Semantic Similarity Measures on Different Ontologies: Survey and a Proposal of Cross Ontology based Similarity Measure

D Jayasri[1], D Manimegalai[2]

[1]Associate Professor, Department of Mathematics
ULTRA college of Engineering & Technology for Women, Madurai, India
jaysri6@yahoo.co.in

[2]Professor and Head, Department of Information Technology
National Engineering College, Kovilpatti, Tamilnadu, India
megalai_nec@yahoo.com

Abstract: ***Semantic similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Semantic-similarity measures quantify Concept Similarities in a given ontology. Typically, it is computed by mapping terms to ontology and by examining their relationships in that ontology. In this paper, a comparative study on different measures such as path based, information content based, feature based and hybrid similarity measures is done for identifying semantically similar concepts in ontology. The focus is on more than one ontology methods since it is interesting than the single ontology and semantic similarities are calculated between terms stemming from different ontologies (WordNet and MeSH in this work). The purpose of this survey is to explore how these similarity computation methods could assist to improve the retrieval effectiveness of Information retrieval models based on Cross Ontology.***

Keywords: Ontology, Semantic similarity measures, information retrieval, retrieval effectiveness

1. Introduction

Semantic similarity relates to computing the similarity between concepts which are not lexically similar. This is an important problem in Natural Language Processing (NLP) and Information Retrieval (IR) research and has received considerable attention in the literature. Several algorithmic approaches for computing semantic similarity have been proposed. Detection of similarity between concepts or entities is possible if they share common attributes or if they are linked with other semantically related entities in ontology [3, 8]. To relate concepts in different ontologies, semantic similarity works by discovering linguistic relationships or affinities between ontological terms across different ontologies [10].

We present a critical evaluation of several semantic similarity approaches using two well known taxonomic hierarchies (or ontologies) namely Word Net and MeSH. Word Net is a controlled vocabulary. The percentage of relevant information we get mainly depends on the semantic similarity matching function we used. So far, there are several semantic similarity methods used which have certain limitations despite the advantages. No one method replaces all the semantic similarity methods. When a new information retrieval system is going to be built, several questions arise related to the semantic similarity matching function to be used. In the last few decades, the number of semantic similarity methods developed is high. This paper discusses the overall view of different similarity measuring methods used to compare and find very similar concepts of ontology and also between two (ontologies). The pros and cons of existing similarity metrics are discussed.

2. Related Work

Issues related to semantic similarity algorithms along with issues related to computing semantic similarity on Word Net and MeSH are discussed below.

2.1. Word Net

Word Net is an on-line lexical reference system developed at Princeton University. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also organized into senses (i.e., corresponding to different meanings of the same term or concept). The synsets (or concepts) are related to other synsets higher or lower in the hierarchy defined by different types of relationships. The most common relationships are the Hyponym/Hypernym (i.e., Is-A relationship), and the Meronym/Holonym (i.e., Part-Of relationship). There are nine noun and several verb Is-A hierarchies (adjectives and adverbs are not organized into Is-A hierarchies).

2.2. MeSH

MeSH (Medical Subject Headings) is a taxonomic hierarchy (ontology) of medical and biological terms (or concepts) suggested by the U.S National Library of Medicine (NLM). MeSH terms are organized in Is-A taxonomies with more general terms (e.g., “chemicals and drugs”) higher in a taxonomy than more specific terms (e.g., “aspirin”). Each MeSH term is described by several properties, the most important of them being the MeSH Heading (MH) (i.e., term name or identifier), Scope Note (i.e., a text description of the term) and Entry Terms (i.e., mostly synonym terms to the MH). In this work, entry terms are treated as synonyms.

2.3. Semantic Similarity

Several methods for determining semantic similarity between terms have been proposed in the literature and some of them have been tested on Word Net. We present an evaluation for a more complete and up-to-date set of methods and we also investigate cross ontology methods. Similar results on MeSH have not been reported in the literature. Similarity measures apply only for nouns (and verbs in Word Net) and for Is-A relationships. Taxonomic properties like commonality, identity and differential properties for adverbs and adjectives do not exist. Semantic similarity methods are classified into following main categories:

1. Edge Counting Methods
2. Information Content Methods
3. Feature Based Methods
4. Hybrid methods

2.3.1 Edge Counting Method

Measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy [5, 6, 3, 4].

2.3.1.1 Path Length Approach

The shortest path length and the weighted shortest path are the two taxonomy based approaches for measuring similarity through inclusion relation.

Shortest Path Length

A simple way to measure semantic similarity in taxonomy is to evaluate the distance between the nodes corresponding to the items being compared. The shorter distance results in high similarity. In [14], shortest path length approach is followed assuming that the number of edges between terms in taxonomy is a measure of conceptual distance between concepts as inEqn [1].

***distRada*(*ci; cj*) = minimal number of edges in a path from *ci* to *cj* [1]**

This method yields good results. Since the paths are restricted to **IS-A** relation, the path lengths corresponds to conceptual distance.

Weighted Shortest Path Length

In this method, weights are assigned to edges. In brief, weighted shortest path measure is a generalization of the shortest path length. Obviously it supports commonality and difference properties.

- Similarity of immediate specialisation
-Similarity of immediate generalisation

P= (p1,….., pn) where,
Pi **ISA** pi+1 or Pi+1 **ISA** pi

For each i with x=p1 and y=pn.

Given a path P=(p1,…..pn), set s(P) to the number of specializations and g(P) to the number of generalizations along the path P as in Eqn [2]:

s(P)=|{i\pi ISA pi+1}|
g(P)=|{i|Pi+1 ISA pi}| [2]

If p1 …pm is all paths connecting x and y, then the degree to which y is similar to x can be defined in Eqn [3]:

simWSP(x,y)=max{s(pj)s(pj)}, j=1,….m **[3]**

The similarity between two concepts x and y, sim(x,y)WSP (weighted Shortest Path) is calculated as the maximal product of weights along the paths between x and y. Hence the weighted shortest path length overcomes the limitations of shortest path length wherein the measure is based on generalization property and achieves identity property.

Depth Relative Scaling

This depth-relative scaling approach [2] defines two edges representing inverse relations for each edge in taxonomy. The weight attached to each relation r is a value in the range [minr; maxr]. The point in the range for a relation r from concept c1 to c2 depends on the number nr of edges of the same type, leaving c1, which is denoted as the type specific fan-out factor:

W (c1→r c2) =maxr-{maxr--minr/nr (c1)}

The two inverse weights are averaged and scaled by depth d of the edge in the overall taxonomy. The distance between adjacent nodes c1 and c2 are computed as:

Dist$_{sussna}$(c1,c2)=w(c1→rc2)+(c1→r'c2)/2d **[4]**

where *r* is the relation that holds between *c*1 and *c*2, and *r'* is its inverse. The semantic distance between two arbitrary concepts *c*1 and *c*2 is computed as the sum of distances between the pairs of adjacent concepts along the shortest path connecting *c*1 and *c*2.

Conceptual Similarity

Wu and Palmer [15], proposed a measure of semantic similarity on the semantic representation of verbs in computer systems and its impact on lexical selection problems in machine translation. Wu and Palmer define *conceptual similarity* between a pair of concepts *c*1 and *c*2 as in Eqn [5]:

Simwu&palmer(c1,c2)=2*N3/N1+N2 [5]

where N1 is the number of nodes on the path from c1 to a concept c3. , denoting the least upper bound of both c1 and c2. N2 is the number of nodes on a path from c2 to c3. N3 is the number of nodes from c3 to the most general concept.

Normalized Path Length

Leacock and Chodorow [3], proposed an approach for measuring semantic similarity as the shortest path using is-a hierarchies for nouns in Word Net. This measure determines the semantic similarity between two synsets (concepts) by finding the shortest path and by scaling using the depth of the taxonomy in Eqn [6]

$$\mathbf{Sim_{Leacock\&Chaodorow}(c1,c2)=-g(Np(c1,c2)/2D)} \quad [6]$$

Np (c1, c2) denotes the shortest path between the synsets (measured in nodes), and D is the maximum depth of the taxonomy.

2.3.2 Information Content Methods:

Measure the difference in information content between two terms as a function of their probability of occurrence in a corpus [6,2,7,1].

In this method rather than counting edges in the shortest path, they select the maximum information content of the least upper bound between two concepts. Resnik [2], argued that a widely acknowledged problem with edge counting approaches was that they typically rely on the notion that edges represent uniform distances. Let C denote the set of concepts in a taxonomy that permits multiple inheritance and associates with each concept c 2 C, the probability p(c) of encountering an instance of concept c. For a pair of concepts c1 and c2, their similarity is given in Eqn [7]:

$$\mathbf{Sim_{Resnik}sim(c1;c2)=max_{c€s(c1,c2)}[-log(c)]} \quad [7]$$

where,S(c1,c2) is a set of least upper bounds in the taxonomy of c1 and c2. p(c) :Monotonically non-decreasing as one moves up in the taxonomy, p(c1) ≤ p(c2), if c1 is a c2.

The similarity between the two words w1 and w2 can be computed as in Eqn [8]:

$$\mathbf{wsim_{Resnik}wsim(w1,w2)=max_{c1;c2}sim(c1,c2);} \quad [8]$$

Where c1 ranges over s(w1) and c2 ranges over s(w2). The major drawback of the information content approach is that they fail to comply with the generalization property due to symmetry.

Page-count-based similarity metrics

Page-count-based metrics use association ratios between words that are computed using their co-occurrence frequency in documents. We use the notation {D} for a set of documents, |D| for document set cardinality, {|D|w} for the set of documents that contains the word w, and {D|w1,w2} for the set of documents that contains both words w1 and w2.

Jaccard and Dice Coefficients

The Jaccard coefficient is a measure for calculating the similarity (or diversity) between sets. The variation of the Jaccard coefficient is defined as

$$\mathbf{J(w1,w2)=D|w1,w2|/|D|w1|+|D|w2|-|D(w1,w2)|} \quad [9]$$

In probabilistic terms, finds the maximum likelihood estimate of the ratio of the probability of finding a document, where words w1 and w2 co-occur over the probability of finding a document where either w1 or w2 occurs. If w1 and w2 are the same word, then the Jaccard coefficient is equal to 1 (absolute semantic similarity). If two words never co-occur in a document collection, then the Jaccard coefficient is 0.

The Dice coefficient is related to the Jaccard coefficient and is computed as in Eqn[10]

$$\mathbf{C(w1,w2)=2|D|w1,w2|/|D|w1|+|D|w2|} \quad [10]$$

Again, the Dice coefficient is equal to 1 if w1 and w2 are identical, and 0 if two words never co-occur.

Mutual information

If we assume that the number of documents indexed by the words w1, and w2 are random
Variables X, Y respectively, then the point wise mutual information (MI) between X and Y measures the mutual dependence between the occurrence of words w1 and w2 [9]. The maximum likelihood estimate of MI is

$$\boldsymbol{I(x,y) = \log \frac{\frac{||D|w_1,w_2|}{|D|}}{\frac{|D|w_1|}{|D|}\ \frac{|Dw_2|}{|D|}}} \quad [11]$$

Mutual information measures the information that variables X and Y share. It quantifies how the knowledge of one variable reduces the uncertainty about the other. For instance, if X and Y are independent, then knowing X does not give any information about Y and the mutual information is 0. For X = Y the knowledge of X provides the value of Y with certainty and the mutual information is 1.

Google-based semantic relatedness

Motivated by Kolmogorov complexity google proposed a page-count-based similarity measure, called the Normalized Google Distance, defined as

$$\boldsymbol{G(w_{1,}w_2\ ) = \frac{\max\{A\}-\log|D|w_1,w_1|}{\log|D|-\min\{A\}},} \quad [12]$$

Where A={log|D|w_1|,log|D|w_2| }

As the semantic similarity between two words increases, the distance computed by (4) decreases. Thus, this metric can be considered as a dissimilarity measure. Note that the metric is also unbounded, ranging from 0 to 1 as given in Eqn [13].

$$\mathbf{G'(w_{1,}w_2=}\boldsymbol{e^{-2G(w_1,w_2)}} \quad [13]$$

Where G (w1, w2) is computed according to (4). Note that the Google-based Semantic Relatedness is bounded taking values between 0 and 1.

2.3.3 Feature-Based Methods:

Measure the similarity between two terms as a function of their properties (e.g., their definitions or "glosses" in WordNet or "scope notes" in MeSH) or based on their relationships to other similar terms in the taxonomy. Common features tend to increase the similarity and (conversely) non-common features tend to diminish the similarity of two concepts [9].

2.3.4 Hybrid methods [10] combine the above ideas: Term similarity is computed by matching synonyms, term neighbourhoods and term features. Term features are further distinguished into parts, functions and attributes and are matched similarly to [9].

An important observation and a desirable property of most semantic similarity methods is that they assign higher similarity to terms which are close together (in terms of path length) and lower in the hierarchy (more specific terms), than to terms which are equally close together but higher in the hierarchy (more general terms). Edge counting and information content methods work by exploiting structure information (i.e., position of terms) and information content of terms in a hierarchy and are best suited for comparing terms from the same ontology. Because the structure and information content of different ontologies are not directly comparable, cross ontology similarity methods usually call for hybrid or feature based approaches. Semantic similarity methods can also be distinguished between:

Single Ontology- similarity methods that assume that the terms, which are compared, are from the same ontology (e.g., WordNet).

Cross Ontology- similarity method which compare terms from different ontology (e.g., WordNet and MeSH)

3. Cross Ontology Semantic Similarity

A recent contribution by Rodriguez [10] proposed a framework for comparing terms stemming from the same or from different ontologies. The similarity between terms *a* and *b* is computed as a weighted sum of similarities between synonym sets (synsets), features and terms neighbourhoods given in Eqn[14]:

$$\mathbf{Sim(a,b)=w.s_{synsets}(a,b)+u.s_{features}(a,b)+v.s_{neighbourhoods}(a,b)} \qquad \mathbf{[14]}$$

with *w*, *u* and *v* denoting the relative importance of the three similarity components. Features are further specialized into "parts", "attributes" and "functions". For example, in Word Net $S_{features}$ is implemented as the matching of terms having the Part-Of relationship. Notice that no Part-Of relationships are defined in MeSH and this term is omitted when this method is applied on MeSH. Assuming that all terms in the neighborhoods of terms *α* and *b* as well as their features (i.e., their corresponding parts, attributes and functions) are also represented by synsets, each similarity component is computed by Tversky [9] as in Eqn [15]

$$\mathbf{S(a,b)} = \frac{|A\cap B|}{|A\cap B|+\gamma(a,b)|A\setminus B|+(1-\gamma(a,b)|B\setminus A|} \ \mathbf{15]}$$

where *A*, *B* denote synsets of terms *a*, *b* and *A\B* denotes the set of terms in A but not in B (the reverse for *B\A*). Parameter *γ (a,b)* is computed as a function of the depth of the terms a and *b* in their taxonomy in Eqn [16]:

$$\gamma(a,b) = \begin{cases} \frac{depth(a)}{depth(a)+depth(b)}, & depth(a) \le depth(b); \\ 1-\frac{depth(a)}{depth(a)+depth(b)}, & depth(a) > depth(b); \end{cases} \qquad [16]$$

X-Similarity relies on matching between synsets and term description sets. The term description sets contain words extracted by parsing term definitions ("glosses" in Word Net or "scope notes" in MeSH). Two terms are similar if their synsets or description sets or, the synsets of the terms in their neighborhood (e.g., more specific and more general terms) are lexically similar. First, we propose replacing Equation 4 by plain set similarity

$$\mathbf{S(a,b)}=\frac{|A\cap B|}{|A\cup B|} \qquad [4_a]$$

where *A* and *B* denote synsets or term description sets. Because not all terms in the neighborhood of a term are connected with the same relationship, we propose that set similarities are computed per relationship type (e.g., Is-A and Part-Of for WordNet and only Is-A for MeSH) as in Eqn [17]

$$S_{neighborhoods}\ \mathbf{(a,b)=max}\ \frac{A_i\cap B_i}{A_i\cup B_i} \qquad \mathbf{[17]}$$

where i denotes relationship type. Eqn. 17 suggests computing the similarity between term neighborhoods, by matching same type relationships between synsets of more specific and of more general terms (i.e., for each term, the union of the synsets of all terms up to the root of each term hierarchy is taken) and by taking their maximum. The above ideas are combined into a single Eqn[18]

$$Sim(a,b) = \begin{cases} 1, & if\ Ssynset(a,b) > 0 \\ max\{S_{neibourhood}(a.b), S_{descriptions}(a,b) & if S_{synset}(a,b) = 0 \end{cases} \qquad [18]$$

$S_{descriptions}$ denotes matching of term description sets. $S_{descriptions}$ and $S_{synsets}$ are computed according to Equation 4_a. Notice that, two terms with at least one common synonym term are 100% similar.

4. Proposed Work

Cross ontology measures compares the words from diverse ontologies such as Word Net and MeSH. The cross ontology approaches often requires hybrid or feature based measures, because the structure and information content between diverse ontologies cannot be compared directly. For instance, two terms are alike if they have same spelling or meaning, or they are related with other terms that are alike. Several intelligent knowledge-based applications have techniques for computing semantic similarity between the terms. Most of the existing semantic similarity measures have used ontology structure as their key source, but they cannot calculate the semantic similarity between words and concepts using several ontologies.

4.1 Extracting set of relevant definitions, features, synsets, and neighbors from both ontologies

In general, ontologies can be distinguished into domain ontologies, representing knowledge of a particular domain, and generic ontologies representing common sense knowledge about the world. There are several examples of general purpose ontologies available including WordNet that attempts to model the lexical knowledge of a native speaker of English. English nouns, verbs, adjectives, and adverbs are organized into synonym sets, called synsets, each representing a concept.

As well, one of the domain specific ontology designed for medical concepts includes MeSH. Based on the relevant input query keyword, the set of appropriate definitions, features (Hypernyms), synset, neighbors (Hyponyms) are extracted from both the ontologies, WordNet and MeSH. The sample XML descriptions about the query keywords from both ontologies with the given bio-medical term are shown below.

Table 1: XML descriptions taken from the Wordnet and MeSH ontology

WordNet: Adenovirus	*MeSH: Rotavirus*
<Term>Adenovirus <Definition>any of a group of viruses including those that in humans cause upper respiratory infections or infectious Pinkeye, </Definition> <Synset> adenovirus, </Synset> <Hypernyms> animal_virus,, </Hypernyms> <Hyponyms> parainfluenza_virus,, </Hyponyms> </Term>	<Term>rotavirus enteritis <Definition>A viral infectious disease that results in inflammation located in stomach and located in intestine, has_material_basis_in Rotavirus, which is transmitted by ingestion of contaminated food or water, or transmitted by fomites. The infection has symptom fever, has symptom vomiting, has symptom diarrhoea, and has symptom abdominal pain. </Definition> <Synset>rotavirus enteritis, Enteritis due to rotavirus (disorder), </Synset> <Hypernyms>Nil </Hypernyms> <Hyponyms> rotavirus enteritis </Hyponyms> </Term>

4.2 Finding Cross ontology measure for the input query

In order to find the cross ontology measure for the input query, we have found out the semantic similarity measures of the extracted feature sets, synsets, neighbourhoods and the definitions of two different ontology's. The similarity between two different terms is computed as a weighted sum of similarities between synonym sets (synsets), features, neighbourhoods and their definitions. Consider the WordNet O_1 and MeSH O_2 ontology's, in which the Query keyword Q consists of Features F , Synsets S , Neighbourhoods N and Definitions D obtained from both the ontology's. In addition, we have combined all the chosen features together in a vector named as A_s . Based on the input query, we have to find out the cross ontology measure for every set of features, synsets, neighbourhoods and definitions obtained from the ontology's. The set of features, synsets, neighbourhoods and definitions obtained from the ontology's O_1 and O_2 are represented as follows,

$$F=\{(f_i^{(1)},f_i^{(2)})\,|\,f_i^{(1)}\in O_1,f_i^{(2)}\in O_2\}\ 1\le i\le m$$

$$S=\{(s_i^{(1)},s_i^{(2)})\,|\,s_i^{(1)}\in O_1,s_i^{(2)}\in O_2\};1\le i\le m$$

$$N=\{(n_i^{(1)},n_i^{(2)})\,|\,n_i^{(1)}\in O_1,n_i^{(2)}\in O_2\};\ 1\le i\le m$$

$$D=\{(d_i^{(1)},d_i^{(2)})\,|\,d_i^{(1)}\in O_1,d_i^{(2)}\in O_2\};1\le i\le m$$

$$A_s=\{F,S,N,D\}$$

The similarity measure $Sim(Q_1,Q_2)$ of the input query keywords Q_1 and Q_2 from ontologies O_1 and O_2 respectively is computed with the aid of the set of features, synsets, neighborhoods and the definitions extracted from both the ontologies. The formula utilized for computing the similarity measure of the corresponding query keyword from the Wordnet and MeSH is given as follows,

$$Sim(Q_1,Q_2)=\sqrt{\frac{\alpha S_f^2(Q_1,Q_2)+\beta S_s^2(Q_1,Q_2)+\gamma S_n^2(Q_1,Q_2)+\delta S_d^2(Q_1,Q_2)}{4}}$$

Where, $\alpha,\beta,\gamma,\delta$ are the set of the similarity parameters and these parameters are identified as shown below.

$$\alpha=\frac{\left|f^{(1)}\cap f^{(2)}\right|+\left|\cup A_s^{(1)}\cap\cup A_s^{(2)}\right|}{\left(f^{(1)}\cap f^{(2)}\right)+\left(s^{(1)}\cap s^{(2)}\right)+\left(n^{(1)}\cap n^{(2)}\right)+\left(d^{(1)}\cap d^{(2)}\right)+\left(\cup A_s^{(1)}\cap\cup A_s^{(2)}\right)}$$

$$\beta=\frac{\left|s^{(1)}\cap s^{(2)}\right|+\left|\cup A_s^{(1)}\cap\cup A_s^{(2)}\right|}{\left(f^{(1)}\cap f^{(2)}\right)+\left(s^{(1)}\cap s^{(2)}\right)+\left(n^{(1)}\cap n^{(2)}\right)+\left(d^{(1)}\cap d^{(2)}\right)+\left(\cup A_s^{(1)}\cap\cup A_s^{(2)}\right)}$$

$$\gamma=\frac{\left|n^{(1)}\cap n^{(2)}\right|+\left|\cup A_s^{(1)}\cap\cup A_s^{(2)}\right|}{\left(f^{(1)}\cap f^{(2)}\right)+\left(s^{(1)}\cap s^{(2)}\right)+\left(n^{(1)}\cap n^{(2)}\right)+\left(d^{(1)}\cap d^{(2)}\right)+\left(\cup A_s^{(1)}\cap\cup A_s^{(2)}\right)}$$

$$\delta=\frac{\left|d^{(1)}\cap d^{(2)}\right|+\left|\cup A_s^{(1)}\cap\cup A_s^{(2)}\right|}{\left(f^{(1)}\cap f^{(2)}\right)+\left(s^{(1)}\cap s^{(2)}\right)+\left(n^{(1)}\cap n^{(2)}\right)+\left(d^{(1)}\cap d^{(2)}\right)+\left(\cup A_s^{(1)}\cap\cup A_s^{(2)}\right)}$$

Also, $S_f(Q_1,Q_2)$, $S_s(Q_1,Q_2)$, $S_n(Q_1,Q_2)$ and $S_d(Q_1,Q_2)$ are the individual similarity measures of the every feature set, synsets, neighbourhoods and definitions respectively. Here, the formula for finding the similarity of every set of terms by means of their common universal set of all terms with features, synsets, neighbourhoods and the definitions is given in detail.

$$S_f(Q_1,Q_2)=\left(\frac{\left|f^{(1)}\cap f^{(2)}\right|}{\left|f^{(1)}\right|*\left|f^{(2)}\right|}\right)+\left(\frac{\left|\sim f^{(1)}\cap\sim f^{(2)}\right|}{\left|\sim f^{(1)}\right|*\left|\sim f^{(2)}\right|}\right)-\left(\frac{\left|f^{(1)}\cap\sim f^{(2)}\right|}{\left|f^{(1)}\right|*\left|\sim f^{(2)}\right|}\right)-\left(\frac{\left|\sim f^{(1)}\cap f^{(2)}\right|}{\left|\sim f^{(1)}\right|*\left|f^{(2)}\right|}\right)$$

$$S_s(Q_1,Q_2)=\left(\frac{\left|s^{(1)}\cap s^{(2)}\right|}{\left|s^{(1)}\right|*\left|s^{(2)}\right|}\right)+\left(\frac{\left|\sim s^{(1)}\cap\sim s^{(2)}\right|}{\left|\sim s^{(1)}\right|*\sim\left|s^{(2)}\right|}\right)-\left(\frac{\left|s^{(1)}\cap\sim s^{(2)}\right|}{\left|s^{(1)}\right|*\left|\sim s^{(2)}\right|}\right)-\left(\frac{\left|\sim s^{(1)}\cap s^{(2)}\right|}{\left|\sim s^{(1)}\right|*\left|s^{(2)}\right|}\right)$$

$$S_n(Q_1,Q_2)=\left(\frac{\left|n^{(1)}\cap n^{(2)}\right|}{\left|n^{(1)}\right|*\left|n^{(2)}\right|}\right)+\left(\frac{\left|\sim n^{(1)}\cap\sim n^{(2)}\right|}{\left|\sim n^{(1)}\right|*\sim\left|n^{(2)}\right|}\right)-\left(\frac{\left|n^{(1)}\cap\sim n^{(2)}\right|}{\left|n^{(1)}\right|*\sim\left|n^{(2)}\right|}\right)-\left(\frac{\left|\sim n^{(1)}\cap n^{(2)}\right|}{\left|\sim n^{(1)}\right|*\left|n^{(2)}\right|}\right)$$

$$S_d(Q_1,Q_2)=\left(\frac{\left|d^{(1)}\cap d^{(2)}\right|}{\left|d^{(1)}\right|*\left|d^{(2)}\right|}\right)+\left(\frac{\left|\sim d^{(1)}\cap\sim d^{(2)}\right|}{\left|\sim d^{(1)}\right|*\sim\left|d^{(2)}\right|}\right)-\left(\frac{\left|d^{(1)}\cap\sim d^{(2)}\right|}{\left|d^{(1)}\right|*\sim\left|d^{(2)}\right|}\right)-\left(\frac{\left|\sim d^{(1)}\cap d^{(2)}\right|}{\left|\sim d^{(1)}\right|*\left|d^{(2)}\right|}\right)$$

5. Evaluation of Semantic Similarity Methods

In the following, we present a comparative evaluation of the similarity methods referred above. All data sets used in the experiments below are available on the Web5. the proposed cross ontology based similarity measure is compared along with the X-similarity measure [42] and the Rodriguez M.A‘s [41] similarity measure. Here, the similarity measures of the [41, 42] are taken from the semantic similarity system intelligence laboratory. They have analyzed by their own similarity measures with the aid of the WordNet and the MeSH ontology terms. In this, some of the medical terms fail to reach the similarity values of the existing ones, in which our proposed cross ontology based similarity measure performs better. Table 3 lists the comparative values obtained by the proposed similarity measure and the existing works.

Table 3: Cross ontology based similarity measure comparison

Query keyword		*X-similarity measure [42]*	*Rodriguez M.A [41]*	*Proposed similarity measure*
WordNet	*MeSH*			
aAdenovirus	Rotavirus	0.16	0.018666667	0.03406453
Anemia	aAppendicitis	0	0	0.02938514816
Pneumonia	Asthma	0.07	0.0119	0.01566590728
Carcinoma	Neoplasm	0.17	0.04	0.0569153419
Hypothyroidism	Hyperthyroidism	0.387	0	0.0871796247
Pain	Ache	1	0.021666667	0.04950827
Dementia	Atopic Dermatitis	0	0	0.044338768
Malaria	Bacterial Pneumonia	0.113	0	0.04502309
Osteoporosis	Patent ductus Arteriosus	0.122	0	0.2681062
Sinusitis	Mental Retardation	0	0	0.07598254
Urinary Tractinfectionnn	Pyelonephritis	0.03	0.01	0.115328473
Iron Deficiency Anemia	Sickle Cell anemia	0.14	0.01166667	0.060882246

6. Conclusions & Future work

In this paper, we have presented an effective cross ontology based similarity measure. As well, the experimentation is carried out with the aid of the PubMed database documents. The performance of the proposed similarity measure is analyzed by means of the two existing cross ontology based similarity measure for different medical terms. We experimented with several semantic similarity methods for computing the conceptual similarity between natural language terms using WordNet and MeSH.

The experimental results indicate that it is possible for these methods to approximate algorithmically the human notion of similarity reaching correlation (with human judgment of similarity) up to 83% for WordNet and up to 74% for MeSH. This work also presents an improved X-Similarity, a novel semantic similarity measure which has been shown to out-perform the state-of-the-art cross ontology matching method [10].

References

[1] Li Y., Bandar Z.A., and McLean D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans. on Knowledge and Data Engineering, 15(4), 871-882.

[2] Resnik O. (1999). Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. Journal of Artificial Intelligence Research, 11, 95-130.

[3] Leacock C. and Chodorow M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet. In: Christiane Fellbaum, editor, An Electronic Lexical Database, pp. 265-283. MIT Press.

[4] Li Y., Bandar Z.A., and McLean D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE

Trans. on Knowledge and Data Engineering, 15(4), 871-882.
[5] Rada R., Mili H., Bicknell E., and Blettner M. (1989). Development and Application of a Metric on Semantic Nets. IEEE Trans. on Systems, Man, and Cybernetics, 19(1), 17-30.
[6] Richardson R., Smeaton A., and Murphy J. (1994). Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words. Tech. Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland.
[7] Miller G. and Charles W.G. (1991). Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, 6, 1-28.
[8] Jiang J.J. and Conrath D.W. (1998). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Intern. Conf. on Research in Computational Linguistics, Taiwan.
[9] Tversky A. (1997). Features of Similarity. Psychological Review, 84(4), 327-352.
[10] Rodriguez M.A. and Egenhofer M.J. (2003). Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Trans. on Knowledge and Data Engineering, 15(2), 442-456.
[11] Seco N., Veale T., and Hayes J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. Tech. report, University College Dublin, Ireland.
[12] D. Bollegala, Y. Matsuo, and M. Ishizuka, “Measuring Semantic Similarity between Words Using Web Search Engines,” Proc. Int’lConf. World Wide Web, pp. 757-766, 2007.
[13] J. Gracia, R. Trillo, M. Espinoza, and E. Mena, “Querying the Web:A Multiontology Disambiguation Method,” Proc. Int’l Conf. WebEng., pp. 241-248, 2006.
[14] Roy Rada, H. Mili, Ellen Bicknell, and M. Blettner. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17{30, January 1989.
[15] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133{138, Morristown, NJ, USA, 1994. Association for Computational Linguistics

Author Profiles

D. Jayasri obtained her Master’s degree in Applied Mathematics from P.S.G College of Technology,Coimbatore, and M.Phil degree from R.E.C,Tiruchirappalli.She worked as Professor and Head in the department of Applied sciences in SETHU institute of Technology, virudhunagar district,for 11 years .Currently she is working as Head, department of Mathematics ,ULTRA college of Engineering & Technology for women, Madurai.She is pursuing her Ph.D in Bharathiar University ,Coimbatore. Her area of specializations includes Webmining, Numerical methods, and Object oriented programming

Dr. D.Manimegalai is Professor and Head, Department of Information Technology, National Engineering College, Tamil nadu, India. She has published more than fifteen research papers in national and International Journal and Conferences. Her area of specializations includes Image Processing, Web mining and Mobile Ad hoc Networks.