algorithm, its working principle and implementation with results for the ASC based measure is explained in [26]. The ASC measure first computes IC value and similarity for all the K common parents is computed using this IC value

$$IC_{ONT}(c) = 1 - [\log(num\_desc(c) + 1) / \log(max_{ONT})] \tag{12}$$

$$Sim_{ASC-Lin}(c1, c2) = \frac{2 * \sum_{k=1}^{n} IC(Ck)}{IC(c1) + IC(c2)} \tag{13}$$

$$Sim_{ASC-JC}(c1, c2) = 1 - \left[\frac{IC(c1) + IC(c2) - 2 * \sum_{k=1}^{n} IC(Ck)}{2}\right] \tag{14}$$

6. Psycholinguistic Evaluation

This evaluation approach has been adopted by many researchers to compare the semantic similarity values. The R&G study made by Rubenstein and Goodenough [29] was related to the relationship between similarity of context and meaning (synonymy). For their study they used 65 pairs of English words which were highly synonymous pairs, completely unrelated pairs and few with less similarity. These 65 pairs of words were rated by 51 human subjects on the scale of 0.0 to 4.0, depending on the word similarity. Their study supported Psycholinguistic Evaluation to be the best approach to measure the quality of any semantic similarity measure against the human judgment values. Pedersen et al[30] also did similar experiment 120 medical terms. Saruladha [27] also applied the Psycholinguistic Evaluation method to compare the new similarity measure against the human assessment values. Kalkowski & Sick[31] also did a Correlation for a Domain Specific Fashion Ontology to compare Similarity Measure with Human Judgment

This work also uses the same Psycholinguistic Evaluation on the basis of above study. We consider concept pairs from both MESH.owl and Human.owl datasets. Both the proposed semantic similarity measures [22, 26] are implemented on these concept

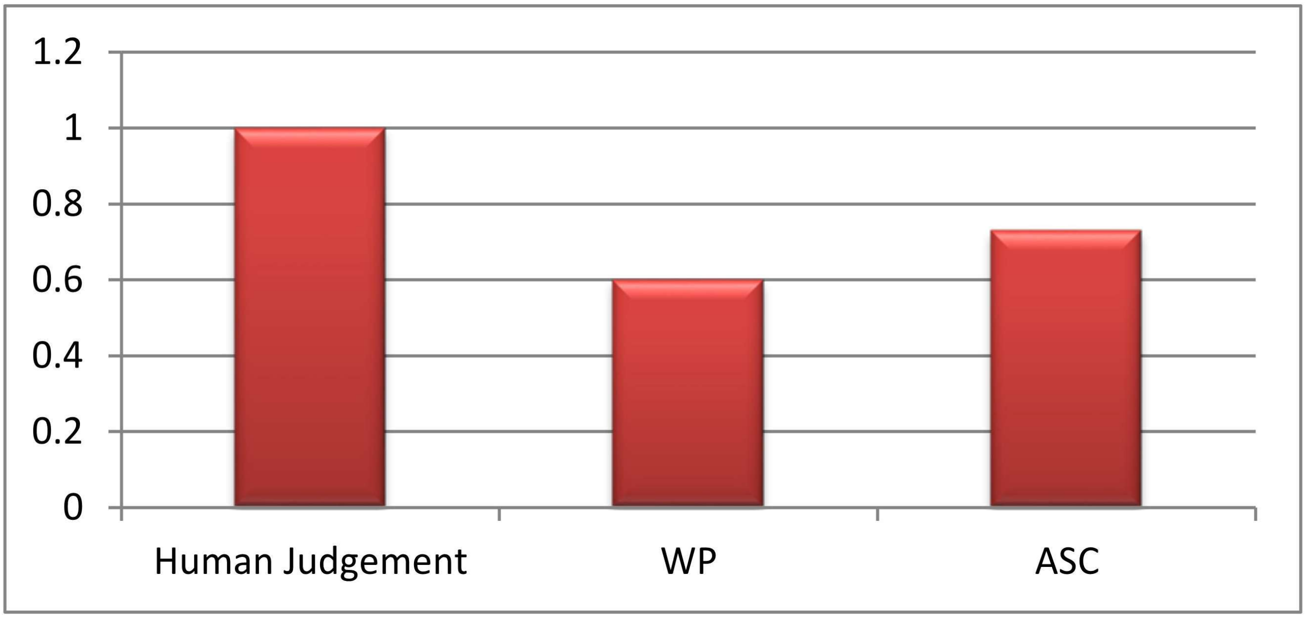

Figure 1. Correlation between Human assessment values with WP measure and the proposed ASC measure

pairs and evaluated against human assessment values. The similarity assessment values for all the pairs of concepts are collected from 40 human respondents including professors and teachers, postgraduate and undergraduate students especially from the faculty of English, Zoology and Medicine. These people were given with a questionnaire having the meanings (semantics) of English terms taken from Online Oxford Dictionary and medical terms were taken from MESH Browser and Wikipedia. The respondents were asked to give the similarity rating between 1 to 4. The similarity value is 4 if the concept pair has highest similarity 1 if the concept pair has no similarity at all. Other possible values can be 1.5, 2, 2.5, 3, and 3.5 depending upon the degree of similarity.

A correlation coefficient value is computed between the human assesment values and the similarity values of the proposed measures. The value that is close to human judgment will be evaluated as the best value. The Human Judgement Values compared with WP: Wu & Palmer and proposed ASC measure is shown in Figure 1.

It is observed from figure 1 that the ASC measure value is close to Human Judgement value than compared to the WP: Wu & Palmer measure. Similarily the Human Judgement Values compared with JC and ASC_JC and LIN and ASC_LIN measure is shown in Figure 2.

Figure 2 also shows that the proposed measures are more close to the Human Judgment values than compared to the existing measures. Thus the Psycholinguistic Evaluation proves the proposed measures to outperform in semantic similarity computations than compared to the existing measure.

7. Conclusions and Future Enhancements

This paper gives a breif on semantic similarity measures and their broad categories and explains the working principle of some major similarity measures in Ontology based and Information content-based measures. The problem of Multiple Subsumption is explained and the proposed semantic similarity measures are discussed. The results of the Psycholinguistic Evaluation of the existing and proposed semantic similarity measures are explained. There also exist the problem of multiple subsumptions in some measures of Feature based and Hybrid category of semantic similarity measures. This work can be extended to this category of measures also.

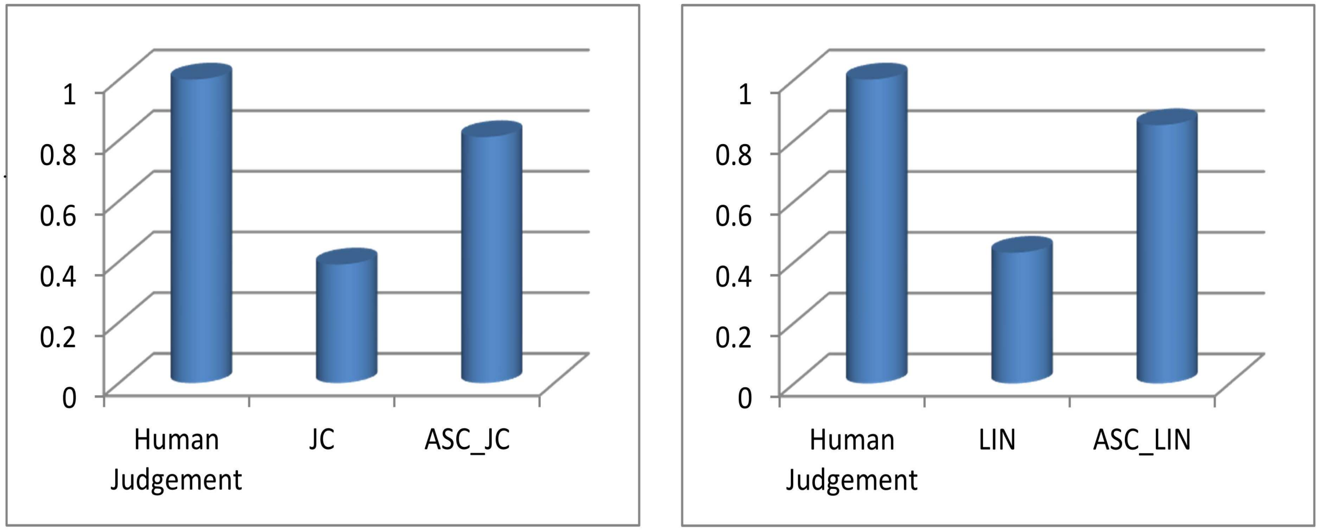

Figure 2. Correlation between Human assessment values with JC, ASC_JC and LIN, ASC_LIN measures

References

[1] Lee, J., Kim, M., Lee, Y. (1993). Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, *Journal of Documen*

tation, 49, 188-207.

[2] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G. M., Milios, E. E. (2006). Information Retrieval by Semantic Similarity, *International Journal on Semantic Web and Information Systems*, 2 (3) 55-73.

[3] Pirro, G., Ruolo, M., Talia, D. (2009). SECCO: On Building Semantic Links in Peer to Peer Networks. *Journal on Data Semantics*, XII: 1-36.

[4] Meilicke, C., Stuckenschmidt, H., Tamilin, A. (2007). Repairing ontology mappings, *In*: Proceedings of AAAI, 1408-1413.

[5] Ravi, S., Rada, M. (2007). *Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity*, *In*: Proceedings of ICSC, 2007.

[6] Lin, D. (1998). An Information-Theoretic Definition of Similarity, *In*: Proceedings of Conference on Machine Learning, 296-304.

[7] Jiang, J., Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, *In*: Proceedings of ROCLING X.

[8] Resnik, P. (1995). Information Content to Evaluate Semantic Similarity in a Taxonomy. *In*: Proceedings of IJCAI, 448-453.

[9] 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. 17-30.

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

[11] Wu, Zhibiao., Palmer, Martha. (1994). Verb semantics and lexical selection. *In*: *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*. Pages. 133-138. 1994.

[12] Slimani, T., Ben Yaghlane, B., Mellouli, K. (2008). A New Similarity Measure based on Edge Counting. World Academy of Science, Engineering and Technology, 23.

[13] Vadivu Ganesan., Rajendran Swaminathan., Thenmozhi, M. (2012). Similarity Measure Based On Edge Counting Using Ontology. *International Journal of Engineering Research and Development*. 3 (3) August, 40-44.

[14] Resnik P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy.14th *International Joint Conference on Artificial Intelligence*, IJCAI 1995, Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc 448-453.

[15] Jiang, J. J., Conrath, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *International Conference on Research in Computational Linguistics*, ROCLING X, Taipei, Taiwan, 19-33.

[16] Lin, D. An Information-Theoretic Definition of Similarity. 15th International Conference on Machine Learning (ICML98), Madison, Wisconsin, USA, Morgan Kaufmann, 296-304.

[17] Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A. (2003). Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation. Bioinformatics, 19 (10) 1275-83.

[18] Seco, N., Veale, T., Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. 16th *Eureopean Conference on Artificial Intelligence*, ECAI 2004.

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

[20] Valerie Cross, Xueheng Hu. (2011). Using Semantic Similarity in Ontology Alignment". Conference: Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011.

[21] Batet, M., Sánchez, D., Valls, A. (2011). An ontology- based measure to compute semantic similarity in biomedicine". Journal of Biomedical Informatics.

[22] Banu, Ayesha., Syeda Sameen Fatima,Khaleel Ur Rahman Khan. A New Ontology-Based Semantic Similarity Measure for Concepts Subsumed by Multiple Super Concepts. *International Journal of Web Applications,* 6 (1) March. Pages 14-22.

[23] Pedersen, T., Pakhomov, S., Patwardhan, S. (2005). Measures of Semantic Similarity and Relatedness in the Medical Domain, University of Minnesota Digital Technology Center Research Report DTC 2005/12.

[24] Hliaoutakis, A., Varelas, G., Petrakis, E.G. M., Milios, E. (2006). MedSearch: A Retrieval System for Medical Information Based

on Semantic Similarity. In: Gonzalo J., Thanos C., Verdejo M.F., Carrasco R.C. (eds) Research and Advanced Technology for Digital Libraries. *ECDL 2006. Lecture Notes in Computer Science*, 4172. Springer, Berlin, Heidelberg.

[25] Miller, G., Charles, W. (1991). Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6 (1991) 1–28.

[26] Ayesha Banu., Syeda Sameen Fatima., Khaleel Ur Rahman Khan. (2015). Information Content Based Semantic Similarity Measure for Concepts Subsumed By Multiple Concepts. IJWA 7 (3). *p* 85-94.

[27] Saruladha, K., Aghila, G., Sajina Raj. (2010). A New Semantic Similarity Metric for Solving Sparse Data Problem in Ontology based Information Retrieval System. *IJCSI International Journal of Computer Science Issues*, 7 (3) 11, May. Pages 40-48.

[28] Rubenstein, H., Goodenough, J.B. (1965). Contextual Correlates of Synonymy. Computational Linguistics. 8, 627-633.

[29] Pedersen,T., Pakhomov, S., Patwardhan, S. (2016). Measures of Semantic Similarity and Relatedness in the Medical Domain, University of Minnesota Digital Technology Center Research Report DTC 2005/12.

[30] Edgar Kalkowski., Bernhard Sick. (2016). Correlation of Ontology-Based Semantic Similarity and Human Judgement for a Domain Specific Fashion Ontology. Chapter Web Engineering. Volume 9671 of the series Lecture Notes in Computer Science pp 207-224. May.