International Journal of Research in Advent Technology, Vol.2, No.3, March 2014
E-ISSN: 2321-9637

A Survey on Semantic Similarity Measure

S. Anitha Elavarasi[1], Dr. J. Akilandeswari[2], K. Menaga[3].
Assistant Professor[1],Professor & Head[2],PG Scholar[3].
Department of Computer Science and Engineering[1,3],
Department of Information Technology[2].
Sona College of Technology[1,2,3].
anishaer@gmail.com[1],akilandeswari@sontech.ac.in[2],menagak91@gmail.com[3].

Abstract-Measuring semantic similarity between concepts is an important problem in web mining and text mining which needs semantic content matching. Semantic similarity has attracted great concern for a long time in artificial intelligence, psychology and cognitive science. Many measures have been proposed. The paper contains a review of the state of art measures including path based measures, information based measures, feature based measures and hybrid measures. The features, performance, advantages, disadvantages and related issues of different measures are discussed. This paper makes a review of semantic similarity measures with various approaches.

Index Term- Semantic Similarity; Path based measure; depth relative measure; information content based measure; hybrid measure; feature based measure.

1. INTRODUCTION

Similarity plays a central role in information management, especially in the context of environment like the semantic web where data may originate from different sources and has to be combined and integrated in a flexible way.

Semantic similarity is a metric over a set of documents, based on the likeliness of their meaning, which refers to similarity between two concepts in a taxonomy or ontology and it is achieved through ontology or taxonomies to define a distance between words or using statistical means. Similarity among concepts is a quantitative measure of information, computed based on the properties of the concepts and their relationships. With the advent of Semantic Web, the semantic similarity measures are becoming important components in Information Extraction (IE), Information Retrieval (IR) and other intelligent knowledge based systems.

Potential application for these measures includes search, knowledge discovery in database and data mining or decision support systems that utilize ontology. Semantic similarity refers to the closeness of two concepts within a given ontology or taxonomy.

2. CLASSIFICATION OF SEMANTIC SIMILARITY MEASURE

The classification of semantic similarity includes similarity measure for single ontology and multiple ontologies. The classification is based on how the semantic similarity measure is quantified. The quantification is either based on the ontological structure or based on the information content.

2.1. Semantic similarity based on single ontology [1]

As in Fig.1 similarity between concepts belonging to single ontology have different approaches such as

- Path length based measure
- Depth relative measure
- Information content based measure
- Hybrid measure
- Feature based measure

Based on the quantifying similarity approaches are used for the semantic measure. Also in some cases both path length based and information content based approaches have been used.

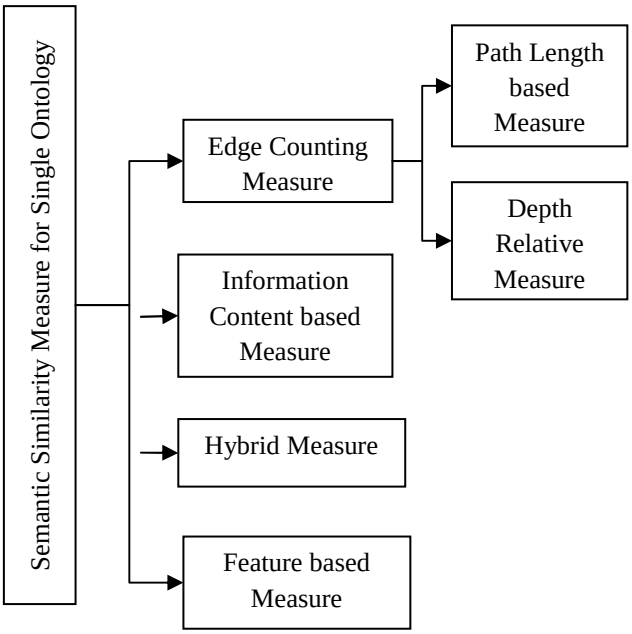

Fig 1. Classification of semantic similarity based on single ontology

2.1.1. Path length based measure

The similarity measurement between concepts is based on the path distance separating the concepts. In this measure the quantification of similarity is based on the ontology or taxonomy structure. In these ontology or taxonomical structure, most predominant relations are connected through is-a type relation. Thus similarity is computed by shortest path and the degree of similarity is determined based on path length. The various path length based similarity measures are,

- Rada Similarity Measure
- Bulskov Measure
- Al-Demonstils Measure

2.1.2. Depth relative measure

The depth relative measure is a shortest path approaches, but it considers the depth of the edges connecting the two concepts in the overall structure of the ontology. It calculates the depth from root to the target concept. The various depth relative measures are

- Wu and Palmer measure
- Sussna measure
- Leacock and Chodorow Similarity measure

2.1.3. Information content based measure

Both the path length and depth relative measure use the knowledge solely captured by ontology to computationally determine the similarity between concepts. In this section the knowledge revealed by corpus is used to augment the information already present in the ontologies or taxonomy. Thus information content based approach is also referred as the corpus based approach or information theoretic based approach. The various information content based measures are

- Resnik Measure
- Lin Measure
- Jiang and Conrath measure

2.1.4. Hybrid measure

Hybrid combines knowledge derived from various sources of information. The major advantage of these approaches is if the knowledge of an information source is inadequate then it may be derived from the alternate information sources. The various hybrid similarity measures are

- Li measure
- Zuber and Faltings measure

2.1.5. Feature based Measure

Feature based approach takes into account the features that are familiar to both concepts and also the specific differentiating features of each concept. Thus the various feature based measures are

- Tversky measure
- Pirro Measure

2.2. Semantic similarity based on multiple ontology [2]

The semantic similarity measures discussed earlier are meant single ontology. Now in recent days with the growing information sources on the web, there is a need for developing measures which will compute similarity among concepts belonging to different ontologies.

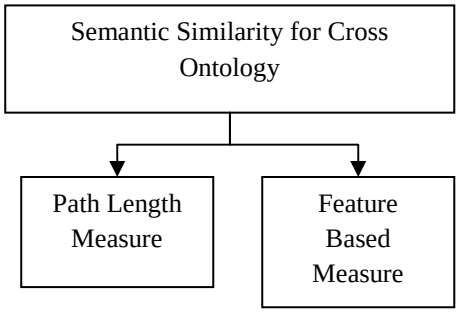

Fig 2. Classification of semantic similarity based on multiple ontology

As in Fig.2 similarity measures between concepts in multiple ontology is classified as

- Path length based measure
- Feature based measure

Cross ontology measures compare the words from different ontology. The cross ontology often requires hybrid or feature based measure, because the structure and information content between diverse ontologies cannot be compared directly. Cross ontology measure includes the following steps:

- Extracting set of relevant definitions, features, synsets and neighbors from both ontology
  - Word matching
  - Feature matching
  - Semantic neighborhood matching
- Finding cross ontology measure for the input query

3. LITERATURE SURVEY

3.1. An ontology based semantic similarity measure for biomedical data- application to radiology reports [3]

A notion of semantic similarity is used in this paper to overcome the limitation of direct concept matching. Consider an example where the concept glioma is extracted from first document and the concept neoplasm is extracted from second document. A direct comparison may result in no relation between two concepts. But glioma is a specialized form of neoplasm in SNOMED ontology. In this case semantic similarity is taken into account. **Edge counting approach** is used for semantic similarity measure. A semantic vector consisting of all parent-child (is-a) relationships are exploited. Semantic similarity is defined as

$$\text{Similarity}(c1, c2) = \frac{1}{d} \qquad \text{Eq. (1)}$$

Where d is the number of nodes in the shortest path between concept nodes c1 and c2. Eq (1) used to set the minimum distance between the ancestor and the seed concept in the document. Using simple weight measure documents are assigned with weight after finding shortest distance.

$$\text{weight} = \frac{1}{1 + \text{shortest distance}} \qquad \text{Eq. (2)}$$

After measuring the weight, semantic similarity measure is defined by cosine measure

$$\text{Similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sum_{i=1}^{n} A_i^2 * \sum_{i=1}^{n} B_i^2} \qquad \text{Eq. (3)}$$

Evaluation in this paper shows that semantic based approach increases the similarity of documents describing the same anatomies.

3.2. A hybrid knowledge based and data driven approach to identifying semantically similar concepts [4]

Quantifying the similarity among concepts is a difficult task, however such similarity is context dependent. A comprehensive method is proposed which computes a similarity score for a concept pair by combining data driven and ontology driven knowledge. Evaluation is done on concepts from SNOMED-CT and on a corpus of clinical notes of patients with chronic kidney disease. By combining information from usage patterns in clinical notes and from ontological structure, concepts that are simply related which are semantically similar are pruned out. Three different metrics are applied when combining data driven and ontology driven approaches. They are note based similarity, ontological similarity and definitional similarity.

Note based similarity measure is computes for all concept pairs which takes Unified Medical Language System(UMLS) concept as input and similarity score defined by cosine measure.

Ontological similarity describes a nivel method for semantic similarity using ontologically defined relationships. SNOMED-CT is taken as a flat terminology and concentrated on edge types rather than the hierarchy method. To assign weights ontological weights for each individual pair wise path following formulae was used

$$\text{Sim}_O = \sum_{e=1}^{E} \frac{\text{Weight}_e}{|E|} - \propto (|E| - 1) \qquad \text{Eq. (4)}$$

E= $\{e_1, e_2, \ldots, e_n\}$ where e_i=edge in path, weight_e=assigned weight for edge e, α=.2

Definitional metric is a measure of lexical commonality between two concepts- a metric widely used in word sense disambiguation.

$$\text{Sim}_D = |(C1 + C2)| - \frac{|C1 + C2|}{\text{Min}(|C1|, |C2|)} \qquad \text{Eq. (5)}$$

The evaluation of all the three methods was calculated on the 794 pairs. The definitional and ontological similarity measures were used and evaluated as secondary metrics. The first evaluation was performed on the note based method alone to assess its individual contribution. Next the average of the note based and definitional method as well as the average of note based and ontological methods were calculated. Finally the average of all the three method was computed to find the threshold on note based similarity.

3.3. Semantic similarity estimation in the biomedical domain: An ontology-based information theoretic perspective [5]

Semantic similarity estimation has been the focus of much research, which has led to the definition of heterogeneous measures using different theoretical principles and knowledge resources in a variety of contexts and application domains. In this paper several of these measures are discussed in addition to other similarity coefficients that may be useful in determining the similarity of sets of terms. In order to make them easier to interpret and improve their applicability and accuracy, a framework is proposed in information theory that allows the measures to be uniformly redefined. SNOMED-CT concepts are used through ontology

IC of a concept is computed by

$$\mathrm{IC}(c) = -\log\left(\frac{\frac{|\mathrm{leaves}(c)|}{|\mathrm{subsumers}(c)} + 1}{\max\_\mathrm{leaves} + 1}\right) \qquad \text{Eq. (6)}$$

With the IC based semantic measure, new ontology based edge counting measures in terms of IC are redefined. To find the distance between concepts in ontology redefined Rada measure is proposed

$$\mathrm{dis}_{\mathrm{rad}}(c_1, c_2) = \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 \times \mathrm{IC}\big(\mathrm{LCS}(c_1, c_2)\big) \qquad \text{Eq. (7)}$$

Also Wu & Palmer measure is redefined as

$$\mathrm{sim}_{w\&p}(c_1, c_2) = \frac{2 \times \mathrm{IC}(\mathrm{LCS}(c_1, c_2)}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)} \qquad \text{Eq. (8)}$$

The proposed framework is based on approximating concept semantics in terms of Information Content (IC). IC is computed in a scalable and efficient manner from the taxonomical knowledge. Correlation values obtained for various semantic measures are analyzed. From the analysis IC-based measures based on intrinsic IC calculation obtain higher accuracy rates than those based on corpora (0.68-0.71 vs. 0.45-0.6 for physician. The evaluation of the proposed measure shows that new measures provide a high degree of accuracy.

3.4. *An ontology based measure to compute semantic similarity in biomedicine [6]*

Several approaches for assessing word similarity by exploiting different knowledge sources have been proposed. Some of those measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data or from medical ontologies. In this paper these approaches are introduced and analyzed in order to determine their advantages and limitations with respect to the considered knowledge bases. Later a new measure based on the exploitation of the taxonomical structure of a biomedical ontology is proposed. SNOMED-CT is used as the input ontology. The similarity between two concepts is defined as

$$\mathrm{sim}(c_1, c_2) = -\log_2 \frac{|T(c_1) \cup T(c_2)| - |T(c_1) \cap T(c_2)|}{|T(c_1) \cup T(c_2)|} \qquad \text{Eq. (9)}$$

Where $T(c_i)=\{c_j \epsilon C | c_j \text{ is superconcept of } c_i\} \cup \{c_i\}$

The proposed similarity measure achieved a level of accuracy similar to corpus based approaches but retaining the low computational complexity and lack of constraints of path based measures. Correlation values obtained for each measure are discussed and the proposed measure attains 0.73 correlations which is higher than the other measures.

3.5. An weighted ontology based semantic similarity algorithm for web service [7]

This paper proposed a weighted ontology based semantic similarity algorithm for web service to support a more automated and various service discovery and rank process, by distinguishing among the potentially useful and the likely irrelevant services and also by ordering the potentially useful ones according to their relevance to the requester's query. Web service matching queries are represented as vectors. Web service ontology is used as input and a part of university ontology is taken to measure web based semantic measure. Similarity distance between a provider service vector p and query service vector q can be computed as the vector inner product:

$$\mathrm{sim}(p, q) = d.q = \sum_{i=1}^{t} w_{id} * w_{iq} \qquad \text{Eq. (10)}$$

In the above equation d is the document vector. w_{id} and w_{iq} are the semantic similarity of interface parameter i, which can be represented as a concept or a term i.e., the similarity of web service can be addressed through calculating the vector inner

product of concept vector. A higher similarity score indicates a closer similarity between the query and retrieved web services. Concept vector similarity is computed by information theory based concept semantic similarity algorithm.

The ontological structure defines the function with a given concept, returns the set of more generic concepts directly linked to c. The set of paths between two concepts c_a and c_b can be defined as

$$\mathrm{Path}(c_a, c_b) = \Big\{(c_1, \dots, c_n) | (c_a = c_1) \wedge (c_b = c_n) \wedge \Big(\forall i: (1 \le i < n) \wedge \big((c_i \in \mathrm{function}(c_i + 1))\big)\Big)\Big\} \quad \text{Eq. (11)}$$

A concept a is an ancestor of a concept c when there is at least one path from a to c:

$$\mathrm{Ancestor}(c) = \{a | \mathrm{Path}(a, c) \neq \phi\} \quad \text{Eq. (12)}$$

The frequency of concept c, Freq(c) can be defined as the number of times that c and all its descendents occur:

$$Freq(c) = \sum \{occur(c_i) | c \in Ancestor(c_i)\} \quad Eq. (13)$$

An estimate for the likelihood concept probabilities of observing an instance of a concept c is

$$\mathrm{Prob}(c) = \frac{\mathrm{Freq}(c)}{N} \quad \text{Eq. (14)}$$

Where N is the total number of all concepts in the corpus. The information content of a concept c can be defined as

$$\mathrm{IC}(c) = -\log(\mathrm{Prob}(c)) \quad \text{Eq. (15)}$$

Based on the similarity probability IC(c), the semantic similarity distance and similarity algorithms are described as

(1) Semantic similarity distance: share(c_1,c_2) and wsim(w_1,w_2)

Semantic similarity measures assume that the similarity between two concepts is related to the extent to which they share information. Shared information between two concepts share(c_1,c_2) can be defined as

$$Share(c_1, c_2) = \max\{IC(a) | a \in sub(c_1, c_2)\} \quad Eq. (16)$$

Where sub(c_1,c_2) is the concepts that subsume both c_1 and c_2. Rather to measure word similarity wsim(w_1,w_2) can be defined as

$$\mathrm{wsim}(w_1, w_2) = \max_{c_1 c_2}[\mathrm{Share}\,(c_1, c_2)] \quad \text{Eq. (17)}$$

Where c_1 ranges over s(w_1) and c_2 ranges over s(w_2)

(2) Share(c_1,c_2) and Wsim(w_1,w_2) based semantic similarity algorithm

Wu & Palmer, Resnik, Jiang and Conrath, Lin, Li and Bandar proposed their semantic similarity algorithms based on the share information and word similarity measure.

Wu and Plamer defined their similarity as

$$\mathrm{Sim}_{wp}(c_1, c_2) = 2 * \frac{N_3}{(N_1 + N_2 + 2 * N_3)} \quad \text{Eq. (18)}$$

Where N_1 and N_2 are the number of is-a links from c_1 and c_2 to their superclass C; N_3 is the number of is-a links from C to the root taxonomy.

Resnik defined their similarity measure as

$$\mathrm{Sim}_{\mathrm{Resnik}} = \mathrm{Share}(c_1, c_2) \quad \text{Eq. (19)}$$

Jiang and Cornath defined their similarity measure as

$$\mathrm{Dist}_{\mathrm{JC}}(c_1, c_2) = \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2 * \mathrm{Share}(c_1, c_2) \quad \text{Eq. (20)}$$

The above equation measures the distance and similarity algorithm is

$$\mathrm{Sim}_{\mathrm{JC}}(c_1, c_2) = \frac{1}{(\mathrm{dist}_{\mathrm{JC}}(c_1, c_2) + 1)} \quad \text{Eq. (21)}$$

Lin defined their similarity measure as

$$\mathrm{Sim}_{\mathrm{Lin}}(c_1, c_2) = \frac{2 * \mathrm{Share}(c_1, c_2)}{(\mathrm{IC}(c_1) + \mathrm{IC}(c_2))} \quad \text{Eq. (22)}$$

Also Resnik proposed a weighted similarity word meaure as

$$WSim_{\propto}(w_1, w_2) = \sum \propto (c_i)[-\log p(c_i)] \quad Eq. (23)$$

The proposed concept is used to support a more automated and reality service discovery process, by distinguishing among the potentially useful and the likely irrelevant services to the developer query.

3.6. An approach for measuring semantic similarity measure between words using multiple information sources [8]

Semantic similarity measure by a number of information sources are described in this paper which consists of structural semantic information from a lexical taxonomy and information content from a corpus. A new measure is proposed to measure semantic similarity which combines information nonlinearly. Experimental evaluation against a benchmark dataset is described which demonstrates that the proposed similarity measure performs well than the existing measure. Thus the proposed similarity measure is

$$S(w_1, w_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \qquad \text{Eq. (24)}$$

Where *l* is the shortest path length between w_1 *and* w_2, *h* is the depth of subsume in the hierarchy semantic nets and *d* is the local semantic density of w_1 *and* w_2. Based on the benchmark dataset optimal parameter for the proposed measure is $\propto$=0.2 and β=0.6. The correlation value of the proposed measure is 0.8914 against Rubenstein-Goodenough's human ratings which has been 0.8484.

3.7. Measuring semantic similarity between biomedical concepts within multiple ontologies [9]

Measuring semantic similarity between biomedical concepts using multiple ontologies is discussed in this paper. MeSH and wordnet ontologies are used as input. Thus proposed measure is based on three features

(1) Cross modified path length between two concepts

(2) A new feature of common specificity of concepts in the ontology.

(3) Local granularity of ontology clusters.

Rules and Assumptions for cross ontology approach are

- The semantic similarity scale system reflects the degree of similarity of pairs of concepts
- Semantic similarity must obey local ontology's similarity rules

Proposed cross ontology semantic similarity approach includes

- Single ontology similarity

Granularity is not considered within single ontology and so length and depth features are used to get semantic distance between two concepts as follows:

$$\text{SemDist}(c_1, c_2) = \log\left((\text{Path} - 1)^{\propto} \times (\text{CSpec})^{\beta} + k\right) \qquad \text{Eq. (25)}$$

$$\text{CSpec}(c_1, c_2) = D - \text{Depth}\left(\text{LCS}(c_1, c_2)\right) \qquad \text{Eq. (26)}$$

Where $\propto$>0 and β>0 are contribution factors of two features (Path and CSpec), k is a constant, Path is the shortest path length between two concept nodes

- Cross ontology semantic similarity

In cross ontology to measure semantic similarity between two concepts (c_1,c_2), there are four cases:

Case 1: Similarity within primary ontology

Using Eqn (25) similarity within single ontology is calculated

Case 2: Cross ontology similarity (Primary & Secondary)

The common specificity feature: Two concepts belonging to two different ontologies are measured using

$$\text{LCS}_n(c_1, c_2) = \text{LCS}(c_1, \text{bridge}_n) \qquad \text{Eq. (27)}$$

The cross-ontology path length feature: The path length between two concept nodes is calculated by adding up two path lengths from each of them to bridge node. Path length between two concepts are defined as

$$\text{Path}(c_1, c_2) = d_1 + \text{PathRate} \times d_2 - 1 \qquad \text{Eq. (28)}$$

Where d_1 and d_2 are the shortest path length between the concept and bridge.

$$PathRate = \frac{2D_1 - 1}{2D_2 - 1} \qquad Eq. (29)$$

Where D_1 and D_2 are the depth of primary and secondary ontologies. Finally the semantic distance between two concept nodes is given as

$$\text{CSpec}_i(c_1, c_2) = D_1 - \text{Depth}\left(\text{LCS}(c_1, \text{Bridge}_i)\right) \qquad \text{Eq. (30)}$$

$$SemDist_i(c_1, c_2) = \log\left((Path_i - 1)^{\propto} \times (CSpec_i)^{\beta} + k\right) \qquad Eq. (31)$$

$$SemDist(c_1, c_2) = MIN_q\{SemDist_q(c_1, c_2)\} \qquad Eq. (32)$$

Case 3: Similarity within single secondary ontology

This case is used when both concepts are in a single secondary ontology. Then semantic distance in this case must be converted into primary ontology as follows:

$$Path(c_1, c_2) = Path(c_1, c_2)_{sec} \times PathRate \qquad Eq.(33)$$

$$CSpec(c_1, c_2) = CSpec(c_1, c_2)_{sec} \times CSpecRate \qquad Eq.(34)$$

$$CSpecRate = \frac{D_1 - 1}{D_2 - 1} \qquad Eq.(35)$$

$$SemDist(c_1, c_2) = \log(Path_i - 1)^{\propto} \times \left(CSpec_i\right)^{\beta} + k) \qquad Eq.(36)$$

Case 4: Similarity within multiple secondary ontologies

In this case, one of the two secondary ontologies act temporarily as primary to calculate the semantic features using case 2 then the semantic similarity is computes using case 3 to scale the feature of primary ontology similarity again.

In single ontology the evaluation is performed with the four measures. Those measures are applied to MeSH and SNOMED-CT. Correlation obtained for MeSH is 0.841 and correlation for SNOMED-CT is 0.726.

In cross ontology the evaluation is made for WordNet and MeSH which result with the correlation of 0.809 and the correlation of WordNet and SNOMED-CT is 0.745.

3.8. Assessment of Semantic Similarity of concepts defined in ontology [11]

This paper proposes a method to determine similarity between concepts defined in ontology. Thus proposed method focuses on the relation between concepts and their semantic relation instead of using ontology definition. Four features are proposed with this system

- Semantic-oriented
- Context-aware
- Granularity-sensible
- Dynamic/adaptive

The proposed method of this paper to determine similarity between two concepts when all features of the concepts are considered is described with two concepts c_i and c_j. in such case concepts consists of two components.

- First component, $sim_1(c_i,c_j)$ represents similarity based on the feature that are shared between two concepts.
- Second component, $sim_2(c_i,c_j)$ is used to determine contributions to the overall similarity from feaetures that are different for both concepts.

To present a formula for assessment of similarity, some quantities are defined. The first component is defined as

$$\mathrm{sim}_1(\mathrm{c_i}, \mathrm{c_j}) = \left|\mathrm{R}(\mathrm{c_i}, \mathrm{c_j})\right| + \sum_{\mathrm{c_k} \epsilon \mathrm{N(ij)}} \left[\max_{\substack{\mathrm{r_i} \epsilon \mathrm{R}(\mathrm{c_i}, \mathrm{c_k}) \\ \mathrm{r_j} \epsilon \mathrm{R}(\mathrm{c_i}, \mathrm{c_k})}} (\mathrm{relationSim}(\mathrm{r_i}, \mathrm{r_j})) \right] \quad \text{Eq. (37)}$$

where |.| represents cardinality of a set. $R(c_i,c_k)$ and $R(c_j,c_k)$ represents set of relation. N(i) denotes set of concepts c_i is connected to in a gien ontology. Also $N(ij)=N_{common}(c_i,c_j)$ is a set of concepts that both c_i and c_j are connected to.

Thus second component is defined as

$$sim_2(c_i, c_j) = \sum_{c_z \epsilon N^0(i)} \left[\max_{\substack{c_y \epsilon N^0(j) \\ r_i \epsilon R(c_i, c_z) \\ r_j \epsilon R(c_j, c_y)}} \begin{pmatrix} \{relationSim(r_i, r_j)\} \\ \oplus \\ \max_{w \epsilon Y} \{sim(c_z, c_w)\} \end{pmatrix} \right] \quad Eq.(38)$$

Where $N_i^0=N(i)-N(ij)=N(j)-N(ij)$ represents unique features of the both concepts.

Finally the similarity between concepts c_i and c_j is defined as

$$\mathrm{sim}(\mathrm{c_i}, \mathrm{c_j}) = \frac{\mathrm{sim}_1(\mathrm{c_i}, \mathrm{c_j}) + \mathrm{sim}_2(\mathrm{c_i}, \mathrm{c_j})}{|\mathrm{N(i)}|} \qquad (39)$$

Using Eqn.39 similarity is obtained. When the features defining each concept are different then obtained similarity is asymmetric.

3.9. Ontology –based semantic similarity: A new feature based approach [12]

In this paper ontology based approaches such as edge counting, feature based approach and measures based on information content are classified and a new ontology-based measure relying on the exploitation of taxonomical features is proposed. In order to semantic distance between concepts, amount of dissimilarity with taxonomical

feature are defined with the sample ontology according to their feature..

The set of taxonomical features describing the concept a is defined in terms of relation ≤ as:

$$\phi(a) = \{c \in C | a \leq c\} \qquad \text{Eq. (40)}$$

where C is the set of concepts of an ontology. A is a term in the taxonomy.

The normalized dissimilarity between a and b according to the taxonomical feature is calculated as:

$$\mathrm{dis}_{\mathrm{norm}}(a,b) = \log_2\left(1 + \frac{|\phi(a)\backslash\phi(b)| + |\phi(b)\backslash\phi(a)|}{|\phi(a)\backslash\phi(b)| + |\phi(b)\backslash\phi(a)| + |\phi(a) \cap \phi(b)|}\right) \qquad \text{Eq.(41)}$$

The generalized dissimilarity measure which is able to deal with polysemic terms is defined as:

$$\mathrm{dis}_{\mathrm{generalized}}(a,b) = \min_{\substack{\forall a' \in A \\ \forall b' \in B}} \mathrm{dis}_{\mathrm{norm}}(a',b') \qquad \text{Eq. (42)}$$

Where A is the set of conceptsfor the term a and equally for the term b.

The evaluation of this measure results in high accuracy. In this measure the set of features is built from the categorization of concepts modeled in ontology. Correlation value for Miller and Charles benchmark is 0.83 and correlation value for Rubenstein and Goodenough benchmark is 0.857.

3.10. Unsupervised Semantic Similarity Computation between terms using web documents[13]

To measure semantic similarity between terms in web documents require metrics such as page counting, ontology, external knowledge, documents to download, web search engine. The proposed algorithm of this paper does not require all these metric instead it requires only context based metric for web documents search. Context based metric requires fixed size of words for feature selection. Thus similarity between words is computed by

$$S^k(w_1, w_2) = \frac{\sum_{i=1}^{N} t_{w_1,i} t_{w_2,i}}{\sqrt{\sum_{i=1}^{N} (t_{w_1,i})^2} \sqrt{\sum_{i=1}^{N} (t_{w_2,i})^2}} \qquad \text{Eq. (43)}$$

$t_{w,i}$ is calculated according to the scheme like binary, term frequency, tf-idf, log tf and so on.

Table 1.Context Feature Weighting Scheme

Scheme	Acronym	$t_{w,i}$ (if $c(v_i) > 0$)
Binary	B	1
Term frequency	TF	$\frac{c(v_i)}{c(w)}$
Add-one TF	TF1	$\frac{c(v_i)+1}{c(w)+\alpha_w}$
Log of TF	LTF	$\frac{\log(c(v_i))}{\log(c(w))}$
Add-one LTF	LTF1	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha_w)}$
TF-inverse document freq.	TFIDF	$\frac{c(v_i)}{c(w)} \log \frac{\lvert D\rvert}{\lvert D\rvert v_i\rvert}$
Log of TFIDF	LTFIDF	$\frac{\log(c(v_i))}{\log(c(w))} \log \frac{\lvert D\rvert}{\lvert D\rvert v_i\rvert}$
Add-one LTFIDF	LTF1IDF	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha_w)} \log \frac{\lvert D\rvert}{\lvert D\rvert v_i\rvert}$

Since w represents word and the feature vector of word is represented as t_w. $c(v_i)$ represents number of occurrence of the term in the document. C(w) represents number of words in the document.

Evaluation is made for Charles-Miller data set and MeSH data set which results in higher correlation with the context feature weighting scheme. Correlation value for Charles-Miller data set using binary scheme is 0.88. Correlation value for MeSH data set using Log of TFIDF is 0.69.

3.11. A review of semantic similarity measure in wordnet [10]

Table 1. Comparison Of Semantic Similarity Approaches

Approach	Principle	Measure	Features	Advantages	Disadvantages
Path Based	Function of path length linking the concepts and the position of the concepts in the taxonomy	Shortest Path	Count of edges between concepts	Simple measure	Two pairs with equal lengths of shortest path will have the same similarity

		Wu & Palmer	Path length to subsume, scaled by subsumer path to root	Simple measure	Two pairs with common lowest common subsume and equal lengths of path will have the same similarity
		L&C	Count of edges between and log smoothing	Simple measure	two pairs with equal lengths of shortest path will have the same similarity
		Li	Non linear function of the shortest path and depth of lowest common subsumer	Simple measure	two pairs with the same lowest common subsumer and equal lengths of shortest path will have the same similarity
IC Based	The more common information two concepts share, the more similar the concepts are	Resnik	IC of lowest common subsume	Simple measure	two pairs with the same lowest common subsumer will have the same similarity
		Lin	IC of lowest common subsumer and the compared concepts	Take the IC of compared concepts into consideration	two pairs with the same summation of IC(c_1) and IC(c_2) will have the same similarity
		Jiang	IC of lowest common subsumer and the compared concepts	Take the IC of compared concepts into consideration	two pairs with the same summation of IC(c_1) and IC(c_2) will have the same similarity
Feature Based	Concepts with more common features and less non-common features are more similar	Tversky	Compare concepts feature	Takes concept feature into consideration	Computational complexity. It can't work well when there is not a complete features set
Hybrid Method	Combine multiple information sources	Zhou	Combines IC and shortest path	Well distinguished different concept pairs	parameter need to be adapted manually.

4. CONCLUSION

This paper describes the basics of semantic similarity measure, classification of single ontology based similarity measure and cross ontology based similarity measure. A brief introduction of various semantic similarity measures are outlined with the survey of various papers. As discussed before, purely ontology based similarity approaches like edge counting measures are advantageous due to their lack of dependency on corpora availability and human pre-processing of data. Also it is possible to increase the accuracy by considering the principles of information theory and properly estimating the IC of concepts.

REFERENCES

[1] Thabet Silmani. Description and evaluation of semantic similarity measures approaches. International Journal of Computer Applications(0975-8887). Volume 80- No.10, October 2013.

[2] Jayasri D and Manimegalai D. Semantic similarity measures on different ontologies: survey and a proposal of cross ontology based similarity measure. International Journal of Science and Research (IJSR), India online ISSN: 2319-7064. Volume 2 Issue 2, February 2013.

[3] Thusitha Mabotuwana *et al.* An ontology-based similarity measure for biomedical data-Application to radiology reports. Journal of Biomedical Informatics; 2013. http://dx.doi.org/10.1016 /j.jbi.2013.06.013

[4] Pivovarov R and Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. Journal of Biomedical Informatics. 2012; 45(3):471–81.

[5] David Sanchez and Montserrat Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics 44 (2011) 749–759. doi:10.1016/j.jbi.2011.03.013

[6] Montserrat Batet, *et al.* An ontology-based measure to compute semantic similarity in biomedicine. Proceedings at Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) Research Group, Department d'Enginyeria Informatics Matemàtiques, Universitat Rovira Virgili, Tarragona, Catalonia, Spain. Journal of Biomedical Informatics 44 (2011): 118–125.

[7] Min Liu, *et al.* An weighted ontology based semantic similarity algorithm for web service. Expert systems with Applications 36 (2009) 12480-12490. Doi: 10.1016/j.eswa.2009.04.034.

[8] Yuhua Li, *et al.* An approach for measuring semantic similarity measure between words using multiple information sources. IEEE transactions on knowledge and data engineering, vol.15, no.4, july/august 2003.

[9] Hisham Al-Mubaid and Hoa 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, no.4, july 2009.

[10] Lingling Meng, *et al.* A review of semantic similarity measure in wordnet. International Journal of Hybrid Information Technology. Vol.6, no.1, January 2013.

[11] Parisa D, *et al.* Assessment of semantic similarity of concepts defined in ontology. Journal of Information Sciences (2013). Doi: 10.1016/j.ins.2013.06.056.

[12] David Sanchez, *et al.* Ontology based semantic similarity: A new feature-based approach. Journal of expert systems with applications 39(2012) 7718-7728.

[13] Elias Losif and Alexandros Potamianos. Unsupervised semantic similarity computation between terms using web documents. IEEE transactions on knowledge engineering, vol.22, no.11, November 2012.