Research Journal of Applied Sciences, Engineering and Technology 8(18): 1923-1932, 2014
DOI:10.19026/rjaset.8.1183
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 04, 2014 Accepted: June 08, 2014 Published: November 15, 2014

Research Article

Semantic Similarity Measurement Methods: The State-of-the-art

Fatmah Nazar Mahmood and Amirah Ismail
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

Abstract: With increasing importance of estimating the semantic similarity between concepts this study tries to highlight some methods used in this area. Similarity measurement between concepts has become a significant component in most intelligent knowledge management applications, especially in fields of Information Extraction (IE) and Information Retrieval (IR). Measuring similarity among concepts has been considered as a quantitative measure of the information; computation of similarity relies on the relations and the properties linked between the concepts in ontology. In this study we have briefly reviewed the main categories of semantic similarity.

Keywords: Feature based measures, hybrid measures, information content-based measures, ontology based measures

INTRODUCTION

Semantic similarity measurement techniques have gained great importance with the advent of Semantic Web (Chaves-González and MartíNez-Gil, 2013). The term semantic similarity indicates the computation of the amount of similarity among the concepts, which does not necessary to be a lexical similarity but it could be a conceptual similarity. Semantic similarity measurements determine similar concepts in a given ontology. Usually, similarity is calculated based on the target terms to ontology and through testing their relations in ontology (Hliaoutakis *et al*., 2006). Detection of semantic similarity relations among concepts or entities might be possible if these concepts are semantically linked or share some common attributes in ontology. The main objective of measuring the similarity among concepts is to provide strong approaches for standardizing the contents and to deliver information over information and communication technology. The functions of semantic similarity matching concepts define the methods of comparing the concepts and display those in a given ontology (Jayasri and Manimegalai, 2013).

Many methods have been proposed to compute the similarity among the concept, where the similarity between any two concepts is calculated to determine the shortest path length connecting these concepts in the taxonomy. If a concept is polysemous, then more than one path may exist between these target concepts. In this case, just the shortest path that connects any two senses of the concepts will be considered for computing similarity. However, the problem with this approach is that, it relies on the notion that all links in taxonomy express a uniform distance.

Resink (1995) has used information content to measure similarity between two concepts in the taxonomy. He defined the similarity between any concepts as the maximum information content that the words belong to. He used Word Net as the taxonomy and used Brown corpus to compute information content.

Li *et al*. (2003) have proposed a similarity measure using shortest path length, depth and local density in the taxonomy. On the other hand, Lin defined the similarity as the common information between two concepts and the information contained in every single concept (Bollegala *et al*., 2011).

The degree of semantic similarity between the concepts is determined according to the meaning shared. This is typically established by analyzing a list of terms and assigning a metric depending on the similarity of their meaning or the concept they represent or express. That means, discovering the similarity between any concepts or entities in a taxonomy is possible if they are linked semantically or share common attributes; basically a floating point number between 0 (total dissimilarity) and 1 (complete similarity) will be provided to mark the presence or absence of similarities (Ren and Bracewell, 2009). This present paper has summarized some popular similarity measuring approaches.

Corresponding Author: Fatmah Nazar Mahmood, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

CLASSIFICATION OF SEMANTIC SIMILARITY MEASURES

Computing the values of similarity between concepts require knowledge sources. Category membership and similarity consider two important aspects of concept matching. The vital use of similarity measurement in context ontologies is determined by knowing how one concept of ontology is related/similar to the concept of another ontology. Depending on the various notions of estimation similarity, the semantic similarity either uses the path distance between concepts or the information content of a concept, as a quantifying measure (McInnes and Pedersen, 2013). The combination of both, the path distance and information content based approaches has been tested in certain contexts. In this direction, whatever may be the knowledge sources, the quantifying measure factor are compulsory to compute similarity. These measures of semantic similarity are useful techniques in many applications, like natural language processing, information retrieval systems and ontology mapping systems. Developing semantic similarity measure is complex task; particularly the one which totally agrees with human assessment of similarity is very difficult to be designed.

Path length based measures: In this method, the quantification of similarity measurement among concepts is determined according to the path distance, which separates the concepts on the taxonomy or ontology structure. In these taxonomic or ontology structure, it is supposed that, the dominant relations that link different concepts is only is-a relations type. In these measures similarity is computed by finding the shortest path between the target concepts (synsets group of synonyms) in the taxonomy. Based on the path distance, the amount of similarity is determined, and generally it will inversely match with the length of the path (Saruladha *et al*., 2010). In the following sections different path length based similarity measures have been explained.

Rada similarity measure: Based on the Quillian spreading activation theory (Quillian, 1968; Rada *et al*. (1989) have defined semantic similarity between concepts. Quillian (1968) spreading activation theory assumes that, the semantic network is organized along the lines of similarity. Quillian (1968) depicts the concepts as points in a multi-dimensional conceptual space. Thus, the distance of conceptual could be easily measured, where the geometric distance between the points representing the concepts. The conceptual distance is used for quantifying the similarity between the concepts. It is a decreasing function of similarity. This means that, whenever the conceptual distance is smaller, the two concepts will be more similar.

Rada *et al*. (1989) have argued that, if activation theory is applied on is-a links type alone, then the shortest paths could give a positive guide of similarity. According to Rada *et al*. (1989) semantic distance among concepts in the taxonomy has been computed by counting the number of edges between them. MeSH (Medical Subject Headings-biomedical ontology) ontology has been used to conduct the experiments. The is-a relation between the concepts of the MeSH ontology is considered for quantifying similarity. The main critique of the edge-counting approach is that, this approach is susceptible to the taxonomy quality that is used.

Let C_1 and C_2 be the two concepts in a is-a semantic net. The conceptual distance between C_1 and C_2 is given by:

$$Distance(C_1, C_2) = Minimum\ number\ of\ edges\ separating\ C_1\ and\ C_2 \quad (1)$$

Rada *et al*. (1989) have used biomedicine domain to evaluate their work in information retrieval tasks. However, this metric has many attractive features because of its mathematical and semantic tractability. Rada *et al*. (1989) have consummated that; the metric of distance could capture the conceptual similarity if it is operated on better semantic nets.

Hirst and St-Onge similarity measure: Hirst and St-Onge (1998) determines the similarity among concepts based on the path distance between two concepts. Hirst and St-Onge (1998) classifies the semantic relations in the WordNet lexical ontology into three main relations as follows: Extra Strong Relations, Strong Relations and Medium Strong Relations. These relations are linked among the noun definitions WordNet. Two concepts C_1 and C_2 of WordNet would have strong relation, if any one of the following conditions is satisfied:

- Two concepts C_1 and C_2 have a common synset.
- If $C_1$and C_2 concepts are connected to two different synsets by horizontal link.
- Any kind of semantic relation should exist between the synset containing the concepts and necessarily one concept should be a compound word that includes the other one.

The strength of relationship between C_1 and C_2 concepts is medium-strong, particularly if there is an admissible path between the concepts. A path would be admissible, if the path consists of less than five links and complies eight patterns defined by Hirst and St-Onge (1998). The path connection weight of two concepts C_1 and C_2 is computed as follows:

$$Weight = c - length(C_1, C_2) - k\ x\ turns(C_1, C_2) \quad (2)$$

where, (candk) are constants, length (C_1, C_2) is the shortest length admissible path connecting C_1 and C_2 synsets and turns (C_1, C_2) represents the number of changes in direction in the shortest admissible path. Difference and identity are the basic properties. The

extra-strong relation in this framework has priority over strong relation. Furthermore, the strong relation has priority over medium-strong relation. If the length of path is longer, then the number of changes in the directions will be more, and thus the weight of the path might be decreased. Mainly, this similarity measure has been developed in the context of a system to automatically detect and correct malapropisms (correct the word that do not fit in a context) using lexical chains which is beneficial in natural language processing applications.

Bulskov measure: The similarity based on definition of Bulskov *et al*. (2002) is the concept inclusion is-a relation for atomic and compound concepts of ontology. The similarity quantification is based on the direction of the concept inclusion. They have defined semantic relations like CHR (Characterized by), CBY (Caused By), WRT (With Respect To) and concept inclusion is-a. It is said that, the concept inclusion is-a axiomatically beholds the strong similarity in the opposite way of inclusion (specialization). Additionally, the inclusion way (generalization) must engage some similarity degree. The distance which reflects similarity is measured by using the length of path, corresponding to the is-a relation. The measure of similarity can be calculated by using specialization $\sigma \in [0, 1]$ and generalization $\gamma \in [0, 1]$ factors as parameters. The path P between nodes C_1 and C_2 (concepts) is multiple and is given by:

$$P\ (C_1, C_2) = (P_1, \ldots., P_m) \tag{3}$$

where, P_i is $-aP_{i+1}$ or P_{i+1} is $-aP_i$ for each i with $C_1 = P_1$ and $C_2 = P_m$. If $P_1 \ldots P_m$ represent all paths connecting C_1 and C_2, then the degree to which C_2 is similar to C_1 can be appointed as follows:

$$Sim(C_1, C_2) = max_{j=1..m}\, \sigma^{s(p_j)} \gamma^{g(p_j)} \tag{4}$$

Along the path P, S (P_j) refers to the number of specialty and G (P_j) refers to the number of generality. The S (P) and G (P) are given by:

$$S\ (P) = |\{i | P_i is - aP_{i+1}\}| \tag{5}$$

$$G\ (P) = |\{i | P_{i+1} is - aP_i\}| \tag{6}$$

According to the Eq. (4), the similarity of concepts C_1 and C_2 is computed as a maximum weight product along the paths between C_1 and C_2. The similarity can be extracted as the product of the weights on the paths. This method better quantifies similarity, where, different weights are assigned to specialization and generalization of a concept inclusion relation. Conceptual querying of ontology was used to test the similarity measure, which has been employed in information retrieval systems applications. The retrieval of certain similarity properties has been determined, which contains the generalization property. Later a fuzzy similarity measure has been defined and an indexing scheme for ontology based information retrieval has been proposed by Bulskov (2006).

To quantify similarity based on the above methods, the path length is taken into consideration. But in these path length approaches, the concepts depth in the taxonomy has not been considered. The following part elaborates the depth relative measures, which consider the depth to quantify the semantic similarity between the concepts.

Depth relative measures: The edge counting suffer from several problems due to their dependence on the edges to the taxonomy for representing a uniform distances. The edge should be increased with the increasing depth, when representing the distance by an edge (Sussna, 1993). The depth approach is basically the shortest path approach, but in this technique, the depth of edge that connects two concepts is considered, to quantify the similarity in the general ontology structure; where, it computes the depth, beginning from the root of taxonomy and ending with the intended concepts. In terms of association there exist two edges, which represent inverse relations in taxonomy (ontology).

Wu and Palmer similarity measure: Wu and Palmer (1994) have measured semantic similarity between the concepts C_1 and C_2 by taking into consideration the depths of concept nodes in the ontology as follows:

$$Sim_{wp}(C_1, C_2) = 2 \times \frac{N_3}{N_1 + N_2 + 2 \times N_3} \tag{7}$$

where, N_1, N_2 and N_3 represent the length between the concepts in the hierarchy. N_1 is the length represents number of nodes in the path from C_1 to C_3, which is the Least Common Super (LCS) concept of C_1 and C_2, N_2 is the length given in number of nodes in the path from C_2 to C_3 and N_3 shows the overall hierarchy depth and it is used as a scaling factor. The formula for Wu and Palmer (1994) measure is rewritten as follows:

$$Sim(C_1, C_2) = \frac{2 \times Depth(LCS(C_1, C_2))}{Depth(C_1) + Depth(C_2)} \tag{8}$$

The LCS node determines the common features sharing of two concept nodes. Furthermore, the semantic distance between two concepts is determined as follows, based on Wu and Palmer (1994) Equation:

$$Dist_{wp}(C_1, C_2) = 1 - Sim_{wp}(C_1, C_2) \tag{9}$$

According of Wu and Palmer (1994), similarity measure only takes into account the depths of concept nodes and skips the most important path length feature Eq. (7) or the contributions of two features are not weighted by Eq. (8). For that motivation combine

strengths of some existing approach measures as well as merge more semantic features for enhancing the computations (Sánchez *et al*., 2012).

Sussna measure: The depth-relative scaling approach has been used by Sussna to measure similarity (Sussna 1993). He said that "there are two edges representing inverse relations associated with each edge in taxonomy". Each relation ris attached by a weight and is a value in the range (min_r-max_r). The point in the range for a relation r from C_1 to C_2 relies on number of edges of the same type, leaving C_1, which is indicated as fan out factor. The fan out factor symbolizes the mitigation of the connotation strength between the source and the target concepts; and considers the possible inconsistency between the two nodes, where, the connotation strength varies from one direction to that in the other direction:

$$W\ (C_1 \rightarrow C_2) = max_r \frac{max_r - min_r}{n_r(C_1)} \quad (10)$$

The two inverse weights are scaled and averaged by the depth of edge d in the overall taxonomy. Based on the observation made by Sussna (1993), the scaling that has the deeper sibling concepts in the tree are more tightly related than those higher in taxonomy. This was also tested by using the noun hierarchy of the WordNet lexical ontology (Scriver, 2006). The distance is computed between adjacent nodes C_1 and C_2 as follows:

$$Dist_{Sussna}(C_1, C_2) = \frac{(W(C_1 \rightarrow C_{2r}) + W(C_2 \rightarrow C_{1r}))}{2d} \quad (11)$$

where, r represents the relation between C_1 and C_2. In other words, the semantic distance is computed between two concepts by summation of the distance between adjacent concepts along the shortest path, linking C_1 and C_2.

Leacock and Chodorow similarity measure: According to Leacock and Chodorow (1998), at first the similarity between two concepts is determined by discovering the shortest path length, which connects these two concepts in the WordNet taxonomy. This identified length of depth comprises the value between 0 and 1, then the similarity is calculated as the negative logarithm of this value (Batet *et al*., 2011). The Equation of Leacock and Chodorow (1998) can be written as follows:

$$sim_{LC}\ (C_1, C_2) = -\ log \frac{length(C_1, C_2)}{2D} \quad (12)$$

Length (C_1, C_2) indicates the length, calculated in nodes, of the shortest path between the concepts C_1 and C_2 and D indicates the maximum depth of hierarchy in the WordNet.

Leacock and Chodorow (1998) measure can be explained with reference to fragment of WordNet given

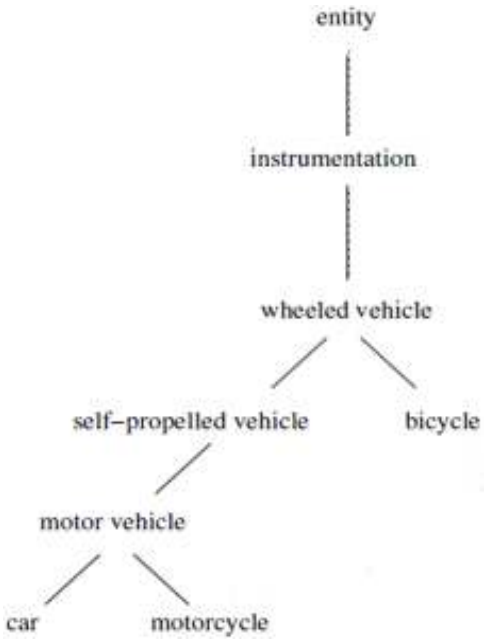

Fig. 1: Fragment of WordNet

in Fig. 1. The shortest taxonomic path between the concepts (motorcycle and bicycle) is:

Motorcycle (Is-A) motor vehicle (Is-A) self-propelled vehicle
(Is-A) wheeled vehicle SUBSUMES bicycle

It is worth noting that, the taxonomic path length differs from the network path length, as just the hypernymy and hyponymy relationships have been taken into consideration. Assume an arbitrary maximum depth of 10 in the WordNet taxonomy, the similarity value between (motorcycle and bicycle) would be calculated as:

$$\begin{aligned} & sim_{LC}(motorcycle, bicycle) \\ & = -\ log \frac{length(motorcycle, bicycle)}{2 \times 10} \\ & = -\ log \frac{5}{20} \\ & = 0.60 \end{aligned}$$

Information content based measures (corpus): Information Content based approaches (IC) are also referred to as information theoretic based approaches or corpus based approaches. The knowledge got by the corpus analysis is used to increase the amount of information that already exists in the taxonomy or ontology. Three measures which incorporate the corpus statistics as an additional and qualitatively different knowledge source have been presented in this section. Usually the notion of Information Content (IC) is used by information based approaches, which can be seen as a quantifying information measure of the concepts expressed. Corpus based approaches generally calculate

the needed IC values by sharing probabilities to every concept in the taxonomy. These probabilities are based on the appearance of concepts in a given corpus (Harispe *et al*., 2013). The information content values of the intermediate concepts in the taxonomy range from 1 to 0. The leaf level concepts of taxonomy will have the information content value as 1, as they are maximally expressed and could not be further differentiated. The information content of the root concept or the most abstract concept is 0. The method of computing information content has been discussed below.

Information content computation: Let us denote the set of concepts by C in a taxonomy, which allows multiple inheritance and associates with every concept $C_i \in C$; the probability P (C_i) of encounters an instance of concept C_i. Pursuing the standard definition from Shannon information theory (Shannon, 1948), the Information Content (IC) of C_i is defined as-log (P (C_i)), where P (C_i) indicate to the probability of the concept appearance in the Brown corpus. All the three information based measures outlined below use the information content as a basis for computing similarity between concepts.

Resnik measure: The similarity according to Resink (1995) depends on the amount of information shared between two concepts. Most Specific Common Abstraction (MSCA) concept gives this shared information that accommodates both the concepts. The concepts similarity is equal to the information content of the most specific common abstraction concept. The concepts will be dissimilar, if there are no common concepts, which mean that, the similarity between the concepts is 0. The measure introduced by Resink is as follows:

$$Sim_{res} = IC(LCS) \tag{13}$$

where, (LCS) is the Least Common Super and the Information Content (IC) is defined as:

$$IC(c) = -\log P(c) \tag{14}$$

And P (c) is the probability of finding an instance of concept c in a large corpus.

Lin similarity measure: Lin (1998) has provided a more general and theoretically stronger basis, depending on the definition of similarity between the concepts than previously provided works. He has stated that, the similarity measures should neither rely on the details of the sources, nor on the domain of application that they use. Lin (1998) has suggested the following three key intuitions about the similarity.

Intuition 1: "The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are".

Intuition 2: "The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are".

Intuition 3: "The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share".

Lin (1998) has claimed that, as there are different ways for capturing the above intuitions, a further set of assumptions are necessitated. Therefore he has proposed a set from six assumptions that can capture these intuitions, and from which, a measure of similarity may be derived. The six assumptions are announced in terms of information theory. In the following assumptions, common (A, B) is a proposition that reports the commonalities of the objects A and B and description (A, B) is a proposition that states what A and B are.

Assumption 1: The commonalities between A and B is computed by: IC (common (A, B)).

Assumption 2: The difference between A and B is computed by: IC (description (A, B)) - IC (common (A, B)).

Assumption 3: The similarity between A and B is a function of the commonalities and differences of A and B. Formally: sim (A, B) = f (IC (common (A, B), IC (description (A, B)).

Assumption 4: Always the similarity between a pair of identical objects is one. sim (A, A) = 1.

Assumption 5: The similarity between a pair of objects with no commonalities is always zero. $\forall y>0$, f (0,y) = 0.

Assumption 6: If the similarity between A and B can be measured by using two independent sets of criteria, then the total similarity is the weighted average of the two similarity values:

$$\forall x1 \le y1, x2 \le y2: f(x1 + x2, y1 + y2) = \frac{y1}{y1+y2} f(x1, y1) + \frac{y2}{y1+y2} f(x2, y2)$$

Lin (1998) proves the following similarity theorem by using the above six assumptions listed:

$$sim_L(A, B) = \frac{\log P(common(A,B))}{\log P(description(A,B))} \tag{15}$$

For applying the above similarity theorem to a conceptual taxonomy, Lin (1998) has used similar reasoning such as Resnik (1995). The concept in a taxonomy that matches the statement of the commonalities between the two concepts C_1 and C_2, is the lowest super-ordinate, denoted lso (C_1, C_2); characterizing similarity of concepts C_1 and C_2 is the union of the two concepts. The Information Content of

the statement (C_1 and C_2) is the sum of the Information Content of C_1 and C_2.

Based on the basic premise of information theory, the Information Content is the negative log of its probability and thus the sum of the Information Content of C_1 and C_2 is: $-\log P(C_1) + -\log P(C_2)$. Replacing by Lin's similarity theorem, will get:

$$sim_L(C_1, C_2) = \frac{2 \times log\, P(lso(C_1, C_2))}{log\, P(C_1) + log\, P(C_2)} \quad (16)$$

Therefore, Lin's measure is considered as the ratio of the information shared in common to the total quantity of information present in those target concepts. It is completely similar the Resnik's measure except that, the Resnik's measure takes into account only the information that is shared by the concepts and does not consider the total amount of information that they represent.

Jiang and Conrath measure: Jiang and Conrath (1997) have endeavored to merge the advantages of path-based approaches and Information Content approaches. They have weighed each edge in order to compensate for the unreliability of the edge distances by linking probabilities, based on the corpus statistics. Jiang and Conrath (1997) approach is similar to Resnik's approach, where, it uses information from both, a text corpus and a conceptual taxonomy. However, Resnik (1995) determines the value of similarity, based on the Information Content of one node (the most informative common subsume), whereas, Jiang and Conrath (1997) use theory of information for determining the weight of every link in a path. They have claimed that, the similarity degree between a parent and its child in the Noun hierarchy of WordNet is proportionate to the probability of encountering the child, given an instance of the parent: P (c| par (c)). Through definition, the quantity P (c | par (c)) is:

$$P(c \mid par(c)) = \frac{P(c \cap par(c))}{P(par(c))} \quad (17)$$

Like Resnik (1995) and Jiang and Conrath (1997) deem each example of a child to be an example of its parent; consequently, P (c ∩ par (c)) = P (c). That is, it is superfluous to need both, a child c and its parent par (c), where each example of c is also represented by an example of par (c). Therefore, the probability Equation of a child, given an example of its parent, can be simplified as:

$$P(c \mid par(c)) = \frac{P(c)}{P(par(c))} \quad (18)$$

The equation of semantic distance has been derived by defining Jiang and Conrath (1997) to the semantic distance between a child c and parent par (c), where the Information Content of the conditional probability of c given par (c) and by following the basic properties of information theory as follows:

$$\begin{aligned} &dist_{JC}(c, par(c)) = -\log P(c|par(c)) \\ &= IC(c \cap par(c)) - IC(par(c)) \\ &= IC(c) - IC(par(c)) \end{aligned} \quad (19)$$

The semantic distance between the concepts (a parent and its child) has been considered as the difference in their Information Content. This seems to be a reasonable conclusion, where the difference in Information Content should reflect the information needed for distinguishing a concept from all of its sibling concepts. For example, if a parent has just a single child, then the conditional probability will be P (c | par (c)) = 1. In this situation, taking negative logarithm gives $dist_{JC} = 0$. If no further information is required to recognize a child from its parent and thus, the semantic distance between them must be equal to zero; they are effectively the same concept.

For computing total semantic distance between any two concepts in the taxonomy, Jiang and Conrath's measure uses the summation of single distances between the nodes in the shortest path. As the shared subsume (referred by lso (C_1, C_2) for the lowest super-ordinate shared by C_1 and C_2) does not have a parent in the path, this node is excepted from the summation. Therefore the semantic distance between any two concepts C_1 and C_2 in the taxonomy is computed as follows:

$$\begin{aligned} &dist_{JC}(c_1, c_2) = \\ &\sum_{c \in path(c_1, c_2) \setminus lso(c_1, c_2)} dist_{JC}(c, path(c)) \end{aligned} \quad (20)$$

Through replacing the expression in Eq. (20) into (21) and by expansion the summation, we get:

$$\begin{aligned} &dist_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times \\ &IC(lso(c_1, c_2)) = 2\, log\, P(lso(c_1, c_2)) - \\ &(log\, P(c_1) + log\, P(c_2)) \end{aligned} \quad (21)$$

Hybrid measures: Hybrid measure approach joins the knowledge come from various sources of information. The main advantage of these approaches is that, if the knowledge of an information source is inadequate, then it may be derived from alternative information sources. Hence, the quality of similarity would be improved, to get better assessment. The hybrid approaches have used the path length, distance, depth length and semantic density of the concepts to measure similarity. In this domain, Li *et al*. (2003) and Schickel-Zuber and Faltings (2007) have offered similarity measure approaches, as explained in following sections.

Li measure: Li *et al*. (2003) have addressed the weakness of Rada's method in terms of counting edge.

When tested for more general semantic nets like WordNet, the results were not good. Depth and local density of the words have been considered in terms of computing similarity in Li measure. The similarity between two words W_1 and W_2, S (W_1, W_2) are:

$$S(W_1, W_2) = F(l, h, d) \quad (22)$$

where l represents the shortest path between the words W_1 and W_2, d is the depth of the subsumed paths in the hierarchy, and d is the local semantic density of the two words W_1 and W_2. Path length and depth have used in this work, which are derived from the lexical database. The local semantic density is calculated from a corpus. Li *et al*. (2003) have determined path length between W_1 and W_2, which are hierarchically organized in a semantic network as shown below:

Case 1: The path length of words W_1 and W_2 is defined as 0 if they are in the same concept.

Case 2: The path length of W_1 and W_2 is defined as 1, if W_1 and W_2 are not in the same concept, but the concepts for W_1 and the concept for W_2 have one or more same words.

Case 3: The actual path length is measured if W_1 and W_2 are not in the same concept.

The path length is modeled as transfer function f_1 (l) which is a monotonically decreasing function and is given by:

$$f_1(l) = e^{-\alpha} l \quad (23)$$

where, α is a constant and the value of f_1 is in the range 0 and 1.

The subsumed depth is derived by counting the levels from the subsumed to the top of the lexical hierarchy. The words at the higher layers of the hierarchy have more general abstract concepts and will be less similar than the words at the lower levels of the hierarchical semantic nets. Thus the transfer function of depth h, f_2 (h) must be a monotonically increasing function and is defined as:

$$f_2(h) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (24)$$

where, β>0 is a smoothing factor. The depth of a word is not considered if β->α, the semantic density as a monotonically increasing function, which uses the information content of words computed by using a Brown corpus:

$$f_3(wsim) = \frac{e^{\lambda wsim(w_1,w_2)} - e^{-\lambda wsim(w_1,w_2)}}{e^{\lambda wsim(w_1,w_2)} + e^{-\lambda wsim(w_1,w_2)}} \quad (25)$$

where, wsim is the information shared by w1 and w2. Thus, the similarity is computed by combining the transfer functions in Eq. (23) to (25) and is given in one Equation as below:

$$Sim_{Li}(C_1, C_2) = \frac{e^{-\alpha l}(e^{\beta h} - e^{-\beta h})}{(e^{\beta h} + e^{-\beta h})} \quad (26)$$

Brown corpus and WordNet are used to compute the similarity and the quality of the semantic similarity is better as opposed the edge counting method and information based approaches.

Ontology Structure based Similarity (OSS) measure: Schickel-Zuber and Faltings (2007) have proposed this method and computed the similarity between two concepts according to the following steps:

- Measuring the Apriori Score (APS) of the concepts which captures the concept information.
- Measuring how much apriori score has been moved T (c) between two concepts.
- The score transfer T (c) transform into a distance measure D (C_1, C_2).

The apriori score for a concept shows how much a concept is chosen in a particular context and is calculated by analyzing the ontology structure. The apriori score of a concept C with n descendants is given below:

$$APS(C) = \frac{1}{(n+2)} \quad (27)$$

The Eq. (27) defines that the apriori score, which beholds the ontology leaves, will have an APS equal to 1/2, which is equal to the mean of a uniform distribution between 0 and 1. On the contrary, the lowest values will be found at the root. This means that, when traversing up in the ontology, the concepts become more general, and thus the APS is reduced. Another important side of this APS is the fact that decreases the difference in score between the concepts when the ontology is traversed up, due to the increasing number of descendants. Resink also uses topology to compute the information content of a concept. The APS share some resemblances with information content. For example, the difference in both, IC and APS reduces, when traversed upwards the ontology. However, some deep differences exist in terms of using a bottom up approach to compute the APS score by considering the differences between the concepts. Resink follows a top down approach to compute IC by considering the commonalities between two concepts. The amount of score that must be transferred is measured by determining the chain of concepts linking the two concepts being compared. The OSS similarity metric based on the transfer score is given as:

$$Sim_{OSS}(C_1, C_2) = 1 - \frac{(log(T(C_1,C_2)))}{(maxD)} \quad (28)$$

where, T (C_1, C_2) is the transfer of score from concept C_1 to C_2 and max D is the maximum distance between any two concepts in the ontology. Based on the information transfer, this measure is defined, which take place between the concepts and is normalized by the taxonomy depth considered. To test this measure, WordNet general ontology and Gene ontology was used.

Feature based measure: The previous part has discussed the similarity measures which use the distance, path length, depth and semantic density of the concept to quantify the similarity. In this part we have discussed another kind of similarity measure, based on the features possessed by a concept. To quantify similarity, according to Lin (1998), the commonalities and distinct characteristics of a concept should be considered.

Feature based approach takes into consideration the common features between two concepts and also the specific different features of every concept. A function of the features, common to C_1 and C_2 is the similarity of a concept C_1 to a concept C_2, which means those in C_1, but not in C_2; and thosse in C_2; but not in C_1. The Tversky (1977) abstract model of similarity (Tversky 1977) is given as:

$$Sim_{tvr}(C_1, C_2) = \alpha.F\big(\varphi(C_1) \cap \varphi(C_2)\big) - \beta.F\big(\varphi(C_1)/\varphi(C_2)\big) - \gamma.F\big(\varphi(C_2)/\varphi(C_1)\big) \quad (29)$$

where, F is a function which represents the salience of a set of features. α, β and γ are parameters which afford the differences in focus on the different components. Similarity model of Tversky (1977) is set-theory based, and according to his model, the similarity is asymmetric. Pirró (2009) has defined concepts features in terms of information theoretic domain and has proposed a similarity measure as discussed below.

Pirró measure: Pirró (2009) has defined similarity measure, based on the feature based approach. The common features and also the different features among the concepts were defined in terms of information theoretic domain. Tversky (1977) formulation of similarity Eq. (29) redefined in terms of information theoretic terms as follows:

$$Sim_{tvr}(C_1, C_2) = 3 * IC\big(MSCA(C_1, C_2) - IC(C_1) - IC(C_2)\big) \quad (30)$$

where, IC (MSCA) quantifies the information content of the most specific common abstraction concept (common characteristic features), IC (C_1) the information content of concept C_1 (distance features of concept C_1) and IC (C_2) the information content of concept C_2 (distance features of concept C_2). Earlier we have discussed in Information Content Based Measures (Corpus) where the information content of concepts could be computed. But this way of calculating information content is corpus dependent. Moreover, the information content of the concept will be assumed as 0, if the concept is not defined in the corpus. In literature, this problem has been addressed as sparse data problem. Furthermore, the corpus dependent information content computation is time-consuming. Thus, Pirró (2009) has independently calculated information content of the concepts in a corpus as proposed by Seco *et al*. (2004).

Pirró (2009) has determined that, when the similarity between identical concepts has been computed by using Resnik's measure, the result yields the information content value of their MSCA and not value corresponding to the maximum of similarity. By taking this into consideration, the similarity has been defined as follows Pirró (2009):

$$Sim_{P\&S}(C_1, C_2) = \begin{cases} Sim_{tvr}(C_1, C_2) & if\, C_1 ! = C_2 \\ 1 & if\, C_1 = C_2 \end{cases} \quad (31)$$

The corpus independent is used by Pirró (2009) to compute the information content, as the information content is calculated based on the relations of concept. Word Net texono my and MeSH (Medical Subject Headings) ontology are used to conduct the experiments.

Other similarity measures: A prototype has developed by Budanitsky (1999) for automatic detection of malapropism. Furthermore, Agirre *et al*. (2009) has proposed a combined distributional approach that mainly addresses the data sparseness in the WordNet taxonomy. Based on the ration based Tversky (1977) feature model, Pirró and Euzenat (2010) have proposed a similarity measure.

Pedersen *et al*. (2007) have used a corpus based context vector approach to quantify the similarity between concepts in SNOMED-CT, a biomedical ontology of UMLS framework (Pedersen *et al*. 2007). This approach is corpus and ontology independent. Maya clinic corpus of medical notes was applied to measure the information content of the biomedical terms. They have adapted the information based methods such as, Conrath, Lin and Jiang and Resnik for biomedical domain. Six measures of similarity have been proposed by Pedersen *et al*. (2007) and three measures of relatedness based on the WordNet lexical databases. MEDLINE (a standard corpus) and MeSH ontology are used by Al-Mubaid and Nguyen (2006) to measure the similarity among biomedical terms within UMLS framework. They have offered a cluster based approach to compute similarity among biomedical concepts (Al-Mubaid and Nguyen, 2006). Furthermore, Nguyen and Al-Mubaid (2006) have proposed an

ontology based measure to calculate similarity among biomedical concepts. Similarity among concepts was measured using multiple information sources (Nguyen and Al-Mubaid, 2006).

CONCLUSION

In this study we had discussed and highlighted some methods, used to measure similarity between the concepts in single ontology. We have identified that, shortest path and depth depends on the distance between the target concepts. Information content measure depends on the amount of properties shared between the two concepts. Basically, hybrid method combine shortest path and information content to improve the similarity measurement; and the last approach we had reviewed was feature measure, which considered the common features and as well as the specific different features between the concepts. The evaluation methodology followed by the researchers had also been presented in this study.

ACKNOWLEDGMENT

The author wishes to thank every parson who help me to complete this study especially my family.

REFERENCES

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova, M. Paşca and A. Soroa, 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. Proceeding of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp: 19-27.

Al-Mubaid, H. and H.A. Nguyen, 2006. A cluster-based approach for semantic similarity in the biomedical domain. Proceeding of 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06), pp: 2713-2717.

Batet, M., D. Sánchez and A. Valls, 2011. An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inform., 44(1): 118-125.

Bollegala, D., Y. Matsuo and M. Ishizuka, 2011. A web search engine-based approach to measure semantic similarity between words. IEEE T. Knowl. Data En., 23(7): 977-990.

Budanitsky, A., 1999. Lexical semantic relatedness and its application in natural language processing. Technical Report CSRG-390, Computer Systems Research Group, University of Toronto, August.

Bulskov, H., 2006. Ontology-based information retrieval. Ph.D. Thesis, Roskilde University, Denmark.

Bulskov, H., R. Knappe and T. Andreasen, 2002. On measuring similarity for conceptual querying. In: Andreasen, T. *et al*. (Eds.), Flexible Query Answering Systems. LNAI 2522, Springer, Berlin, Heidelberg, pp: 100-111.

Chaves-González, J.M. and J. MartíNez-Gil, 2013. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowl-Based Syst., 37: 62-69.

Harispe, S., D. Sánchez, S. Ranwez, S. Janaqi and J. Montmain, 2013. A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. J. Biomed. Inform., 48: 38-53

Hirst, G. and D. St-Onge, 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In: WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, pp: 305-332.

Hliaoutakis, A., G. Varelas, E. Voutsakis, E.G.M. Petrakis and E. Milios, 2006. Information retrieval by semantic similarity. Int. J. Semant. Web Inf., 2(3): 55-73.

Jayasri, D. and D. Manimegalai, 2013. Semantic similarity measures on different ontologies: Survey and a proposal of cross ontology based similarity measure. Int. J. Sci. Res., 2(2): 455-461.

Jiang, J.J. and D.W. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy Proceeding of International Conference Research on Computational Linguistics (ROCLING X), Taiwan.

Leacock, C. and M. Chodorow, 1998. Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database, 49(2): 265-283.

Li, Y., Z.A. Bandar and D. McLean, 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE T. Knowl. Data En., 15(4): 871-882.

Lin, D., 1998. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning (ICML, 1998), pp: 296-304.

McInnes, B.T. and T. Pedersen, 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. J. Biomed. Inform., 46(6): 1116-1124.

Nguyen, H.A., 2006. New semantic similarity techniques of concepts applied in the biomedical domain and WordNet. M.S. Thesis, University of Houston, USA.

Pedersen, T., S.V.S. Pakhomov, S. Patwardhan and C.G. Chute, 2007. Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform., 40(3): 288-299.

Pirró, G., 2009. A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng., 68(11): 1289-1308.

Pirró, G. and J. Euzenat, 2010. A feature and information theoretic framework for semantic similarity and relatedness. Proceeding of the 9th International Semantic Web Conference (ISWC'10), 6496: 615-630.

Quillian, M.R., 1968. Semantic Memory. In: Minsky, M. (Ed.), Semantic Information Processing. MIT Press, Cambridge.

Rada, R., H. Mili, E. Bicknell and M. Bletner, 1989. Development and application of a metric on semantic nets. IEEE T. Syst. Man Cyb., 19(1): 17-30.

Ren, F. and D.B. Bracewell, 2009. Advanced information retrieval. Electron. Notes Theor. Comput. Sci., 225: 303-317.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy Proceeding of the 14th International Joint Conference on Artificial Intelligence, pp: 448-453.

Sánchez, D., M. Batet, D. Isern and A. Valls, 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl., 39(9): 7718-7728.

Saruladha, K., G. Aghila and S. Raj, 2010. A survey of semantic similarity methods for ontology based information retrieval. Proceeding of 2nd International Conference on Machine Learning and Computing (ICMLC, 2010), pp: 297-301.

Schickel-Zuber, V. and B. Faltings, 2007. OSS: A semantic similarity function based on hierarchical ontologies." Proceeding of the 20th International Joint Conference on Artifical Intelligence (IJCAI, 2007), 7: 551-556.

Scriver, A.D., 2006. Semantic distance in wordnet: A simplified and improved measure of semantic relatedness. M.A. Thesis, University of Waterloo, Canada.

Seco, N., T. Veale and J. Hayes, 2004. An intrinsic information content metric for semantic similarity in WordNet. Proceeding of the 16th European Conference on Artificial Intelligence (ECAI, 2004). John Wiley, Valencia, Spain.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J., 27: 623-656.

Sussna, M., 1993. Word sense disambiguation for free-text indexing using a massive semantic network. Proceeding of the 2nd International Conference on Information and Knowledge Management. ACM, New York, USA, pp: 67-74.

Tversky, A., 1977. Features of similarity." Psychol. Rev., 84(4): 327. Retrieved from: http://shodhganga.inflibnet.ac.in/bitstream/10603/5300/10/10_chapter%202.pdf.

Wu, Z. and M.S. Palmer, 1994. Verb semantics and lexical selection with incomplete and imperfect judgements. Proceeding of the 15th ACM International Conference on Information and Knowledge management (CIKM'06). ACM Press, New York, pp: 102-111.