Ontology Driven Information Retrieval System: An Application to extract information about Dementia Care.

http://isit990.azurewebsites.net/

Submitted By

5035077 | Nabin Thapa

5151892 | Karan Manandhar

5189524 | Santosh Tiwari

Submitted to

Associate Prof. Ping Yu

Abstract

Nowadays, the amount of information available has reached such a high level that traditional keyword based search engines find it more and more difficult to effectively provide users with meaningful and relevant information from the internet. The effectiveness of a keyword based information retrieval system decreases as the volume of information on the internet increases. The concept of semantic web has been around for more than 15 years and could be a solution to this problem. This report will look at the design and implementation of an ontology based information retrieval system which aims at providing relevant and more meaningful results to users based on an ontology in the domain of dementia.

Introduction & Background

As of writing this, there are more than 4.8 billion pages on the internet (Worldwidewebsize.com, 2017). This gives us an idea of the amount of information that is available on the internet. Traditional search engines have helped users for years to retrieve information from this vast universe of information as per the query of the user. The keyword based approach to information retrieval from the internet is however not as effective as the users expect it to be largely because of the lack of semantics and context in the queries provided to the information retrieval system (Xu, Zhang and Niu, 2008; Fan and Li, 2006). One major issue with keyword based approach is that it fails to provide any semantic relationship between keywords and has no understanding of the intended meaning of the words and phrases Amudaria and Sasirekha, 2011). Therefore, it is important to bring a paradigm shift in the information retrieval system to make it more effective.

In 2001, Tim Berners Lee introduced the concept of Semantic Web which is meant to allow reasoning and inference capabilities to be added to the existing otherwise descriptive resources present on the internet. Semantic web is actually a standard incorporated by the W3C to promote common data formats and exchange protocols on the web. The concept of semantic web was very simple. The web so far was only understandable by humans. The goal of the semantic web project was to make the information in the internet be understandable by machines to infer and create knowledge base for artificial intelligence (W3.org, 2014).

Medical professionals and care service providers have benefited immensely from the information present in the internet for both research and treatment purposes. However, due to the lack of effectiveness in the information retrieval mechanism of the traditional search engines, medical practitioners have often struggled to find helpful information for their queries. Although the information is available on the internet but the sheer volume of it makes it very difficult to find the information that is wanted. This is largely because of the silo nature of information on the internet which means that the information on the webpages exists without the knowledge of information on other pages on the web. The information is linked but only in the navigational sense not with respect to the meaning of the information. The semantic web introduces a approach where data on the internet can be linked using various tools and frameworks such as RDF (Resource Description Framework), OWL (Web Ontology Langauge), SPARQL (SPARQL Protocol and RDF Query Language), etc.

Ontologies are frequently used in information retrieval being their main applications the expansion of queries, semantic indexing of documents and the organization of search results. Ontologies provide lexical items, allow conceptual normalization and provide different types of relations (Jimeno-Yepes, Berlanga-Llavori and Rebholz-Schuhmann, 2010). Ontology is one of the foundation for achieving what Tim Berners-Lee's vision of the semantic web where information on internet is linked from a source to any other source and is understood by computers so that they can perform increasingly sophisticated tasks on the user's behalf.

We believe that an ontology based approach for a information retrieval system for the domain of dementia care will be more effective in helping dementia care service providers gather information which is critical for the treatment of dementia patient. While most of the current work on ontology based information, retrieval involves a process where the user query keywords are

transformed through some of Natural Language Processing method and an ontological meaning is derived from the query, the approach followed in this report is slightly different in the sense that the user cannot simple type in any random keywords but can only develop a query from the existing triples in the ontology.

This report introduces an information retrieval system where the user can visually build semantic queries from a given ontology domain and convert that into SPARQL query which will be executed against an ontology graph. This will be more effective in producing results than the traditional keyword based approach.

Similar Work

WikiArt

WikiArt is an ontology based information retrieval system, not only able to integrate three different types of source of information, a database, a wiki and an ontology but also efficient to generate thematic paths automatically in order to browse contents. The common ontology is used to describes features of art images required to retrieved. Contents are organised in relational database and interaction with users is managed by agent. Where contents are supply by the agent on the analysis of query composed by tags. each and every concept in WikiArt has relation with other concepts in a distinct ontology. The browsing of ontology is explained by the helps of general exploration patterns. The core of WIkiArt is wiki which contains more than 15000 artworks and more than 4000 artists in a very large relational database. Three kinds of users, visitors, editors and administration can browser WikiArt internet and access level is limited by access control technique. The WikiArt database is mapped in to WikiArt.owl ontology is known as OWL-DL profile. Where basic concept is entity tables in which column are known as datatype property and has no primary keys. Those columns have primary keys are defined as object properties. WikiArt is able to purpose thematic paths about specific subject. Where thematic paths defined as the sequences of subjects mostly in time line order. In general ontology browsing can be compared with graph browsing. In which concepts are nodes and relations are arc in the graph. When there is no order among relations then system choose randomly. The WikiArt is a system based on rule and consist of three components namely, the pattern matcher and the agenda, the working memory and the execution engine. In a nutshell, WikiArt is an expert system that provides an ontology about arts. However, it is not semantic wiki because ontology has not employed to tag semantically.

PIBAS FedSPARQL

The availability of chemical, biological and pharmacological data is huge however, these data are in isolate form and it hard to queried together in a straight forward direction. The application of semantic technologies making possible to create the link and mapping among the dataset. Semantic method links all the dataset like as a single and linked into network. As a result, search can be carried out across the dataset. According to this article PIBAS FedSPARQL application has developed that use the semantic technology which allowing for researcher to search across the vast array of data sources. The PIBAS FedSPARQL is detect similar items from various URIs on the base of text mining algorithm and named entities that is applied in vector space model. This application is unique because it can search similar data items because system construct and run federated SPARQL query in multiple data sources such Bio2RDF, Chem2Bio2RDF are known as global initiatives. The initial federated SPAQRL query is created and executed on the basis of input topic, keyword, subtopic and template to obtain data. Moreover, users can choose appropriate data source as per their interest and exploit RDF structure that enhance the query results. To sum up, the proposed system is considered flexible, which allows execution of queries in broad range and similar data items detection algorithm is also useful because that suggest the new data sources. Moreover, cost optimization for new experiment is also important aspects of this application.

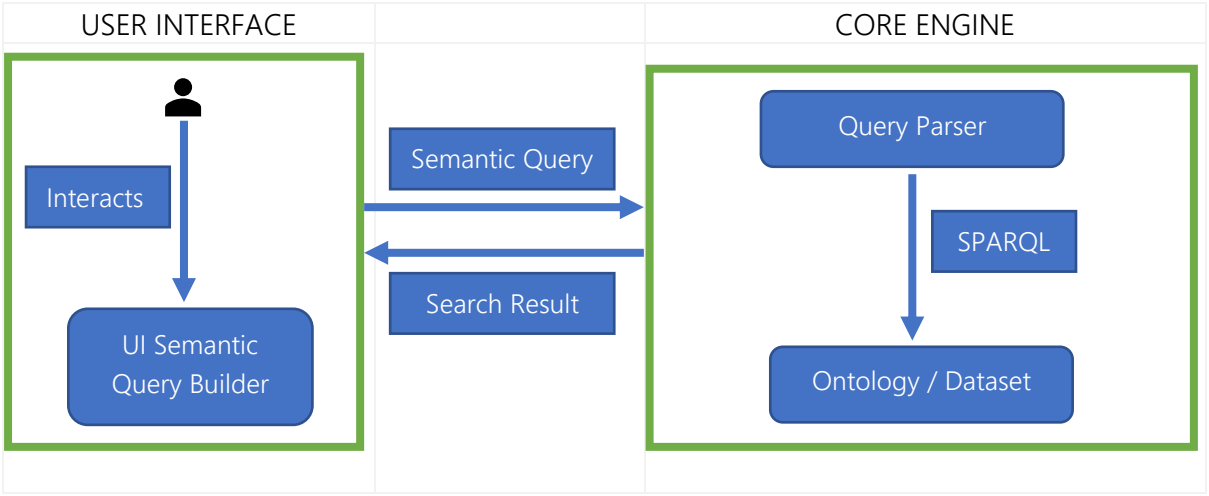
Ontology Based Information Retrieval System for Academic Library

The information retrieval system is inevitable part of current search engine that based on keywords, however, which results the large amount of information to users. So, it is hard for user to understand which information is important and essential as per their requirement. This problem could be solved

by the new web architecture known as semantic web. Semantic web is a vision the idea of having data on the web defined and linked in such a way. This can be used by the machined not only for the purpose of display but also for automation integration as well as reuse across different application. In this article, ontology based information retrieval system for academic library has proposed. That includes the development of ontological database for storage of domain specific language, SPARQL query from user input query, semantic search of NL query and retrieve the related answer from the domain specific ontology. The ontology is divided into three parts, ontology capture, ontology coding as well as possible integration with existing ontology. In this purposed system, the user interface is created where user input the query in natural language. Then it further processed by standford parser then ontotriples are constructed by ontology. After that SPARQL query is formed, which then fired on knowledge based and find appropriate RDF triples using Jena semantic web Framework. The WordNet is a lexical database for English language, which help to find the answer by using the synonyms of entered letters. In a nutshell, sample output for query after parsing by standford parser and then receive subject, predicate and object by using extraction algorithm. Moreover, SPARQL query is generated for these triples to get the relevant answer by the use of Jena API.

Overview of the system

The core concept behind the proposed system is that the user will be able to generate semantic queries within the user interface from an ontology which will then be parsed by the system and converted into a SPARQL query that will executed against the ontology and the output of the SPARQL query will be displayed to the user as the search result.



The user is presented with a web application where the user can build a semantic query using the semantic query builder. The user submits the final semantic query which is then posted to the server where the query is parsed ad converted to a SPARQL query using a query parser and the obtained SPARQL query is finally executed against an ontology. The retrieved result from the execution is returned to the user as a search result query.

The Ontology

The ontology is an adapted version of the one produced by Associate Professor Ping Yu at the University of Wollongong, Australia. The ontology has been designed for the domain of dementia care and attention has been provided to maintain a high degree of quality of the ontology. The ontology has [CLASS_COUNT] classes, [AXIOMS_COUNT] axioms, [PROPERTY_COUNT] properties, and [INDIVIDUAL COUNT] individuals. Each class has a corresponding individual which allows other individuals to reference to a class with an object property. This is referred to as Punning (W3.org, 2008).

The ontology basically has two parts one for the dementia care and the other one for the Web Resources. The dementia care part of the ontology contains all the classes representing the dementia domain such as symptoms, types of dementia, impact, risk factors, etc. While the Web resources part classes such as websites and webpages to represent the domain of world wide web as a source of information for the system. The webpages are the primary source of information for the system. The ontology has individuals from 3 different websites. Namely,

- Alz.org
- Fightdementia.org.au
- Alzheimers.org.uk

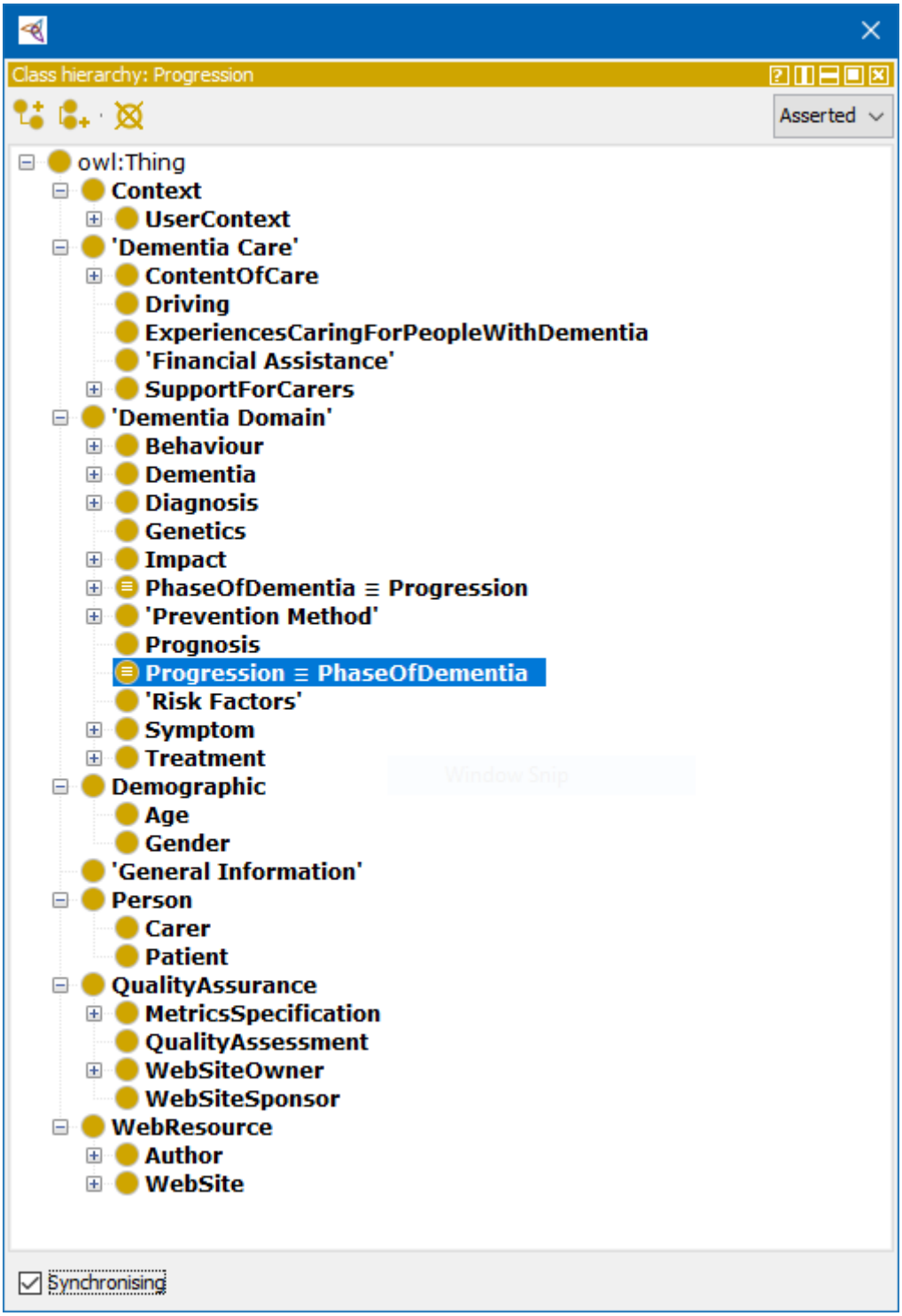


Figure: Screenshot of classes in the ontology inside Protege

Classes represent the concepts in a domain (Noy, 2004). For this system, domain is dementia care and the above figure shows all the classes with their hierarchies in the ontology.

The websites are represented as individual of type Website and each website has at least 30 webpages represented as individuals of type Webpage. These webpages form the source of information. Any information available on the webpage is assigned a 'has definition' property to

the corresponding individual of the type of information. For example, in the screenshot below, we can see that the selected individual "Webpage_AlzOrg_Why_Get_Checked" is an individual of

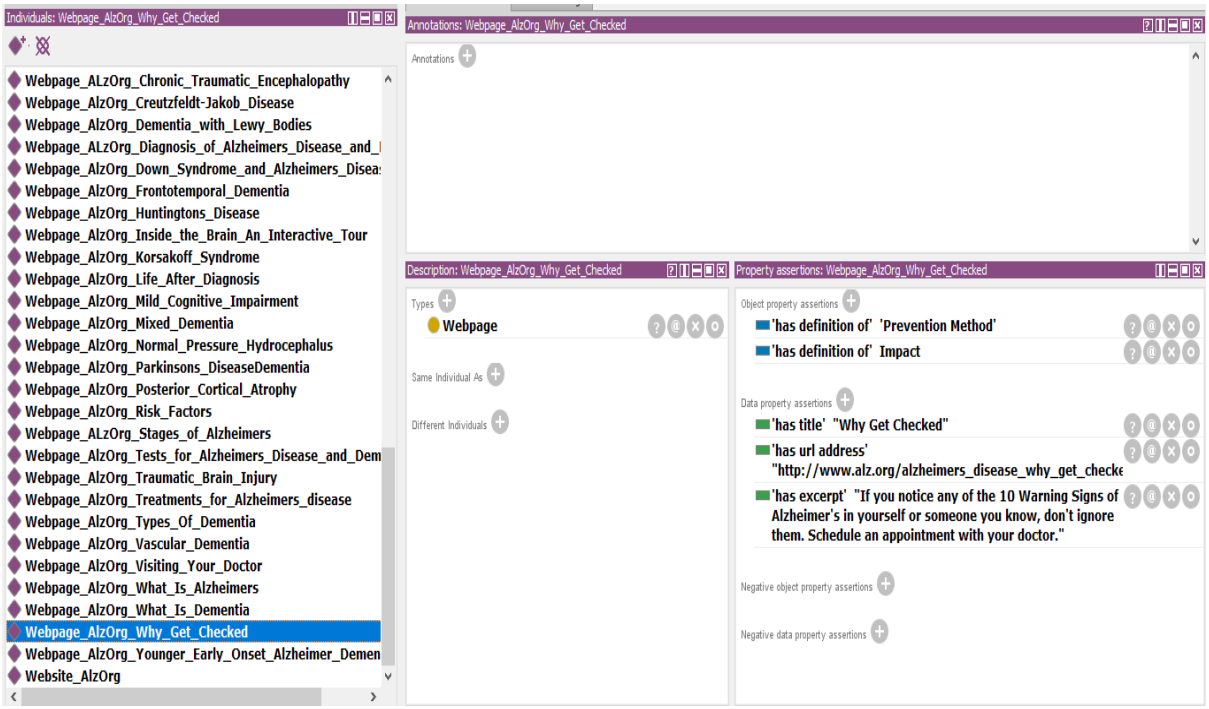


Figure: A screenshot of an individual tab inside protege

type Webpage and is linked with "Prevention Method" and "Impact" with an object property called "has definition of" and also has three data properties to provide values for the title, URL and a brief excerpt of the content of the page. This allows the system to list out all the webpages where the information requested by the user is available.

Properties can be used to state relationships between individuals or from individuals to data values (McGuinness and Harmelen, 2004). The following screenshot shows all the asserted object properties that is being used in the ontology to establish relationship between different concepts represented by the classes.

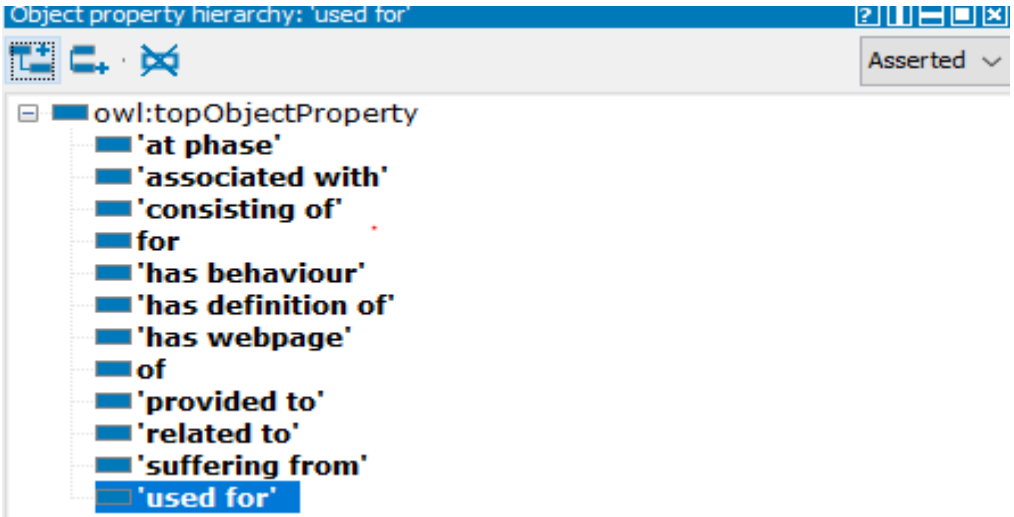


Figure: A screenshot of the Object Properties tab inside Protégé showing all the object properties used in the ontology.

Similarly, data properties relate individuals to literal values which can be of type strings, numbers, dates, etc. Following screenshot shows the data properties used in the system.

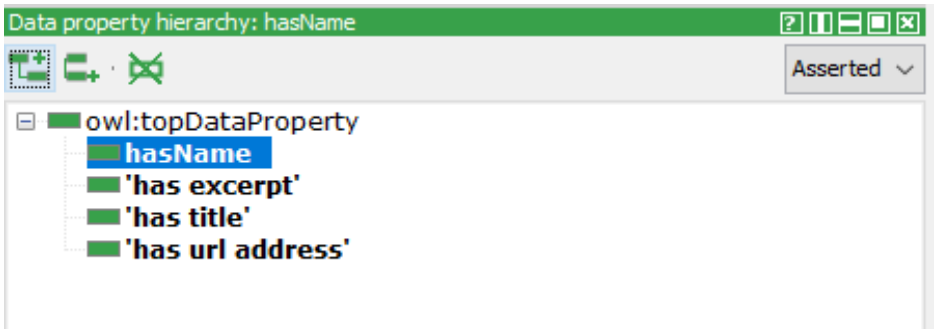


Figure: A screenshot of the Data Property tab inside Protégé showing all the data properties used in the ontology.

The following table provides the domains, range and description of the properties used in the ontology.

Property Name	Type	Domain	Range	Description
atPhase	Object Property	*Dementia*	*PhaseOfDementia*	used to relate Dementia with its phase
associatedWith	Object Property	*Symptom*	*Dementia, TypesOfDementia*	used to relate symptoms with the disease dementia
consistingOf	Object Property	*Symptom*	*All subclasses of Symptom*	used to relate symptoms to particular types of symptoms
for	Object Property	*Symptom, PreventionMethod, Treatment*	*Dementia, PhaseOfDemtnia*	a generic property used to represent a for relation between two entitites
hasBehaviour	Object Property	*Symptom*	*Behaviour*	used to show relationship between symptom and behaviour
hasDefinitionOf	Object Property	*Webpage*	*Dementia*	used to show relation between a webpage and anything under dementia domain
hasWebpage	Object Property	*Website*	*Webpage*	used to relate a website to a webpage if the webpage is part of the website

of	Object Property	*PreventionMethod, RiskFactors, Impact, Prognosis, Diagnosis, Symptom, Treatment*	*Dementia, PhaseOfDementia*	a generic property to relate two concepts with an semantic for relation
providedTo	Object Property	*FinancialAssistance*	*Person*	used to relate financial assistance being provided to a person and its subclasses
relatedTo	Object Property	*RiskFactors*	*Dementia*	used to relate Risk Factors associated with dementia or its subclasses
sufferingFrom	Object Property	*Patient*	*Dementia*	used to relate relation between a patient and dementia or its subclasses
usedFor	Object Property	*Treatment*	*Dementia*	used to relate relation between treatment and its subclasses to dementia and its subclasses.

The Implementation

The role of the system is to read the RDF data which is stored in the Ontology and present it to the user.

Terminologies and Assertions

The schema of the ontology consisting of classes and properties makes up the terminologies and the instances represented by individuals in protégé make up the assertions or the data for our knowledge base.

The information which is presented to the user is stored as an instance of owl:Class Webpage. So all the individuals that have a type of Webpage is the source of information for the user. For example a web page in the internet (found here) can be represented by an individual of type "Webpage" and can be attributed with a "hasDefinition" property to provide relation to any information that is present in the page. For example, in the following screenshot we can see

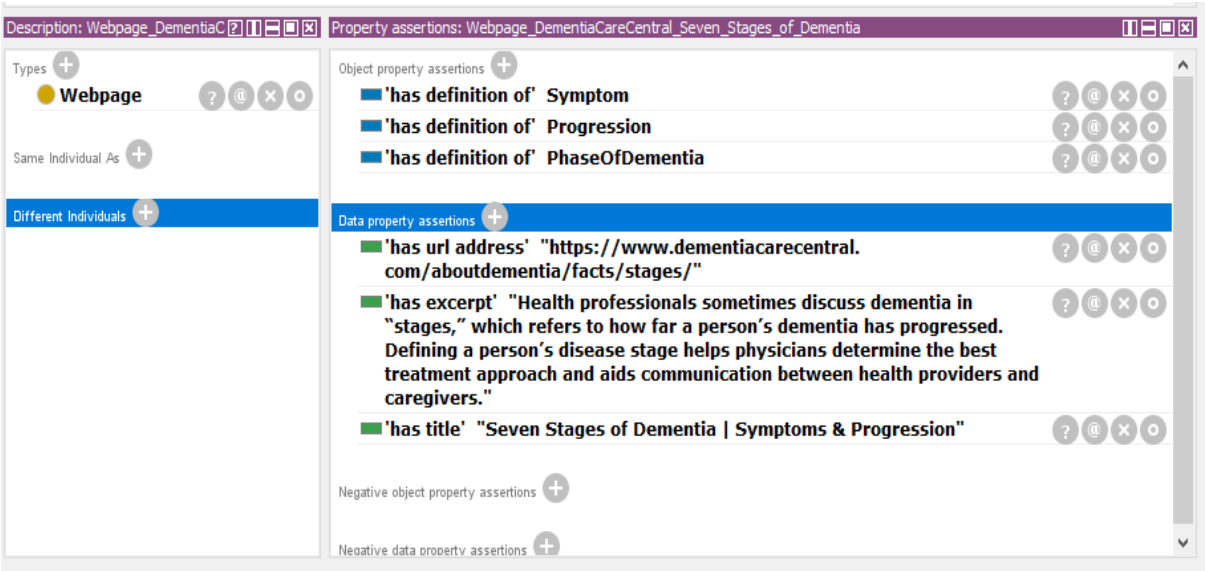


Figure : Screenshot of descriptions and properties of individual "Webpage_DementiaCareCentral_Seven_Stages_of_Dementia"

That the individual "Webpage_DementiaCareCentral_Seven_Stages_of_Dementia" with a type Webpage has definition of Symptom, Progression, Phase of Dementia which is represented by the object property "hasDefinition" connecting to the respective classes. Similarly, the data properties such as "has url address", "has excerpt", "has title", respectively provide literal values to the url, excerpt and title of the webpage.

Query Builder

When the user starts typing anything on the browser, the system first retrieves the list of all classes in the ontology from which the user can select a class/concept. After the selection the user is now provided with all the properties that have rdf:domain of the class that has been selected previously. After selecting a property from the list the user is now presented with all the classes/concepts that are the rdf:range of the previously selected object property. This forms a single triple.

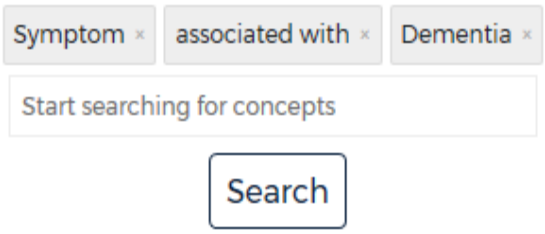


Figure: Query created by the user containing 1 triple

Like for example in the above figure, Symptom is the first selected class. The property "associated with" is an object property which has Symptom as a rdf:domain. Finally Dementia is the class that is the rdf:range of the object property "associated with". In this way any number of triples can be created as a semantic query. The above query can be further expanded to generate a query similar to the following.

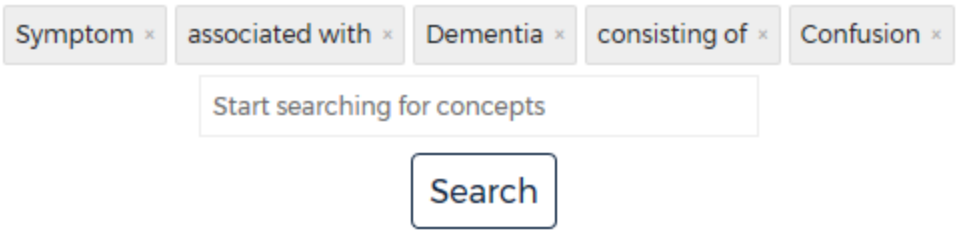


Figure: Query created by the user containing 2 triples

Query Parser

When the user clicks on the Search button the query generated by the user is then sent to the server to be translated to a SPARQL query. This is done by going through each class in the semantic query sent by the user and checking if it has been defined by any webpages. All those webpages that have definitions of all the selected classes are returned as the search result.

In the following figure, the resulting SPARQL query from the user's semantic query can be seen. The following figure shows the actual individual which contains all the information requested by the user and hence is returned to the user as search result.

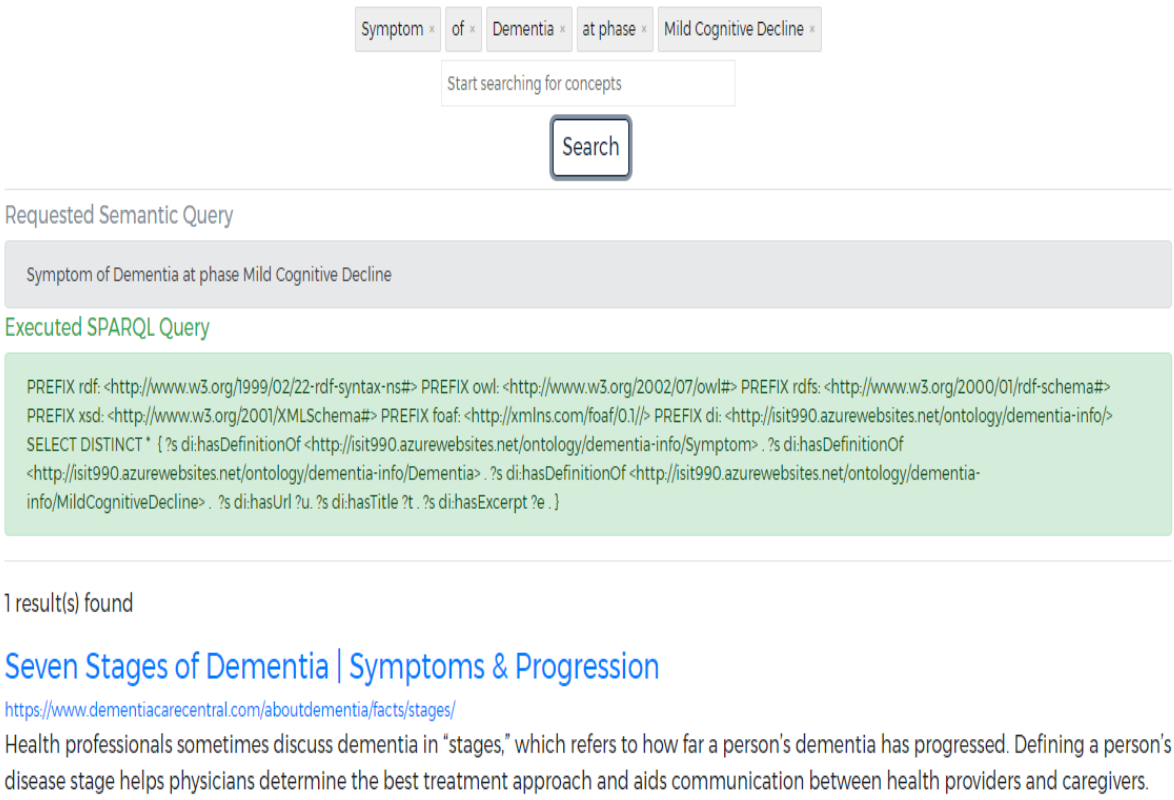


Figure: Example search query

The result displayed in the above query is the following individual in the ontology.

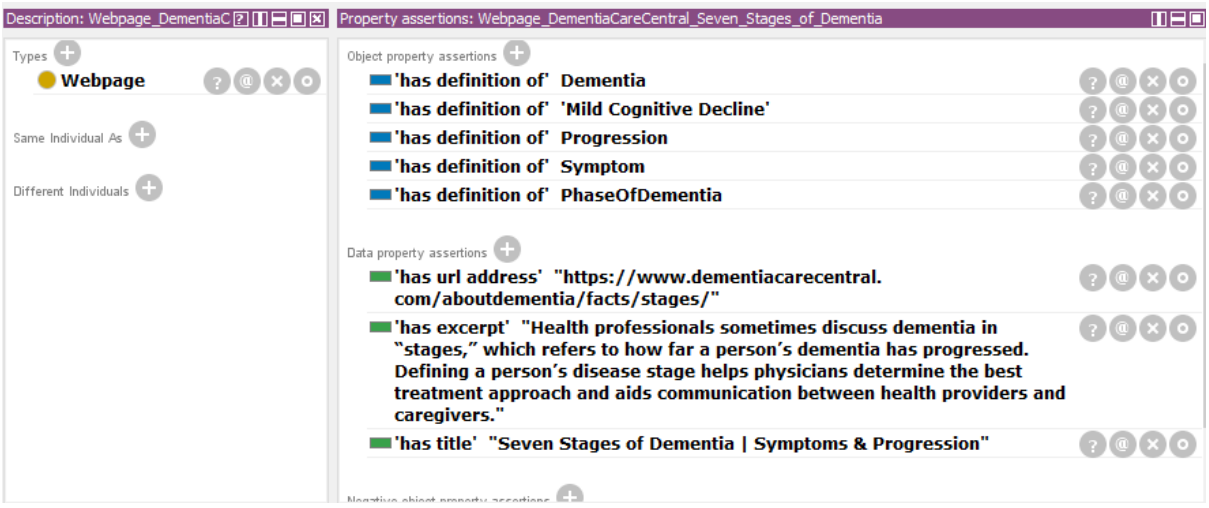


Figure: Individual used for given search query

As it can be seen, that the given individual has definition of all the concepts that were selected by the user and hence is returned to the user as the search result.

Limitations

The system does deliver a promising way to retrieve information from internet using ontology. However, the system also has various limitation that needs to be addressed. One of the main characteristics of an ontology is its ability to apply logic using reasoning capability. However, in the current implementation of the system, SPARQL queries only provide, information of whatever explicitly provided RDF data is present in the ontology. It is not able to take any implicit assertions into account while returning results to the user.

Secondly, the data being collected was done manually by the three group members of the project from three different websites which proved to no very effective and efficient. An automated crawler needs to be developer which will automatically generate the individuals from webpages belonging to any website. A system needs to be developed to achieve such a mechanism.

Lastly, the user of the system also needs to have some general idea about the schema of the ontology in order for the user to generate queries using the front-end query builder which can prove to be challenge for the system to be usable as users are not very keen on learning any new system.

Conclusion

In this report we looked at the design and implementation of an ontology based information retrieval system. The system is able to allow users to dynamically generate simple triple based semantic search queries which is in turn converted to SPARQL query which is executed against the RDF data present in an ontology which provides information to the users as a search result. The system can be helpful to create complex queries with long semantic meaning that can outperform traditional search engines that are based on simple keyword search strategy.

References

Amudaria, S. and Sasirekha, S. (2011). Improving the precision ratio using semantic based search. 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies.

Jimeno-Yepes, A., Berlanga-Llavori, R. and Rebholz-Schuhmann, D. (2010). Ontology refinement for improved information retrieval. Information Processing & Management, 46(4), pp.426-435.

Fan, L. and Li, B. (2006). A Hybrid Model of Image Retrieval Based on Ontology Technology and Probabilistic Ranking. 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06).

McGuinness, D.L. and Van Harmelen, F., 2004. OWL web ontology language overview. W3C recommendation, 10(10), p.2004.

Noy, N. (2004). Ontology Development 101. [ebook] Standford University. Available at: https://protege.stanford.edu/conference/2004/slides/Ontology101_tutorial.pdf [Accessed 18 Oct. 2017].

Ross, G. (2005). An introduction to Tim Berners-Lee's Semantic Web. [online] TechRepublic. Available at: http://www.techrepublic.com/article/an-introduction-to-tim-berners-lees-semantic-web/ [Accessed 8 Oct. 2017].

Tim Berners-Lee, J. H. a. O. L., 2001. The Semantic Web. Scientific American, , 2841(5), p.

Xu, X., Zhang, F. and Niu, Z. (2008). An Ontology-Based Query System for Digital Libraries. 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application.

W3.org. (2008). Punning - OWL. [online] Available at: https://www.w3.org/2007/OWL/wiki/Punning [Accessed 18 Oct. 2017].

W3.org. (2014). Semantic Web Standards. [online] Available at: https://www.w3.org/2001/sw/wiki/Main_Page [Accessed 8 Oct. 2017].

Worldwidewebsize.com. (2017). WorldWideWebSize.com | The size of the World Wide Web (The Internet). [online] Available at: http://www.worldwidewebsize.com/ [Accessed 11 Aug. 2017].