IJCTA, 9(25), 2016, pp. 163-179
© International Science Press

ISSN: 0974-5572

Estimation of Semantic Similarity between Concepts using Multiple Ontologies (Wordnet and Mesh) FOR Biomedical DATA

B. Shobhana* and R. Radhakrishnan**

ABSTRACT

The majority of the intelligent knowledge-based applications include elements for the purpose of measuring semantic similarity stuck between terms. A lot of the available semantic similarity measures that make use of ontology structure as their chief source cannot determine semantic similarity among terms and concepts by means of multiple ontologies. Hence, this research looks for a new approach to determine the semantic similarity among the biomedical concepts by means of multiple ontologies. In this work, a novel resemblance assess is proposed and it is combining both super concept of the assessed ideas and their general certainity feature. This feature takes the deepness of the Least Common Subsumer (LCS) of two ideas and the deepness of the ontology by the way to attain further semantic evidence. The similarity in the midst of two concepts is a weighted sum total of the resemblances of the two features among them. The features are Data content similarity and Presentation style similarity. The weight is calculated by using rank aggregation criteria. The similarity is measured based on some rules and assumptions. The similarity measure is based on Information Content (IC) and context vector. Subsequently, the proposed measure was assessed relative to human specialists' ratings, and evaluated against the existing schemes of biomedical terms by means of two standard ontologies (WordNet and MeSH). WordNet was used as primary general ontology and MeSH was used as secondary ontology. The investigational outcomes proved the effectiveness of the proposed scheme, and demonstrated that proposed similarity measure provides the most excellent complete outcomes of correlation with individual ratings.

Keywords: semantic similarity, super concepts, evaluated concepts, least common subsumer, information content, context vector, data content similarity and presentation style similarity.

1. INTRODUCTION

Semantic similarity methods helps to approximate the similarity among concepts, and take part a significant responsibility in a range of text processing responsibilities, together with document classification, information retrieval, information extraction, automatic spelling error recognition, correction systems and word sense disambiguation.

Similarity schemes employed in the field of biomedical area can be approximately partitioned into knowledge grounded and distributional grounded schemes [1]. Knowledge based schemes makes use of pre-existing knowledge sources, together with taxonomies, dictionaries and semantic networks. Along with the knowledge based schemes to which much attempt in the biomedical domain has been granted are schemes that make use of the taxonomic arrangement of a biomedical terms for the purpose of computing similarity; these comprise path determining measures and intrinsic Information Content (IC) measures. These distributional schemes make use of the distribution of thoughts inside a corpus along with a knowledge foundation for the purpose of computing similarity; these comprise quantity IC and context vector approaches.

* Research Scholar, Anna University, Chennai, Tamilnadu, India, *Email: sobhanab.scholar@gmail.com*

** Principal, Vidhya Mandhir Institute of Technology, Perundurai, Tamilnadu, India.

IC is a kind of determination of idea specificity and it is normally approximated from concept frequencies inside a corpus (corpus IC). On the contrary, intrinsic IC is an estimation of IC calculated from the configuration of taxonomy. Since it does not depend on a corpus, knowledge grounded schemes are extremely useful to calculate than distributional schemes. On the other hand, it is indistinct whether knowledge based schemes are as precise as distributional schemes: assessments in the biomedical area that evaluate these schemes were indecisive.

In the field of biomedical, semantic certainity methods have been calculated on the Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT); Medical Subject Headings (MeSH); and moreover the Unified Medical Language System (UMLS), a compendium of biomedical source terminology that encompass SNOMED CT and MeSH [2]. The most part of these assessments were carried out on a undersized standard of 29 SNOMED CT idea pairs. Medical Subject Headings (MeSH), [3], is one amongst the resource vocabularies employed in UMLS. MeSH comprises in relation to 15 high-level categories, and every category is partitioned into subclassification and allocated a letter: A signifies Anatomy, B signifies Organisms and C represents Diseases, and so on.

Petrakis et al. [4] expanded the previous scheme in accordance with the matching among synonym sets and concept glosses derived from WordNet (that is words mined through the process of parsing concept definitions) or scope notes obtained from MeSH. Once each term to evaluate is independently matched to a concept of all ontologies, they are considered similar when their synonyms (i.e. 10 different labels similar to a particular ontological concept) and moreover the glosses, and those of the notions in their neighbourhood are lexically comparable. The similarity is determined by taking the maximum similarity value acquired through comparing synonyms and glosses per separate by means of the Jaccard coefficient [5].

In this paper, a new similarity measure is formulated and it is combining both super concepts of the estimated concepts and their common specificity feature. This characteristic believes the deepness of the LCS of two ideas and the deepness of the ontology for the purpose of obtaining more semantic confirmation. Then the measure of comparable two concepts is calculates as aweighted total of the certainity of the two features among them. The features are Data content similarity and Presentation style similarity. The weight of similarity among two ideas is intended by using aggregation criteria. The similarity is measured based on some rules and assumptions. Then the similarity measure is based on Information Content (IC) and context vector. The WordNet is used as input ontology. Then the proposed method similarity is evaluated and compares to other existing methods.

This work is prearranged as follows: Section 2 provides a general conception of semantic resemblance and related method, biomedical knowledge foundations and some assessments. Section 3 provides anexisting computation for semantic similarity. In section 4 discusses proposed a scheme for semantic certainity computations and the approaches utilized to assess them. Section 5 presented the outcomes of estimation on semantic similarity process. Section 6 concludes the paper and discussed future work.

2. BACKGROUND STUDY

In the field of biomedical domain, an investigation by Pedersen et al. that exploited a huge medical corpus for the purpose of estimating the concept distributions confirmed that distributional ways perform better than taxonomy grounded path discovery measures. The investigation conducted by Sanchez et al. proved that knowledge grounded intrinsic IC calculates perform improved than distributional measures [2]. On the other hand, methodological diverses in this latter examination prevent a straight evaluation among knowledge basedgrounded and distributional supported measures. Assessments of similarity estimations here applied private, institution-explicit corpora of clinical comments [6]. Huge corpora of clinical comments are not openly accessible because of the compassion of the information controlled in that, and smaller, openly accessible corpora possibly will bias idea frequency evaluates [13]. On the other hand, it is easy to make use of an openly accessible biomedical corpus like MEDLINE, which includes more than 19 million abstracts from biomedical journals [7]. But, assessments of context vector measures in accordance with 300,000 MEDLINE conceptuals showed worst recital than measures in accordance with a clinical corpus. Through a generously proportioned subset of MEDLINE might overcome this complication, however dispensation this corpus indicates a technical dispute that might be too expensive for several applications. In addition, for the purpose of computing corpus IC-based measures, text has to be mapped with concepts. Computerized idea mapping faults possibly will bias idea frequency add up, unconstructively impacting the accurateness of corpus IC grounded measures.

Semantic certainity and related evaluations even though technically they indicate to several concepts of associated, the words similarity and correlated are typically utilized interchangeably. This is incooperated because of the idea that these methods are implemented to the similar categories of text processing responsibilities and assessed on the similar benchmarks [8]. Semantic similarity is a category of semantic correlated, specifically taxonomic relatedness, for instance, Lovastatin is-a Statin [9]. The semantic relatedness can indicate to non-taxonomic associated antonymy, meronymy (part-of), frequent and added practical associations (e.g. treated-by) [10].

The knowledge dependent semantic similarity evaluation comprise random walk, path recognition, and intrinsic IC dependent measures. These kinds of measures produce a thought graph grounded on a taxonomy or semantic system wherein vertices indicate ideas and edges indicate semantic associations. Path recognition and intrinsic IC dependent measures make use of taxonomies, that is, an acyclic, directed idea graph wherein edges indicate taxonomic associations. Taxonomyappropriate for utilizing with semantic similarity methods can be obtained from a knowledge grounded by means of taking a separation of hierarchical semantic associations, and eliminating associations that persuade cycles. Ideas that are simplifications of further concepts are taken as parents or hypernyms; requirements of a idea are indicated to as offspring or hyponyms.

Path finding dependent semantic similarity evaluation calculate similarity as a function of the extent of the shortest path stuck among two concepts. One drawback of path recognition measures is that they provide equivalent weight to the complete associations. Information Content (IC) dependent measures make an attempt to accurate this through weighting edges dependent on IC, a determination of idea specificity [11]. Associations among particular concepts, e.g. Lovastatin is-a Statin, must be weighted more greatly than associations among general concepts, e.g. Lovastatin is-an Enyme Inhibitor. Intrinsic IC dependent measures work out the IC of ideas from the taxonomic configuration. The assumption for this scheme is that the taxonomic configuration is prearranged in a significant manner, like ideas with several hyponyms and a small amount of hypernyms have lesser IC [12].

Random walk evaluation figure out the relatedness among a brace of concepts by means of accidental walks on a idea graph [13]. In opposition to path decision and intrinsic IC methods, the random walk evaluates can make use of graphs that include undirected edges, non-taxonomic associations, and cycles. In place of determining associated as a process of the shortest pathway among ideas, random walk schemes determine the complete connectivity among concepts. Here theyconcentrated on the personalized PageRank (PPR) algorithm that accomplished high-tech recital on common language semantic resemblance standards, however has not assessed on biomedical semantic similarity responsibilities. For a particular concept, PPR produces a totality vector that stands for its linking to further concepts. The relatedness among a brace of ideas is described as the cosine of angle stuck among their score vectors.

Distributional dependent measures make use of a area corpus in combination with a information source; these comprise corpus IC and context vector evaluation. The Corpus IC dependent schemes are equivalent to intrinsic IC grounded schemes, however approximate the IC of a idea from its allocation in a corpus. The context vector evaluation of semantic association are in accordance with the supposition that words that become visible in comparable contexts are associated [14]. This scheme initiates by means of generating word vectors from a corpus that indicate word co-occurrence. Subsequently, descriptor stipulations for a idea are obtained from a information source like a glossary or thesaurus, and can be additionally developed to comprise descriptor stipulations from associated concepts. The word vectors equivalent to a concept's descriptor stipulations are subsequently aggregated for the purpose of constructing a context vector. The similarity among a brace of concepts is described as the cosine of angle among the context vectors.

Recent schemes try to find terminologically-equivalent concepts among different ontologies. In the scheme formulated by Saruladha et al. [15], it computes semantic similarity amongst biomedical ontologies in accordance with an information-theoretic viewpoint of the Tversky's measure [16]. The sum of commonality that is present among ideas c_1 and c_2 is pointed out through the IC of their LCS (i.e. LCS(c_1,c_2)), at the same time their differences are conceived as IC(c1) and IC(c_2). By taking that both ontologies are united by means of a new imaginary root node, the LCS(c_1,c_2) is obtained through matching the set of subsumers of c_1 in the first ontology and the collection of subsumers of c_2 in the second ontology through a terminological matching. The IC of concepts is calculated in a basic approach from the knowledge modeled in the ontology, with the intention of avoiding dependency on corpora availability. The deepness of both concepts in the ontology is calculated in the similar manner as Rodriguez and Egenhofer [17].

Al-Mubaid and Nguyen [18] formulated a scheme for the purpose of assessing similarity of concepts among multiple biomedical ontologies. It effectively evaluates term pairs by means of a similarity measure, which integrates the path distance and common specificity of the correlated ontological concepts. The general specificity of two ideas is computed through subtracting the deepness of the LCS from the deepness of the taxonomic branch to which they fit in. Since, the path length gives absolute values, the scheme based on the selection of a predetermined primary ontology (the remaining are taken as less important) to which certainity values are normalized.

3. EXISTING MEASURES FOR COMPUTING SEMANTIC SIMILARITY

This section discusses the general features shared with the information units fits to the identical concept, and the entire part can be repeatedly attained. The IC and Context Vector measures are discussed as follows.

3.1. Data Content (DC)

Here, the data elements or text nodes comprises similar concept habitually share definite keywords. The information units related to the investigation field where the user comes into a search state typically include the exploration keywords. Text nodes that enclose information units of the similar concept typically have the identical file name.

3.2. Presentation Style (PS)

This characteristic explains how a information unit is exhibited on the system. It includes six style features: basis face, font color, font size, manuscript decoration (underline, strike, etc.), font weight and italic format.

3.3. Path Finding Measures

Here concentrated on the Path [19], Leacock & Chodorow (LCH) [20], and Wu & Palmer [21] path recognition evaluates that are dependent on the shortest path unravelling ideas. Conside$p$ = path (C_1, C_2), the quantity of nodes in the unswerving path splitting two concepts,C_1 and C_2. The shortest path among two concepts navigates their Least Common Subsumer(lcs(C_1, C_2)), i.e. their nearby common parent. The depth(depth (c)) of a idea is described as the amount of nodes in the pathway to the origin of the taxonomy; and indicates the utmost profundity of taxonomy.

Path describes the similarity among two concepts basically as the converse of the extent of the path unravelling them:

$$\mathrm{sim}_{\mathrm{path}}\,(C_1, C_2) = 1/\mathrm{p} \tag{1}$$

LCH fully depends on the fraction of path length to deepness, however executes a logarithmic scaling. In the beginning, LCH was given as follows

$$sim_{lch}^{unscaled}\left(C_1, C_2\right) = -\log\left(\frac{p}{2d}\right) = \log\left(2d\right) - \log\left(p\right) \tag{2}$$

At the same time as proposed range LCH to the unit period by means of dividing by log(2d). Dividing through a stable value has no influence on the spearman association together with benchmarks: the associated position of notion pair similarities continue to be same.

$$sim_{wp}^{unscaled}\left(C_1, C_2\right) = \frac{2 \times depth\left(lcs\left(C_1, C_2\right)\right)}{path\left(C_1, lcs\left(C_1, C_2\right)\right) + path\left(C_2, lcs\left(C_1, C_2\right)\right) + 2 \times d} \tag{3}$$

Wu & Palmer balances the deepness of the LCS through the extent of the path among two concepts:

One major complication with this description that the identity of a idea with itself is below 1 (if $C_1 = C_2$, then path (C_1, lcs(C_1, C_2)) + path (c_2, lcs(C_1, C_2)) = 2). As an alternative, here adopted the description of Wu & Palmer employed in the Natural Language Toolkit:

$$sim_{wp}\left(C_1, C_2\right) = \frac{2 \times depth\left(lcs\left(C_1, C_2\right)\right)}{p - 1 + 2 \times depth\left(lcs\left(C_1, C_2\right)\right)} \tag{4}$$

Based on this, if $C_1 = C_2$, then $p - 1 = 0$, and the similarity determination evaluates to 1.

3.4. Measures based on Information Content

Resnik [22] formulated to balance the taxonomical arrangement of ontology with the data distribution of ideas assessed in input corpora. Here also utilized the concept of IC, by means of connected form probabilities to every idea in the taxonomy, calculated from their incidences in a particular corpus. IC of a term is calculated in accordance with the negative log of its probability of occurrence, $p(a)$ (15). In this way, uncommon words are considered more useful than common ones.

$$\mathrm{IC}(a) = -\log \mathrm{P}(a) \tag{5}$$

In accordance with Resnik, semantic similarity completely based on the quantity of shared information among two terms, a dimension which is indicated with their LCS in ontology. Two terms are maximally unrelated when a LCS does not present (that is, in terms of edge-counting, it might not expected to discover a path connecting them). If not, their similarity is calculated as the IC of LCS.

$$simres\,(a, b) = \mathrm{IC}(LCS(a, b)) \tag{6}$$

A small complication in the Resnik's metric is that any brace of terms comprising the similar LCS outcomes in accurately the similar semantic similarity. Both Lin [23] and Jiang and Conrath [24] extended Resnik's work through taking the IC of each of the assessed terms.

Lin formulated that the similarity among two terms is supposed to be measured as the ratio among the quantity of details required to state their commonality and the details required to completely describe them. At the same time, as a corollary of this theorem, this measure takes, commonality in the similar manner as Resnik's scheme and, in contrast, the IC of every concept alone.

$$sim_{lin}\left(a,b\right)=\frac{2\times sim_{res}\left(a,b\right)}{\left(IC\left(a\right)+IC\left(b\right)\right)} \tag{7}$$

The measure formulated by Jiang and Conrath is completely based on quantifying, in certain manner, the extent of the taxonomical links as the dissimilarity among the IC of a notion and its respective subsumer. While comparing term pairs, they calculate their distance by means of subtracting the total IC of every term alone from the IC of particular LCS (18).

$$dis_{j\&c}\left(a,b\right)=\left(IC\left(a\right)+IC\left(b\right)\right)-2\times sim_{res}\left(a,b\right) \tag{8}$$

It is significant to point out that IC-based measures require to behave appropriately, that the possibility of appearance *p* monotonically raises as one progresses up in the taxonomy (i.e., $\forall$ ci | cj is hypernym of ci = > *p*(*ci*) $\leq$ *p*(*cj*)). This is accomplished by means of computing *p*(*a*) as the probability of meeting any instance of in the particular corpus. In actual fact, every individual occasion of any noun in the corpus is added up as an incidence of every taxonomic class containing it [28].

$$p\left(a\right)=\frac{\sum_{w\in W\left(a\right)}count\left(w\right)}{N} \tag{9}$$

where W(a) indicates the group of nouns within the corpus whose senses are subsumed through , and signifies the overall number of nouns in the corpus.

Accordingly, a precise computation of concept probabilities needs an appropriate disambiguation and annotation of every noun initiated in the corpus. This procedure is typically done manually in order to guarantee the appropriateness of the tagging, applicability and hampering the scalability of this scheme with huge corpora.

Furthermore, when either the taxonomy or the corpus transformations, re-computations are required to be recursively implemented for the affected concepts. As a result, it is essential to carry out a manual and time consuming investigation of corpora and resultant probabilities would depend on the size and temperament of input corpora. Besides, the background taxonomy have to be as absolute as possible (i.e., it should comprise most of the specializations of all concepts) with the aim of providing consistent results. Partial taxonomies with a restricted scope possibly will not be appropriate for this purpose.

By taking the drawbacks of IC-based schemes because of their dependency on corpora, a few authors attempted to intrinsically derive IC values from ontology. These works depends on the supposition that the taxonomic arrangement of ontologies like WordNet is put in order in a significant manner, in accordance with the rule of cognitive saliency [25]. This confirms that humans specialized concepts when they require to distinguish them from previously existing ones. As a result, concepts with several hyponyms (i.e., specializations) are extremely common and offer less details than the concepts in the place of leaves of the hierarchy. In accordance with the Information Theory viewpoint, they take that abstract ontological concepts come into view most likely in a corpus since they subsume several other ones. In this way, the chance of emergence of a concept (that is, the IC) is approximated as a function of the quantity of hyponyms and/or their associated deepness in the taxonomy.

Seco et al., [26] and Pirró and Seco [27] based IC computations on the amount of hyponyms. In view of the fact that, hypo(a) the amount of hyponyms of the concept a and max_nodes the amount of hyponyms of the root node, they calculate IC of a concept in the following manner (10):

$$IC_{\mathrm{sec}o}\left(a\right)=1-\frac{\log\left(hupo\left(a\right)+1\right)}{\log\left(max\_nodes\right)} \tag{10}$$

The denominator guarantees that IC values are normalized in the limit [0...1]. This scheme only takes hyponyms of a particular concept in the taxonomy; as a result, ideas with the similar amount of hyponyms however diverse degrees of generality come out to be equally comparable. With the aim of dealing with this complication effectively, and the similar method as for edge-counting evaluations, Zhou et al., [28] formulated to balance hyponym-based IC computation with the associated deepness of each concept in the taxonomy. The IC of a concept is found as given below:

$$IC_{zhou}(a)=k\left(1-\frac{\log(hupo(a)+1)}{\log(max\_nodes)}\right)+(1-k)\left(\frac{\log(depth(a))}{\log(max\_depth)}\right) \tag{11}$$

Besides hypo and max-, which has the similar meaning as eq. 20, depth(a) corresponds to the deepness of the concept in the taxonomy and max_depth indicates the maximum deepness of the taxonomy. The factor fine-tunes the weight of the two features engaged in the IC assessment. Here used k = 0.5.

3.5. Context Vector Measures of Semantic Relatedness

Patwardhan and Pedersen [10] formulated a determination of semantic relatedness that points out a idea with a context vector which carries out more stretchy than other similarity measurements, because the information resource for context vectors is an untreated corpus of manuscript and the ideas do not necessitate to be associated through a pathway of associations in ontology. They put together gloss vectors in proportion to every idea in WordNet by means of the cooccurrence details together with the WordNet descriptions. During the experimentations, the glosses appears to enclose content well-off terms. They would differentiate a variety of concepts much enhanced than text haggard from additional generic corpus when authors prefer the WordNet glosses. Furthermore, the WordNet glosses can be measured as a corpus of contexts comprising of around 1.4 million words. Subsequently, the gloss vector measure got the maximum correlation concerning human judgment by means of different benchmarks.

Gloss vectors for all concepts in WordNet can be computed in this manner. The associations of two ideas is then determined as the cosine of the normalized gloss vectors matching to the two ideas:

$$related_{vector}(c_1,c_2)=\cos\left(angle(\vec{v}_1,\vec{v}_2)\right) \tag{12}$$

Where c_1 and c_2 are the two given concepts, $\vec{v}_1$ and $\vec{v}_2 \rightarrow$ gloss vectors matching to the ideas and returns the angle between vectors. Using vector products, the above relatedness formula can be rewritten as:

$$related_{vector}(c_1,c_2)=\frac{\vec{v}_1\cdot\vec{v}_2}{|v_1||v_2|} \tag{13}$$

The measure of semantic association grounded on WordNet and MeSH glosses, which is enhanced with information from a huge corpus of text. However, it should be pointed out that this measure is not dependent on WordNet and MeSH glosses, and can be employed with any representation of concepts (such as dictionary definitions), with co-occurrence count up from several corpus.

4. PROPOSED SEMANTIC SIMILARITY MEASURE

In this section, the new semantic similarity measure is proposed for multiple ontologies in biomedical domain. The measures are discussed and semantic similarity is evaluated using WordNet and Mesh.

4.1. System Overview

Fig 1 shows that overview of proposed system. The multiple ontologies of WordNet and MeSH are applied for medical data set to calculate the similarity among ideas. Subsequently, the general specificity characteristic

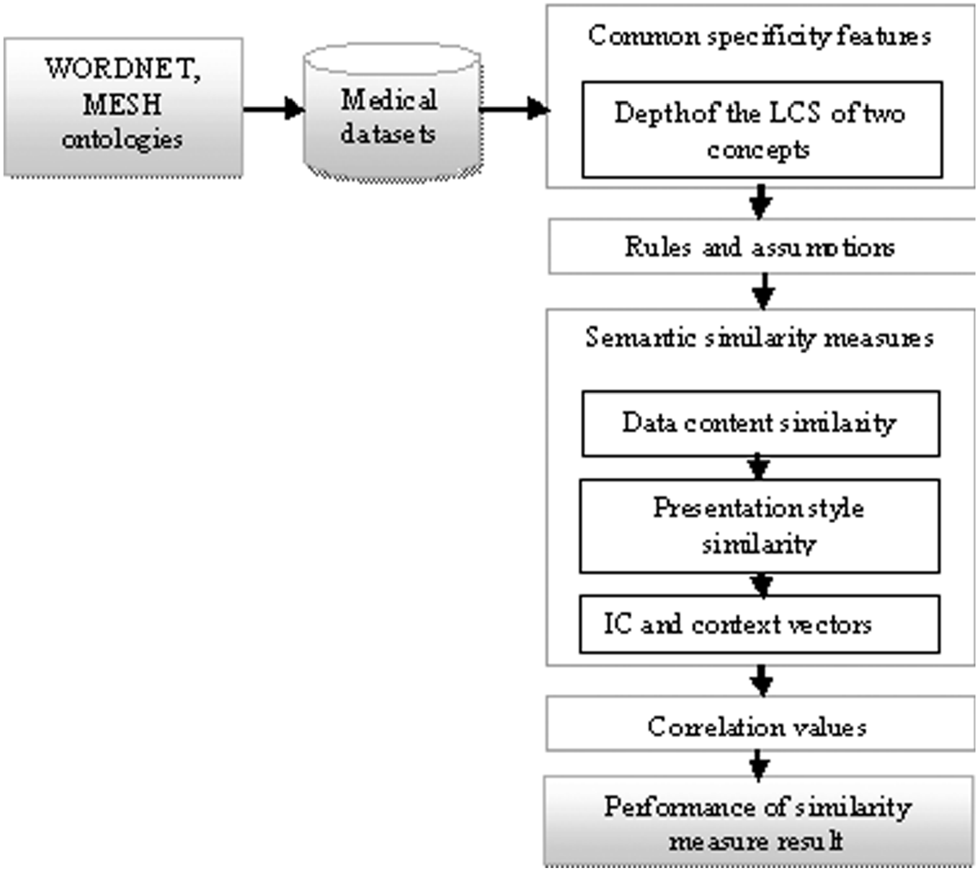
WORDNET, MESH ontologies
Medical datasets
Common specificity features
Depthof the LCS of two concepts
Rules and assumptions
Semantic similarity measures
Data content similarity
Presentation style similarity
IC and context vectors
Correlation values
Performance of similarity measure result

Figure 1: Architecture of proposed system

takes the deepness of the LCS of two ideas and the deepness of the ontology for obtained added semantic confirmation. The similarity measures are based on the rules and assumption. Similarity measures are IC and context vector used to calculate semantic association among concepts. Then calculated the correlation values and discussed the similarity result.

4.2. Proposed Measure for Computing the Semantic Similarity

Ahead of examining the particulars of the anticipated approach, regulations and suppositions are presented to be fulfilled in the anticipated measure.

4.3. Rules and Assumptions

Discussed above all the semantic characteristics are combined in one determination in an efficient and reasonable technique. This intuitive rules and suppositions are summed up as follows:

Rule R1: The semantic certainity scale scheme demonstrates the limit range of association of pairs of ideas like in WordNet and mesh ontology. This rule makes sure that the charting of concept 1 to concept 2 does not weaken the resemblance scale of any concepts.

Rule R2: The semantic association is required to follow local concept's identity rule as given below:

Rule R2.1: The lesser the distance among two idea nodes in the hierarchy tree, then added are related.

Rule R2.2: Subordinate level braces of ideas nodes are semantically nearer than superior level pairs.

Rule R2.3: The utmost similarity is accomplished at the time of two idea nodes are the similar nodes in the hierarchy tree.

Ahead of presenting the particulars of the anticipated determination, all the assumptions are provided clearly regarding the semantic law function:

Assumption A1: Logarithm computations are the universal rule of semantic distance.

Exponential-decay functions are universal rule of stimulus oversimplification for psychological sciences. At this point, make use of logarithm (opposite of exponentiation) for semantic distance. It is to be observed that non-linear combination scheme is the optimum scheme for the purpose of combining semantic features. Rule: R2.3 demonstrates that if the two concept nodes are the similar node (that is they are the same or equal), the semantic linearity have to accomplish highest similarity irrespective of added features, therefore, it is necessary to make use of non-linear scheme for the purpose of combining the features. As a result, it requires another assumption.

Assumption A2: The Non-linear function is the common combination rule of semantic similarity characteristics.

Here, the most common similarity among two concepts (or two transcript nodes) C_1 and C_2 indicates a weighted total of the identities of the two features among them, i.e.:

$$Sim\left(C_1,\ C_2\right) = w_1 * \left(Sim\ C\left(c_1,\ c_2\right) + w2 * Sim\ P\left(c_1,\ c_2\right)\right) \tag{14}$$

4.4. Data content similarity (SimC)

It is Cosine similarity among the expression frequency vectors of C_1 and C_2:

$$simC\left(c_1, c_2\right) = \frac{V_{C_1} \cdot V_{C_2}}{\left|V_{C_1}\right| * \left|V_{C_2}\right|} \tag{15}$$

Where V_d indicates the occurrence vector of terms within the concept C, || V_d ||indicates the extent of V_d, and the numerator is the internal product of two given vectors.

4.5. Presentation style similarity (SimP)

It indicates the typical style Feature Scores (FS) over the entire six arrangement style Features (F) among C_1 and C_2:

$$simP\left(c_1, c_2\right) = \sum_{i=1}^{6} FS_{i/6} \tag{16}$$

Where FSi → score of the style feature and it is distinct by $FS_i = 1F_i^{C_1} = F_i^{C_2}$ and $FS_i = 0$ or else, and → ith style feature of information unit d.

4.6. Common Specificity Feature

Here, a novel customized measure for the purpose of semantic identity by means of combining both the super concepts of the assessed ideas and general specificity characteristic which canconfine further semantic indication. This determination can accomplish improved recital than erstwhile measures, since it is completely grounded on arrangement and continue their simplicity. The contributions of the entire noncommon super concepts of the assessed ideas are considered and all non common super concepts in place of the least path extent to imprison added semantic indication for similarity. Moreover, in proposed measure, the general certainity features of the idea nodes scaled through the deepness of the idea nodes and the deepness of the LCS. As a result, noncommon super concepts also united by considering with general specificity with the objective of obtaining higher semantic data for calculating similarity among two concepts. Consider c_i stand for th concept of ontology. Subsequently, $N(c_i)$ is defined as the collection of the entire super concepts of c_i including c_i itself. As a result, the amount of noncommon super concepts can be determined as follows:

$$Noncomsub\left(c_1,c_2\right)=\left|N\left(c_1\right)\cup N\left(c_2\right)\right|-\left|N\left(c_1\right)\cap N\left(c_2\right)\right| \tag{17}$$

At this point, the value can be a sign of the path length of the two ideas. LCS node of two ideas C_1 and C_2 establishes the common specificity of C_1 and C_2 in the group. So the determination of two ideas are computed through the process of finding the deepness of their LCS node and subsequently scaling this deepness as follows:

$$ComSpec\left(c_1,c_2\right)=D-depth\left(LCS\left(c_1,c_2\right)\right) \tag{18}$$

Where D indicates the deepness of the ontology and the ComSpace feature decides the general specificity of two assessed concepts. The lesser the value of two ideas, the additional details they split, and as a result the more comparable they are.

Subsequently, logarithm task of NonComSub and ComSpec is used to indicate semantic remoteness which is completely contrary to semantic resemblance. As a result, the semantic distance among concepts c_1 and c_2 is given as follows:

$$\mathrm{SemD}\left(\mathrm{c1},\mathrm{c2}\right)=\log\left(\mathrm{NonComSub}\times\mathrm{ComSpec}+1\right) \tag{19}$$

It is to be mentioned that whichever concept can be evaluated with itself.

4.7. Footrule similarity

In this type of similarity attained less accuracy. So the proposed system is concentrated the ranking process. The ranking is provided based on most preferred concept in a universe of U concepts. The ranking is done by using Borda's method. This method relies on the absolute position of concepts in the ranked lists, rather than their relative rankings, and is a typical indication of the class of positional ranking schemes. This scheme is normally efficient on the basis of computational perspective, however no positional scheme can create rankings guaranteed to satisfy the Condorcet criterion [30]. Borda's scheme allocates ranks to items in accordance with an overall Borda score, which is computed on every list by means of showing that the most preferred concept in a universe of U concepts gets |U| points, the next gets |U| – 1 points, and so on.

More formally, Borda's scheme on a set of overall rankings R is calculated as given below. For each concept C_1 and list $r_k \in R$, let B_{rk} (C_1) equal the number of items C_2 in r_k such that $r_k(C_2) > r_k(C_2)$. The total Borda score for the concept C_1 is given by $Bt(C_1) = \sum_{r\in R}Br(C_1)$. Ranks are assigned by sorting scores B_t from maximum to minimum, with the maximum score receiving the low rank.

When R comprises incomplete rankings, one proposal is to allocate any "leftover" score from a particular list evenly among the remaining unranked concepts in the list. Specifically, for a list r_k, where $| r_k | = | U | -$ d, compute the B_{rk} (C_1) as usual for all concepts $i \in r_k$, and assign $B_{rk}\left(C_2\right)=\frac{\left(d+1\right)^2-\left(d+1\right)}{2d}$ for all items $j \in r_k$. The Borda scores B_t and rankings are assigned as above.

Based on the above ranking process the similarity is measured between two concepts C_1 and C_2. In this proposed similarity is based on the footrule similarity. The expansion of the Spearman footrule distance by means of including the result of a similarity function described on concepts in U. This measure is described in concepts of a generic similarity function. In actual fact, the option of a specific similarity measure for domain is of huge significance, as well as the progress of similarity measures is an active area in the data mining and pattern recognition communities. At this point, take the similarity measure as specified, and presume that the similarity scores returned are significant inside the particular domain.

The footrule similarity distance among two (maybe partial) ranked lists σ and r, provided a similarity function s(.,.) is given as:

$$F_{sim}\left(\sigma,r,s\left(c_1,c_2\right)\right)=\sum\nolimits_{C_1\in\sigma\lambda|r}\sum\nolimits_{C_2\in r\lambda|\sigma} sim\left(C_1,C_2\right)\left|\sigma_{\lambda|r}\left(C_1\right)-r_{\lambda|\sigma}\left(C_2\right)\right| \tag{20}$$

Specifically, the footrule similarity distance is computed on similarity projections of σ and r. The difference in ranks for items in these resultant lists is weighted through the power of the similarity.

This determination can be functional to the entire concepts. Subsequently, the most common similarity between these semantic similarity and footrule similarityis a weighted total of the similarities:

$$Sim\left(C_1,C_2\right)=w_1*\left(SemD\left(c1,c2\right)+w2*F_{sim}\left(\sigma,r,s\left(c_1,c_2\right)\right)\right) \tag{21}$$

4.8. Experiments and Results

In this segment, the proposed method is evaluated and results are compared with existing measures.

4.9. Datasets

There are denial typical human rating data sets for the evaluation of semantic identity in the field of biomedical domain. In this system, carried out an experiment similar to the one proposed. Two ontologies are utilized as background information: WordNet and MeSH. WordNet is a lexical database that expresses and organizes over 100,000 general concepts, which are semantically organized in an ontological manner. WordNet includes English words (verbs, nouns, adverbs, and adjectives) that are associated to group of cognitive synonyms (synsets), every one expressing a different concept. Synsets are related through the conceptual-semantic and lexical associations like synonymy, hypernymy (subclass-of), meronymy (part-of), etc. WordNet version 2 is employed in the proposed tests as it is the most ordinary version employed in the associated works. The Medical Subject Headings (MeSH) includes a chain of command of medical and biological stipulations described through the U.S National Library of Medicine. This classification was primarily generated to catalogue books and additional library materials, and in order to index articles for addition in health-associated databases together with MEDLINE. In MeSH tree, there are 16 basic categories, with more than 22,000 concepts. The latest 2011 MeSH XML files used and are available for download.

The reason for combining a general-purpose ontology like WordNet with a domain-specific one, such as MeSH is to consider a least favorable evaluation scenario. Recall both ontologies have been intended with considerably diverse scopes. Hence, the modeled subsumers and taxonomical trees tend to be significantly different for a pair of similar terms. On one hand, the dissimilarity in scope are reproduced by the idea that only 5.4% of WordNet concepts appear in MeSH with the same textual label (including synonyms). This will affect the accuracy of methods based solely on terminological matching's, showing the advantages of a semantically-grounded method.

The proposed system use the benchmarks of Hliaoutakis et al.[29] and Pedersen et al.[10] for evaluation. For the first one, the ratings taken and is given by the 3 physicians, 9 medical coders and the average of both. Owing to the idea that these standards are meant to assess mono-ontology identity measures, most of the word pairs can be found both in MeSH and in WordNet. In fact, 20 out of the 30 word pairs of the benchmark of Pedersen et al. and 35 out of 36 word pairs of the Hliaoutakis et al. benchmark can be found in both ontologies (tables I and II). On one hand, only those pairs are considered that can be found in both ontologies in order to compare the identity precision obtained when using a unique (and semantically coherent) information source with the huge demanding multi-ontology situation. Nonetheless, to assess the multi-ontology situation, one of the two terms of each pair is considered to be in WordNet (regardless it can also be found in MeSH), whereas the second term is considered to be in MeSH (regardless being found in WordNet), as done in [13]. In this way, a appropriate LCS among the two ontologies must be exposed to facilitate the similarity evaluation. Tables I and II show which terms of each benchmark were evaluated in which ontology.

Because of space limitation, Table I includes medical terminology pairs with averaged identity scores of specialists. The standard correlation between physicians is 0.68, and among specialist is 0.78. Since, the specialists are above the physicians, and the agreement among specialists (0.78) is exceeding the correlation among physicians (0.68), presumes that the specialists' rating attainment are more consistent than the physicians' evaluation scores, and as a result specialists' scores is utilized in this experiments.

The subsequent dataset (Dataset 2) comprises 36 biomedical (WordNet and MeSH) terminology pairs [3]. In case of the human rates in this particular data set are the standard assessed scores of consistent doctors. Table II includes medical term pairs with averaged similarity scores of specialists.

For all arrangements and standards, as similarity evaluation, used the existing path-based functions and IC and Context vector measures. Correlation values with human judgments for the diverse process, evaluations, standards and ontology amalgamations are summarized in Table III.

Table 1
Medical Term Pairs with Averaged Identity Scores of Specialists.

Word Net term	*MeSH term*	*Physician ratings*	*Coder ratings*
Renal failure	Kidney failure	4.0	4.0
Myocardium	Heart	3.3	3.0
Infarct	Stroke	3.0	2.8
Abortion	Miscarriage	3.0	3.3
Schizophrenia	Delusion	3.0	2.2
Adenocarcinoma	Metastasis	2.7	1.8
Stenosis	Calcification	2.7	2.0
Diarrhoea	Stomach cramps	2.3	1.3
Atrial fibrillation	Mitral stenosis	2.3	1.3
Rheumatoid arthritis	Lupus	2.0	1.1
Osteoarthritis	Carpal tunnel syndrome	2.0	1.1
Hypertension	Diabetes mellitus	2.0	1.0
Acne	Syringe	2.0	1.0
Antibiotic	Allergy	1.7	1.2
Multiple sclerosis	Psychosis	1.0	1.0
Appendicitis	Osteoporosis	1.0	1.0
Xerostomia	Alcoholic cirrhosis	1.0	1.0
Peptic ulcer disease	Myopia	1.0	1.0
Cellulitis	Depression	1.0	1.0
Hyperlipidaemia	Metastasis	1.0	1.0

Table 2
Medical Term Pairs with Averaged Identical Rates of Specialist Haul Out from the Hliaoutakis et al [35].

WordNet term	*MeSH term*	*Expert ratings*
Appendicitis	Anemia	0.031
Dementia	Atopic Dermatitis	0.060
Otitis Media	Infantile Colic	0.156
Vaccines	Immunity	0.593
Hypothyroidism	Hyperthyroidism	0.406
Lactose Intolerance	Irritable Bowel Syndrome	0.468
Urinary Tract Infection	Pyelonephritis	0.656
Sepsis Neonatal	Jaundice	0.187
Anemia Deficiency	Anemia	0.437
Psychology	Cognitive Science	0.593
Adenovirus	Rotavirus	0.437
Migraine	Headache	0.718
Myocardial Infarction	Myocardial Ischemia	0.750
Hepatitis B	Hepatitis C	0.562
Carcinoma	Neoplasm	0.750
Pulmonary Stenosis	Aortic Stenosis	0.531
Breast Feeding	Lactation	0.843
Antibiotics	Antibacterial Agents	0.937
Seizures	Convulsions	0.843
Ache	Pain	0.875
Chicken Pox	Varicella	0.968

Table 3
Correlations Transversely Evaluates for Physicians, Coders, and both on 29 Pairs of Dataset 1 and Correlations Across Evaluations for Human on 34 Pairs of Dataset 2 Utilization Word Net and MeSH Ontologies

Measure	*Dataset 1*			*Dataset 2*
	Physicians	*Coders*	*Both*	*(Word Net + Mesh)*
Path length	0.36	0.51	0.48	0.586
Leacock and Chodorow	0.35	0.50	0.47	0.677
Wu and Palmer	N/A	0.29	N/A	0.686
Li et al.	N/A	0.37	N/A	0.694
Choi and Kim	N/A	0.15	N/A	0.440
Al-Mubaid and Nguyen (SemDistA&N)	N/A	0.66	N/A	0.735
Montserrat Batet et al. (SimB&S)	0.60	0.79	0.73	N/A
Resnik	0.45	0.62	0.55	N/A
Lin	0.60	0.75	0.69	N/A
Jiang and Conrath	0.45	0.62	0.55	N/A
Context vector (1m notes, diagnostic section)	0.84	0.75	0.76	N/A
Context vector (100,000 notes, diagnostic section)	0.56	0.59	0.60	N/A
Context vector (100,000 notes, all sections)	0.41	0.53	0.51	N/A
Fengqin Yang et al.(SNOMED CT)	0.67	0.77	0.75	0.774
Proposed measure	0.69	0.79	0.78	0.78

4.10. Discussion

Analyzing the results shown in table III, several conclusions may be drawn. Primarily, one observes dissimilarity in correlation values for the two benchmarks when evaluating word pairs in MeSH. For Data set 1, there are29out of 30conceptpairs in WordNet +MeSH and the typical correlation among physicians are 0.69 at the same time the typical correlation between medical coders is 0.79.

The assessment outcomes demonstrate that the path-based resemblance method acquire lesser correlations than 0.36 and 0.66 in case of physicians and coders, correspondingly. It points out that the accurateness of path-based measures is inadequate. Moreover these measures make use of the least amount of path length without taking numerous inheritances, which grounds on much helpful semantic confirmation to be disregarded. In case of the IC-based measures, they progress the outcomes of mainly measures in accordance with path length normally with the maximum value 0.78 for coders and 0.60 for physicians. With regard to the context vector determine, there are possibilities for four scenarios with varying corpus size and corpus assortment. Based on the results, the finest correlations are 0.84 in case of the physicians and 0.78 in case of the coders beneath the circumstance of 1 million notes linking only the investigative segment. Specifically, the correlation rate together with the context vector determination is beyond proposed only when 1 million observations are utilized to produce the vectors. In addition, the data corpus employed to generate vectors was built by means of physicians of the similar formalizing and interpreting information. As a result, the context vector measure sturdily based on the quantity and quality of the background corpus. This measure can exhibit improved performance only in the specific situation and domain.

The anticipated determiatione obtains advanced correlation standards than the entire IC-based measures exposed in Table 3. Based on Table III, it is to be noted that the correlation standards for coders are regularly greater for physicians apart from the context vector evaluation which clarifies that medical coders' standards with added pretraining are added consistent than physicians' standards. Thus, several similarity measures compare similarity in opposition to identical scores of medical coders to attain much enhanced correlations. In case of Dataset 2, find 34 of 36 idea pairs in WordNet+Mesh. The proposed measure is evaluated against the other structure-based similarity measures in accordance with human scores. The correlation of Fengqin Yang et al. measure is found to be 0.774 which is beyond other measures. Nevertheless, the correlation value attained through anticipated measure is 0.78. The qualified outcome (0.78 versus 0.774) proves that the anticipated measure is outperform other consideration shown in Table III.

4.11. Performance evaluation

Performance evaluation of the proposed approach is conducted based on classification context scenario. Precision, Recall, Accuracy and F1 - Score plays an important role in this proposed performance. The performance of proposed WordNet+Mesh simailarity measure is compared to existing higher similarity method of Al Fengqin Yang et al.

Precision measure is considered pedestal on the formula

$$\text{Precision} = \frac{tp}{tp + fp} \tag{22}$$

Recall is calculated based on the formula

$$\text{Recall} = \frac{tp}{tp + fp} \tag{23}$$

Accuracy is calculated based on the formula

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{24}$$

Where tp – True Positive (Correct result)
tn – True Negative (Correct absence of result)
fp – False Positive (Unexpected Result)
fn – False Negative (Missing result)

F-Measure is calculated based on the formula

$$F = 2.\frac{precision.recall}{precision + recall} \tag{25}$$

The simulation results for the evaluation of the proposed approach against various performancemeasures like Precision, Recall, Accuracy and F-Measure.

4.12. Accuracy

The proposed similarity measure using Wordnet and Mesh produced better accuracy rate shown in Fig. 2 which is much greater accuracy results than existing similarity method such as Fengqin Yang et al. based similarity measure. When the number of concepts increases the accuracy of the result is increases. This approach produces high accuracy rate when compared to existing system.

The proposed similarity measure using Wordnet and Mesh produced better precision and recall rate shown in Fig. 3 and fig. 4. The results are much greater precision and recall results compared than existing similarity method such as Fengqin Yang et al. based similarity measure.

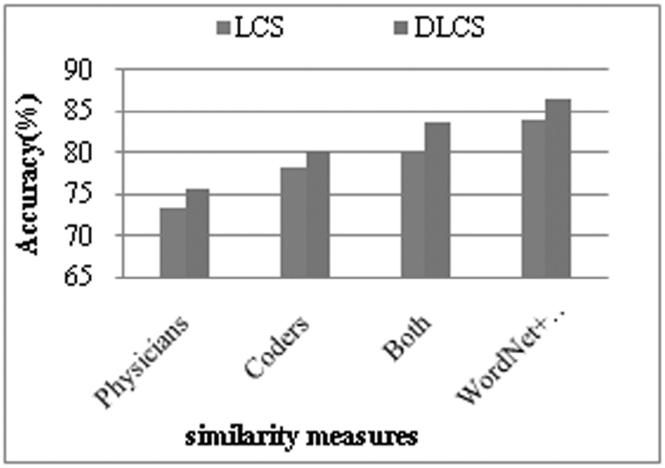

Figure 2: Accuracy Comparison of Similarity Measures

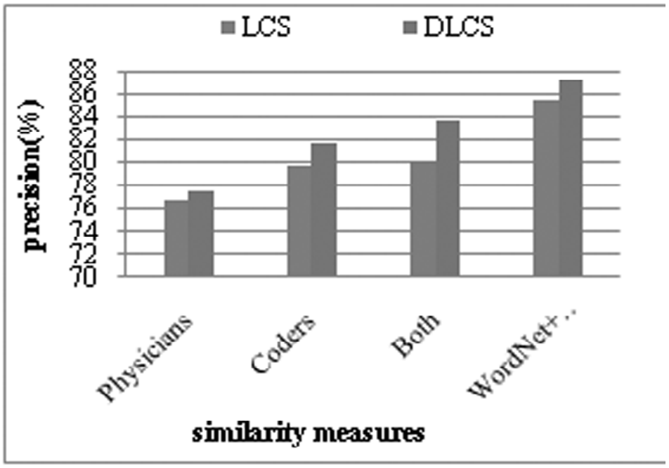

Figure 3: Precision Comparison of Similarity Measures

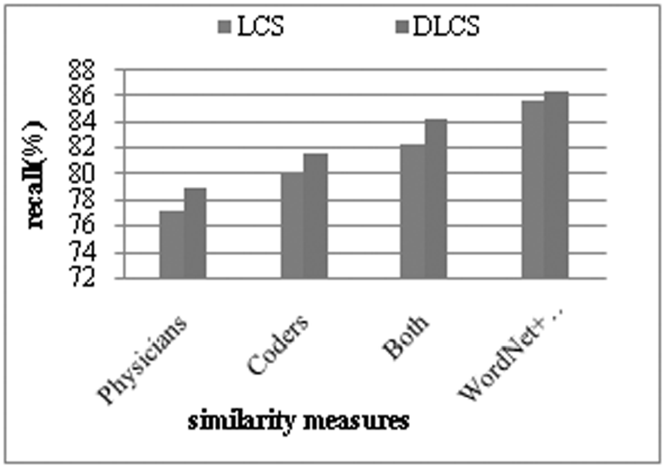

Figure 4: Recall Comparison of Similarity Measures

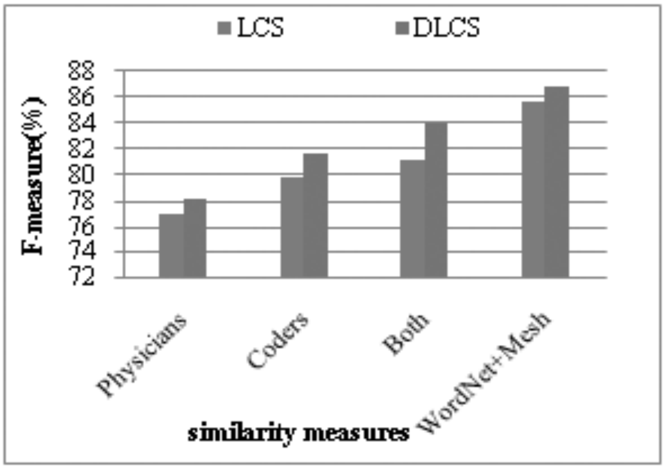

Figure 5: F-Measure Comparison of Similarity Measures

F1- measure is defined as the harmonic mean of accuracy and recall. It signifies that the identical measure performs well with admiration to both precision (P) and recall (R). The proposed similarity measure using Wordnet and Mesh produced high F1-measure shown in Fig. 5 which is higher than the existing similarity method such as Fengqin Yang et al. based similarity measure. When the number of concepts increases the F1-measure of the result is increases. This approach produces high F1-measure rate when compared to existing system.

5. CONCLUSION AND FUTURE WORK

In this work, a new approach to facilitate the resemblance measurement transversely manifold ontologies is presented. Since most ontology-grounded methods spotlight the identical evaluation on the LCS of the compared terms, proposed method attempts to determine a LCS between numerous ontologies that precisely signifies the commonalities among terminology. Proposed approach is based on semantic similarity in the background ontologies. The evaluation, based on well-known standards of biomedical terminology and widely used ontologies, has demonstrated an enlargement in the identity precision when the LCS is measured by proposed methods, in evaluation with associated works. As a result, the accuracies obtained in the multi-ontology scenario (WordNet and MeSH). Hence, proposed process would be helpful when dealt with multi-ontology similarity situations (concepts belonging to distinct ontologies), although mono-ontology similarity would be chosen by its precision, simplicity and efficiency when both concepts appears same in ontology. In the future, to assess other ontology-grounded resemblance measures in the multi-ontology situation by applying proposed method. And also expand the evaluation of proposed method to other domains and ontologies.

REFERENCES

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, “A study on similarity and relatedness using distributional and WordNet-based approaches”, The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of Human Language Technologies Boulder, Colorado, pp. 19–27. 2009.

[2] D. Sánchez, and M. Batet, “Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective”, Journal of Biomedical Information, Vol. 44. pp. 749–759, 2011.

[3] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, and G. Melton, “Semantic similarity and relatedness between clinical terms: an experimental study”, AMIA Annu Symp Proc, pp. 572–576, 2010.

[4] E.G.M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, “X-Similarity:Computing Semantic Similarity between Concepts from Different Ontologies”, Journal of Digital Information Management, Vol. 4. pp. 233-237, 2006.

[5] P. Jaccard, “Distribution of the alpine flora in the dranse’s basin and some neighbouring regions”, Bulletin de la Soc. Vaudoise Sci. Nat, Vol. 37. pp. 241-272, 1901.

[6] Y. Liu, B.T. McInnes, T. Pedersen, G. Melton-Meaux, and S. Pakhomov, “Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora”, Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Miami, Florida, USA, ACM; UMLS and WordNet;, pp. 363–372, 2012.

[7] S.T. Wu, H. Liu, D. Li, C. Tao, M.A. Musen, C.G. Chute, and N.H. Shah, “Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis”, Journal of American Medical Information Association, Vol. 19. Pp. 149–156, 2012.

[8] MEDLINE Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[9] E. Agirre, M. Cuadros, G. Rigau and A. Soroa, “Exploring Knowledge Bases for Similarity,” In LREC, 2010.

[10] R. Rada, H. Mili, E. Bicknell and M. Blettner, “Development and application of a metric on semantic nets”, Man and Cybernetics, IEEE Transactions on Systems, Vol. 19. pp. 17–30, 1989.

[11] A. Budanitsky, and G. Hirst, “Evaluating Word Net-based Measures of Lexical Semantic Relatedness”, Computational Linguistics, Vol. 32. pp. 13–47, 2006.

[12] V.N. Garla and C. Brandt, “Semantic similarity in the biomedical domain: an evaluation across knowledge sources”, BMC bioinformatics, Vol. 13. No. 1. pp. 261. 2012.

[13] T. Hughes, and D. Ramage, “Lexical Semantic Relatedness with Random Graph Walks”, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) Prague, Czech Republic: Association for Computational Linguistics, pp. 581–589, 2007.

[14] S. Patwardhan, “Using Word Net-based context vectors to estimate the semantic relatedness of concepts”, Proceedings of the EACL, pp. 1–8, 2006.

[15] K. Saruladha, G. Aghila and A. Bhuvaneswary, “Computation of Semantic Similarity among Cross Ontological Concepts for Biomedical Domain”, Journal of Computing, Vol. 2. pp. 111-118, 2010.

[16] A. Tversky, “Features of Similarity”, Psycological Review, Vol. 84. pp. 327-352, 1977.

[17] M.A. Rodríguez, and M.J. Egenhofer, “Determining semantic similarity among entity classes from different ontologies”, IEEE Transactions on Knowledge and Data Engineering, Vol. 15. pp. 442–456, 2003.

[18] H. Al-Mubaid and A. Nguyen, “Measuring Semantic Similarity between Biomedical Concepts within Multiple Ontologies”, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 39. pp. 389-398, 2009.

[19] T. Pedersen, S.V. Pakhomov, S. Patwardhan and C.G. Chute, “Measures of semantic similarity and relatedness in the biomedical domain,” Journal of biomedical informatics, Vol. 40. No. 3. pp. 288-299, 2007.

[20] C. Leacock and M. Chodorow, “Combining local context with WordNet similarity for word sense identification”, WordNet: A Lexical Reference System and its Application. 1998.

[21] Z. Wu and M. Palmer, “Verbs semantics and lexical selection”, Association for Computational Linguistics, Proceedings of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico, pp. 133–138, 1994.

[22] P. Resnik, “Using Information Content to Evalutate Semantic Similarity in a Taxonomy”, In C.S. Mellish (Ed.), 14th International Joint Conference on Artificial Intelligence, IJCAI Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., Vol. 1. pp. 448-453, 1995.

[23] D. Lin, “An Information-Theoretic Definition of Similarity”, Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., pp. 296–304, 1998.

[24] J.J. Jiang and D.W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”, In International Conference on Research in Computational Linguistics, ROCLING X, Taiwan, pp. 19-33, 1997.

[25] A. Blank, “Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology”, In R.Eckardt, K. von Heusinger & C. Schwarze (Eds.), Words and Concepts in Time: towards Diachronic Cognitive Onomasiology Berlin, Germany: Mouton de Gruyter, pp. 37-66, 2003.

[26] N. Seco, T. Veale and J. Hayes, “An Intrinsic Information Content Metric for Semantic Similarity in WordNet”, In R. López de Mántaras & L. Saitta (Eds.), 16th Eureopean Conference on Artificial Intelligence, ECAI 2004, including Prestigious Applicants of Intelligent Systems, PAIS Valencia, Spain: IOS Press, pp. 1089-1090, 2004.

[27] G. Pirró, and N. Seco, “Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content”, In R. Meersman & Z. Tari (Eds.), OTM 2008 Confederated International Conferences CoopIS, DOA, GADA, IS, and ODBASE Monterrey, Mexico: Springer Berlin / Heidelberg, Vol. 5332. pp. 1271-1288, 2008.

[28] Z. Zhou, Y. Wang and J. Gu, “A New Model of Information Content for Semantic Similarity in WordNet”, In S.S. Yau, C. Lee and Y.C. Chung (Eds.), Second International Conference on Future Generation Communication and Networking Symposia, FGCNS, Sanya, Hainan Island, China: IEEE Computer Society, pp. 85-89. 2008.

[29] A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis and E.E. Milios, “Information Retrieval by Semantic Similarity”, International Journal on Semantic Web and Information Systems, Vol. 2. pp. 55-73, 2006.

[30] H.P. Young, and A. Levenglick, “A Consistent Extension of Condorcet’s Election Principle”, SIAM Journal on Applied Mathematics, Vol. 35. No. 2, pp. 285–300, 1978.