CHAPTER 3

NEW INFORMATION CONTENT BASED SEMANTIC SIMILARITY MEASURES

3.1 INTRODUCTION

Recent investigations reveal that semantic similarity measures have gained a lot of importance in the field of Artificial Intelligence. The semantic representation of the information sources is encoded using ontologies. The identification of semantically related concepts belonging to different ontologies plays an important role in information retrieval tasks and information integration systems. Most of the existing semantic similarity approaches are based on the ontology structure. On the other hand, information content based approaches quantify information of concepts by means of the statistical information derived from the corpus. However these corpus dependent approaches suffer from sparse data problem. Few approaches quantify similarity, based on the features of the concept. The work based on the combination of the above mentioned methods is attempted by few researchers (Seco et al. 2004) (Pirró 2009). In this research work, a corpus independent approach known as New Information Content (NIC) has been proposed which quantifies the similarity based on the ontology structure and the concept relations.

This Chapter analyzes the pros and cons of the various information content based similarity measures in terms of the experimental investigations and the general properties. It also outlines the NIC measure proposed in this research work. The NIC measure is defined based on the hierarchical concept relations is-a and part-of relations. Further the dominant concepts representing the semantic content of the documents are captured through the NIC based semantic similarity approaches. These computationally identified dominant concepts are used to
index documents and the improvement in retrieval has been experimentally studied.

3.2 PROPERTIES OF SEMANTIC SIMILARITY MEASURES

The general properties of semantic similarity perceived by Lin are universally applicable to arbitrary objects and have been theoretically justified (Lin 1998). They are not restricted to particular application domains, resources or a specific knowledge source. Generally the properties to be adhered by the similarity measures are classified into structure specific properties and retrieval specific properties. The basic properties defined by Lin include Commonality, Difference and Identity property. On the other hand in the information retrieval context Bulskov has defined the retrieval specific property, generalization (Bulskov 2006). Further he has defined structural specific properties which include depth and multiple path property. The following section elucidates on the basic properties, retrieval specific properties and structure specific properties.

3.2.1 Basic Properties

3.2.1.1 Commonality Property

The determination of similarity is based on the common characteristics possessed by the concepts being assessed. The similarity between $C_1$ and $C_2$ is based on the common characteristics shared by the concepts $C_1$ and $C_2$. The more common characteristics they share the more similar they are.

3.2.1.2 Difference Property

The determination of similarity is based on the distinct unique characteristics possessed by the concepts being assessed. The similarity between concepts $C_1$ and $C_2$ is based on distinct characteristics possessed by concepts $C_1$ and $C_2$. The more distinct characteristics they have, the less similar they are.
3.2.1.3 Identity Property

The determination of similarity is based on alike characteristics possessed by the concepts being assessed. The similarity between concepts $C_1$ and $C_2$ is based on the likeness of the characteristics possessed by concepts $C_1$ and $C_2$. The more alike characteristics they have, the more similar they are.

3.2.2 Retrieval Specific Property

3.2.2.1 Generalization Property

Tversky has mentioned that the similarity measures do not possess symmetry property (Tversky 1977). He quotes several examples mentioning that the sentences “the son is like father” and the sentence “father is like son” and “the person resembles portrait” and “portrait resembles person” are not same. Hence based on this, Bulskov defined this generalization property (Bulskov 2006). This property indicates that the similarity measure cannot be symmetrical (asymmetry) and the quantification of similarity in the direction of the inclusion (generalization) should be significantly higher than the similarity in the opposite direction of inclusion (specialization).

3.2.3 Structure Specific Properties

3.2.3.1 Depth Property

The similarity measure between two concepts $C_1$ and $C_2$ can be quantified based on the location of the concepts in taxonomy. Bulskov states that the distance represented by an edge is influenced by the depth of the location of the edge in the taxonomy (Bulskov 2006). Hence the depth property influences the degree of similarity.

3.2.3.2 Multiple Path Property

The aforementioned properties are based on the assumption that the concepts have only one super ordinate. But if the hierarchy allows multiple
inheritance then the number of paths connecting the concepts and the length of these paths influence the similarity. As most of the ontologies allow multiple inheritance, this multiple path property could not be ignored and due consideration has to be given in computing similarity between concepts.

Apart from the general properties of similarity, Lin has mentioned the general assumptions to capture the intuition of similarity (Lin 1998). The intuitions specify that the commonality and differences among concepts could capture similarity. Based on these intuitions the similarity theorem was defined by Lin.

3.3 DEFINITION OF INFORMATION CONTENT (IC)

In a multidimensional concept space, a node represents a unique concept consisting of certain amount of information. An edge connecting two concepts in a concept space represents the direct association between the two concepts. The information shared in common, quantifies the extent to which the concepts are similar. In a hierarchical concept space, the shared information is possessed by the common concept subsuming the two concepts. This super concept should be the first common concept subsuming both the concepts, upward in the hierarchy. The similarity value between the two concepts is defined as the information content value of the common concept. In information theoretic domain, generally the value of the information content of a concept is quantified by estimating the probability of its occurrence in a large text corpus.

Shannon was the first to devise a mathematical definition of the information concept (Shannon 1948). According to him, the measure of information is given in bits (binary digits). However, information according to Shannon does not relate to the qualitative nature of the data, but confines itself to one particular aspect that is of special significance for its technological transmission and storage. Shannon completely ignores whether a text is meaningful, comprehensible, correct, incorrect or meaningless.
The information content of a single item of information (an item of information in this context merely means a symbol, character, syllable, or word) is a measure of the uncertainty existing prior to its reception. Since the probability of its occurrence may only assume values between 0 and 1, the numerical value of the information content is always positive. The information content of a plurality of items of information (for example, characters) is the summation of the information content of the individual items (according to the summation condition). Based on the information theory (Ross 1976), the Information Content (IC) of a concept C can be quantified as follows.

\[ IC(C) = -\log(p(C)) \]  \hspace{1cm} (3.1)

where \( p(C) \) is the probability of encountering an instance of concept C in the test corpus assuming that the concepts are not uniformly distributed. This way of computing information content is borrowed from the information coding theory which decides how many bits would be required to code messages.

Usually the concepts are arranged as nodes in a hierarchical structure. In this hierarchical structure, a concept subsumes the concepts which are lower in the hierarchy. The information content of the concepts of a hierarchy is computed by counting the occurrence of the concept in a standard corpus. It is implicit that \( p(C) \) is monotonic as one moves up the hierarchy. As the concept node’s probability increases, its information content or its information expressiveness decreases. If there is a unique top node in the hierarchy, then its probability is 1, hence its information content is 0. The \( p(C) \) in Equation (3.1) is computed using Equation (3.2) and Equation (3.3).

\[ freq(C) = \sum_{n \in words(C)} count(n) \]  \hspace{1cm} (3.2)

where \( words(C) \) is the set of words subsumed by the concept C and the concept probabilities are computed as shown in Equation (3.3).
\[ p(C) = \frac{\text{freq}(C)}{N} \]  

(3.3)

where \( N \) is the total number of nouns in the standard test corpus. The standard Brown corpus is used by most of the researchers as the test corpus. The standard Brown corpus is discussed in Section 3.5.

As the information content is computed using probabilities, it can be integrated with many kinds of knowledge representations such as first order logic and semantic networks. The similarity measures could be applied to various domains of natural language processing such as word sense disambiguation, malapropism detection and information retrieval.

### 3.4 IC BASED SEMANTIC SIMILARITY DEFINITIONS

**Definition 1: Resnik Definition of Similarity**

Given two concepts \( C_1 \) and \( C_2 \), the similarity between the concepts \( C_1 \) and \( C_2 \) is defined as the maximum information content of the super concept which subsumes the concepts \( C_1 \) and \( C_2 \).

According to Resnik, the similarity depends on the amount of information two concepts share in common (Resnik 1995). This shared information is given by the Most Specific Common Abstraction (MSCA) concept that subsumes the concepts. If the MSCA concept does not exist, then the two concepts are maximally dissimilar, otherwise the shared information is equal to the IC value of their MSCA. Resnik’s formula is modeled as follows:

\[ \text{Sim}_{res}(C_1, C_2) = \max_{C \in S(C_1, C_2)} \{IC(C)\} \]  

(3.4)

where \( S(C_1, C_2) \) is the set of concepts that subsumes \( C_1 \) and \( C_2 \). The IC value is obtained by considering negative log likelihood

\[ IC(C) = -\log(p(C)) \]  

(3.5)
where \( C \) is a concept in the ontology and \( p(C) \) is the probability of encountering \( C \) in a standard corpus. The IC value is monotonically decreasing from the leaves of the taxonomy to its roots. In fact, the concept corresponding to the root node of the is-a hierarchy has the maximum frequency count, since it includes the frequency counts of every other concept in the hierarchy. Resnik used WordNet as a hierarchical knowledge source and the occurrences of the WordNet concepts in the standard Brown Corpus to compute the information content of the concepts using Equation (3.5).

**Definition 2:** Lin’s Definition of similarity

*Given two concepts \( C_1 \) and \( C_2 \) the similarity of concepts \( C_1 \) and \( C_2 \) is measured as a ratio of the amount of information needed to state the commonality of the concepts \( C_1 \) and \( C_2 \) and the information needed to describe fully what concepts \( C_1 \) and \( C_2 \) are.*

This is formally defined as

\[
\text{Sim}_{Lin}(C_1, C_2) = \frac{\log P(\text{common}(C_1, C_2))}{\log P(\text{description}(C_1, C_2))}
\] (3.6)

where \( \text{common} \ (C_1, C_2) \) is a proposition that states the commonality of the concepts \( C_1 \) and \( C_2 \), and \( \text{description} \ (C_1, C_2) \) is a proposition that states specific description about \( C_1 \) and \( C_2 \). In order to compute the similarity, Lin extended the Resnik method of the information content (Resnik 1995). According to the basic premise of information theory, the information content of a message is the negative log of its probability, and therefore the sum of the information content of \( C_1 \) and \( C_2 \) is \((-\log P(C_1) + -\log P(C_2))\). Lin’s measure is therefore the ratio of the information shared in common i.e. \( P(\text{common} \ (C_1, C_2)) \) to the total amount of information possessed by the concepts. \( P(\text{common}) \) is the information content of the common concept subsuming concepts \( C_1 \) and \( C_2 \).
**Definition 3: Jiang’s definition of Similarity**

Given two concepts $C_1$ and $C_2$, the semantic similarity between any such two nodes is the difference of their semantic mass (information content) if they are on the same axis. If the concepts are on different axes then the similarity of the concepts $C_1$ and $C_2$ is defined as the addition of the information content of the concepts $C_1$ and $C_2$ calculated from each node to a common concept node.

In Jiang and Conrath measure, the information content is added as decision factor to the edges connecting the concepts (Jiang and Conrath 1997). In order to compensate for the unreliability of the edge distances, Jiang and Conrath weigh each edge by associating probabilities based on corpus statistics and also consider the link strength of each concept.

In order to compute the total semantic distance between any two concepts in the taxonomy, Jiang and Conrath’s measure uses the sum of the individual distances between the nodes in the shortest path and the amount of information which the two concepts shares in common and is denoted by $Sim_{res}(C_1, C_2)$. The semantic similarity of Jiang and Conrath (Seco 2005) between any two concepts $C_1$ and $C_2$ in the taxonomy is defined as

\[
Sim_{j&c}(C_1, C_2) = \frac{IC(C_1) + IC(C_2) - 2 * Sim_{res}(C_1, C_2)}{2}
\]  

(3.7)

Where $IC$ is computed using Equation (3.1) and $Sim_{res}(C_1, C_2)$ is computed using Equation (3.4).

These similarity computations are used in matching query terms and document terms in information retrieval systems. In ontology mapping systems these are used for finding correspondences between the ontology entities. But all the above computational approaches are corpus based. As reported in the literature, the corpus may not incorporate all the concepts in the real world. If the concepts are not found in a corpus, then the information content could not be quantified. These missing concepts in the corpus are addressed as sparse data...
problem, in the literature. This sparse data problem produces adverse effects in the information retrieval systems by reducing the overall precision of the system. Moreover the semantic similarity computation is dependent on the corpus being used. If the corpus size is large, the computation of IC becomes time consuming. The sparse data problem is discussed with a well known corpus called Brown corpus in the following section.

### 3.5 Sparse Data Problem

Information content of a concept quantifies the amount of specific information it can express. Generally, the IC value of the frequently occurring word is less than the less frequently occurring word. The test corpus considered for computing information content should be adequate in content to the terms of the considered ontology. The Brown corpus and the WordNet light weight ontology of Princeton University is widely used by most of the researchers. Information content of the WordNet concepts is derived based on the Brown corpus. The Brown corpus is a collection of various generes updated from 1961 to 1979. The various Generes of Brown corpus is shown in Table 3.1. As mentioned earlier the information content based approaches compute information content value based on the frequency of occurrences of WordNet concept (noun) in Brown corpus. But parts of speech tags from the Brown corpus do not correspond to the hierarchical structure of the WordNet concepts i.e. there is no meaningful mapping between parts of speech of Brown corpus and WordNet concepts. Out of the 1,55,000 unique words, only 37,000 words exist in the Brown corpus. The significant gap of missing concepts in Brown corpus leads to generation of zero information content value for the missing concepts (nearly 75% of words of WordNet may have information content value as zero). This missing concept in the corpus which leads to zero information content value is termed as sparse data problem.
Table 3.1 Genres of Brown Corpus

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Brown Corpus Genres</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Press: reportage (Political Sports, Society, Spot News, Financial, Cultural)</td>
</tr>
<tr>
<td>2</td>
<td>Press: Editorial (Institutional Daily, Personal, Letters to the Editor)</td>
</tr>
<tr>
<td>3</td>
<td>Press: Reviews (Theatre, Books, Music, Dance)</td>
</tr>
<tr>
<td>4</td>
<td>Religion (Books, Periodicals, Tracts)</td>
</tr>
<tr>
<td>5</td>
<td>Skill and Hobbies (Books, Periodicals)</td>
</tr>
<tr>
<td>6</td>
<td>Popular Lore (Books, Periodicals)</td>
</tr>
<tr>
<td>7</td>
<td>Belles-Letters (Books, Periodicals)</td>
</tr>
<tr>
<td>8</td>
<td>Miscellaneous: US Government &amp; House Organs (Government Documents, Foundation Reports, College Catalog, Industry House organ)</td>
</tr>
<tr>
<td>9</td>
<td>Learned (Natural Sciences, Medicine, Mathematics, Social and Behavioral Sciences, Political Science, Law, Education, Humanities, Technology and Engineering)</td>
</tr>
</tbody>
</table>

Table 3.2 Missing Concepts in Brown Corpus

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>Nouns Missing in Brown Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Soccer</td>
</tr>
<tr>
<td>2</td>
<td>Fruitcake</td>
</tr>
<tr>
<td>3</td>
<td>CPU</td>
</tr>
<tr>
<td>4</td>
<td>Autograph</td>
</tr>
<tr>
<td>5</td>
<td>Serf</td>
</tr>
<tr>
<td>6</td>
<td>Slave</td>
</tr>
<tr>
<td>7</td>
<td>Hockey</td>
</tr>
<tr>
<td>8</td>
<td>World Trade Centre</td>
</tr>
</tbody>
</table>

Some of the missing concepts in Brown corpus are reported in Table 3.2 to illustrate that Brown corpus suffers from sparse data problem.
The reason for the existence of sparse data problem is shift in the vocabulary used over the past 30 years. When Brown corpus was created the term football was widely used and due to shift in vocabulary people now use the word Soccer instead of football. Similarly, the term "World Trade Centre” is not found as the WTC building was erected after a long time the Brown corpus was created.

One possible solution to handle the sparse data problem is to use a large corpus which is rich in concepts to compute information content. But, it is mentioned that the corpus larger than Brown corpus also suffers from sparse data problem. Villavicencio in her work on identification of verb particle construction used in natural language applications, reported that the British National Corpus (BNC) also suffers from data sparseness or sparse data problem (Villavicencio 2003). British National Corpus (BNC) is a 100 million word corpus containing samples of written text from a wide variety of sources. It includes texts from newspapers, journals, books, and many other sources. Apart from this the process of identifying whether a concept is present in a corpus or not is time consuming. This is because the corpus has to be tagged and this tagging itself consumes lot of time. Hence it is evident that the data sparseness or the sparse data problem should be addressed.

3.6 EFFECT OF SPARSE DATA PROBLEM IN IC BASED SEMANTIC SIMILARITY MEASURES

The effect of sparse data problem forces the corpus based (IC based) semantic similarity measures to produce a zero value. The Jiang and Conrath method and Lin method are both sensitive to this issue (Jiang and Conrath 1997) (Lin 1998). If one of the concepts has a frequency of zero, the Jiang and Conrath measure and Lin measure will return a zero. If both the concepts occur with a frequency of zero and if the frequency of the most common abstraction concept is also zero then both the IC based approaches will return a zero or infinity in the continuous domain.
Resnik produces a zero value, when the most common subsuming concept is not found in the Brown corpus. Another important problem with Resnik method is that it does not return a similarity value of 1 for the identical concepts. Resnik returns the similarity value of identical concepts as the information content value of the most common subsuming concept which is not acceptable as the similarity value of the identical concepts should be returned as 1.

Even if the corpus is adequate containing all concepts of real world, the process of computing the information content is still time consuming. Thus to address the sparse data problem, the information content computation should be corpus independent. This work defines a New Information Content (NIC) measure which quantifies the informativeness of the concept based on the relations a concept shares with other concepts when organized in an ontology or taxonomy. The following section discusses about the NIC measure.

3.7 NEW INFORMATION CONTENT (NIC) MEASURE

The New Information Content (NIC) measure is an extension of the Seco et al. information content measure (Seco et al. 2004). The NIC measure is computed based on the relations defined in the WordNet ontology. WordNet ontology is a complete lexical taxonomy where all the lexemes are represented and has a rich population of concepts due to the periodic updating. Hence it is used as the only statistical resource eliminating the need for external statistical resources like Brown corpus. Since the ontology or the taxonomy relations are used for quantification of information content the severity of the sparse data problem is expected to be minimal assuming that the taxonomy or ontology is complete and contains all the real world concepts. This method of obtaining information content values rests on the assumption that the WordNet taxonomy is organized according to the principle of cognitive saliency as stated by Seco et al.

The theory of cognitive saliency (Zavaracky 2003) states that human beings create new concepts when there is a need to differentiate from those concepts that already exist. As there is a concordance between the theory of
cognitive saliency and human way of categorizing concepts, WordNet becomes a reliable source of knowledge to define NIC.

The NIC measure uses the hyponym relations \((is-a\) relations) and meronym relations \((Part-of\) relations) defined in WordNet to quantify the informativeness (Jiang and Conrath 1997) of concepts. For example a hyponym relation could be “bicycle is a vehicle” and a meronym relation could be “a bicycle has wheels”.

This NIC measure rests on the assumption that the taxonomic structure of WordNet is organized in a meaningful and principled way. The hyponymic relations and the meronymic relations are the salient relations which are used to categorize the concepts. A concept with many hyponyms has many different features and hence tends to convey less information about the concept. This agrees well with the probabilistic based information content quantification i.e. the frequency of a concept in a corpus is inversely proportional to the expressiveness of the concept. Hence if a concept is defined with more relations, the differentiation is more and the informativeness of a concept expressed is less.

The topmost concept or the root concept in the taxonomy is the most abstract concept and hence has an information content value of 0. The root concept is further differentiated into sub concepts by the hyponym relations and the meronym relations. The information content of the concepts which are defined at the last level of the taxonomy is maximally differentiated and hence all the concepts at the leaf level of the taxonomy have an information content value of 1. In other words, the leaf node concepts are the most specific concepts defined in the taxonomy. Hence the information they express is maximal. Hence with these intuitions, the information content of a concept could be quantified by the concept relations. As in most of the taxonomies the hyponym and meronym relations are defined, it is appropriate to quantify the NIC measure using the hyponym and meronym relations. Based on this the IC measure is redefined and discussed in detail in the following section.
3.7.1 Definition of New Information Content (NIC) measure

As mentioned in the previous section the new information content is defined using the hyponym and meronym relations.

**Definition 4:** Given a concept $C$ belonging to ontology $O$, it is defined as a function of hyponym and meronym relations normalized by the maximum number of concepts in the taxonomy.

$$\text{NIC}(C) = 1 - \frac{\log(\text{hypo}(C) + \text{mero}(C) + 1))}{\log(\text{max con})} \tag{3.8}$$

The new measure is expressed by two functions namely hypo($C$) and mero($C$) which returns number of hyponyms and the number of meronyms of concept $C$ respectively. The maxcon denotes the maximum number of concepts in the taxonomy. It normalizes the NIC value and hence the NIC values fall in the range of [0..1]. Similar to the corpus based computation of IC, the information content value of the concept decreases monotonically from the leaves of the taxonomy to the root.

In contrast to the information content in Equation (3.1), the NIC computation is purely based on the concept relations defined in taxonomy and hence is corpus independent. The NIC value of the root concept is 0 as it is in the most abstract level.

The leaf node concepts of the taxonomy have the NIC value as 1 as they are maximally differentiated. The information content values of the intermediate nodes range between 0 and 1. The NIC captures semantic information of the concepts.

Though most of the knowledge sources have their concepts defined with hyponyms, some of the taxonomies do not possess meronymic relations. The other type of relations like role-of, inverse-role-of etc is defined. Hence a generalized quantification of information content is necessary for taxonomies
with diversified relations. Therefore the New Information Content measure (NIC) is generalized and quantified by considering all the relations defined in the taxonomy and is given below.

\[
GNIC(C) = 1 - \frac{\log \left( \sum r_i + 1 \right)}{\log \text{max con}}
\]

(3.9)

where \( r_i \) refers to number of the distinct relations defined in the taxonomy.

Assuming that the taxonomy is complete, the strength of the \( GNIC \) measure is that, the information content value of a concept will rarely become zero and hence it will serve as a better solution to the sparse data problem. It also guarantees that any concept included in a hierarchy/lexicon will have a value for the information and rarely the information content of concept may become zero.

The effect of the NIC measure could be appreciated only when this measure is used in the above discussed information content based semantic similarity measures Resnik, Lin and Jiang and Conrath methods. Based on NIC, these methods are redefined and the redefined NIC based similarity measures are discussed in the following section.

3.8 NIC BASED SIMILARITY MEASURES

The definitions by Resnik, Jiang and Conrath and Lin for the similarity measures are redefined in order to incorporate the concept relations in the information content computation.

**Definition 5: NIC Resnik Definition of Similarity**

*Given two concepts \( C_1 \) and \( C_2 \), the similarity between the concepts \( C_1 \) and \( C_2 \) is defined as the maximum information content of the super concept which subsumes the concepts \( C_1 \) and \( C_2 \) where the information content of the super concept is quantified as the function of the concept relations scaled by the maximum number of concepts in the taxonomy.*
The Resnik measure has been redefined by using the NIC measure and is given below.

\[ Sim_{\text{Resnik}}(C_1, C_2) = \max_{\text{c\textit{super}(C_1, C_2)}}(\text{NIC}(C)) \]  

(3.10)

where \( C \) is the most common abstraction concept or the common concept which subsumes the concepts \( C_1 \) and \( C_2 \).

**Definition 6: NIC Jiang & Conrath Definition of Similarity**

Given two concepts \( C_1 \) and \( C_2 \), the semantic similarity between any such two nodes is the difference of their information if they are on the same axis in a conceptual space. If the concepts are on different axes then the similarity of the concepts \( C_1 \) and \( C_2 \) is defined as the addition of the information content of the concepts \( C_1 \) and \( C_2 \) calculated from each node to a common node. The information content of the concepts is computed as a function of concept relations scaled by the maximum number of concepts in the taxonomy.

The Jiang and Conrath (NIC\(_{\text{J&C}}\)) measure is redefined by using the NIC measure as follows.

\[ Sim_{\text{NIC}_{\text{J&C}}}(C_1, C_2) = \frac{(\text{NIC}(C_1)) + (\text{NIC}(C_2)) - (2 \times Sim_{\text{Resnik}}(C_1, C_2))}{2} \]  

(3.11)

where \( \text{NIC}(C_1) \) and \( \text{NIC}(C_2) \) is computed using Equation (3.8) and \( Sim_{\text{NIC}_{\text{J&C}}} \) is computed using Equation (3.10).

Similarly the Lin similarity measure is redefined as follows.

**Definition 7: NIC Lin Definition of Similarity**

Given two concepts \( C_1 \) and \( C_2 \) the similarity of concepts \( C_1 \) and \( C_2 \) is measured as a ratio of the amount of information needed to state the
commonality of the concepts $C_1$ and $C_2$ and the information needed to describe fully what concepts $C_1$ and $C_2$ are. The commonality and the description of each concept are quantified as a function of concept relations scaled by the maximum number of concepts in the taxonomy.

\[
Sim_{NIC_{Lin}}(C_1, C_2) = \frac{(2 \times Sim_{ResTran}(C_1, C_2))}{(NIC(C_1) + NIC(C_2))}
\]  

(3.12)

In this work, the similarity measures for the redefined information content based approaches $Sim_{NIC_{ResTran}}$, $Sim_{ NIC_{IC}}$, $Sim_{NIC_{Lin}}$ were evaluated using the standard R&G dataset (Appendix I). The computational values were correlated against human judgements. The results of the NIC based approaches were compared against IC based approaches and the results are reported in Chapter 5. As mentioned earlier the effect of these similarity measures are evaluated for information retrieval tasks. The results of this experiment are also reported in Chapter 5.

So far the information content based approaches have been used to quantify the information content of a concept. The next section discusses how these similarity measures could be used for quantifying the representativeness of the documents.

3.9 INFORMATION CONTENT BASED APPROACHES FOR SELECTION OF DOMINANT CONCEPTS

In information retrieval processes, the major problem is to determine the specific content of the documents. In traditional information retrieval (IR) systems the keywords of the documents are used for retrieving the documents. The keywords do not capture the semantic content of the document. As the semantic content is not reflected by the keywords the retrieval performance of the IR systems degrades due to retrieval of less precise and irrelevant documents. Hence an alternate and better approach is required to identify the key concepts reflecting the semantic content of the documents.
One way of determining the specific content of the document is by annotating the pages with semantic tags (Fensel et al. 1998) (Benjamins et al. 1998) (Luke et al. 1996) (Martin and Eklund 1999). Though this could help to improve the retrieval effectiveness, frequent modifications or new annotations are to be done when the documents get modified. Moreover these annotations are carried out manually or semi automatically (Kahan et al. 2001). Further this process is time consuming and could be carried out only by domain specialists Heflin et al. , and Sowa describes an annotation process using ontology, where knowledge is represented as conceptual graphs (Heflin et al. 1999) (Sowa 1984). O’Hara et al. use hierarchical heuristics to semi automatically disambiguate word senses (O’Hara et al. 1998). Lin uses automatically extracted terms to characterize the documents (Lin 1998). Loh et al. propose a combined approach mixing the natural language techniques and knowledge discovery techniques to extract automatically the information from the documents (Loh et al. 2000).

As an alternate approach to semi automatic or manual annotation of documents in determining the specific content of the documents, Desmontils and Jacquin have proposed a method to create a content based index to retrieve documents (Desmontils and Jacquin 2002). In the process of building the content based index, Desmontils utilizes path based semantic similarity measure. He has also mentioned that the information content based semantic similarity approaches may quantify the relative importance of a concept in a page in a better and efficient way. Based on this context, an attempt has been made in this thesis to build content based index utilizing the information content based similarity measures. The index terms of the content based index are the dominant concepts representing the content of the document. The dominant or candidate concepts are those concepts which have strong relationships with other concepts of the documents. The candidate concepts of a document is determined or identified by the information content based similarity measures.

The semantic content of the document is captured by computing the representativeness of the concepts in the documents. The representativeness of
the concept quantifies the relative importance of the concept to the document. The representativeness is computed by considering the statistical factor and the semantic importance of the concept. The dominance of the concept in a document is determined by three factors. viz., the context in which the concept occurs in the document, the number of occurrences of the concept in the document and the relationships the concept has with the other concepts of the documents. Hence this representativeness attributes the statistical factor by considering the frequency of the concept in a particular context in the document (weighted frequency) and the semantic importance by capturing the similarity of the concept with other concepts in the document (cumulative similarity). The semantic importance of the concept is quantified using the similarity measures. The weighted frequency and the representativeness of the concept are computed as given in Equation (3.13) and Equation (3.15).

**Definition 8: Weighted frequency**

The weighted frequency $F(C_i)$ of the concept $C$ is defined as the normalized sum of the weights of the different contexts $(M)$ associated with $P$ concept occurrences and as given below.

$$F(C_i) = \frac{P(C_i)}{\max_{k=1..n} P(C_k)}$$  \hspace{1cm} \text{(3.13)}$$

$$P(C_i) = \sum_{j=1}^{p} (M_{i,j})$$  \hspace{1cm} \text{(3.14)}$$

where $M_{i,j}$ is the context weight associated with the $j$th occurrence of the concept $C_i$.

**Definition 9: Representativeness**

The Representativeness $R$ of the concept $C$ belonging to a document is a linear combination of the weighted frequency $F(C_i)$ of the concept $C$ and the cumulative similarity $\text{csim}(C_i)$ of the concept.
$R(C_i) = \frac{\alpha F(C_i) + \beta \text{csim}(C_i)}{\alpha + \beta}$ \hspace{1cm} (3.15)

where $\alpha$ and $\beta$ are constants and $\text{csim} (C_i)$ is the cumulative sum of the similarity values of the concept $C_i$ with every other concept $C_j \ (j \neq i)$ in the documents. In this work the cumulative similarity is computed using the new information content based approaches $\text{Sim}_{\text{nac}_{\text{wc,unk}}}$, $\text{Sim}_{\text{nac}_{\text{wc}}}$, $\text{Sim}_{\text{nac}_{\text{ic}}}$.

The cumulative similarity for each concept of the document is computed. The concepts having the highest cumulative scores are considered to be the dominant concepts and they are the representative terms of the documents. These dominant concepts are identified by the new information content based measures and the same is used to build content based index.

In this work, the redefined similarity measures have been implemented in information retrieval domain. The dominant concepts computationally identified representing the documents are used as index terms to retrieve the documents.

The following section outlines the algorithms for mapping of documents to ontologies ($\text{Ontomap}$), computing similarity ($\text{computesim}$) and for generation of dominant concepts ($\text{Domgen}$).

In the identification of the dominant concepts, the context in which a concept occurs is not ignored. For each context a different weight is assigned and based on the number of occurrences of the concept in a particular context, the weighted frequency of the concept is computed.
// Algorithm Ontomap - Mapping documents into set of ontologies

Construct a word list $W$ (dominant concepts) from a set of documents $D$.

$W : \{ w \mid w \text{ in document} \}$

// loop through all words in the document that is indexed

for $i = 1$ to $|W|$

// Creation of concept list from multiple ontologies

for $k = 1$ to $|O|$

$L = \{ c \in O \}$

for $m = 1$ to $|L|$

// concepts $c$, where the stem of the concept name or the stem of a synonym of the concept equals the stem of the word

if $(w = \text{name}(c) \text{ or } w = \text{synonyms}(c)$ then

$C = C \cup \{ c \}$

next $m$

next $k$

next $i$

Figure 3.1 Algorithm Ontomap
For the identification of the dominant concepts the documents are preprocessed. The preprocessing includes the stop words removal, lemmatization and extraction of well formed terms. The weighted frequencies of the terms are computed. The algorithm *Ontomap* outlines the steps required to build the mapping of the ontological concepts to the concepts of the documents.

The *computesim* algorithm is used to evaluate the similarity values computed using the new information content based measures. The algorithm Domgen outlines the procedure for identifying the dominant concepts of the documents using new information content similarity measures. The *Ontomap* algorithm builds the content based index by mapping the dominant concepts (identified through dominant concept algorithm) to the concepts of the ontology. Further this content based index was used for retrieval of documents. The discussion of the results and performance improvement in the information retrieval task is reported in Chapter 5 which elaborately discusses the evaluation of the effectiveness of the similarity measures.

This Chapter discussed the basic properties of the similarity measures, and the corpus based information content based approaches. Further the sparse data problem prevalent in corpus based information content approaches was discussed. A new information content measure which is corpus independent was proposed. The corpus based similarity approaches Resnik, Lin and Jiang and Conrath approaches were redefined using the proposed new information content measure.

The methodology for using the NIC based similarity measures for capturing the semantic content of the documents was also discussed. The content based index thus created was used for improving the precision of the information retrieval tasks. The experimental evaluation of the proposed NIC measure is discussed in Chapter 5.
Algorithm computesim (C1, C2)
// Computation of similarity for concept pairs of standard data set or terms pairs from the documents
For each concept C of the Concept pair
{
    Extract relations (Hyponyms, meronyms) of the concept using an external knowledge source (/*WordNet is used*/)
    // compute new information content (NIC(c))
    \[ NIC(C) = 1 - \frac{\log(\text{hyp}(c) + \text{mer}(c))}{\log(\text{max}\ con)} \]
    For relations other than hyponym and meronyms
    // Compute new information content using GNIC
    \[ GNIC(C) = 1 - \frac{\sum_{i=1}^{n} (r_i + 1)}{\log(\text{max}\ con)} \]
} End For
For each Concept pair (c1, c2) in the standard data set or document
{
    // Compute similarity of concept pairs c1 and c2 by Resnik method
    \[ Sim_{\text{Resnik}}(C_1, C_2) = \max_{c_{\text{comp}} \in \{C_1, C_2\}} (NIC(C)) \]
    // Compute similarity of concept pairs c1 and c2 by Lin method
    \[ Sim_{\text{Lin}}(C_1, C_2) = \frac{2 \cdot Sim_{\text{Resnik}}(C_1, C_2)}{(NIC(C_1) + NIC(C_2))} \]
    // Compute similarity of concept pairs c1 and c2 by Jiang and Conrath method
    \[ Sim_{\text{J&C}}(C_1, C_2) = \frac{(NIC(C_1)) + (NIC(C_2)) - (2 \cdot Sim_{\text{Resnik}}(C_1, C_2))}{2} \]
} End For
// Collecting human judgements
{ for each concept pair of the standard R&G data set, collect human judgements and for each document dominant concepts collect human judgements}
Check human judgement integrity
Eliminate human judgements of the concept pairs which are irrelevant
Compute correlation between human judgements and computational approaches NIC_{Resnik}, NIC_{Lin} and NIC_{J&C}
End

Figure 3.2 Algorithm Computesim
Algorithm Domgen // computational approaches to identify dominant concepts in the documents

// D – collection of documents in a particular domain

For each document Di in the collection

Preprocess (Di) // Remove stop word Terms from the Documents
Lemmatize (Di) // reduce Terms to their root word
Posttag(Di) // Annotate Document with parts of speech tags
Extract concepts from documents
For each concept of the Documents

Compute weighted frequency $F(C_i)$

$$F(C_i) = \frac{P(C_i)}{\max_{k=1..n} P(C_k)}$$

$$P(C_i) = \sum_{i=1}^{p} (M_{i,j})$$

End for

For each concept of the document

Compute cumulative similarity (csim) using Algorithm computesim

End for

For each concept C of the Document

// Compute Representativeness R(C) of the concept where

$\alpha = \{1, 2\}$ and $\beta = 1$

$$R(C_i) = \frac{(\alpha F(C_i) + \beta csim(C_i))}{(\alpha + \beta)}$$

End for

Eliminate concepts having representativeness lesser than threshold
Assign remaining concepts as representative dominant concepts of the Document.

End for

End

Figure 3.3 Algorithm Domgen

The proposed cross ontology semantic similarity measures are elaborately discussed in the next Chapter.