IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 Volume 2, Issue 8 (August 2012), PP 21-32
www.iosrjen.org

Associations among entities related to Alzheimer Disease using Ontology based approach

R.Porkodi[1] , Dr.B.LShivakumar[2]
[1]*Assistant Professor, Department of Computer Science, Bharathair University, Coimbatore, Tamilnadu, India*
[2]*Associate Professor, Department of Computer Applications,SriRamakrishna Engineering College, Coimbatore Tamilnadu, India, porkodi_r76@yahoo.co.in*[1], *blshiva@yahoo.com*[2]

Abstract: - *Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases whereas Text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. The research aspect in Text mining includes Information Extraction, Information Search and Retrieval, Text Document Categorization (Supervised Classification), Text Document Clustering (Unsupervised Classification)and Text Summarization. This research work focuses on Information Extraction and mining of Association rules in which the Information Extraction is used to identify and extract textual information of specific types of entities and their relationships, and the mining of Association rules is nothing but extraction of semantic knowledge among entities that are extracted by Information Extraction phase. The goal of our research is to design information extraction approach to extract entities from Medline abstracts and to mine association rules among extracted entities. This work uses Bioinformatics databases Gene Ontology, MeSH ontology and Medline abstracts related to a particular interest particularly abstracts related to Alzheimer Disease.*

Keywords: *Information Extraction, Gene name, Protein name, Regular Expression, Medline abstracts, Ontology, Alzheimer disease.*

I. INTRODUCTION

The extraction of desired information directly from biological literature is a challenging problem in text mining and Natural Language Processing (NLP). Many biomedical information sources have been developed and used in extraction process. Most of the text mining methods use vector space model to represent a document and the vector space model represents documents as feature vector of terms that contained in it. Each feature vector contains term weights and the similarity among documents is computed using various similarity measures. The vector space model not considered the semantic relations of terms in documents. The ontology approach represents an effective knowledge representation within controlled vocabulary. The Wordnet ontology[1] is a lexical database for general English covering most of the general English concepts. In biomedical domain, the Unified Medical Language System (UMLS) framework [2] includes much biomedical ontologies.

This proposed approach in this chapter integrates the two popular ontologies Gene Ontology (GO) [3] and Medical Sub Heading (MeSH) [4] by mapping with Gene details used to extract significant associations among entities from Medline abstracts. The main objective of the integration of two ontologies is to provide better semantic relations among the entities in Medline abstracts. The gene products or GO annotation terms of genes are referred from GO in order to find out the associations of gene products related to genes of a particular interest. The integrated ontology consists of four main concepts that are

- MeSH concepts
- Human Genes which includes Alzheimer disease
- GO annotation terms of genes
- Proteins that causes Alzheimer disease

The mapping between above concept includes

- Mapping between MeSH and Gene concepts
- Mapping between MeSH and Protein concepts
- Mapping between Gene and GO annotation concepts

These mapping are transitive and inverse, thus it provides associations in any aspect not only from direct mapping.

Alzheimer's disease (AD) is the most common cause of progressive decline of cognitive function in aged humans, and it is characterized by the presence of numerous senile plaques and neurofibrillary tangles accompanied by neuronal loss. The Medline abstracts in data set also contain abstracts that related to Alzheimer disease. This approach used to provide semantic associations among GO annotation terms, genes, proteins, etc. with additional set of relations which are not provided by the vector based approach.

The mapping of concepts in the proposed ontology is specified in such a way that the MeSH concepts related to Alzheimer disease are linked with Genes that causes the disease and also with the possible proteins that are responsible for Alzheimer disease, and then gene concepts are mapped with corresponding GO annotations in order to identify functionality of genes i.e., the GO terms are helpful in the identification of exact biological processes that consists of set of molecular functions which are activated in a particular place or cellular component in a human body are responsible for inhibiting the particular disease.

This approach is used to discover significant meaningful associations or relationships between proteins, genes and GO annoatations of genes related to Alzheimer disease. The integrated biomedical ontology has been validated in all three aspects such as structural, syntactic and semantic validation measures. This approach proves that ontology approach is the best ever and provides better semantic relationships among entities than the vector based approach.

The integrated ontology is developed in protégé tool which is the famous tool for designing ontology. The protégé tool provides facilities such as visualization of concepts which clearly show the semantic relations of a concept, query some results based on concepts, object properties and data properties, etc.

This approach overcomes the limitations in the first approach in which the first approach did not provide better semantic relationships and had ambiguity in extracted information. For example, the genes and proteins in Medline abstracts may have synonyms and those synonyms may appear in any other Medline abstract in our data set. If this is the case, the first approach fails to treat both are same, it treats both genes are different genes and may not produce meaningful associations. In addition to that the ontology based approach provides many relations among entities such as synonyms information about the particular gene/protein, GO annotation of particular gene, type of gene, list of applicable members of gene, gene identifier, other database references of gene, chromosome information, equivalent gene, super class of a particular gene class or concept, etc. This information is used to express relationships or semantic similarities among concepts or entities. Thus this approach provides better and meaningful semantic relations among entities.

The integrated ontology is developed to accumulate information from different repositories for better information extraction. The approach integrates the popular two ontologies GO ontology and MeSH ontology, in addition to that, gene details are added to integrated ontology and linking between GO, MeSH and gene is established using set of object properties and data properties. This ontology extracts semantic meanings and relations among entities in a better way than the vector based approach.

The paper is organized as follows: Review of literature relate to this work is presented in section 2. In section 3, the ontology based framework is elaborately discussed. The experimental design and results discussion is presented in section 4. Section 5 discusses on validating the proposed ontology and finally, this paper is concluded in section 6.

II. RELATED WORK

A lot of NLP based works have been reported for the past decades related to concept extraction [5], association rule discovery [6, 7] and extracting relationships among various concepts [8, 9]. Many approaches have been developed for extracting significant associations and interactions among various biological entities [8, 9, and 10] and discovering protein-disease associations. However, these approaches have not been produced promising results, due to inconsistencies prevailed in gene names. Related to gene names extractions, paper [11] has presented the extraction of gene names from articles' titles and abstracts and identified genes related to colon cancer disease. The paper [8] has presented a statistical approach for discovering group of genes related to breast cancer disease. In paper [9], author constructed a relationships network among biomedical entities which are extracted from Medline abstracts.

In paper [12] the authors proposed new text mining approach which utilizes the concept of expectation, evidence a Z-score in determining significant associations between genes and Alzheimer disease. In paper [13], researchers expressed the method using association and functional relationship discovery algorithm in extracting gene relations from Medline abstracts.

Recent works have been reported that ontology is a useful tool to improve the performance of any text mining tasks such as text clustering and association rule mining. In text clustering the paper [14] uses conceptual features that are extracted from text using ontology and prove that ontology could improve the performance of text clustering. The paper [15] shows the case study on the integration of biomedical information in to ontology. In paper [16] author proposed bio ontology methodology and compared this with other bio-ontologies. The limitations and benefits of GO ontology are expressed in paper [17]. The author had studied the strength and limitation of biomedicine ontologies based on its text and concept representation [18]

This paper presents a new ontology that integrates the famous two ontologies such as Gene Ontology (GO) and MeSH by adding of semantic mappings or relations between GO terms, Gene names and MESH keywords related to a particular disease (Alzheimer disease). Finally the integrated ontology has validated based on syntactic, structural and semantic validation measures in order to prove its correctness and validity.

III. ASSOCIATIONS AMONG ENTITIES USING ONTOLOGY BASED APPROACH

The framework proposed in this chapter is shown in Figure 1 that integrates the two popular ontologies GO and MeSH by mapping with Gene details to extract significant associations among concepts from Medline abstracts. The main objective of the integration of two ontologies is to provide semantic relations among concepts related to genes in Medline abstracts. The GO annotations of genes are referred from GO to find out the associations of gene products. The integrated ontology consists of MeSH concepts related to Alzheimer disease, linking of Alzheimer disease MeSH concepts to proteins that cause this Alzheimer disease, linking of Alzheimer disease proteins to genes that inhibit Alzheimer disease and finally linking of Alzheimer disease genes to respective GO annotations in which we identify the exact biological processes that consists of set of molecular functions which are taking place in a cell or cellular component in which the disease can be found.

The proposed framework presented in this chapter uses many functions such as tokenization, stop words removal, stemming, extraction of entities, concept checker, concept adder, etc. These functions play an important role in the proposed framework. The above specified functions are explained in subsequent section and the pseudo code for the proposed research in this chapter is given below.

A. Preprocessing

The preprocessing is the first step to transform Medline abstracts, which typically are strings of characters into a suitable representation.

1. Tokenization

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.

2. Stop words and Verbs Removal

The Medline abstracts contains protein names, gene symbols and other words. The other words that contains stop words and verbs are removed by using the stop word list and verb list downloaded from NCBI. The token string is the combination of all the stop words, verbs, protein names, gene symbols and other words. From the token string we removed the stop words and verbs by using the stop word and verb list. After removal of stop words and verbs we ended up with tokens.

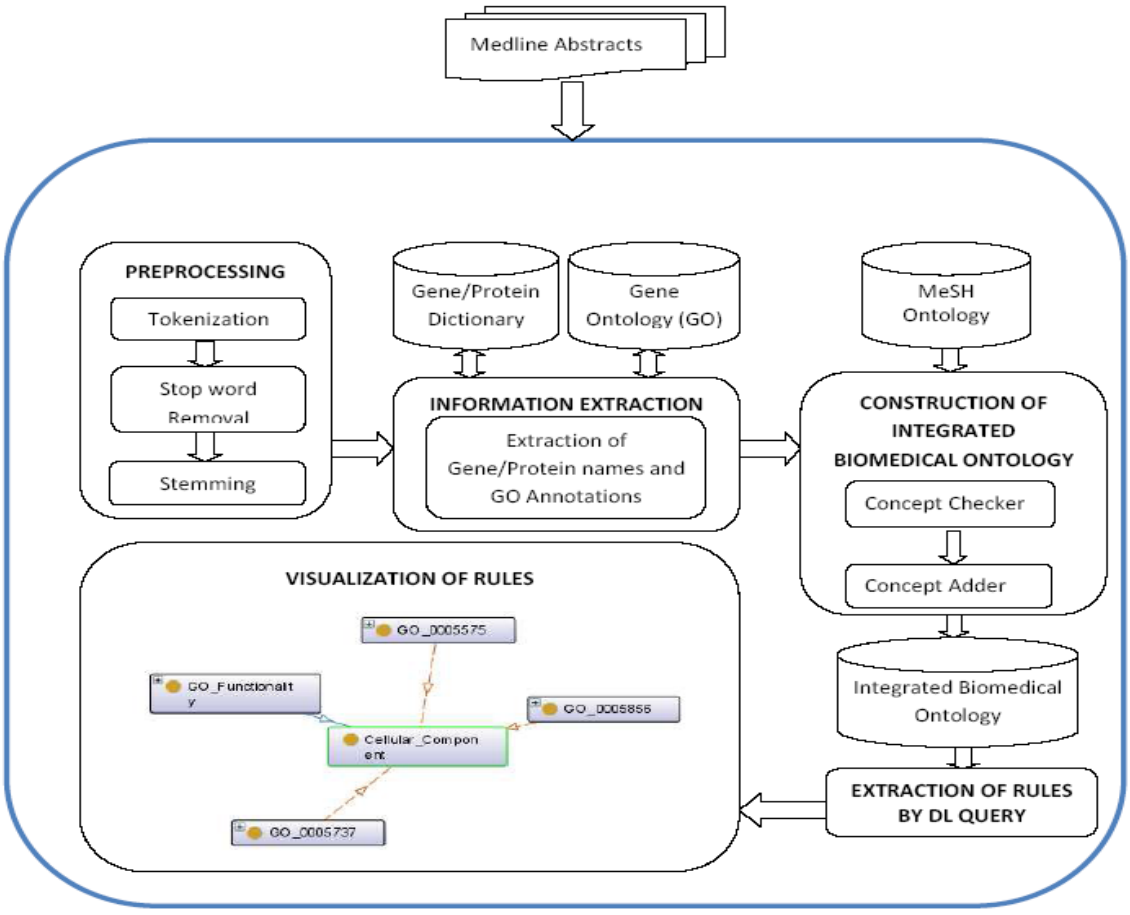

Figure 1 Associations among Entities using Ontology based Approach

3. Stemming

Stemming means the process of suffix removal to generate word stems which is used to improve the extraction process.

```
Procedure Entity_Associations_Ontology( )
// extract associations among entities in Medline abstracts using Ontology based approach
Let MA_Dataset be the list of Medline abstracts of particular interest
Let stop_words[ ] be list of stop words
Let verb_words[ ] be list of verbs
Initialize ontology with main concept Thing
i ←0
For each abs in MA_Dataset do
        Str_token[ ] ←Tokenization(abs)
        //remove stop words and verbs from set of tokens
        Str_token_arr [ ]← Remove(Str_token, stop_word, very_words)
        Train the str_token_arr using GP_Dictionary and n-gram approach
        concept_set ←extract entities using above step
        concept_arr ←distinct(concept_set)
        //add
        For each concept in concept_arr do
                If conept is a Gene then
                        //Check the concept is in Gene main concept of ontology
                        If  !(concept € ontology) then
                                //check the concept Gene occurs as synonyms in any other
                        //concept in the ontology
                                If  !(checkforsynonyms(concept)) then
                                //add concept into Gene main concept of ontology
                                Ontology ←concept
                                //compute GO annotations for entity
                                GO_annotation_arr[ ] ← GO(Gene)
                                For each go in GO_annotation_arr do
                                        Ontology ←go
                                         GeneGOMapping(concept, go)
                                end for each
                        end if
                                end if
                else if concept is a Protein then
                        If  !(concept € ontology) then
                                Ontology ←concept
                                GeneProteinMapping(concept)
                        end if
                end if
        end for each
add other relevant details specified in section 6.4
using set of object properties and data properties provide mapping among concepts
end for each
Visualize associations using Ontgraf and DLquery
End procedure
```

Pseudo code for Association among entities using Ontology Approach

IV. CONSTRUCTION OF INTEGRATED ONTOLOGY

A. Concept Checker and Adder

The proposed ontology consist of four different category of information that are

- Gene names with all associated details
- GO annotations for Genes
- Protein names
- MeSH Terms related to Alzheimer Disease

The Gene/Protein names collected from the previous phase are checked by Concept checker module for its availability. Two types of checking are done by this module;

- Direct searching in which the Gene/Protein name is checked with the concepts that are available.
- If it is not available, indirect checking is performed in which the synonyms of the given concepts are checked to identify whether the current entity is a synonym of already existing concept.

The current entity is the synonym of any concept; do not include it into ontology. The current entity is not found in ontology either as direct or indirect there is no need to add it into ontology. If it is not available add it into ontology based on its category. The concepts or classes added in this integrated ontology are shown in Table 1 and the same is created in protégé tool is shown in Figure 2. The MeSH terms related to Alzheimer disease are collected form information extraction phase and added to proposed ontology as concepts based on it hierarchy as shown in Figure 3.

B. Semantic analysis & Concept mapping

After adding concepts into ontology, check whether the currently added concept is gene or protein, if it is either gene or protein then find out the equivalent concept. This equivalence is computed from the details which are available in NCBI source. If the equivalent concept is found in the ontology, map the currently added concept and equivalent concept using equivalent relation. Some of the other relations can also be implemented for the currently added concepts using data properties and object properties. Another checking is to find out the ancestors GO terms from Gene Ontology and add all those into ontology under GO annotation category if the current concept is GO and also add all relevant information to it using data properties and object properties. Some of the important object properties and data properties created in the integrated ontology shown in Table 6.3 and 6.4. Finally, the mappings among Gene, Protein, GO annotations and MeSH terms are established using the data properties and object properties mentioned in Table 2 and 3. For example, the particular Gene concept is mapped with its associated GO terms by 'has_go object property. Similarly all concepts are mapped.

Table 1 Main Concepts in Proposed Ontology

Class/Concept	***Description***
GO_0003674	*This is the base class for all molecular function. All molecular functions are the subclass of GO_0003674 and the respective molecular function class for a gene is mapped with a particular gene.*
GO_0005575	*This is the base class for all cellular components. All cellular components are the subclass of GO_0005575 and the respective cellular component for a gene is mapped with a particular gene.*
GO_0008150	*This is the base class for all biological process. All biological process are the subclass of GO_0008150 and the respective biological process for a gene is mapped with a particular gene.*
GO_Functionality	*This class specifies the three GO functionalities as class such as cellular component, molecular function and biological process and these classes are mapped with the above three classes.*
Gene	*This specifies all human genes.*
Gene_Type	*This specifies the gene type for a gene such as protein coding, pseudo coding and unknown.*
MeSH	*This specifies the MeSH keywords that include proteins, disease level, etc. for Alzheimer disease.*

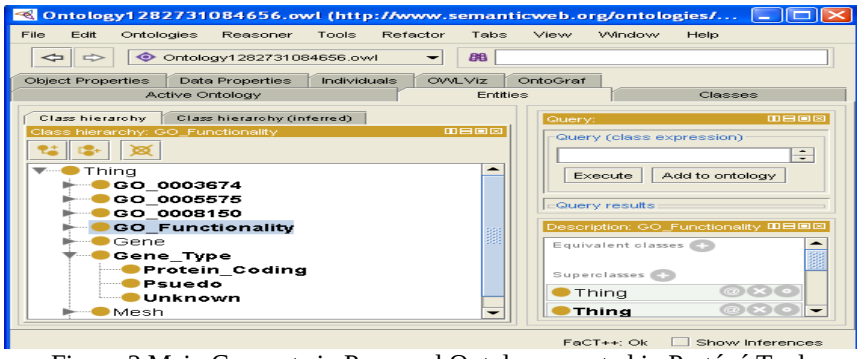

Figure 2 Main Concepts in Proposed Ontology created in Protégé Tool

Table 2 List of Object Properties in Proposed Ontology

Object properties	***Description***
belongs_to	*This property is used to map GO classes.*
curated_GO_References	*This property is used to map genes referred in Pub Med literature*
has_gene	*This is used to relate gene with GO*
has_genetype_as	*The gene types protein coding, pseudo coding and unknown is mapped with genes.*
has_go	*This is used to all applicable GO concepts are mapped with gene at all functional level.*
has_inducing_protien	*This is used to map disease with respective proteins.*
has _synonym	*This is used to specify all possible synonyms for a gene.*
Inhibits	*This property is used to map gene with disease.*
is_found_in	*This is used to map disease with gene and it has transitive relationship with inhibits property.*

Table 3 List of Data Properties in Proposed Ontology

Data Properties	***Description***
gene_annotations	*This property is used to specify different data base reference for a particular gene such as ENSENBL, HNGC, HPRD, MIM and UNIPROT.*
gene_descriptions	*This is used to specify the alternate name and full name for gene.*
has_gene_id	*This is used to specify the gene identifier for gene*
is_in_chromosome	*This property is used to specify the chromosome map location and chromosome number.*

The above mentioned object properties and data properties are created and assigned to concepts in the integrated biomedical ontology are specified in protégé tool shown in Figure 3 and.4. The object properties are properties that have value such as only, some, etc. The data properties are properties that have values of type string, char, int and float.

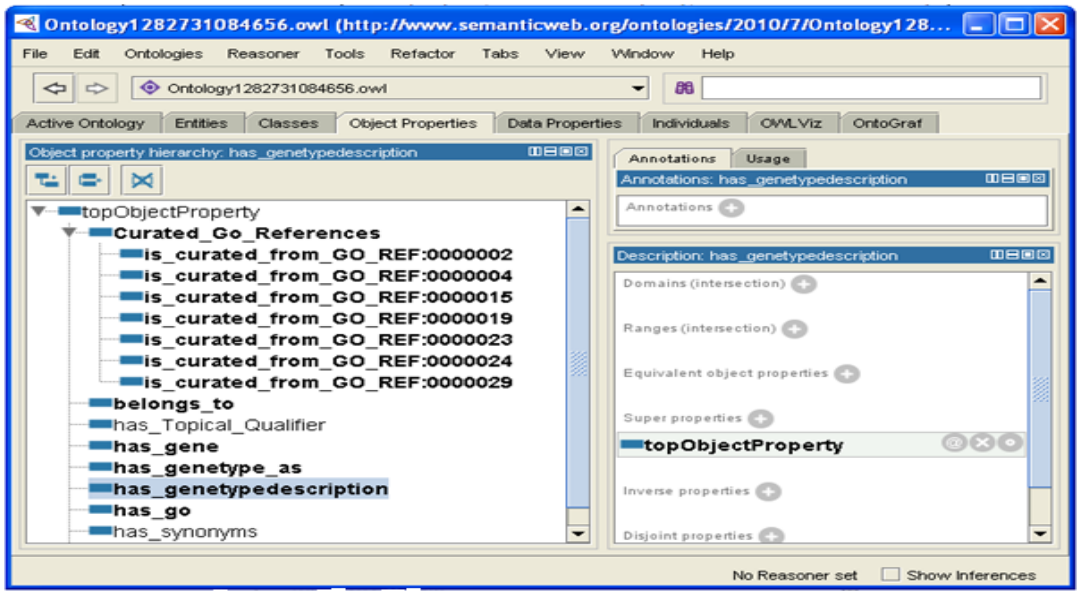

Figure 3 Object Properties created in Protégé Tool

IV. EXTRACTIONS OF ASSOCIATIONS & RESULTS AND DISCUSSION

The integrated ontology consists of three different main concepts or classes which are GO term functionalities, all human genes with or without related to a particular disease (in this ontology all human genes with or without related to Alzheimer disease are considered) and MeSH terms related to a particular disease. The sub classes for GO term functionalities class are all possible GO functionalities for the genes that are added into the ontology and the Gene main class consists of all human genes with or without related to a particular disease. The subclasses created to MeSH class includes all disease branches in which Alzheimer disease is derived, amino acids, peptides and proteins related to a particular disease is shown in Figure 5.

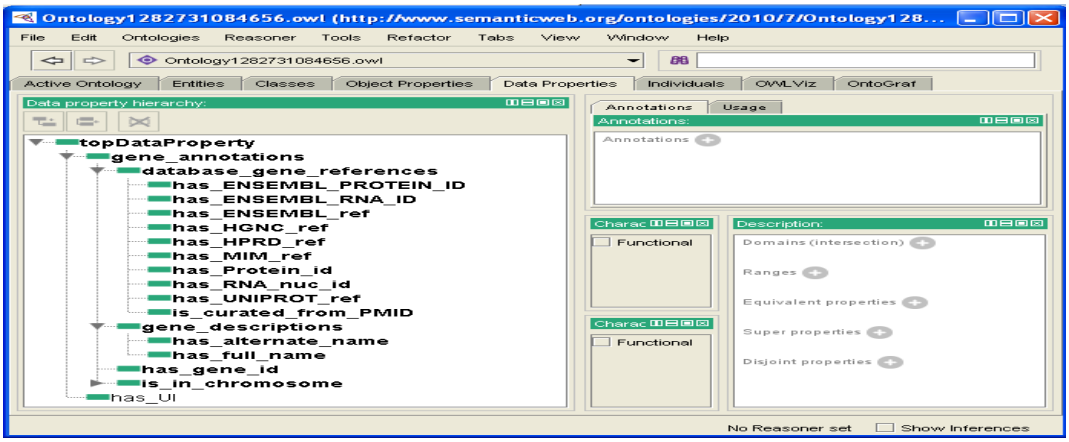

Figure 4 Data Properties created in Protégé Tool

The integrated ontology provides all types of information related to genes, protein, GO annotations and other details are called direct mapping and also provides association through indirect mapping. The indirect mapping provides associations such as relations between GO annotation terms with its ancestors, it descendents and also extract association using transitive mapping and inverse mapping. The ontology can be manipulated in different ways in which the most important manipulation techniques are using OntGraf and DL query. The structural evaluation is necessary for ontology to verify the consistency, if it is not structurally evaluated, it may produce some wrong or inconsistent results when information from the ontology is extracted. The visualization of concepts with its semantic relations is experimented using OntGraf tool.

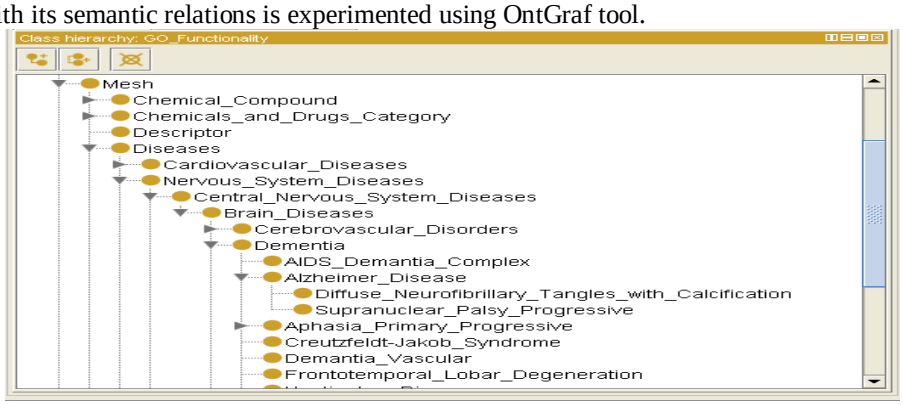

Figure 5 Mesh Hierarchies in Integrated Biomedical Ontology

The DL Query is a powerful and easy-to-use feature for extracting information from a classified ontology. The query language (class expression) is based on the Manchester OWL syntax, a user friendly syntax for OWL DL that is fundamentally based on collecting all information about a particular class, property, or individual into a single constructed, called a frame. The DL query returns the query result in form of classes with its associated information such as super classes, ancestor classes, equivalent classes, descendant classes, subclasses and individuals. This is an effective tool to retrieve any kind of semantic related information from the given ontology. Some of the information retrieval queries and results are shown in below figures. The Figure 6 shows the extraction of human genes related to Alzheimer Disease using the DL query.

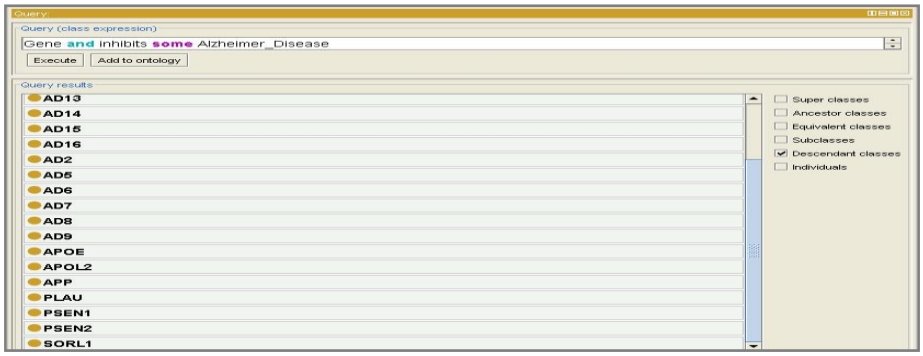

Figure 6 Genes Extracted by DL query "Gene and inhibits some Alzheimer_Disease"

The Figure 7 shows the extraction of protein names that inducing Alzheimer disease. In the screenshot the terms Proteins and Alzheimer_Disease are concepts in the ontology and these two classes are mapped with the object property '*has_inducing_disease'*. The result of this query extracted all proteins that are inducing Alzheimer disease and that also extract information such as super classes, ancestor classes, equivalent classes, descendant classes, subclasses and individuals of the extracted proteins.

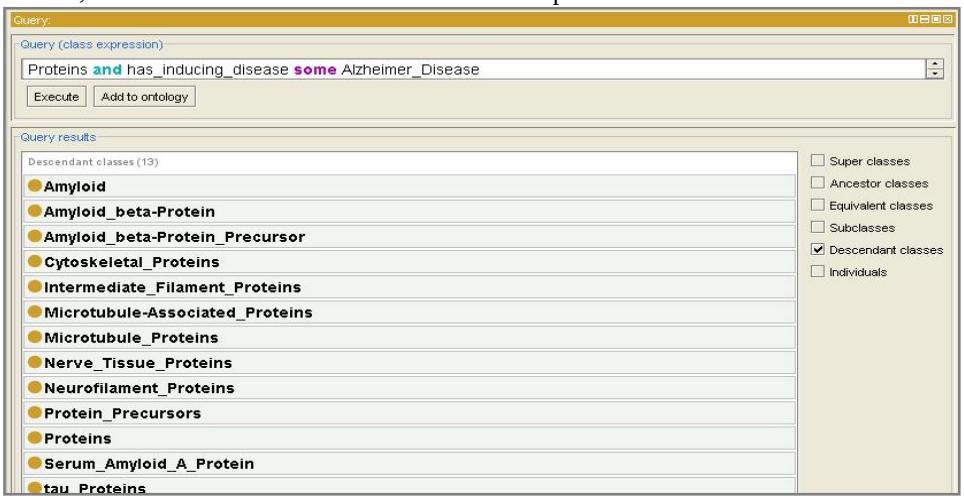

Figure 7 Proteins Extracted by DL query "Proteins and has_inducing_disease some Alzheimer_Disease

Finally the Figure 8 shows the extraction of human genes that inhibits Alzheimer disease in particular chromosome level, in our data set there are 3 genes (gene identifiers mapped with genes are shown) that inhibits Alzheimer disease in chromosome level "10".

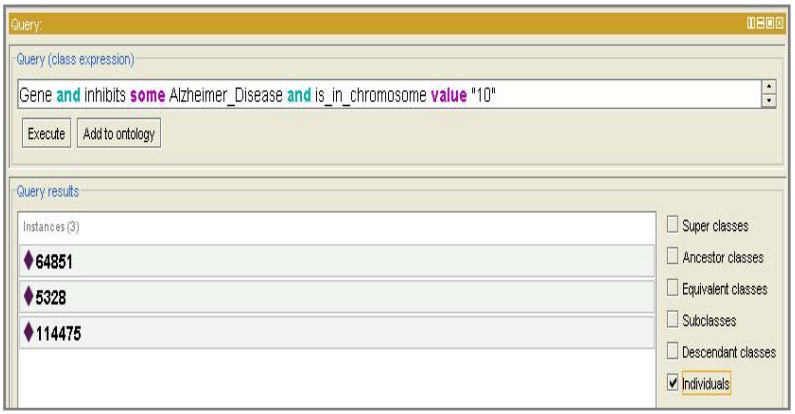

Figure 8 Mapping Identifiers for Genes Extracted by DL query

This query contains two concepts such as Gene and Alzheimer_Disease and two properties namely inhibits and is_in_chromosome in which the first one is object property that has no value and the later one is data property that has value 10.

The extracted information can also be viewed using graphical visualizer OntGraf which is a plug-in in protege tool. The visualization of concepts with its semantic relations is experimented using OntGraf tool. The main concepts or classes in the integrated ontology are shown in Figure 9. This figure provides many semantic relations among main concepts included in our proposed ontology, in which Thing is a main concept that has 8 direct subclasses namely *Gene, Gene_Type, Mesh, Go_Functionality, Obsolete_Class, GO_0008150, GO_0003674 and GO_0005575*. The main concept Thing is linked with its direct subclasses using blue color link and the subclasses of direct subclass links are represented with yellow color. In the figure, yellow link is used to express relationship between Mesh direct subclass and its subclasses *Cellular_Component, Molecular_Function and Biological_Process.*

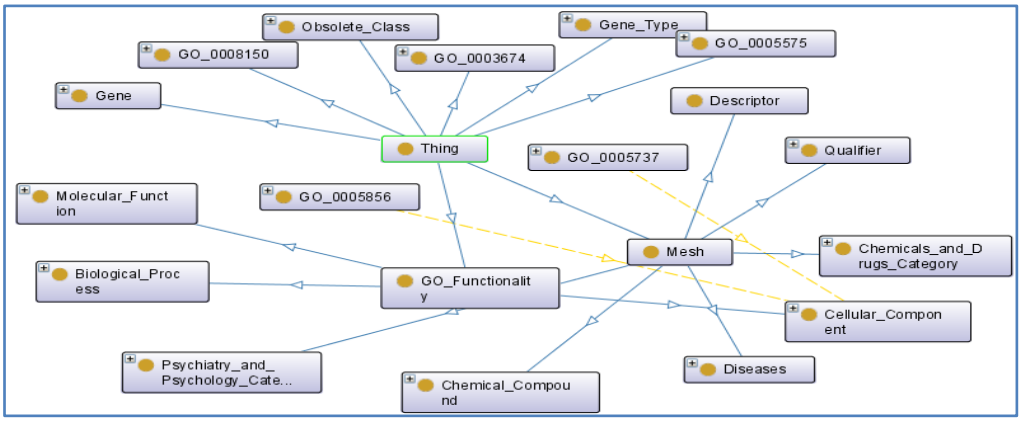

Figure 9 Overview of Thing concept

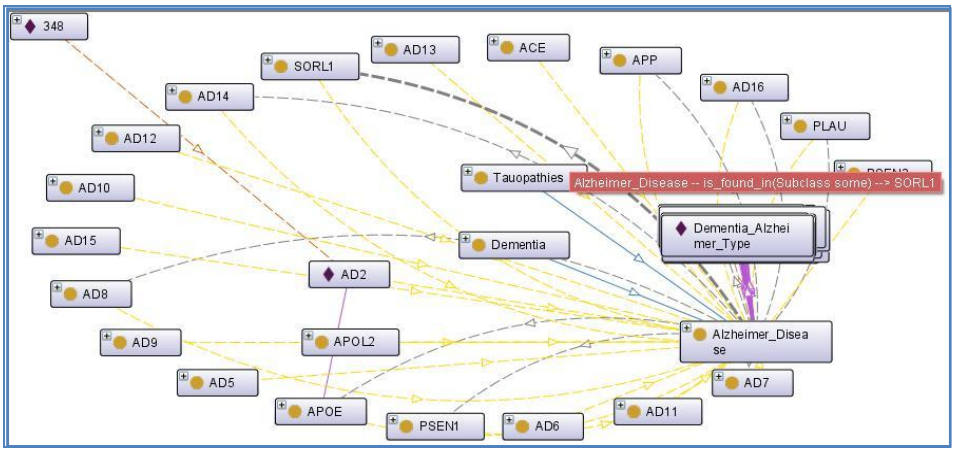

Figure 10 Genes Related to Alzheimer Disease

The Figure 10 shows the association among genes that causes Alzheimer_Disease, in which many associations are extracted such as genes that causes the particular disease, synonyms of extracted genes, gene identifier that has link with different biomedical data bases such as ENSEMBL, HGNC, HPRD, MIM, UNIPROT and PubMed, MeSH keywords, Proteins that inhibits Alzheimer disease, chromosome level in which the particular gene activates to cause Alzheimer, etc.

The Figure 11 shows the association among proteins that inhibits Alzheimer_Disease. The type of proteins related to Alzheimer disease are *cytoskeletal proteins, nerve tissue proteins, protein precursors and amyloid*. The *amyloid* and protein precursors are further derived into the actual protein '*amyloid beta protein precursor*' that inhibits Alzheimer disease. Another protein that inhibits Alzheimer disease is '*tau protein*' which has the immediate parent microtubule associated protein that in turn derived from *microtubule proteins* which is hierarically derived from *cytoskeletal proteins.*

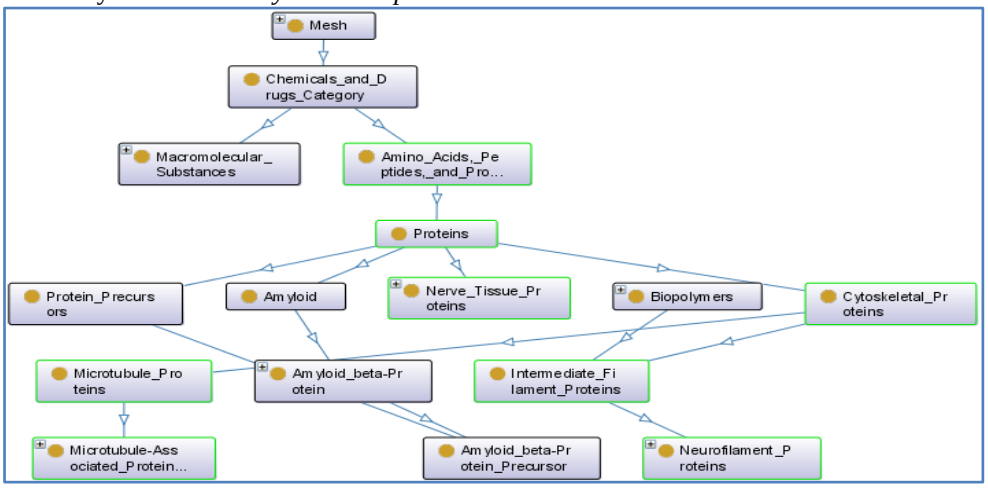

Figure 11 Associations among Proteins that inhibits Alzheimer disease

Finally, the associations among entities can be obtained by processing the concerned RDF/XML using any .NET based language. The inner level associations which are not possible in GO-Keywords approach are identified by processing the created ontology of type RDF/XML. Some of the important and in depth associations are observed while processing the RDF/XML file which are explained in next coming paragraphs.

The biological process '*Oxidation Reduction*' in gene ADH1B in the proposed ontology provides the new semantic relation stating that this process has more chance to co-occur with the biological process '*Histone acetyle transferase complex*' which is not identified by vector based approach.

The rules identified in vector based approach stating that same biological process has more chance co-occur with the molecular function '*Protein binding'*. In ontology approach the inner level associations of protein binding stating that the molecular functions namely *transmembrane receptor activity, complement receptor activity, signal transducer activity and complement component receptor activity* are implicitly associated with '*Oxidation reduction*'. In the same way this approach identified more inner level associations for any particular process or function.

The Figure 12 shows the associations between the biological process '*oxidation reduction*' in the genes *ADH5, ADH7, ADH1A, ADH1B* and *A2M* with other biological processes and molecular functions which includes direct associations and inner level associations.

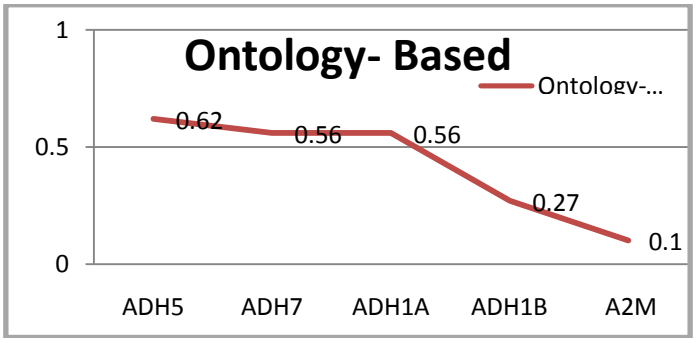

Figure 12 Associations among oxidation reduction with other biological processes and molecular functions

The biological process '*Oxidation Reduction*' is the dominating biochemical process takes place in human body irrespective of genes. Hence, it is decided to identify and extract the associations between this with other molecular functions and biological processes. In the vector based approach fails to bring more semantic association, where as in this approach it has been handled in a better way and it produces better and inner level associations by referring different levels of concepts. Thus this approach takes association more than one level. In previous approach, the rules identifies the biological process '*oxidation Reduction*' has more than 70% chance to co-occur with the other biological processes namely *Retinal metabolic process, Fatty acid metabolic process, Quinine metabolic process, alcohol catabolic process, lipid catabolic process and nitrsyl catabolic process*. But in ontology approach for the same biological process '*oxidation Reduction*' has associations with the above said processes and also has more inner level associations with other biological processes and molecular functions. The semantic associations between oxidation reduction and other molecular functions and biological processes of different genes by two approaches are clearly shown in Figure 13

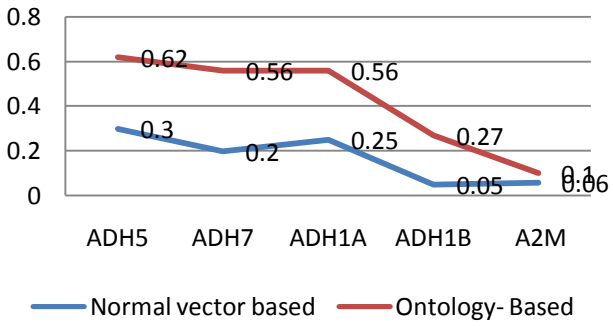

Figure 13 Association among entities using two approaches

V. VALIDATING ONTOLOGY

The integrated ontology has to be validated to check the correctness. This section explores the evaluation methods used to validate our proposed ontology framework. The proposed ontology is syntactically verified for its consistency using FACT ++ reasoner available in protégé tool. The next validation method is semantic validation; the semantic validation of the ontology is verified by the domain experts. This ontology is

validated by domain experts in biological field. Another evaluation method to validate the ontology is structural validation. The structural validation is performed by the different metrics defined in paper [25] that are class match measure, density measure, betweeness measure and semantic similarity measure. The ontology concept is ranked based on the total score of all the four metrics. The weights are assigned based on the concept representation so that the overall score lies between 0 and 1.

Class Match Measure (CMM) – This measure evaluates the ontology for the specified concepts. The specified concepts are searched in the ontology to determine the occurrence of it. If it occurs directly as a concept, the maximum weight will be given for the specified concept. If it partially occurs as instances of any class, then the 50% of maximum weight may be assigned. The CMM evaluates the concepts either as exact match or partial match found in the ontology.

Density Measure (DEM) - The DEM evaluates the ontology based on the degree of richness of attributes of a specified concept and includes the details of subclasses, inner attributes, siblings and relations with other classes in the ontology. The weight may be assigned based on the degree of richness of attributes of a concept.

Betweenness Measure (BEM) – This measure evaluates the ontology based on centrality of a specified concept in the ontology. The centrality of a concept is computed using the count of shortest path between the specified concept and other concepts in the ontology. Based on the shortest path, weight may be assigned.

Semantic Similarity Measure (SSM) – The SSM evaluates the ontology based on the proximity of classes in the ontology the specified concept matches, that is the count of links the specified concept has to map with the existing concepts in the ontology.

The 4 corpuses are framed from our integrated ontology for the validation and each corpus consists of concepts of ontology and its important properties. Each corpus is the superset of the previous one. The corpus C1 consists of main concepts that include more subclasses and have rich relations or links with other sub concepts. The corpus C2 consists of sub concepts in C1 and other concepts in the ontology. The corpus C3 is the subset of C1 and has concepts in C1 and two important properties related to those concepts. The corpus C4 contains concepts in C1 and three important properties related to those concepts. All 4 corpuses framed from the ontology shown in Table 6.5 and the overall score is computed as follows from the above mentioned measures. Let O be the set of corpuses framed from the proposed ontology; Let w_i be a weight factor and M be the different similarity metrics such as CMM, DEM, SSM and BEM.

$$Score(o \in O) = \sum_{i=1}^{4} w_i \frac{M[i]}{\max_{1 \le j \le |O|} M[j]}$$

From the overall score it is found that the corpus C1 has the maximum score as it considered concepts as direct match. The score may be lesser when the concept with partial match is found. The corpus C4 is found to have less score and ranked as 4 due to the DEM and BEM measure score values. The DEM and BEM measure gives lowest score, because concepts in C2 have no related inner attributes and links with other concepts in C2. The corpus C3 is found to have second highest score due to the CMM and SSM score values, since the concepts in C2 are direct concepts and have good number of links among concepts in C3. The corpus C2 is found to have third highest score due to the CMM and SSM score values and also C4 is the sub set of C3.

All the four metrics are provided with equal weights and we found that some of the corpuses may produce low score due to the DEM and BEM measures. The DEM and BEM score values may be increased when we use different weights. In our proposed ontology, we found that the concepts and its relations are linked correctly and further some of the missing relations may be added in future as to produce more promising results.

Table 6.5 Overall Scores and Ranks

Corpus(constructed from Ontology)	**Score**	**Rank**
C1	0.79	1
C2	0.42	3
C3	0.46	2
C4	0.34	4

Finally the class match measure produces high score when there is an exact match found in the ontology. This score may decrease when there is a partial match found in the ontology. The density measure score found to be good when more relations exists among concepts. The betweenness measure found to be good when the concepts related with more other concepts in the ontology. The semantic similarity measure is found to be good when the concept have more synonyms and its relations. The overall score and ranks are plotted in a line graph shown in Figure 6.14.

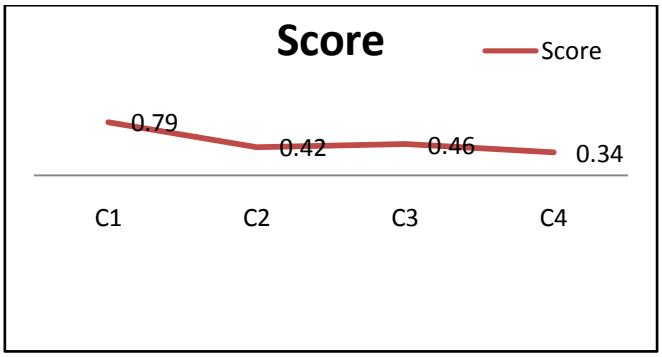

Figure 6.14 Corpus Vs Similarity Metric Score

VI. CONCLUSION

We studied almost all biomedical ontologies and identified all their merits and demerits. In consideration with this in mind the integrated ontology has proposed by accumulating the essential features represented in all specified ontologies. The integrated ontology is implemented in protégé tool that consists of three main concepts namely GO term functionalities, all human genes with or without related to Alzheimer Disease and MeSH terms related to the same disease. This frame work also addresses the problems of GO ontology, in which all information are given as annotations and that are not directly accessible by the user, because information of these kind are given as http links. The MeSH ontology represents the entry terms for the particular term and associated links in various repositories such as Pub Med, Medline, MIM, etc. In the proposed work, all associated information of GO functionalities, genes are specified directly in our ontology, not as links. This ontology gives all possible inner level semantic relations applicable for all concepts defined in the ontology. The proposed framework offers two types of associations that provide direct associations provide associations among gene, protein, MeSH terms and GO annotation terms. Indirect associations provide many semantic associations that include associations between gene and geneid, gene and gene type, GO with GO_functionality, gene with all data bases in which the particular gene is cited. The other indirect relations are parent gene, equivalent gene, common members of two different genes, synonyms,etc. The ontology is also evaluated for its correctness and validity using various metrics. In the results of the experiments, it is found that the ontology is modeled correctly by providing necessary concepts and relations. The ontology may be further improved by adding more relations to the existing concepts of gene and MeSH to get a higher score.

Acknowledgment

This work was performed as part of the Minor Research Project, which is supported and funded by University Grants Commission, New Delhi, India.

REFERENCES

[1] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, pp. 39–41, 1995.
[2] Kohli J. Genetic nomenclature and gene list of the fission yeast Schizosaccharomyces pombe. Curr Genet 1987;11(8):575-89.
[3] http://www.geneontology.org/
[4] http://www.ncbi.nlm.nih.gov/mesh/ meshhome.html
[5] Uramoto, N., H. Matsuzawa, T. Nagano, A. Murami and H. Takeuchi, 2004. A text-mining system for knowledge discovery from biomedical documents.
[6] Hristovski, D., J. Stare, B. Peterlin and S. Dzeroski, 2001. Supporting discovery in medicine by association rule mining in Medline and UMLS. Proc. MedInfo Conf., London, England, Sep. 2-5, 10: 1344-1348.
[7]. Creighton, C. and S. Hanash, 2003. Mining gene expression databases for association rules. Bioinformatics, 19-1: 79-86.
[8] Wren, J.D., R. Bekeredjian, J.A. Stewart, R.V. Shohet and H.R. Garner, 2004. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20: 3.
[9] Adamic, L.A., D. Wilkinson, B.A. Huberman and E. Adar, 2002. A literature based method for identifying gene-disease connections. IEEE Computer Soc. Bioinformatics Conf.
[10] Palakal, M., M. Stephens, S. Mukhopadhay, R. Raje and S. Rhodes, 2002. A Multi-level Text Mining Method to Extract Biological Relationships. Proc. IEEE Computer Soc. Bioinformatics (CSB) Conf., pp: 97-108.
[11] Wilkinson, D.M. and B.A. Huberman, 2004. A method for finding communities of related genes. Proc. Natl. Acad. Sci. U.S.A., 101 Suppl. 1: 5241- 5248.
[12] Hisham Al-Mubaid and Rajit K Singh, 2005. A New Text Mining Approach for Finding Protein-to-Disease Associations, American Journal of Biochemistry and Biotechnology 1 (3): 145-152, ISSN 1553-3668.
[13] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, 2001. Detecting Gene Relations From Medline Abstracts, Pacific Symposium on Biocomputing 6:483-496.
[14] A. Hotho,A.Maedche and S.Staab, "Ontology-based text document clustering"[A], *Proc. of the Conf. on Intelligent Information Systems[C]*,2003.
[15] Paulo Gottgtroy1, Prof. Nik Kasabov1, Stephen MacDonell1, 2004. An ontology driven approach for knowledge discovery in Biomedicine.
[16] Robert Stevebs et.al., "Ontology based knowledge representation for bioinformatics", published in briefings in Bioinformatics, 2000.
[17] Barry Smith, et. Al., "The Ontology of the Gene Ontology", Proceedings of AMIA Symposium 2003.
[18] Olivier Corby et.al., "Searching the Semantic Web: Approximate Query processing based on bio ontologies". Published in IEEE Computer Society, 2006.