CHAPTER 4

CROSS ONTOLOGY SEMANTIC SIMILARITY (COSS) MEASURES

4.1 INTRODUCTION

Recent investigations in information retrieval and data integration have emphasized the use of ontologies and similarity functions as a mechanism for comparing objects that can be retrieved and integrated across heterogeneous repositories. In this context, ontology is a type of knowledge base that describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. In environments with multiple information systems, independent systems may have their own intended model and their own ontologies. In such systems, the similarity of concepts of one ontology needs to be matched with the concept of another ontology.

The general approach to data integration and concept matching in such environments is to map the local terms of the distinct ontologies onto a single shared ontology. Once this shared ontology is constructed, the semantic similarity approaches applicable for single ontology (which have been discussed earlier in Chapter 2) could be employed to compute similarity among concepts belonging to different ontologies. The construction of shared ontology is feasible and is not impractical. But it is a challenging task to maintain the integrated shared ontology when the local ontologies are updated. Hence an alternate approach of devising computational models which compute semantic similarity among concepts belonging to heterogeneous independent ontologies without creating apriori shared ontology is required.
Few researchers have proposed similarity measures to compute cross ontology concept similarity. A path based approach was proposed by Al-Mubaid and Nguyen and a feature based approach by Rodriguez and Egenhofer. However these similarity values correlate less with human judgments.

Further, the literature indicates that importance must be given to attempt information content based approaches for computing semantic similarity among multiple ontologies and it would be efficient if similarity is computed without constructing the apriori a shared ontology. Therefore in this direction an attempt to propose information content based approaches for measuring cross ontology concept similarity has been made. Hence the alternate six computational measures presented in this chapter computes semantic similarity among concepts belonging to heterogeneous independent ontologies without creating apriori a shared ontology by using the information content which is corpus independent.

In this chapter, the generic models of similarity, the existing approaches for measuring cross ontological similarity are presented. The shortcomings of the existing cross ontology similarity measures are mentioned. In addition to this, this chapter also highlights how these shortcomings have been overcome to suit the requirements for measuring semantic similarity among concepts belonging to multiple ontologies.

Further, this chapter elaborates on the six measures proposed for measuring cross ontology concept similarity. The Tversky psychological model (Tversky 1977) is used as the base model for devising the COSS measure (Cross Ontology Semantic Similarity) which is discussed in Section 4.7. Moreover, the design of variant of COSS measure RDCOSS is detailed in Section 4.8. The Pirró measure used for measuring single ontology concept similarity is suitably modified to measure the cross ontology concept similarity. This modified Pirró measure (RPCOSS) is discussed in Section 4.9.

The single ontology information content based approaches Resnik, Lin and Jiang and Conrath measures discussed in the previous chapter have been adapted
for measuring semantic similarity among concepts belonging to multiple ontologies (Resnik 1995) (Lin 1998) (Jiang and Conrath 1997). These modified measures RRCOSS, RLCOSS and RJCROSS are discussed in Section 4.10.

4.2 GENERIC MODELS OF SIMILARITY

Goldstone and Son report different types of similarity models based on the psychological studies (Goldstone and Son 2005). These models have profound impact in statistics, automatic pattern recognition by machines, data mining and information retrieval tasks. The major five similarity models are Geometric model, feature based model, alignment based model, logic based model and Transformational model. Each of these models has been discussed in the following sections.

4.2.1 Geometric Model

Geometric model (Torgerson 1965) of similarity is one of the most influenced approaches to similarity among the existing approaches. These models are exemplified by non metric multidimensional scaling models (Shepard 1962). The multidimensional scaling model represents similarity between entities in terms of geometric model that consists of a set of points embedded in an ‘n’ dimensional space. The similarity between a pair of objects is taken to be inversely related to the distance between two objects. Richardson worked to determine the dimensions that human subjects used in their judgements to quantify similarity (Richardson 1938). Multidimensional scaling models have also been used in mathematical and computational models of cognitive processes.

4.2.1.1 A layered framework for Similarity based on geometric model

Ehrig et al. have proposed a similarity approach to measure similarity among and within ontologies (Ehrig et al. 2005). A layered framework has been proposed which computes similarity between instances and concepts of different entities of the ontology.
**Definition 1**: Ehrig definition of similarity

A similarity function is defined as a real valued function $\text{Sim}(E_1, E_2) \rightarrow [0, 1]$ on a set $S$ measuring degree of similarity between two Entities of $S$. The Similarity function $\text{Sim}(E_1, E_2)$ ought to obey reflexivity and symmetry property

For all $E_1, E_2 \in S$ it holds that,

$\text{Sim}(E_1, E_1) = 1$ (reflexivity property) and

$\text{Sim}(E_1, E_2) = \text{Sim}(E_2, E_1)$ (Symmetry property)

**Definition 2**: Ehrig definition of similarity for multiple ontologies

Given two ontologies $O_i$ and $O_j$, the Similarity function is defined as a real valued function $\text{Sim}(O_i(E_i), O_j(E_j)) : (E_i \times E_j) \rightarrow [0,1]$ such that $E_i, E_j$ are defined by concepts ($C$), datatypes ($D$), relations ($R$), attributes ($A$), instances ($I$), data values ($DV$). The similarity function computes similarity between concepts, datatypes, relations, instances, attributes and data values of the entities $E_i$ and $E_j$.

4.2.2 Feature based contrast model (Tversky Psychological Model)

Feature Models were brought into prominence by Tversky (1977). He has mentioned in his work, that subjective assessments of similarity do not always satisfy the assumptions of geometric models of similarity. The assumptions made by the geometric model include minimality, symmetry and triangle inequality. The minimality property states that all objects are similar to it. The Symmetry property states that the distance from point A to point B is equal to the distance between point B and point A. The triangle inequality states that the distance between two points A and B cannot be more than the distance between point A and the third point C plus the distance between point C and point B. Tversky proved that these three assumptions were not true for similarity assessments. Tversky has reported the difficulty in describing the objects which possess a large number of features. The feature models overcome the drawback of geometric model as the geometric
models do not have the capability of predicting that the common features of the objects would increase the similarity.

Hence because of these potential problems of the geometric model, Tversky proposed a feature based contrast model to assess similarity. Tversky suggested a contrast model to characterize similarity based on weighing common features and distinctive features.

In this model entities are represented as collection of features. Tversky presented an abstract model of similarity, based on set theory that takes into account the features that are common to two concepts and also the differentiating features specific to each. According to Tversky, the similarity of a concept $C_1$ to a concept $C_2$ is a linear function of the features common to $C_1$ and $C_2$ and the distinct features of $C_1$ (the features in $C_1$ but not in $C_2$) and the distinct features of $C_2$ (the features in $C_2$ but not in $C_1$). This is shown in Figure (4.1).

![Figure 4.1 Tversky Feature Based Contrast Model](image)

**Definition 3: Tversky Definition of Similarity**

Similarity of the concepts $C_1$ and $C_2$ is defined as a linear combination of a measure of the common features of concepts $C_1$ and $C_2$ and the distinctive features of the Concept $C_1$ and Concept $C_2$. This is formalized as given below.

\[
\text{Sim}_{tvr}(C_1, C_2) = \alpha F(\psi(C_1) \cap (C_2)) - \beta F(\psi(C_1) \cup (C_2)) - \gamma F(\psi(C_2) \cup (C_1))
\]

(4.1)
where α, β and γ are constants. The feature defined in this model may be a property, stimuli, pattern or a characteristic of an item being compared. The item being compared could be a molecule, scene, concept, stimulus etc. This contrast model predicts asymmetric similarity i.e. \( \text{Sim}(C_1, C_2) \neq \text{Sim}(C_2, C_1) \). This similarity property is a property required by applications like information retrieval in which the information retrieval systems retrieve documents specific to the query.

The contrast model also account for non mirroring between similarity and difference judgments. The common features in similarity is hypothesized to receive more weight in similarity than difference judgments and the distinctive features in difference judgments receive relatively more weight. In an experiment conducted by Tversky, the concept pair, “East Germany and West Germany” and the concept pair, “Ceylon and Nepal” were compared. The interesting fact is that the number of similar features and number of distinct features were more for both of these pairs. Usually the similarity value should report that the concept pairs are similar or the concept pairs are different or dissimilar. But for certain concept pairs the similarity values may be simultaneously perceived to be more similar and more different from each other.

Thus instead of a linear combination of common and distinct features, a ratio based model was proposed by Tversky.

**Definition 4: Tversky’s Ratio based Definition of Similarity**

It is defined as the ratio of the common features of the concepts \( C_1 \) and \( C_2 \) to the distinct features of the concepts \( C_1 \) and \( C_2 \) scaled by the relative importance of the concepts.

\[
S(C_1, C_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(C_1, C_2)|C_1 \cap C_2| + (1 - \alpha(C_1, C_2))|C_2 / C_1|}
\] (4.2)
where the intersecting features of the concepts are quantified by cardinality of \( C_1 \cap C_2 \) and the distinct features by the cardinality of the features in \( C_1 \) not in \( C_2 \) and the features in \( C_2 \) not in \( C_1 \).

The ratio based formulation of Tversky is normalized and it is a non linear combination of common and distinct features. Moreover, this similarity function is not forced to satisfy the metric properties (symmetry, minimality and triangle inequality).

The COSS measure proposed in this chapter is designed based on the Tversky ratio based similarity model. This is because COSS measure was designed for use in applications like information retrieval where the similarity between query and document is asymmetrical. Hence the proposed COSS measure is a non linear measure of similarity. The ratio based formulation of Tversky model was also used by Rodriguez and Egenhofer for assessing similarity among entity classes. But the similarity measure did not correlate well with human judgements when it was used for measuring cross ontology concept similarity. The Rodriguez and Egenhofer measure is discussed in the next section.

4.2.2.1 Rodriguez and Egenhofer Feature based measure

According to Rodriguez and Egenhofer, a concept is considered as an entity class (Rodriguez and Egenhofer 2003). This measure was proposed to determine the semantic similarity among entity classes of different ontologies. This measure computes similarity among entity classes by finding the similarity between the synonym sets of the entity classes, similarity between the distinguishing features of the entity classes and the similarity between semantic neighbourhoods of the entity classes.

The similarity function between entity classes is then the weighted aggregation of the similarity among the three specification components (synonym sets, distinguishing features and semantic neighbourhoods). Hence the similarity between entity classes of the ontology \( p \) and of ontology \( q \), is given as
where \( S_w \) measure the similarity between synonym sets, \( S_u \) measure similarity between features, and \( S_n \) measure similarity between semantic neighborhoods and \( W_w, W_u \) and \( W_n \) are the respective weights of the similarity of each specification component.

The similarity function \( S_w \) of Equation (4.3) determines the number of common words and different words in the synonym sets and this word matching is done using Equation (4.2).

Feature matching (\( S_u \)) in Equation (4.3) is used to find similarity between the distinguishing features of the entity class. The distinguishing features of the entity class are the parts defined in the entity class, functions of the entity class and the attributes of the entity class. Hence the similarity among the distinguishing features needs to compute similarity among the parts (\( S_p \)), functions (\( S_f \)) and the attributes (\( S_a \)) of the entity classes and it is defined in Equation (4.4). These weights \( W_p, W_f \) and \( W_a \) in Equation (4.4) are considered to be equal and it is experimentally fixed as 0.33. The equal weights signify that the three distinguishing features are considered equally important to quantify similarity. The \( S_p, S_f \) and \( S_a \) are computed using Equation (4.2) and \( \alpha \) (\( C_1, C_2 \)) is the relative importance of the non common characteristics and it takes value in the range \([0..1]\). This \( \alpha \) (\( C_1, C_2 \)) is used in the proposed COSS measure to compute the relative importance of the concept in the ontology.

\[
S_u(c_i^p, c_i^q) = W_p S_p(c_i^p, c_i^q) + W_f S_f(c_i^p, c_i^q) + W_a S_a(c_i^p, c_i^q) 
\]

The semantic neighbourhood matching \( S_n \) in Equation (4.3) is computed using Equation (4.5). The Semantic-neighborhood matching (\( S_n \)) compares entity classes \( C_1 \) and \( C_2 \) of ontologies \( p \) and \( q \) with radius \( r \), respectively, which is a function of the cardinality (\(|N|\)) of the semantic neighborhoods (\( N \)) and the
approximate cardinality of the set intersection ($\cap_n$) between these semantic neighborhoods is given by

\[
S(C_i, C_j) = \frac{|P \cap_n C_i \cap_n C_j|}{|P \cap_n C_i \cap_n C_j| + (1 - a(C_i, C_j))|P \cap_n C_i \cap_n C_j|}
\] (4.5)

This method of computing similarity among entity classes has been experimented with SDTS (Spatial Data Transfer Standard) and WS (An ontology created from SDTS and WordNET) and with WordNet ontology and WS ontology. The correlation between the similarity measure and human judgements was very low and was 0.37 and 0.71 respectively. However this measure correlated well with human judgements, when experimented with instances of the same ontology (WS and WS). The correlation was reported as 0.97. The other generic similarity models are discussed in the next section.

4.2.3 Alignment based model of similarity

Relational structure commonalities play an important role in similarity (Gentner and Markman 1994). He has also reported that assessment of similarity by human subjects entail structural alignment and mapping. The network model, geometric model and feature models represent knowledge in an unstructured way. In contrast to the other similarity models, the alignment model represents knowledge in a structured way. The alignment models represent concepts with their parts, describe parts separately with features, relate parts to wholes, and relate concepts to other concepts. Similarity Interaction Activation Model (SIAM) is an alignment based similarity model (Goldstone and Medin 1994). This was developed for measuring similarity among spatial scenes which are described by roles, components and features.

4.2.4 Transformational model

So far in all the similarity models, the concepts are described by the properties and relations. But in transformational models the concept descriptions
are indirectly specified as a set of transformations. The transformational model assesses similarity by transforming one object to another. These transformational models specify similarity by specifying the possible transformational operations. In a transformational model the similarity among the concept is more, if the number of transformation operations is less. The main problem in transformational models is to find suitable set of transformations. These models have been applied to well known domains of perceptual stimuli such as chain of alphabetic strings and chain of filled and unfilled circles. The transformation is based on perceptual attributes only. Operations (Goldstone and son 2005) like rotation, reflection, translation and dilation have been used to define transformations. But still the perceptual transformations are not available to describe relation of conceptual stimuli.

4.2.5 Logic based model

Apart from the above mentioned models, logic based computational models are proposed to measure similarity. They have gained more importance in semantic interoperability with the increased development of ontologies. Janowitz has proposed a similarity measure SIM-DL for measuring similarity among geospatial ontologies (Janowicz 2006a, 2006b). Concepts are specified based on primitive concepts, roles, and language constructors such as union, intersection and existential quantification. This similarity satisfies asymmetry property and it is a context aware similarity measure. Because this similarity measure has asymmetry property it is suitable for use in information retrieval systems where the similarity is determined by the overlap between query concept definitions and concept definitions of documents.

From the discussion made above it is evident that the geometric model has its own limitations. The Alignment and Transformational models are more suited for spatial domain and perceptional stimuli based domains. Logic models and feature models have been used for measuring similarity among concepts belonging to single ontology. Only limited work has been reported (Path based approach discussed in next section) so far for measuring semantic similarity among concepts belonging to multiple ontologies. Few measures are reported in the literature
which measure cross ontology concept similarity by integrating the ontologies. The next section discusses the current state of art in measuring cross ontological semantic similarity.

4.3 PATH BASED MEASURE

Al-Mubaid and Nguyen proposed a new ontology structure based measure for measuring semantic similarity among concepts in multiple ontologies. The MeSH (Medical Subject Headings) and SNOMED-CT (Systemized Nomenclature of Medicine Clinical Term) ontologies within the framework of Unified Medical Language System (UMLS) was used for measuring concept similarity (Al-Mubaid and Nguyen 2009).

According to Al-Mubaid and Nguyen, the semantic similarity between cross ontological concepts is measured by considering one ontology as primary and another as secondary ontology. For measuring cross ontology similarity of concepts, path length feature, depth and the granularity of ontologies is taken into consideration. From Al-Mubaid and Nguyen point of view, the concepts being compared belong to the same primary ontology (Case 1), concepts belong to primary ontology and secondary ontology (Case 2), both concepts belong to secondary ontology (Case 3) and concepts belong to multiple secondary ontologies (Case 4). The process of how similarity is measured under the above four cases is discussed below.

Case 1: In the case of single ontology, the path length and depth features are used to get semantic distance of two concepts as follows

\[
SemDist(C_1, C_2) = \log((path - 1)^{\alpha} \cdot (Spec)^{\beta} + k)
\]  

(4.6)

where \(\alpha, \beta\) and \(k\) are constants.
**Case 2:** By considering more than one ontology, the semantic similarity measure is based on the three following features

- Common specificity of concepts in the ontology (CSpec)
- Cross Modified Path Length between two concepts
- A local granularity of ontology clusters

The CSpec ($C_1$, $C_2$) quantifies the commonalities that exist among the concepts. In measuring cross ontology concept semantic similarity, the common specificity feature between two concepts $C_1$ and $C_2$ is calculated by finding the least common subsuming concept (LCS), its depth and the depth of the ontology. Hence the computation of CSpec ($C_1$, $C_2$) is as given below.

$$CSpec(C_1, C_2) = D - \text{Depth}(LCS(C_1, C_2))$$  \hspace{1cm} (4.7)

If the CSpec value is less, the concepts have more shared information and hence the commonality between the two concepts is more. In this case, two concepts are belonging to two different ontologies identified as primary and secondary ontologies. Hence, the secondary ontology is connected to the primary ontology by joining the common nodes called the bridge nodes.

The LCS node of two concept nodes ($C_1$, $C_2$) is measured by considering the LCS of the first node $C_1$ in primary ontology and the bridge node,

$$LCS(C_1, C_2) = LCS(C_1, bridge)$$  \hspace{1cm} (4.8)

The part of path length in secondary ontology is then scaled into primary ontology’s scale of path feature. Thus the path length between two concept nodes is calculated by adding up two path lengths from each of the concept nodes to bridge node. The path length feature is computed as $d_1 + d_2 - 1$ where $d_1 = d(C_1, \text{bridge})$ and $d_2 = d(C_2, \text{bridge})$, where $d(C, \text{bridge})$ is the shortest path length from
$C_1$ to the bridge node, and $d(C_2, \text{bridge})$ is the shortest path between $C_2$ and bridge node.

In order to scale the path length and CSpec features in the secondary ontology to the primary ontology, the granularity ratio of the primary ontology over the secondary ontology is considered and the scaled path length feature is given by

$$PathRate = \frac{(2D_1 - 1)}{(2D_2 - 1)}$$  \hspace{1cm} (4.9)

where $D_1$ and $D_2$ represent depth of the primary and secondary ontology. Similarly the Common specificity CSpec ($C_1, C_2$) is scaled to the depth of the primary ontology.

Since there may be many bridge nodes between two concepts, there can be more than one path length i.e. $\{\text{path}_i\}$ and the semantic distance, $\text{SemDist}$, between two concept nodes is given as follows.

$$\text{CSpec}_i(C_1, C_2) = D_1 - \text{Depth}(\text{LCS}(C_1, \text{Bridge}_i))$$  \hspace{1cm} (4.10)

$$\text{SemDist}(C_1, C_2) = \log((\text{path}_i - 1)^a * (\text{CSpec}_i)^\beta + K)$$  \hspace{1cm} (4.11)

**Case 3:** In the case of similarity within single secondary ontology, both the concepts are present in the secondary ontologies. Therefore the semantic distance between these concepts of the secondary ontology can be computed only when the Path ($C_1, C_2$) and CSpec ($C_1, C_2$) are translated to primary ontology scales. The scaled Path ($C_1, C_2$) and CSpec ($C_1, C_2$) features of the secondary ontology are given by

$$Path(C_1, C_2) = \text{path}(C_1, C_2)_{\text{secondary}} \ast PathRate$$  \hspace{1cm} (4.12)
\[ C_{Spec}(C_1, C_2) = C_{Spec}(C_1, C_2)_{\text{secondary}} \times C_{SpecRate} \] (4.13)

where \( \text{path} (C_1, C_2)_{\text{secondary}} \) is the shortest path between \( C_1 \) and \( C_2 \) in the secondary ontology and \( C_{Spec} (C_1, C_2)_{\text{secondary}} \) is computed based on the secondary ontology using Equation (4.7). The Semantic distance between the concepts in the secondary ontology is given by

\[ \text{SemDist}(C_1, C_2) = \log((\alpha \times (\text{path} - 1)^{\alpha} \times (C_{Spec})^\beta + k) \] (4.14)

**Case 4:** In the case of similarity within multiple secondary ontologies, the two concepts are in two different secondary ontologies in which one among the two acts as a primary ontology (temporarily) and the other acts as a secondary ontology. The secondary ontology which has more number of concepts will be chosen as the primary ontology. Then the semantic similarity between two concepts can be calculated by considering the path length and CSpec feature of the secondary ontology after scaling up their values to the primary ontology.

The path length measure was tested for biomedical ontologies SNOMED-CT and MeSH. This measure was compared against the human judgements. In this measure commonality among the concepts were quantified by identifying bridge nodes. The identification of the bridge nodes is a time consuming process. But still this similarity measure correlated well with the human judgments when compared to the Rodriguez and Egenhofer measure.

In the Rodriguez and Egenhofer measure and in Al-Mubaid measure the ontologies were not integrated. Some of the researchers compute similarity among concepts belonging to different ontologies by integrating the ontologies. The ontology integration approaches for measuring cross ontology concept similarity is discussed in the next section.
4.4 ONTOLOGY INTEGRATION APPROACHES

The thought of integrating ontologies to measure cross ontology concept similarity is practically possible with lesser number of ontologies. The major difficulty in this approach is maintaining the consistency of the ontologies. Another difficulty is managing the heterogeneity of the ontologies. This problem is very similar to the database heterogeneity problem ONIONS (Gangemi 1998) use a set of generic ontologies to integrate the ontologies. OBSERVER (Mena et al. 1996) (Kashyap and Sheth 1998) (Mena et al. 2000) ontology based system uses the terminological relations to map the non translated terms of the user ontology to terms in a target ontology. The translation process is recursive and it consists of substituting non translated terms with the intersection of their immediate parents or the union of the immediate children. A semi automatic process for mapping terms from source to target ontology by using hyponym relations and synonyms of the terms is also proposed. (Bergamashi et al. 1998)

Once the ontologies are integrated, the similarity measures are applied to measure similarity between concepts as done in a single ontology. Therefore in general the current methods of computing similarity of concepts belonging to different ontologies is based on apriori integration of local ontologies through a top level ontology or through terminological relations. But the Cross Ontology Semantic Similarity measure (COSS) discussed in this chapter is a computational approach which compares concepts belonging to different ontologies without integrating the ontologies.

4.5 LIMITATIONS OF CURRENT STATE OF ART OF CROSS ONTOLOGY SIMILARITY MEASURES

The Rodrigues and Egenhofer measure computes semantic similarity among cross ontological concepts and has been evaluated using the concepts derived from the WordNet ontology, WS (A portion of WordNet ontology) and SDTS spatial knowledge source. The correlation coefficient achieved using SDTS and WS ontologies is 0.37 and between WordNet and WS is 0.71. Moreover the
approaches do not take into consideration the relations shared by the compared concepts.

The approach followed by Al-Mubaid and Nguyen used biomedical ontologies and has followed the distance measure to quantify semantic similarity. The correlation coefficient achieved is better when compared to Rodriguez and Egenhofer measure.

These limited approaches for computing semantic similarity among cross ontology concepts has given the motivation to develop cross ontology concept similarity measures using information content based approaches. There is a need for analysing the information content based approaches for information retrieval and ontology mapping applications. This thesis focuses on adapting the single ontology information content based approaches for computing semantic similarity among cross ontological concepts. In addition to this, a new measure is devised where the Tversky feature based ratio model has been mapped to information theoretic domain. The Tversky’s model is also referred in literature as Tversky index. The variants of Tversky index, the dice index or dice coefficient have been mapped to the information theoretic domain to measure cross ontology concept similarity.

4.6 MAPPING OF TVERSKY FEATURE BASED PSYCHOLOGICAL MODEL TO INFORMATION THEORETIC DOMAIN

Earlier, as it is mentioned in chapter 2, Pirró similarity measure is based on Tversky psychological model and it has been mapped to information theoretic domain to measure similarity among concepts in a single ontology (Pirró 2009). The results of this computational measure were appreciable when tested with WordNet knowledge source. However, this measure has not been used for measuring similarity among cross ontology concepts. Hence, this thesis work attempts to map the ratio based Tversky psychological model to information theoretic domain to measure cross ontology concept similarity.
In information theoretic domain, the common characteristics among concepts are quantified by the information content of the most specific common abstraction concept. The most specific abstraction concept is the common concept which subsumes the concepts being compared. The distinct characteristics of the concept are quantified by the information content of the concept. Hence while mapping the Tversky ratio based model to information theoretic domain the problem of finding the most specific common abstraction concept should be given importance. Once the most specific common abstraction concept is found, the intersecting common features of $C_1$ and $C_2$ could be mapped to the information content of the most specific common abstraction concept.

In a multiple ontology scenario, finding the most common abstraction concept is a tedious task. This is because the concepts being compared do not belong to the same ontology. The methods mentioned in Section 4.4 may be followed to solve this problem. But it is clear that integrating the ontologies is a tedious and time consuming process. Further maintaining the consistency of the ontologies is a difficult task.

Hence, keeping the above mentioned issues in mind the design of the proposed COSS measure is explained in the next section.

**4.7 DESIGN PRINCIPLES OF COSS SIMILARITY MEASURES**

The following principles were kept in mind for designing the new semantic similarity measures for ontological concepts belonging to two different ontologies.

- The COSS measure should be based on human psychological models as all of the existing semantic similarity measures are evaluated against human judgments.
- The COSS measure should be based on Information Content measures as in the literature most of the IC based measures achieve highest correlations against human judgments.
• The Information Content (IC) calculation should be corpus independent as corpus dependent IC calculations are time-consuming and require tagged corpora.
• The identification of the Most Specific Common Abstraction concept should be done without constructing a-priori a shared ontology.
• The depth property of the semantic similarity should not be ignored as the specificity of the concept increases with the depth.

The COSS measure is based on the Tversky’s feature based Psychological model. Based on the variant of Tversky model namely Dice index has been adapted to measure semantic similarity among cross ontology concepts. The COSS measure is discussed here in detail.

The COSS similarity measure considers a corpus independent information content based similarity computation to assess asymmetric similarity between concepts belonging to multiple ontologies. The problem of finding the most specific common abstraction concept is solved by connecting the ontologies through a virtual root. This virtual root paves a way for finding the most specific common abstraction concept which subsumes concepts belonging to different ontologies. It also eliminates the complexity of constructing apriori a shared ontology.

**Definition 5: COSS Definition of Similarity**

*The Similarity between Cross ontological concepts is defined as the ratio of the common characteristics shared by the compared concepts to the distinct characteristics quantified by the position of the concepts in the primary and secondary ontology.*

The COSS measure considers the ontology with more number of concepts as primary ontology and ontology with less number of concepts as secondary ontology.
As mentioned earlier the common characteristics according to Tversky are formulated as \( C_1 \cap C_2 \) and the unique characteristics of the Concept \( C_1 \) are quantified as \( C_1 / C_2 \) and the unique characteristics of \( C_2 \) as \( C_2 / C_1 \).

In the information theoretic domain, the common characteristics are quantified as the information content of most specific common abstraction concept. The information content of most specific common abstraction concept is denoted as \( \text{IC}(\text{MSCA}(C)) \). This MSCA concept subsumes the concepts \( C_1 \) (belonging to the primary ontology) and \( C_2 \) (belonging to secondary ontology). The IC (MSCA(C)) has to be computed based on two ontologies as \( C_1 \) belongs to \( O_1 \) (primary ontology) and \( C_2 \) belongs to \( O_2 \) (secondary ontology).

The problem of finding most specific common abstraction concept of concepts belonging to primary and secondary ontology could be solved by following the methods given below.

- Constructing a apriori shared ontology
- Integrating the ontologies.

As the methods mentioned above are time consuming an alternate way of computing the MSCA of the concepts is devised. The primary and secondary ontologies are connected through a virtual root (VR) called “Anything”. Then starting from the virtual root, the primary and the secondary ontologies are traversed to identify the most common abstraction concept subsuming the concepts \( C_1 \) and \( C_2 \). As most of the ontologies support multiple inheritance, there is a possibility of having many most specific abstraction concepts. In such cases, most specific common abstraction concept at the deepest level will be considered as the MSCA.

The information content of the MSCA concept is computed using Equation (4.17). It quantifies the common characteristics of the concepts belonging to primary and secondary ontology. The quantification of the MSCA concept is based on the hyponym relations possessed by the concepts.
The unique characteristics of the compared concepts are determined by Equation (4.18). IC (C1) is equivalent to unique characteristics of C1 alone and excludes the characteristics of C2 (|C1/Ca|) and IC (C2) is equivalent to unique characteristics of C2 not present in C1 (|C2/C1|). Hence the common characteristics and unique characteristics are defined in information theoretic terms by using Equations (4.17) and (4.18) respectively. Now the information theoretic terms defined above are mapped to the Tversky ratio based model Equation (4.2) and the Cross Ontology Semantic Similarity measure (COSS) is defined as

\[
\text{Sim}_{\text{COSS}}(C_1, C_2) = \frac{\text{IC(MSCA}(C))}{\text{IC(MSCA}(C)) + \alpha(C_1, C_2)(\text{IC}(C_1) + (1 - \alpha(C_1, C_2))(\text{IC}(C_2))].}
\]

where IC (MSCA) is computed using Equation (4.17) and IC (C1) and IC (C2) is calculated using Equation (4.18). The relative importance of the concepts in the ontology is quantified by \(\alpha(C_1, C_2)\). This is computed using Equation (4.16).

\[
\alpha(C_1, C_2) = \begin{cases} 
\frac{\text{Depth}(C_1^{\alpha_1})}{\text{Depth}(C_1^{\alpha_1}) + \text{Depth}(C_2^{\alpha_2})} & \text{if} \text{ depth}(C_1) \leq \text{ depth}(C_2) \\
1 - \frac{\text{Depth}(C_1^{\alpha_1})}{\text{Depth}(C_1^{\alpha_1}) + \text{Depth}(C_2^{\alpha_2})} & \text{if} \text{ depth}(C_1) > \text{ depth}(C_2) 
\end{cases}
\]

In the COSS measure, \(\alpha\) is a function that determines the relative importance of the concept in the ontology and is defined using depth of the concepts C1 and C2 in the ontology. The value of \(\alpha\) was experimentally determined using the SNOMED-CT and MeSH ontologies, and found to be within the range of 0 to 0.9. For example, let O1 be a primary ontology and O2 be a secondary ontology.

Case 1:

Let C1 of O1 and C2 of O2 be the most abstract concepts then depth (C1) =1 and depth (C2) =1 which satisfies the first condition of Equation (4.16). Hence \(\alpha = 0.5\)
Case 2:

Let \( C_1 \) of \( O_1 \) and \( C_2 \) of \( O_2 \) be at the same depth. Let the concepts be at depth of 5, then \( \alpha = 0.5 \)

Case 3:

Let \( C_1 \) of \( O_1 \) be at the depth of 1 (more abstract level) and \( C_2 \) of \( O_2 \) be at the depth of 8 (more specific level) then \( \alpha = 0.1 \) and \( 1 - \alpha = 0.9 \).

For various values of depths the \( \alpha \) (\( C_1, C_2 \)) is calculated. Further if the value of \( \alpha \) (\( C_1, C_2 \)) = 0.5, it means that the unique characteristics of \( C_1 \) and that of \( C_2 \) are given equal importance. Hence for different concepts depending on the position at which they are defined in the ontology the relative importance will vary. Further the unique characteristics are quantified based on the concept relations a concept has in the respective ontology. The formal definition of information content of most specific common abstraction concept is given below.

**Definition 6:** Information content for most specific common abstraction concept

It is defined as the function of the maximum number of hyponyms possessed by the common concept in both the primary and secondary ontologies scaled by the maximum number of concepts in the primary ontology.

The IC (MSCA) is computed as follows.

\[
IC(MSCA(C)) = 1 - \frac{\log(\max(hypo(O_1(C_1), O_2(C_2)))+1)}{\log(\max_{\text{con}})}
\]

(4.17)

where function \( \max (hypo (O_1 (C_1), O_2 (C_2))) \) returns maximum hyponymy of the concept. \( \max_{\text{con}} \) returns maximum concepts of the primary ontology. The common characteristics are quantified by computing IC (MSCA).

The unique characteristics of the compared concepts are to be determined in information theoretic terms and this quantification of the unique characteristics
is quantified by using the information content definition stated by Seco. (Seco et al. 2004). The definition of information content of a concept is given by

\[ IC(C) = 1 - \frac{\log(\text{hypo}(C) + 1)}{\log(\text{max}_{\text{con}})} \]  

(4.18)

where the function hypo returns the number of hyponyms of a given concept C and \( \text{max}_{\text{con}} \) represents total number of concepts in the considered taxonomy.

The similarity between two cross ontology concepts is defined by five rules based on the position of the concepts in the taxonomy.

**Rule 1:** If \( C_1 \) in primary ontology is deeper when compared to \( C_2 \) in secondary ontology then the similarity value found between \( C_1 \) and \( C_2 \) is maximum.

**Rule 2:** If \( C_1 \) in primary ontology is at a level higher than concept \( C_2 \) in secondary ontology then the similarity value found between concepts \( C_1 \) and \( C_2 \) is minimum.

**Rule 3:** If the MSCA concept is present in the sub-root of both the ontologies then the similarity value results in zero.

**Rule 4:** If the two concepts are present in the same level then it may or may not increase the similarity values depending upon the number of hyponym possessed by two ontologies.

**Rule 5:** If both the concepts \( C_1 \) and \( C_2 \) are present at the most abstract levels or as sub-roots of both ontologies then it results in very less similarity value.
The effectiveness of COSS measure was evaluated using the biomedical data sets used by Al-Mubaid and Nguyen. The Tversky model defined in Equation (4.2) is reduced to Dice index when the $\alpha$ and $(1-\alpha)$ are 0.5 and when the scaling is done using the unique features alone. Hence the Dice index was mapped suitably to information theoretic domain and the definition of similarity using Dice Index is discussed in the next section.

4.8 VARIANT OF TVERSKY SIMILARITY INDEX (RDCOSS)

The dice index is a variation of Tversky similarity index.

**Definition 7**: Dice Definition of similarity

The similarity between $A$ and $B$ is defined as a ratio between twice the shared information of $A$ and $B$ to the information of the combined set.

$$S(A, B) = \frac{|A \cap B|}{\frac{1}{2}(|A|+|B|)}$$  

(4.19)

The shared information is quantified as $(|A|\cap|B|)$ and the information over the combined set as $(|A|+|B|)$. The Dice index without any changes cannot be adapted for measuring the cross ontology concept similarity. This is because such the Dice index insists on giving equal importance to the unique characteristics of the two concepts being compared. The concepts will have equal importance only when they are defined at the same level in both the primary and secondary ontologies. This could be a specific case and could not be same for concepts defined at different levels of the primary and secondary ontology.

If one concept is at the abstract level and another is at the deepest level (near leaf nodes) then $\alpha$ value could be 0.9. As this work is focused on adapting Dice index for computing semantic similarity among concepts belonging to multiple ontologies, it was experimentally found that $\alpha=\beta=0.5$ could not be true for all the compared concepts. Hence the Dice index definition is redefined to suit the
scenario required to measure cross ontology concept similarity. The Redefined Dice index is discussed below.

**Definition 8: Redefined Dice index (RDCOSS) of similarity**

The similarity between concepts A and B is defined as a ratio between the common features to the summation of the distinct features and is given by

\[
S(A, B) = \frac{|A \cap B|}{\alpha(|A|) + \beta(|B|)}
\]

(4.20)

The common features \(A \cap B\) is defined in information theoretic terms using Equation (4.17). The information over the combined set is quantified as the arithmetic addition of information content of individual concepts. The information content of the individual concepts is quantified using Equation (4.18). Thus the Refined Dice index is mapped to the information theoretic domain. Hence the RDCOSS similarity measure for measuring semantic similarity between concepts belonging to cross ontologies is given by

\[
Sim_{RDCOSS}(C_1, C_2) = \frac{IC(MSCA(C))}{\alpha(|IC(C_1)|) + \beta(|IC(C_2)|)}
\]

(4.21)

where \(\alpha\) is computed using Equation (4.16) and \(\beta\) is computed as \((1-\alpha)\). The \(\alpha\) ranges in the value of \([0..0.9]\).

The Pirró measure defined for measuring concept similarity in a single ontology is also based on the linear model of Tversky. This Pirró measure has been redefined to measure cross ontology concept similarity. The Redefined Pirró measure is discussed in the following section.
4.9 REFINED PIRRÓ MEASURE (RPCOSS)

The Pirró (Pirró 2009) feature based similarity measure is based on the Tversky Feature based linear model. The Pirró measure for measuring single ontology concept similarity is given by

\[ Sim_{p&k} (C_1, C_2) = 3IC(MSCA(C_1, C_2)) – IC(C_1) – IC(C_2) \] (4.22)

Similar to single ontology Pirró measure the Tversky model is mapped to information theoretic domain to measure the cross ontology concept similarity. Hence, the Equation (4.22) is redefined as

\[ Sim_{rpcoss} = 3*IC(MSCA((O_1, C_1), (O_2, C_2)) – IC(O_1, C_1) – IC(O_2, C_2) \] (4.23)

where IC (MSCA ((O_1, C_1), (O_2, C_2))) is computed using Equation (4.17) where (O_1, C_1) is the concept belonging to the primary ontology and (O_2, C2) is the concept belonging to the secondary ontology. The IC(MSCA((O_1,C_1),(O_2,C_2))) is the information content of the most specific common abstraction concept subsuming the concepts C1 and C2. The IC(O_1,C_1) and IC(O_2,C_2) is computed using Equation (4.18).

4.10 REFINED INFORMATION CONTENT BASED MEASURES FOR CROSS ONTOLOGY SIMILARITY

The semantic similarity measures such as Resnik, Jiang and Conrath and Lin use information content (IC) value to compute similarity between two concepts in the ontology (Resnik 1995) (Jiang and Conrath 1997) (Lin 1998). The main drawbacks in these measures are

- They are corpus dependent
- They require analysis of the corpus which is time consuming
- They have not been used for computing semantic similarity for cross ontological concepts.
The single ontology information content based approaches Resnik, Jiang
and Conrath and Lin have been refined as RRCOSS, RJCOSS and RLCOSS so as
to compute cross ontological concept similarity.

4.10.1 Refined Resnik Measure (RRCOSS)

The existing Resnik (1995) measure for single ontology is given by

\[
Sim_{res}(C_1, C_2) = \max_{C \in S(C_1, C_2)} \left[ IC(C) \right]
\]  

(4.24)

The above definition of Resnik is refined to measure concept similarity
among ontologies. The Refined Resnik Cross ontology semantic similarity
measure (RRCOSS) for multiple ontologies is defined as given below.

\[
Sim_{RRCOSS}(C_1, C_2) = \max_{C \in S(O_1, O_2)} \left[ IC(C) \right]
\]  

(4.25)

where C in Equation (4.25) is the most specific abstraction concept and it’s
IC is calculated by using Equation (4.17). The C_1 and C_2 concepts belong to
primary and secondary ontology respectively. The similarity value of the concepts.
C_1 and C_2 is the information content of the most specific common abstraction
concept C.

4.10.2 Refined Lin Measure (RLCOSS)

The existing Lin measure for single ontology is given by

\[
Sim_{Lin}(C_1, C_2) = 2^* \frac{IC(MSCA(C_1, C_2))}{IC(C_1)+IC(C_2)}
\]  

(4.26)

The Equation (4.26) is modified suitably to measure similarity among
c and C_2 are belonging to multiple ontologies and is defined as follows:

\[
Sim_{RLCOSS}(C_1, C_2) = 2^* \frac{IC(MSCA((O_1, C_1), (O_2, C_2)))}{IC(C_1)+IC(C_2)}
\]  

(4.27)
where IC(MSCA((O₁,C₁),(O₂,C₂))) is computed using Equation (4.17) where (O₁,C₁) is the concept belonging to the primary ontology and (O₂,C₂) is the concept belonging to the secondary ontology. The IC(MSCA((O₁,C₁),(O₂,C₂))) is the most specific common abstraction concept subsuming the concepts C₁ and C₂. The IC(O₁,C₁) and IC(O₂,C₂) is computed using Equation (4.18).

4.10.3 Refined Jiang and Conrath measure (RJCOSS)

The Jiang and Conrath measure is used to measure similarity among concepts in a single ontology. This measure has been redefined as given in Equation (4.28), to measure similarity among cross ontological concepts.

\[
\text{Sim}_{JC}(C_1, C_2) = \frac{IC(C_1) + IC(C_2) - Min[IC(C_1), IC(C_2)]}{IC(C)}
\]

(4.28)

where C is the most specific common abstraction concept and its IC is computed using Equation (4.17), where C₁ is the concept belonging to the primary ontology and C₂ is the concept belonging to the secondary ontology. The IC (C₁) and IC (C₂) are computed using Equation (4.18).

4.11 CrossSim ALGORITHM

Let O₁, O₂,….Oₙ be multiple ontologies available in UMLS framework. Among the available ontologies designate one ontology as primary and the other as secondary based on the granularity of the concepts they possess. Let O₁ (C₁) and O₂ (C₂) be the concepts belonging to the corresponding ontologies and r₁ and r₂ are the root nodes of the primary and secondary ontologies. Create a virtual root (VR) (as shown in Figure 4.2) which connects the root nodes r₁ and r₂ to VR.

The concept descriptions of the concepts belonging to the primary ontology and secondary ontology are created and are stored in the form of XML files. Each concept has a unique id with respect to the ontology they belong to. The concept description has the relations associated with the concepts, their position in the
respective ontology and the synonym set of the concepts. These concept
descriptions serve as input to the algorithm. The steps followed to compute the
similarity among cross ontology concept similarity is outlined in Figure 4.3. The
proposed similarity measures were tested for information retrieval applications.
The COSS measure was used to find correspondences between various categories
of SNOMED-CT and MeSH ontologies.

In Figure 4.2, the way in which the similarities are calculated for a concept
pair is illustrated and it is explained below. The semantic similarity between a
concept pair (Urinary tract infection, Pyelonephritis) can be calculated by first
connecting the virtual root to the subcategory Clinical finding in SNOMED-CT
and the subcategory Disease in MeSH. The ancestor list of Urinary tract infection
is being matched with the ancestor list of Pyelonephritis. The urological disease is
the common ancestor and the information content of the common ancestor
concept “urological disease” is computed. The Disease concept which is at the
most abstract level is also a common concept. But as it is at the most abstract level,
the concept at the deepest level is considered. The relative importance of the
concepts urinary tract infection and pyelonephritis is computed using Equation
(4.16). The information content computation of the concept “urological disease” is
computed using Equation (4.17). The algorithm CrossSim outlined in Figure 4.3
describes steps followed for computing the semantic similarity among concepts
belonging to multiple ontologies.

The correctness of the algorithm is verified by correlating the computed
similarity ratings against similarity judgments made by human subjects. The
human judgments for the concepts are collected from the experts who have domain
knowledge about the concepts. However, the reliable human judgments alone are
taken into consideration by checking the users’ integrity (Step 12 of the CrossSim
Algorithm).
Figure 4.2 Connecting Independent Ontologies through Virtual Root

The COSS measures discussed in this chapter is evaluated for the standard biomedical datasets. The correlation of these computational approaches COSS, RDCOSS, RRCOSS, RLCOSS, RJCOSS and RPCOSS are correlated against the human judgements and the results are reported in Chapter 5.
CS_Score Algorithm (CrossSim(C1,C2))

//CS_Score represents Cross ontology Similarity Score; C1 and C2 represents concepts belonging to different ontologies.
1. Let C1 & C2 be two concepts from O1 and O2 ontologies respectively.
2. Let S1 and S2 be the set of all ancestors subsuming the concepts C1 and C2 which belong to O1 and O2 ontologies respectively.
3. Compare S1 and S2 ancestor sets until common ancestor is found.
4. Compute the information content of the Most Specific Common Abstraction concept // the common ancestor found in step3.

\[ IC(MSCA(c)) = 1 - \frac{\log(max(hypO(C_1),O_2(C_2)))+1}{\log(max_{con})} \]
5. Calculate Information content for specific concepts C1 and C2.

\[ IC(C) = 1 - \frac{\log(hypO(C)+1)}{\log(max_{con})} \]
6. Compute the relative importance of the concepts by considering both the ontologies using Equation (4.16).
7. Compute cross ontology concept similarity using Simcoss(C1,C2) // Let Simcoss(C1,C2) be the semantic similarity function used for computing similarity between cross ontological concepts C1 and C2 . calculated as

\[ Sim_{CROSS}(C_1,C_2) = \frac{IC(MSCA(C))}{IC(MSCA(C)) + \alpha(C_1,C_2) . (IC(C_1) + (1-\alpha(C_1,C_2) . (IC(C_2)))} \]

//Regardless of depth, if there exist matched synonym set of the C1 and C2 concepts belonging to two different ontologies then it results twice the value obtained by SimCROSS because both unique concepts contributes equal amount of shared information to the similarity value between concepts belonging to two different ontologies.//
8. Calculate semantic similarity for proposed Dice index (RDCOSS) based IC measure using Equation (4.21) and RPCOSS using Equation 4.23.
9. Calculate semantic similarity for Proposed IC measures (RRCOSS & RLCOSS and RJCOSS) using Equation(4.25) Equation(4.27) and Equation (4.28).
10. Collect human judgements for which similarity rating is to be calculated.
11. Check User Integrity by a rating coefficient (i.e., Rc) defined as

\[ R_c = \sum_{i=0}^{n} c_i - avg_i \]
12. Eliminates human judgment which are incorrect using Rc
14. Compare the performance of the proposed measures.

Figure 4.3 CrossSim Algorithm
4.12 Information retrieval using COSS measures

The similarity measures play an important role in information retrieval. Hence the proposed similarity measures have been evaluated for information retrieval applications. In the context of information retrieval, the similarity between the document terms and the query terms are measured using the similarity measures proposed in this chapter.

In this thesis, to evaluate the proposed cross ontology measures the SNOMED-CT and MeSH biomedical ontologies were used. A collection of MEDLINE articles related to biomedical domain was used as a corpus to test the information retrieval applications. The retrieval performance is evaluated by computing the Precision, Recall, Fallout and F-measure parameters and the results are discussed in chapter 5.