

Design and Development of Integrated Biomedical Ontology for Information Extraction from Medline Abstracts

Dr.B.LShivakumar¹, R.Porkodi²

¹Professor and Head, Department of Computer Applications, SNR Sons College, Coimbatore -6

²Assistant Professor, Department of Computer Science, Bharathair University, Coimbatore – 46

Abstract—Due to the ever-increasing amount of scientific articles in the bio-medical domain, Information Extraction in Text Mining has been recognized as one of the key technologies for future bio-medical research. The Information extraction from these biomedical domains plays a vital role in bioinformatics field. Thus, bioinformatics researchers extend their work in both information extraction and construction of biomedical knowledge sources. In knowledge extraction, researchers are involved to develop efficient and effective technique by combining Natural Language Processing (NLP) and text mining techniques to find out and extract information and significant associations among the extracted information. In the other side bioinformatics researchers are busy with the construction of knowledge sources or repositories related to biomedical domain which simplify the work of the researchers in knowledge extraction process. This paper presents a semiautomatic framework that integrates the well-known two ontologies Gene Ontology (GO) and Medical Subject Heading (MeSH) ontology by adding of semantic mappings or relations between GO terms, Gene names and MESH keywords related to a particular disease (Alzheimer disease). The integrated ontology has validated in all three aspects such as structural, syntactic and semantic validation measures. This framework is used to discover significant associations or relationships between proteins and genes related to Alzheimer disease that are extracted from Medline abstracts.

Keywords—Ontology, Alzheimer Disease, GO, MeSH, Stemming, Tagging.

I. INTRODUCTION

NCBI is a fast growing knowledge source for bioinformatics community, which has Medline database [1] that currently contains over 15 million citations of biological abstracts and it is growing by more than 40,000 abstracts per month. To extract 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 text mining methods use vector space model to represent a document. The vector space model represents a document as a feature vector of terms contained in it. Each feature vector contains term weights and similarity between documents is computed using various similarity measures. This approach not considered the semantic relations of terms in documents. The ontology approach represents an effective knowledge representation within controlled vocabulary. The Wordnet ontology [14] is a lexical database for general English covering most of the general English concepts. In biomedical domain, the Unified Medical Language System (UMLS) framework [13] includes much biomedical ontology.

This paper integrates the Gene ontology (GO), Medical Subject Headings (MeSH) and all human genes which include genes that cause Alzheimer disease in human. 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 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 the some results based on the concepts, object properties and data properties, etc.

The paper is organized as follows: Review of literature related to this work is presented in section 2. In section 3, brief introduction on biomedical ontologies are presented and section 4 the ontology based framework is elaborately discussed. The experimental design and results discussion is presented in section 5. 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 [2], association rule discovery [3, 4] and extracting relationships among various concepts [5, 6]. Many approaches have been developed for extracting significant associations and interactions among various biological entities [5, 6, and 7] 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 [8] has presented the extraction of gene names from articles' titles and abstracts and identified genes related to colon cancer disease. The paper [6] has presented a statistical approach for discovering group of genes related to breast cancer disease. In paper [5], author constructed a relationships network among biomedical entities which are extracted from Medline abstracts.

In paper [9] 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 [10], 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 [11] uses conceptual features that are extracted from text using ontology and prove that ontology could improve the performance of text clustering. The paper [12] shows the case study on the integration of biomedical information in to ontology. In paper [21] author proposed bio ontology methodology and compared this with other bio-ontologies. The limitations and benefits of GO ontology are expressed in paper [22]. The author had studied the strength and limitation of biomedicine ontologies based on its text and concept representation [23]

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. BIOMEDICAL ONTOLOGIES

The Molecular Biology Ontology (MBO) [16] was the first attempt to begin to define the entities in the domain to promote consistent interpretation across resources. A second phase saw the adoption of ontology by the biological community itself. Pre-eminent among these is the Gene Ontology (GO) [15]. The Microarray Gene Expression Data (MGED) ontology [18] provides a vocabulary for describing a biological sample used in an experiment, the treatment that the sample receives in the experiment and the microarray chip technology used in the experiment. The Functional Genomics Ontology (FUGO) [17] is another type of ontology in the field of bioinformatics. The next popular ontology MeSH thesaurus is the NLM's controlled vocabulary for subject indexing in MEDLINE. It is structured in a hierarchy of descriptors, with each descriptor including a set of concepts, and each concept itself containing a set of terms, which are synonyms and lexical variants.

The next coming sections give a brief explanation on Gene Ontology (GO) and MeSH ontology that are integrated in our work.

3.1 GO

Biological knowledge is most often represented in 'bio-ontologies' that are formal representations of knowledge areas in which the essential concepts are combined with properties that describe relationships between concepts. Bio-ontologies are constructed according to textual descriptions of biological activities. One of the most popular bio-ontology is Gene Ontology (GO) [15] that contains more than 18 thousands terms. The GO ontology is a controlled vocabulary of gene and protein roles in cells, addressing the need for consistent description of gene products. This is mainly used in almost all biological researches and to predict the gene functions based on patterns of annotation. The GO describes the molecular function of a gene product, the biological process in which the gene product participates, and the cellular component where the gene product can be found.

3.2 MeSH

Medical Subject Headings (MeSH) [24] is another popular ontology designed by the National Library of Medicine which mainly consists of the controlled vocabulary and a MeSH Tree. The controlled vocabulary contains several different types of terms, such as Descriptor, Qualifiers, Scope note, Tree number and Entry terms. Descriptor terms are main concepts or main headings. Entry terms are the synonyms or the related terms to descriptors. For example, "Amyloid beta-Protein Precursor" as a descriptor has the following entry terms "Amyloid A4 Protein Precursor", "Amyloid beta Precursor Protein", "Amyloid Protein Precursor", etc. MeSH descriptors are organized in a MeSH Tree, which can be seen as a MeSH Concept Hierarchy. In the MeSH Tree there are 15 categories (e.g. category A for anatomic terms) and each category is further divided into subcategories.

For example, the MeSH tree structure of Alzheimer Disease is shown in Fig. 1. For each subcategory, corresponding descriptors are hierarchically arranged from most general to most specific. In addition to its ontology role, MeSH descriptors are originally used to index MEDLINE articles.

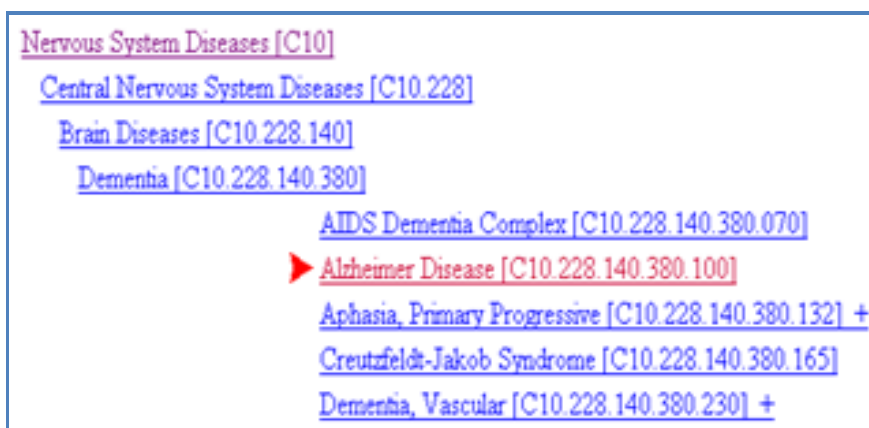


Fig. 1 MeSH Tree Structure for Alzheimer Disease

IV. PROPOSED FRAMEWORK

The proposed framework shown in Fig. 2 integrates the two popular ontologies GO and MeSH by mapping with Gene details used to extract significant associations among concepts from Medline abstracts related to Alzheimer disease. The main objective of the integration of two ontologies is making use of semantic relations among the concepts in Medline abstracts using MeSH ontology terms. The gene products of genes are referred from GO in order to find out the associations of gene products related to Alzheimer disease genes. 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 gene products in which we identify the exact molecular functions that result in Alzheimer disease, biological processes in which the gene product participates to result in Alzheimer disease and the cellular component where the gene product of Alzheimer disease can be found. The components of this framework are explained below.

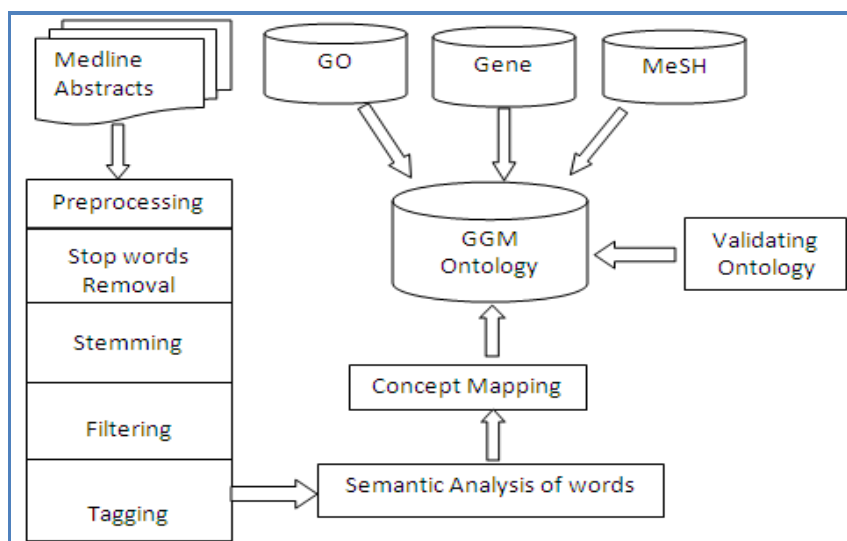


Fig. 2 The Proposed Ontology Framework

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

- a. **Removal of stop-words:** The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). Removal of stop-words improves clustering results [19].
- b. **Stemming:** By word stemming it means the process of suffix removal to generate word stems. The Porter stemmer [20] which is a well-known algorithm is used for this task.
- c. **Filtering:** Domain vocabulary V in ontology is used for filtering. By filtering, document is considered with related domain words (term). It can reduce the documents dimensions. The filtering task used in our work filters the documents related to Alzheimer disease.
- d. **Tagging:** The concepts in Medline abstracts are identified using Genia tagger [26] and the identified concepts are mapped with concepts related to the categories specified in the proposed ontology. The categories used in the proposed ontology are gene, MeSH and GO.

e. Semantic analysis & Concept mapping:

Class/Concept	Description
GO_0003674	This is the base class for 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 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 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.

Table 1. Main Concepts in Proposed Ontology

After preprocessing, the extracted concepts in Medline abstracts are analyzed in terms of semantic meaning and added to ontology, if it is not available. The concepts or classes used in this integrated ontology are shown in Table 1. The first step for adding concept is to find out the equivalent concept from the ontology and add the concept and possible semantic relations which includes object properties and data properties into ontology, if it is not found. Some of the important object properties and data properties created in the integrated ontology shown in Table 2 and 3.

For example, the gene name “A2MP1” is extracted from Medline abstract and this term or concept is to be added to the ontology. If there is an equivalent gene “A2M” is already in the ontology, add “A2MP1” to ontology and assign is-a relationship along with other possible properties such as has_go, has_synonym, has_genotype_as, has_inducing_protien, inhibits, etc. between “A2M” and “A2MP1”.

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_genotype_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 2. List of Object 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 ENSEMBL, 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.

Table 3. List of Data Properties in Proposed Ontology

V. EXPERIMENTAL DESIGN & RESULTS

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 for MeSH class includes all disease branches in which Alzheimer disease is derived, amino acids, peptides and proteins related to a particular disease. The integrated ontology provides all types of information related to GO terms, genes and a particular disease details.

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 results or inconsistent results when we manipulate information from the ontology. The visualization of concepts with its semantic relations is experimented using OntGraf tool. The important subclasses for MeSH main concepts or classes are shown in Fig. 3. The visualization of human genes related to Alzheimer disease is shown in Fig. 4. The visualization of proteins that are inducing Alzheimer disease is shown in Fig. 5.

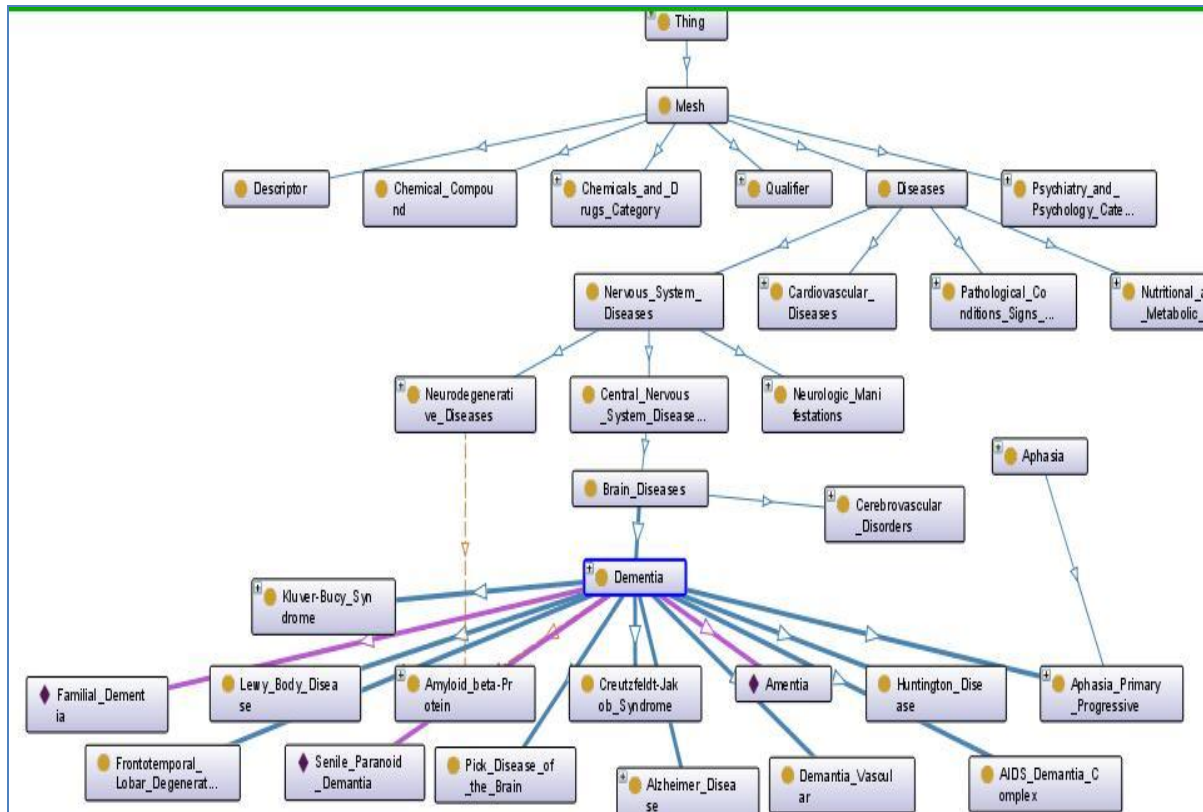


Fig. 3 The Overview of MeSH Concept

Another way to manipulate the ontology is using DL query tool. 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 Fig. 6 shows the extraction of Alzheimer Disease

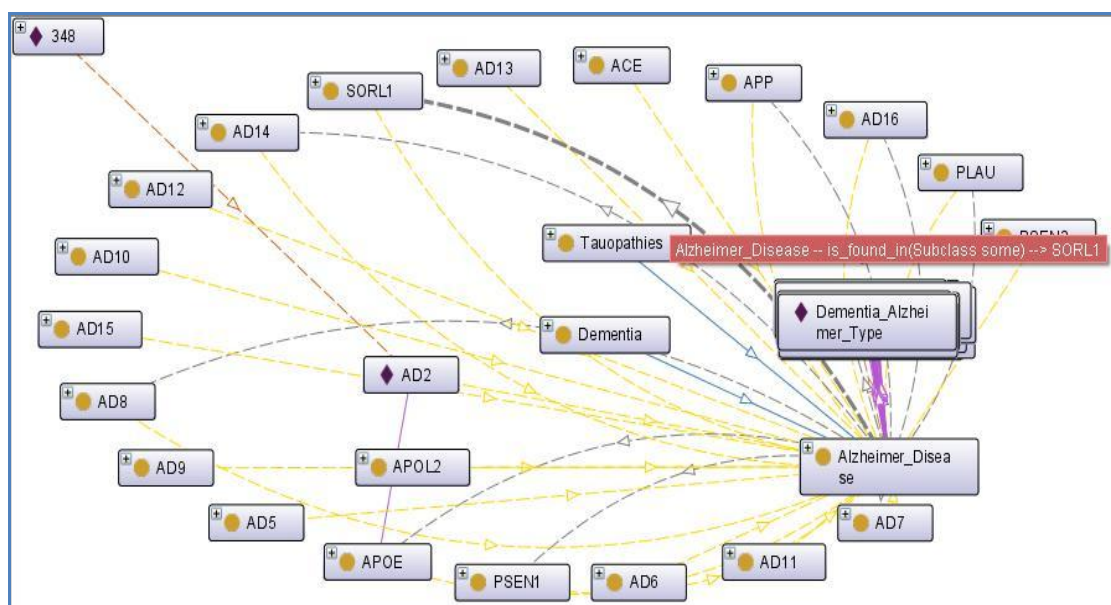


Fig. 4 Genes Related to Alzheimer Disease

related human genes with the respective DL query and the Fig. 7 shows the extraction of protein names that inducing Alzheimer disease. Finally the Fig. 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 in Fig. 8) that inhibits Alzheimer disease in chromosome level “10”.

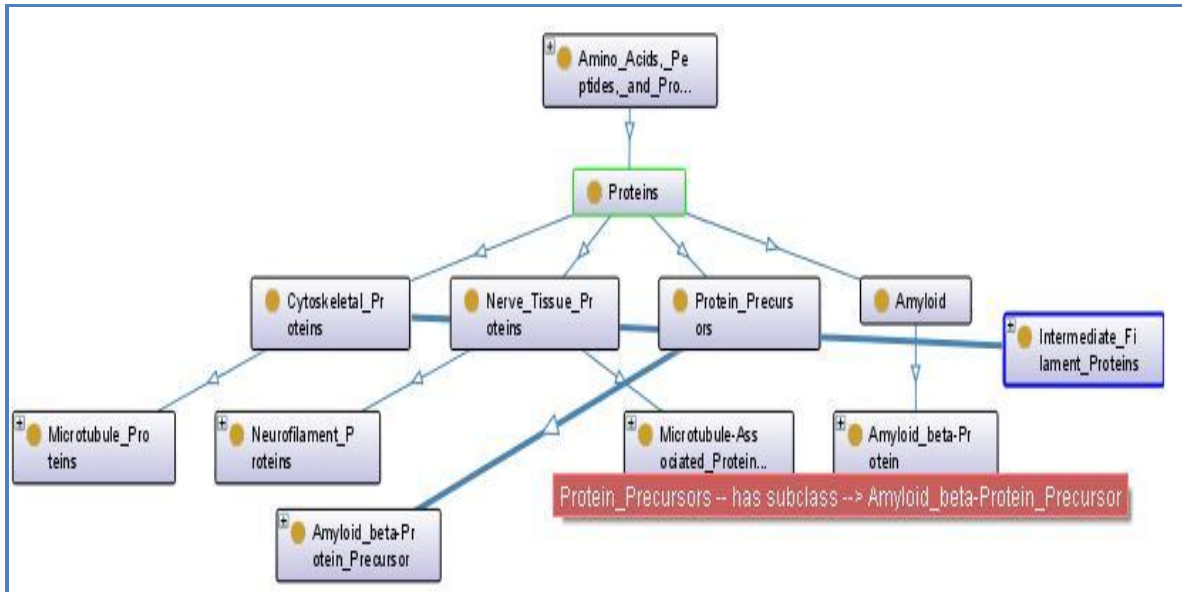


Fig. 5 Proteins Related to Alzheimer Disease

Fig. 6 Genes Extracted by DL query “Gene and inhibits some Alzheimer_Disease”



Fig. 7 Proteins Extracted by DL query “Proteins and has_inducing_disease some Alzheimer_Disease

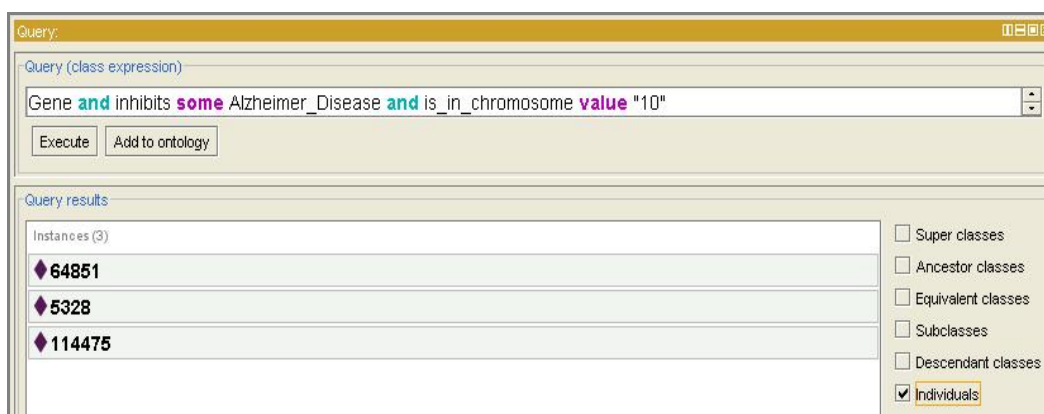


Fig. 8 Mapping Identifiers for Genes Extracted by DL query

5.1 Validating the Ontology

The integrated ontology has to be validated to check the correctness. This section explores the evaluation methods used for validating 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 and 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, betweenness 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 and the weights are assigned in such a way 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.

5.2 Dataset

We framed 4 corpuses from our integrated ontology to validate it 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 4 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.

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

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 C2 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 C4 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.

Corpus(constructed from Ontology)	Score	Rank
C1	0.79	1
C2	0.39	4
C3	0.46	2
C4	0.41	3

Table 4. Overall Scores and Ranks for Corpuses

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 corpuses Vs metrics score is represented in a bar chart is shown in Fig. 9.

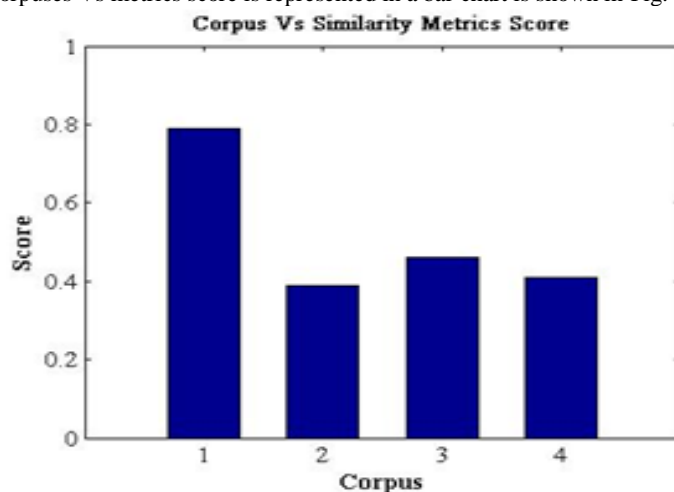


Fig. 9 The Bar chart of Corpus Vs Similarity Metric Score

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 our work, all associated information of GO functionalities, genes are specified directly in our ontology, not as links. This ontology gives all possible semantic relations applicable for all concepts defined in the ontology. The ontology is also evaluated for its correctness and validity using various metrics. In the results of the experiments, we 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.

Further the integrated ontology to be used in association rule mining to extract the significant associations among the proteins, genes that are related to Alzheimer Disease. Instead of using simple vector space model to calculate the term frequency and inverse document frequency from the Medline abstracts, we decide to use ontology approach to consider the semantic relationships of terms that appear in the Medline abstracts may give better results.

VII. 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]. NCBI PubMed, <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>
- [2]. Uramoto, N., H. Matsuzawa, T. Nagano, A. Murami and H. Takeuchi, 2004. A text-mining system for knowledge discovery from biomedical documents.
- [3]. 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.
- [4]. Creighton, C. and S. Hanash, 2003. Mining gene expression databases for association rules. Bioinformatics, 19-1: 79-86.
- [5]. Wren, J.D., R. Bekeradjian, J.A. Stewart, R.V. Shohet and H.R. Garner, 2004. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20: 3.
- [6]. 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.
- [7]. Palakal, M., M. Stephens, S. Mukhopadhyay, 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.
- [8]. 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.
- [9]. 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.
- [10]. M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, 2001. Detecting Gene Relations From Medline Abstracts, Pacific Symposium on Biocomputing 6:483-496.
- [11]. A. Hotho, A. Maedche and S. Staab, "Ontology-based text document clustering"[A], *Proc. of the Conf. on Intelligent Information Systems[CJ]*, 2003.
- [12]. Paulo Gottgroy¹, Prof. Nik Kasabov¹, Stephen MacDonell¹, 2004. **An ontology driven approach for knowledge discovery in Biomedicine.**
- [13]. R. Kleinsorge, C. Tilley, and J. Willis. (2000). Unified Medical Language System (UMLS) Basics [Online]. Available: http://www.nlm.nih.gov/research/umls/pdf/UMLS_Basics.pdf.
- [14]. G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, pp. 39-41, 1995.
- [15]. <http://www.geneontology.org/>
- [16]. Schulze-Kremer S. Adding semantics to genome databases: Towards ontology for molecular biology. In: Proceedings of the Fifth International Conference for Intelligent Systems for Molecular Biology Conference (ISMB), 1997; pp. 272-5.
- [17]. <http://www.fugo.org>
- [18]. Whetzel PL, Parkinson H, Causton HC, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006;22:866-73.
- [19]. Mark Sinka and David Corne, "A Large Benchmark Dataset for Web Document Clustering", *In Soft Computing Systems: Design, Management And Application*, Vol.87 of Frontiers in Artificial Intelligence and Applications, pages 881-890, 2002.
- [20]. M.F.Porter, "An Algorithm for Suffix Stripping", *Program* 14(3), July 1980, pp.130-137.
- [21]. Robert Stevebs et.al., "Ontology based knowledge representation for bioinformatics", published in briefings in Bioinformatics, 2000.
- [22]. Barry Smith, et. Al., "The Ontology of the Gene Ontology", Proceedings of AMIA Symposium 2003.

- [23]. Olivier Corby et.al., “Searching the Semantic Web: Approximate Query processing based on bio ontologies”. Published in IEEE Computer Society, 2006.
- [24]. <http://www.ncbi.nlm.nih.gov/mesh/meshhome.html>
- [25]. Amal Zouaq, Roger Nkambou, “Building Domain Ontologies from Text for Educational purposes”. IEEE Transactions on Learning Technologies, Vol.1, No.1, Jan-Mar 2008.
- [26]. <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>