ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

# Semantic Ontology Based Querying On Biomedical Concepts

[1] Thayyaba Khatoon Mohammed [2] Dr.A.Govardhan
[1] Research Scholar [2] Professor
Dept. of CSE, Jawaharlal Nehru Technological University

**Abstract: - Biomedical knowledge resources, such as terminologies and on tologies, are important for community-based annotation and sharing of data. Creating and maintaining these resources is challenging given the rapid growth of scientific knowledge. Hence query processing on biomedical data become very slow which increases complexity in performance and efficiency query processing on biomedical data and also there is a lack of tools to ease the integration and ontology based semantic queries in biomedical databases, which are often annotated with ontology concepts. We have developed a system called ONTOBIO, which tells about semantic query engine that provides semantic reasoning and query processing, and translates the queries into ontology repository operations on biomedical data. The system provides an interface for executing queries and bio ontologies which organize taxonomies information. This paper addresses about attributes and describes relationships between biomedical concepts. The system is also compatible for integration and deployable with databases and description logic based ontologies. ONTOBIO integrates the two schemes by Bio medical concepts using a set of semantic rules. We have applied our method to aknowledge base of autism phenotype definitions, which are modeled using the web ontology language, and its rule language, SWRL.**

**Key words: OWL2 RL, Semantic web, SWRL, RDBMS.**

## I. INTRODUCTION

The emergence of information and communication technologies has drastically changed biomedical scientific processes. Experimental data and results today are easy to share and repurpose thanks to the Web and public application programming interfaces (APIs) enabling connection to databases containing such information. As a consequence, the variety of biomedical data available in the public domain is now very diverse and ranges from genomic-level high-throughput data to molecular aging studies to published research articles. The paradox of such an expansion isthat biomedical researchers now faces the problem of extracting the specific data they need. Measures must be taken to prevent this problem from worsening as data repositories grow fast1. Biomedical researchers have turned to ontologism and terminologies to describe their data and turn it into structured and formalized knowledge.

In order to develop integrative translational bioinformatics approaches to interpret these datasets, there is a strong and pressing need to be able to identify all experiments that study a particular disease. A key query dimension for such integrative studies is the sample, along with a gene or protein name. For example, a researcher studying the allelic variations in a gene would want to know all the pathways that are affected by that gene, the drugs whose effects could be modulated by the allelic variations in the gene, and any disease that could be caused by the gene, and the clinical trials that have studied drugs or diseases related to that gene. The knowledge needed to study such questions is available in public biomedical resources; the problem is finding that information.

Biomedical ontologies are based on meta logics. Biomedical concepts can be expressed using different standards such as OWL,RDF. A notable feature of the OWL 2 RL is that the profile is designed to be implementable using a rule-based OWL reasoner.

The specification document contains a set of first order implication rules of rules that can be used to implement such a reasoner. The rules are described in terms of an RDF serialization of an OWL ontology.
Reasoning with the OWL 2 RL profile is polynomial with respect to the size of the ontology.

## II. RELATED WORK

Ontology represents concepts and relations between concepts in forms that can be interpreted by both humans and machines. Much biomedical ontology has been developed in the past for different domains. Most of these ontologies are included in the NCBO BioPortal [3, 5], which provides the abilities to browse, search and visualize ontologies as well as to comment on, and create

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

mappings for ontologies. LexGrid [13] is a framework for representing, storing, and querying biomedical terminologies, and often used by ontology repositories as the backend for managing ontologies and providing query services. caDSR is a database and a set of APIs and tools to create, edit, control, deploy, and find common data elements (CDEs) for use by metadata consumers. Many applications or databases are providing semantic annotations to the data by linking data to ontology concepts. For example, NCI Annotation and Image Markup project [8] are Pathology Analytical Imaging Standards project provide semantic enabled models to support semantic interoperability.

In the last few decades much research has-been done in the field of biomedical informatics and voluminous amount of research data has been collected in the fields of clinical research, biomedical research, life sciences, gene research, patient records, clinical trials etc. Simultaneously various biomedical tools have been developed to perform wide range of function like data mining, data management, data collection etc. This in turn has forced scientists to analyze and structure the knowledge to make further inferences from the present knowledge. A survey conducted showed that DNA sequence databases have been doubling for every 18 months. Most of the existing databases have overlapped data as they are built independent to each other. Database systems today are facing the task of serving ever increasing amounts of data from growing complex user community which is getting more and more demanding. All the researchers need almost the same data but with different meaning or context. This makes semantics very important for this domain.

***Ontology***

Ontology - "***science of being***" - typically has different meanings in different contexts. The word originated in philosophy where several philosophers-from Aristotle (4th century) to Leibniz (1646-1716),and more recently the 19th Century majorontologists like Bolzano, Brentano, Husserl and Frege have provided criteria for distinguishing between different kind of objects and the relations among them. The objects can be both concrete and abstract.

In the late 20th century, Artificial Intelligence (AI) adopted the term and began using it in the sense of a "***specification of a conceptualization***" in the context of knowledge and data sharing.

Ontology is "hierarchal structuring of knowledge about concepts by sub-classing them according to their properties and qualities". It can also be defined as "a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and the relationships between them" .Ontology gives the description of concepts and the relations that can exist between them. The concept is very important for data sharing and knowledge representation.

Ontology can be classified according to level of detailed knowledge they provide. ***Upper Ontologies*** provides very generic knowledge with low domain specific knowledge. For example, *Disease* ontology is upper ontology compatible for any biomedical domain. ***General ontologies*** represent knowledge at an intermediate level ofdetail independently of a specific task. ***Domainontologies*** represent knowledge about a particular part of the world, such as medicine, and should reflect the underlying reality through a theory of the domain represented. For example, Gene Ontology, Finally, ontologies designed for specific tasks are called ***application ontologies*.**

## III. LITERATURE SURVEY

This chapter starts with study of literature, provides a background for the research work. The following section consists of the related work that was found closer to our intended work.

Kato et. al., [18] developed an e-health system based on ontology and Semantic Web technologies to help the people to find suitable Thai herbs to cure diseases. Fan et.al., [19] in order to improve the level of knowledge sharing, a plant domain knowledge model based on ontology is discussed in this work. Zhou, et.al,[20] proposed a method that described an ontology model of the plants domain knowledge using related botany knowledge. The effectiveness of the proposed approach is demonstrated by experiments on a Traditional Chinese Medicine text corpus. Dezheng, et.al., [21] constructed the Traditional Chinese Medicine knowledge network into graph by using information from ontology, then adopted centrality algorithm to analyze and process

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

this graph, and finally mined valuable medicine knowledge.

Tran, et. al., [17] proposed two new algorithms called "semantic elements extracting algorithm" and "new semantic elements learning algorithm" for health care semantic words extraction and ontology enhancement. The algorithms were used to extract pairs of concepts and names of diseases in health care information domain from web pages and to get summary information of documents with respect to all extracted semantic words.

Luo, et. al., [22]presented Med Search, a specialized medical Web search engine, to address these challenges. Med Search uses several key techniques to improve its usability and the quality of search results. First, it accepts queries of extended length and reforms long queries into shorter queries by extracting a subset of important and representative words. Miyao, et.al.

Lim et al [15, 16] summarize major problems and challenges of supporting semantic queries in relational databases, such as graph based queries and vagueness of queries, and propose query-by-example (QBE) based semi-automatic approach to solve the problems. In SciPort project [17], semantic enabled authoring and queries are provided to link specific ontologies such as RadLex with structured data, through providing RESTful Web Service based interfaces. Extending to additional ontologies, however, needs development of new RESTful APIs for each ontology repository. Early work in [18,19,20] tried to support semantic operations in RDBMS through user-defined functions. Our work provides an effective generic framework and can support multiple different ontologies. Semantic Web based knowledge management has been an active research area [21]. DBOntoLink takes a middle layer based approach and is extended on existing database management systems and query languages.

## IV. FRAME WORK

There is lack of a unifying framework for user assessment of biomedical ontologies. Current evaluation models provide knowledge engineers, domain experts and ontology modelers the means to evaluate an ontology based largely on its design. The research makes a contribution by providing the design of a flexible framework to support user evaluation of biomedical ontologiesin a dynamic environment. Concepts from systems theory and multi criteria and formative evaluation approaches are exploited and integrated into a new novel approach that provides for a flexible tool design. The design offers the flexibility to relate an ontology to changing user perspectives when assessing and selecting an ontology. It also provides for feedback to enable users determine requirements for improving on existing models. The framework shall help to elicit new requirements for iteratively and incrementally extending and modifying existing biomedical ontologies to suit changing user needs and accommodate new types of data. This shall facilitate extending and modifying existing ontology for reuse and avoid the huge effort of starting or building entirely new ontologies in terms of time, effort and domain specific knowledge. The model design is flexible, generic and can be applied to evaluations in other domains with dynamic environments.

## V. ONTOLOGY QUERY OPTIMIZATION TECHNIUES

Many algorithms are available to perform this task for optimization of querying from biomedical data. The semantic search algorithms are one which search for the words in medical data.

*Algorithm ( semantic search )*

**Input : Search := { $x_1, x_2, x_3, x_4 \ldots X_n$ }**
**Output : Set O := { semantic retrieval }**
**Procedure :**
**Let Set I := Φ , Set P := Φ**
**for every keyword set $Search_l$ in web database**

**if Search ∩ $Search_l \neq \Phi$ then add $Search_l$ to I;**
**Let I := {$search_1, search_2, search_3 \ldots search_n$}**
**P:= $search_1 \cup search_1 \cup search_1 \cup \ldots. search_n$**
**C := count number of keywords in Set P**

**$O(P_l)$ :={ Occurrence, Prioritise Keysets with the probability of occurrence }**

The above algorithm illustrates the semantic searching of words on the web. As we know present scenario of searching depends on heavy parsing algorithms in order to yield results as computer cannot understand the meaning of the documents. Here we have retrieved all the keysets related to the search query in the database and the union of all the keysets are taken in another set. The probability of occurrence was found and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

according to that priority was given to it as a perfect search result.

***Example:***

Search Query= { "Indian Universities" }
Let Keyword sets obtained were :
Search1 = { "Indian", "Courses", "AIU", "Universities", "UGC", "Top", "Colleges", "States"}
Search2 = { "Indian", "UGC","AIU", "List", "Exams", "Top", "Universities"}
Search3 = { "Universities", "UGC", "Colleges", "Top", "Indian", "Ranking" }
Search4 = { "Indian", "States", "AIU", "Universities", "Questions", "Ranking "}
Search5 = { "Indian", "Courses", "Ranking", "UGC", "States", "Universities"}
Now we have a set P in which union of all the keysets are taken
P = { "Indian", "Universities", "Courses", "AIU", "UGC", "Top", "Colleges", "States", "List", "Exams", "Ranking", "Questions" }
Now as we see Indian Universities occur in every search hence the comparison is made with reference to these two keywords.
Occurrence (Pi) = { ϕ,ϕ, 2, 3, 4, 3, 2, 2, 1, 1, 3, 1 }

## VI. EXPERIMENTAL EVALUATION

In this section, we evaluate the proposed method for discovering semantic associations and detecting errors in biomedical ontologies on an *electronic health records* dataset. We first describe the dataset and then present the evaluation results.

In this evaluation, we analyze the electronic health records of real patients. The clinical note data are from Stanford Hospital's Clinical Data Warehouse (STRIDE). These records archive over 17-years worth of data comprising of 1.6 million patients, 15 million encounters, 25 million coded ICD9diagnoses, and a combination of pathology, radiology, and transcription reports totaling over 9 million clinical notes(i.e., unstructured text). We obtained the set of drugs and diseases for each patient's clinical note by using a new tool, the *Annotator Workflow*, developed at the National Center for Biomedical Ontology (NCBO), which annotates clinical text from electronic health record systems and extracts disease and drug mentions from the electronic health records. For this study, we specifically configured the workflow to use a subset of the NCBO ontology library that are most relevant to clinical domains. The resulting set of ontologies contains 1 million sub sumption ("is a") statements. vectorize texts and turned them into a bag-of-word representation, from which an RDF bipartite graph is constructed, including 148 million RDF statements for the data. Any how we are not going in depth with the evaluation tools but providing summary of them as well approaches of evaluation and metrics of ontology evaluations in the following table.1,2,3.

***Table. 1. Ontology Evaluation Tools Compared***

| Criteria | ODEval | OntoManager | OntoQA | AKTiveRank | OntoMetric |
|---|---|---|---|---|---|
| Users | Ontology developers | Domain experts, administrators., analysts | Knowledge engineers | Knowledge engineers [KE] | Knowledge engineers |
| Purpose | Ontology content evaluation | Fit ontology to domain requirements | Ontology ranking | Ontology ranking | Fit ontology to domain task |
| Evaluation type | Structure evaluation | Structure & user evaluation | Structure & knowledge base | Structure evaluation | Structure, cost, methodology, language, software |
| Input | Unpublished ontology | Ontology in use | User Search terms | User Search terms | Ontology in use |
| Metric s used | Consistency, redundancy | User needs | Schema and instance metrics | Class density, semantic similarity, centrality measure | Multi-level tree [MTC] of 160 characteristics |
| Output | Well designed ontology | Ontology fitting domain requirements | Ranked ontology | Ranked ontology | Ontology fit to project |
| Validation | Not defined | Not defined | Expert users | Tool result to Questionnaire assessment | Not defined |
| User involvement | High | High | Low | Low | High |

Existing approaches to ontology evaluation use various contexts and conduct evaluation at different levels of complexity. A taxonomy of evaluation approaches based on type and purpose that adopts levels of vocabulary, taxonomy, semantic relations, application, syntax, structure and design is provided by Brank *et al.* (2005). This level based taxonomy categorizes existing ontology evaluation approaches as follows:

***Table.2.Approaches and levels of Ontology Evaluation***

| | Evaluation approach | | | | |
|---|---|---|---|---|---|
| Ontology Characteristic or Evaluation level | Golden standard | Application based | Data or corpus driven | Human assessment | Semiotic based |
| Lexical or vocabulary | ✓ | ✓ | ✓ | ✓ | X |
| Hierarchy | ✓ | ✓ | ✓ | ✓ | ✓ |
| Semantic relations | ✓ | ✓ | ✓ | ✓ | ✓ |
| Content application | ✓ | ✓ | X | ✓ | X |
| Syntactic | ✓ | X | ✓ | ✓ | ✓ |
| Architecture and design | X | X | X | ✓ | X |
| Granularity, process and function | X | X | X | X | X |

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 8, August 2016

***Table 3.Metrics for Ontology Evaluation***

| Metric used | What is measured | Purpose or indicator of | Evaluation tool |
|---|---|---|---|
| Relational density | Ratio of inheritance relations | Diversity of relationships | OntoQA |
| Schema deepness | Subclasses per class | Class distribution in taxonomy | OntoQA |
| Class utilization | Percent populated classes | Richness of knowledge base | OntoQA |
| Cohesion | No. of connected components | Connectedness in knowledge base | OntoQA |
| Relational richness | Ratio of class relations | Suitability for use | OntoQA |
| Class match measure | Class to search term measure | Concept coverage | AktiveRank |
| Density measure | Class measure | Information content | AktiveRank |
| Betweenness measure | Class measure | Centrality of classes | AktiveRank |
| Semantic similarity | Class to search term measure | Closeness of classes | AktiveRank |
| Connectedness | Relations between classes | Level of integration | No tool defined |

## V. CONCLUSION

While ontologies are proliferating in biomedical domains, most biomedical data are available as structured data managed in relational DBMS or XML DBMS. Using ontologies to enrich the semantics of data for queries and interoperability is becoming increasingly important, but there is a lack of tools to ease the integration and use of ontologies in databases using standard query languages or query interfaces. **ONTOBIO** integrates the two schemes by Bio medical concepts using a set of semantic rules. We have applied our method to acknowledge base of autism phenotype definitions, which are modeled using the web ontology language, and its rule language, SWRL. The system provides an interface for executing queries and bio ontologies which organize taxonomies information. This paper addresses about attributes and describes relationships between biomedical concepts. The system is also compatible for integration and deployable with databases and description logic based ontologies. Semantically annotated biomedical databases thus can be easily extended with powerful and expressive semantic enabled queries with DBOntoLink to use major ontologies hosted at NCBO Bio Portal, with high query efficiency achieved through caching management.

### *Future Work*

We are unable to explain when more semantic operations can be extended in the future, such as complex semantic operations. This could be implemented by extending current set of semantic operators, and providing corresponding translation in the semantic query engine and UDFs in the database. Another future work is to extend the system to support more database management systems. Since different DBMSs have their own UDF frameworks, we can port current UDFs to databases such as Oracle, PostgreSQL and mySQL by implementing the UDF interfaces, where the internal Java codes can be reused across DBMSs.

## REFERENCES

[1] Cimino JJ. and Zhu X. The practical impact of ontologies on biomedical informatics. Methods Inf Med 45 Suppl 1 2006:124–35.

[2] Rubin DL, Shah NH. andNoy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2007.

[3] NCBO BioPortal. bioportal.bioontology.org/. (accessedJul 2013).

[4] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of semantic distance," Comput. Linguistics, vol. 32, no. 1, pp. 13–47, 2006.

[5] J. E. Caviedes and J. J. Cimino, "Towards the development of a conceptual distance metric for the UMLS," J. Biomed. Inf., vol. 37, no. 2, pp. 77–85,2004.

[6] A. Hliaoutakis, "Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline," Master's thesis,Tech. Univ. Crete, Chani´a, Crete, 2005.

[7] Caviedesa, J.E., Cimino, J.J.: Towards the development of a conceptual distance metricfor the UMLS. Biomedical Informatics 37(2), 77–85 (2004).

[8] Ashburner, M., Sim, I., Hute, C.G., Solbrig, H., Storey, M.A., Smith, B., Day-Richter, J.,Noy, N.F., Musen, M.A.: National Center for Biomedical Ontology: AdvancingBiomedicine through Structured Organization of Scientific Knowledge. OMICS AJournal of Integrative Biology 10(2), 185–198 (2006)Ontology-based Queries over Cancer DataAlejandra Gonz´alez-Beltr´an1,2, Ben Tagger1, and Anthony Finkelstein.

[9] Enabling Ontology Based Semantic Queries In Biomedical Database Systems,Shuai Zheng.

[10] Semantic Data Integration Environment for Biomedical Research, (Ontological Framework for e-

Science & Imaginizer)V. Astakhov, B. Sanders, Jeffrey S. Grethe, E. Ross, D. Little, A. Gupta, T. Astakhova.

[11] An Evaluation Framework for Large-ScaleOntology-based Biomedical Data IntegratedSystemsby, Gilbert Maiga.

[12] Biomedical Data Management: a ProposalFrameworkDouglas TEODOROa,1, Rémy CHOQUETb, Emilie PASCHEa, Julien GOBEILLa,c,

[13] Mining Biomedical Ontologies and Data Using RDFHypergraphsHaishan Liu, Dejing Dou.

[14] Current status of ontologies in Biomedical and Clinical Informatics,Rishi Kanth Saripalle.

[15]http://www.immunetolerance.org/sites/files/BMIR-2007-1245.pdf

[16]http://www.ncbi.nlm.nih.gov/pubmed/23404054

[17]http://www.sersc.org/journals/IJDTA/vol4_no2/2.pdf

[18]http://sophocles.gws.uky.edu/~ben/Interface.html.

[19] H. G. Molina et al., The TSIMMIS Approach to Mediation: Data Models and Languages. Journal ofIntelligent Information Systems **8** (1997), 117-132.

[20] S. B. Davidson, C. Overton, V. Tannen and L. Wong, BioKleisli: A Digital Library for BiomedicalResearchers. International Journal of Digital Libraries **1** (1997), 36-53.

[21] C. A. Goble et al., Transparent access to multiple bioinformatics information sources. IBM SYSTEMSJOURNAL **40** (2001), 532-551.

[22] Y. Arens, C. A. Knoblock and C. Hsu, Query Processing in the SIMS Information Mediator. AdvancedPlanning Technology, Austin Tate, AAAI Press, Menlo Park, CA (1996).

[23] A. D. Preece et al., The KRAFT Architecture for Knowledge Fusion and Transformation. Proceedings of the 19th SGES Int. Conf. on Knowledge-based Systems and Applied Artificial Intelligence (1999).

[24] T. Etzold and P. Argos, SRS – an indexing and retrieval tool for flat file data libraries. COMPUTERAPPLICATIONS IN THE BIOSCIENCES **9** (1993), 49-57.

[25] G. D. Schuler et al., Entrez: Molecular biology database and retrieval system. COMP. METHODS FORMACROMOLECULAR SEQUENCE ANALYSIS **266** (1996), 141-162.

[26] P. Kersey L. Bower, L. Morris et al., Integr8 and Genome Reviews: integrated views of completegenomes and proteomes. NUCLEIC ACIDS RESEARCH **33** (2005), D297-D302.

[27] C. Lovis et al., DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Tech Inform **136** (2008), 641-6