Ontology based Semantic Search Engine for Cancer

Syam Raj B S
Anna University
JayShriram Group of Institution
Dharapuram Road, Avinashipalayam
Tamil Nadu India

Sarumathi S
Anna University
JayShriram Group of Institution
Dharapuram Road, Avinashipalayam
Tamil Nadu India

ABSTRACT
The idea is to analyze the knowledge about the real world and then create a standard upon stabled rules and relation types to translate the human (natural) language in a machine and human readable language. For that it need to classify and organize data such as text, pictures, videos or database entries in a system with logical connections between data representing the knowledge shared by people. The Ontology provides a framework for the development of Semantic Web and Artificial Intelligence. Here Medical Knowledge Engineering is the Key. This paper deals with the Medical Knowledge Base to build an ontological structure. In this paper Medical Knowledge about cancer is been combined with the semantic web search engine. Based on the introduction of ontology theory, the author uses Protege 2000 of Stanford, the construction and maintenance tool of ontology, designed and completed Medical Knowledge based on Ontology and all details about cancer, cancer categories, its cause, symptoms etc. The system also learned from this details and new details from the searching process. The improvement and learning process is achieved by comparing the details with some knowledge organization systems. Knowledge acquisition in semantic web is done by RDF explorer. RDF scheme defines relationship and those relationship make the searching in a different level.

General Terms
Semantic web vision, RDF Schema Micro-format etc

Keywords
Semantic web**,** Cancer Protege, RDF,RDFS,OWL.

1. INTRODUCTION

1.1 Semantic web vision
After the invention of the World Wide Web, Tim Berners-Lee proposes the Semantic Web. The Semantic Web simply means the web of meaning. In the World Wide Web information is presented in natural human language which is not rich enough to convey formal meaning and therefore it is not machine processable. This current web contains millions and millions of resources such as HTML files, documents, images and graphics, and media files. These resources contain huge amounts of information scattered in various web pages and documents. The current web is a web of documents and understandable only to humans. This makes information retrieval processes very hard; humans alone cannot deal with this huge amount of resources on the web. Software agents or machines could help in this process but a difficulty arises from the fact that machines do not understand human language. Trying to make machines act as humans is a very complex task and needs a lot of training. The idea of the Semantic Web was introduced mainly to solve the problem that content on the current web is intended only for human consumption. The basic idea of the Semantic Web is to give information a well-defined meaning, thus better enabling agents and people to work in cooperation [1]. W3C states [2] "**The Semantic Web is about two things. It is about common formats for interchange of data, where on the original Web we only had interchange of documents. Also it is about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing**". This paper simply says that the Semantic Web is a web of data rather than a web of documents. Semantic Web is about two things: It is about common formats for interchange of data, as opposed to documents. This data is well-defined so that agents will fully comprehend the semantics of the data. Also it is about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one knowledge base, and then move through an unending set of knowledge bases which are linked by being about the same or related domains. The Semantic Web can be viewed as a mesh of information linked up in such a way as to be easily processable by machines, on a global scale. It is an efficient way of representing data on the World Wide Web, or as a globally linked database. The challenge of the Semantic Web was to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web. As stated by Berners-Lee [1] "**Making the language for the rules as expressive as needed to allow the Web to reason as widely as desired**". The Semantic Web uses RDF (Resources Description Framework) to represent information. Each piece of information on the Semantic Web is called a resource and each resource is uniquely identified. Information about resources is represented as a Directed Graph of triples (Subject, Predicate, Object) also called RDF statements. Knowledge on the Semantic Web is stored in an ontology. The ontology holds both the data and metadata, which enables understanding the semantics. The Semantic Web structure enables not only combining Semantic Web statements to create larger pieces of information but also the ability to infer new information based on the rules defined in its ontology. Figure 1 shows the different layers of the Semantic Web. The three upper layers are still under construction and are not final yet. Logic or reasoning is one of the major important issues for Semantic Web and it is an important design issue when creating a Semantic Web agent.

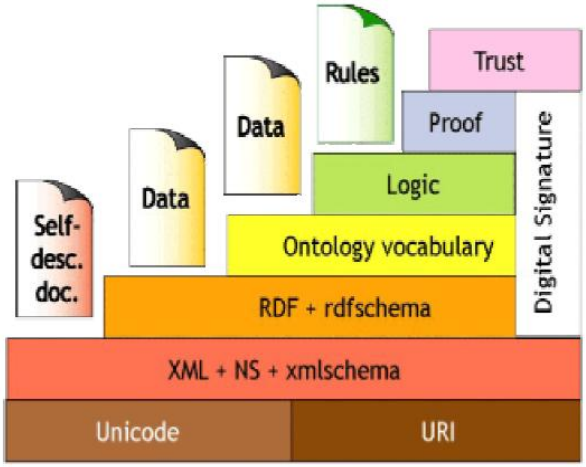

Figure: 1 Semantic Web layers [5]

1.2 WWW TO SEMANTIC WEB

Semantic Web is not intended to replace the existing web but rather it is an extension of the current Web. The current web depends on visual representation of information through HTML tags. This visual representation makes information clear for humans to understand but very difficult for machines to understand and process. For example to emphasize something it could be in a different font or color. Some form of extraction is required to strip off the information part from the presentation part. Other techniques are used to infer meaning from this information; this leads to an increased complexity in the agents dealing with the World Wide Web. Another problem with the current web is the fact that different terms are used to represent the same meaning, for example in a shopping site could refer to the shopping cart as cart, while another would refer to it as shopping basket or basket for short, yet another site could refer to it as shopping bag. All these words refer to the same meaning or the same semantics, which is very obvious to humans while it is unknown to software agents. These agents have to be explicitly informed that the previous terms are all the same. Another example comes from the fact that the web is multi lingual; an English shopping website would use the word price to refer to an items price, while a Dutch website would use the word prijs, a French website would use the word prix, a Spanish site would use the word precio, an Italian site would use the word prezzo, and an Arabic site would use the word الثمن . An agent that is looking for a product and comparing prices to retrieve a list of the cheapest sites would have to be familiar with these terms. These are just a sample of languages that exist on the web while there are many more. The Semantic Web targets solving these problems by providing not only the data but also metadata that describes explicitly what this data means. This form of data annotation makes an agent understand the semantics behind the data and thus allows for better interpretation between data and gents and allows for better inter agent communication and collaboration. As stated by Berners-Lee [4], "this notion of being able to semantically link various resources (documents, images, people, concepts, etc) is an important one. With this it can begin to move from the current Web of simple hyperlinks to a more expressive semantically rich Web, a Web where it can incrementally add meaning and express a whole new set of relationships (hasLocation, worksFor, isAuthorOf, hasSubjectOf, dependsOn, etc) among resources, making explicit the particular contextual relationships that are implicit in the current Web. This will open new doors for effective information integration, management and automated services". The Semantic Web promises a solution in which the web becomes one big knowledge base and everyone has access to it. In order for this to happen there should be supporting technology that allows for such annotation in a formal and unified syntax, such annotation are RDF/RDFS and OWL which are standards set by the W3C. Also reasoning on the Semantic Web promises for more intelligence in services provided by the web such as personalized notifying agents, search agents, personalized search agents, e-learning and many other applications where agents would pull the information and process it having a better understanding of its meaning.

1.3 ONTOLOGY

The term Ontology has its roots in the philosophical domain. In order to understand the basic structure of the world and the study of existence, the word ontology has been connected with a branch of metaphysics. The problem is that the philosophical definition of ontology is not easy to port to the scientific domain. Therefore Dunwoodie [2007] uses an intelligible definition of ontology:"An ontology is a detailed model/picture/schema of a slice of reality which is based on the facts that know about that reality. This model /picture /schema is a description of some of the things and some of the relationships between the things that are known about that reality" [11]. In Helfin [2004], the term "Ontology" is defined as following:"Ontology defines the terms used to describe and represent an area of knowledge". These ontologies can be shared by different applications, people and databases within a domain.

A domain can be an area of knowledge, like medicine or a specific subject area. The definitions of ontologies are machine readable and they describe basic concepts in the domain and the relations between them. The knowledge, which is encoded in ontologies, is reusable due to the fact that the encoded knowledge can span different domains. Ontologies are able to specify the following kinds of concepts, which enable the description of almost every knowledge:

- Classes (things)
- Relationships between things
- Properties (attributes) of things

There are many motivations for developing and using ontologies:

- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

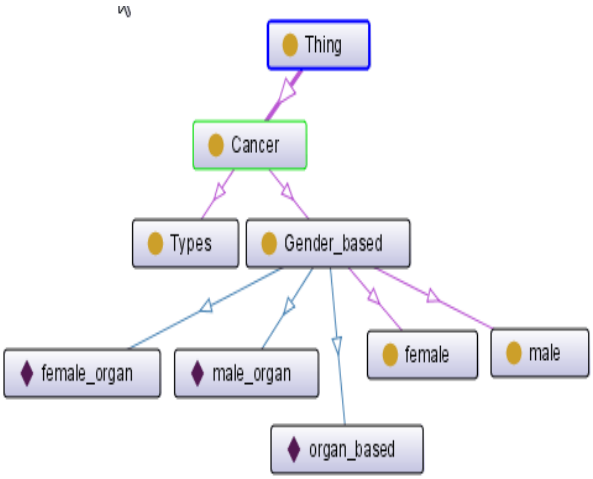

Figure: 2 Cancer Detection Ontology

1.4 RDF SCHEMA (RDFS)

RDF is a metadata language that does not provide special vocabulary for describing the resources. It is often essential to be able to describe more of a subject than saying it is a resource. Some form of classification for these resources is often required to be able to be able to provide a more precise and correct mapping of the world. The basic idea behind Semantic Web is to provide meaning of resources, as defined in the Knowledge Representation domain, "knowledge is descriptive and can be expressed in a declarative form" [16]. The formalization of knowledge in declarative form begins with a conceptualization. This formalization includes the objects presumed or hypothesized to exist in the world. This is why RDF schema (RDFS) was introduced as a language that provides formal conceptualization of the world. RDF Schema semantically extends RDF to enable us to talk about classes of resources, and the properties that will be used with them. The RDF schema defines the terms that will be used in RDF statements and gives specific meanings to them. It provides mechanisms for describing groups of related resources and the relationships between these resources. Meaning in RDF is expressed through reference to the schema. RDFS consists of a collection of RDF resources that can be used to describe properties of other RDF resources this makes it a simple ontology language which allows more capture of semantics than just pure RDF. The most important resources described in RDFS are:

- **Classes:** RDFS deals with classes through the term **rdfs:Class** which defines a class. RDFS allows for creating hierarchies of classes in which a class is defined as the subclass of another class using **the rdfs:subClassOf property**. RDFS also allows for creating instances of a class; that is data that are of the type of that class. **The rdf:type property** may be used to state that a resource is an instance of a class. RDFS defines a special class **rdfs:Literal** which is the class of literal values such as strings and integers, **rdfs:Literal** is an instance of **rdfs:Class** [6, 9]

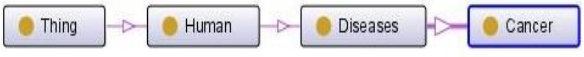

Figure :3 Example RDF representation

1.5 SEMANTIC CONTENT IN OWL

OWL (Web Ontology Language) was introduced to provide richer vocabulary than RDFS. OWL semantically extends RDF/RDFS which means that OWL ontology is an

RDF graph. A formal semantics describes the meaning of knowledge precisely and rich semantics describes fine grained knowledge. OWL was introduced by W3C to provide a richer ontology language that allows for: a well-defined syntax, efficient reasoning support, a formal semantics and sufficient expressive power [7]. The main content of OWL ontology is carried in its axioms and facts, which provide information about classes, properties, and individuals in the ontology. There are a several major capabilities that OWL adds to RDF and RDFS. The first is the ability to create local range restrictions. In RDFS, a property is allowed to have only one class as its range while in many cases there is a need defining more restrictions on the property range restrictions [5].

2. SEARCHING THE SEMANTIC WEBSITES

As the Semantic Web grows in size there will be an increasing need of searching for information. Users will want to perform search queries and expect Semantic Web documents that best match their query as results, in analogy to WWW search; only the expected precision of Semantic Web search should be better due to the better understanding of the search terms. It must be noted that Semantic Web documents are different than conventional web documents, information viewed in Semantic Websites are what the developer wants to present but the semantics behind the presentation is what matters, while with web documents the presentation s simply the way of formatting the looks of the information. For example a computer shopping site in Dutch, Arabic, and English would share the terms from a computer shopping ontology, while the terms are presented in the three different languages but their semantics is the same because they have a common source of semantics. A search engine searching for certain computer specifications would perform its search and return the result based on the user preferred presentation. In this section it will show the analogy between Semantic Web search and WWW search, the requirements of Semantic Web search and the expectations, and finally give some examples of current semantic search agents

3. SEMANTIC WEB TOOLS

Semantic Web development tools are essential for rapid and easy development of Semantic Websites. This chapter introduces some of the existing Semantic Web tools. There is a need for tools to support the development of ontologies and knowledge bases. Also storage tools to store and manipulate triples and finally tools to support query on the stored information. It must also be noted that there is a trend for standardization so these tools should offer standardized formats of RDF/RDFS and OWL. Quite a number of tools exist, such as

- **Ontology Editors: Protégé1**
- **Large Triple stores: Sesame**
- **Full development environments and servers: RDF gateway.**

Ontology editors should be able to validate the consistency of the edited ontology and report errors. Triple store on the other hand should be capable of storing and processing large amounts of triples and perform queries in the least time possible and allow for querying multiple ontologies. Now will

discuss two of these tools in more details namely: Protégé and Sesame.

3.1 Protégé

The first step one needs to take when developing a Semantic Website is to identify and ontology editor to use. In this work Protégé used which is a free, open source ontology editor and knowledge-base framework. It allows visualization of ontologies and is extended by a large number of plug-ins [19]. The Protégé editor supports two main ways of modeling ontologies [20]:

- The Protégé-Frames editor enables users to build and populate ontologies that are frame based, in accordance with the Open Knowledge Base Connectivity protocol 28 (OKBC). In this model, ontology consists of a set of classes organized in a hierarchy to represent a domain's salient concepts, a set of slots associated to classes to describe their properties and relationships, and a set of instances of those classes.

- The Protégé-OWL editor enables users to build ontologies for the Semantic Web, in particular in the W3C's Web Ontology Language (OWL). Protégé in OWL mode enables the control of level of complexity it uses. It provides support for using OWL Lite , OWL DL and OWL full. It also allows for usage of RDF and RDF/RDFS Protégé provides a highly scalable database back-end, allowing users to create ontologies with hundreds of thousands of classes. Protégé also supports saving an ontology in various formats namely RDF/XML N-triple N3. Protégé also has support for namespace usage and ontology reuse through allowing ontology import. When an ontology imports another ontology, all of the class, property and individual definitions that are in the imported ontology are available for use in the importing ontology, Protégé allows for building on these classes and properties and are shown in the ontology editor in read only mode. Protégé also has support for reasoning, users can download the reasoner they would like to use and configure Protégé to use it. This is possible because Protégé is a DIG aware application which enables it to use any DIG29 (DL Implementation Group) reasoner such as: RACER, FaCT++ or any other DIG compliant reasoned [21]. In this work Protégé has used to model RDF/RDFS ontology via the Protégé Protégé-OWL editor. It used the import feature of Protégé to investigate the usage of sharing and reuse among ontologies. Ontoviz [22] visualization plug-in was used to visualize the created ontology

3.2 Sesame

Sesame is an open source RDF database or RDF store (also called RDF repository) with support for RDF Schema inference and querying. It supports both HTTP access and SOAP access for applications to manipulate the RDF/RDFS data remotely or from the web. Internally it supports persistent storage of millions of triples, to do that it needs a scalable repository which is built on top of a DBMS. To be general and not dependent on a certain DBMS, all DBMS specific code is concentrated in a single architectural layer of Sesame called the Storage and Inference Layer. Sesame also supports not using a certain DBMS for the storage of the triples, in this case they are stored persistently in a file and Sesame deals with them as one big memory object, i.e. all operations are done in memory. This latter approach obviously decreases Sesame's performance a little but adds more flexibility to its usage [23]

4. RESULT

The Cancer ontology has given expressive power and syntactic interoperability. The Universal expressive power is high when compared to first and second generation web technologies[4].Since it is not possible to anticipate all potential uses, Ontology have enough expressive power to express any form of data. The ontology is created using protégé. Another feature is Support for Syntactic Interoperability. By syntactic interoperability, mean how easy it is to read the data and get a representation that can be exploited by applications. For example, software components like parsers or query APIs should be as reusable as possible among different applications. Syntactic interoperability is high when the parsers and APIs needed to manipulate the data are readily available.

Browser extensions are very useful if the Linked Data you are working with is not available on the public Internet. Browser-based Linked Data viewers are generally limited by the size of the local graph they can maintain. Hosted browsers often have greater computational resources for managing a graph cache, but they are limited to data sources that are publicly available

5. CONCLUSIONS

Right now the semantic web techniques cannot replace a human. He still must validate all the results that a computer generates. Still the human is the one to formally define concepts, things, and events, real live and presented in a machine-understandable form. Even if the vision about the Web of trust can be still far way, we pointed out the important steps already achieved. Our present work Cancer Ontology covered almost all the life cycle for a semantic application.

This includes:

[1] Cancer and its Property Definition.

[2] Cancer Ontology implementation through.

[3] URI, XML, RDF, RDFs, OWL.

[4] Discovery of new semantic communities in cancer ontology

[5] Browse the Machine Processable OWL data through Ontology Browsers (Data Link Explorer, Manchester Ontology Browser) and DL query execution (SPARQL, RIFs).

We hope that all the above steps contributed to an extent in which the inferred knowledge about Cancer is presented in Machine Understandable as well as human understandable semantic form.

6. REFERENCES

[1] Tim Berners-Lee, Jim Hendler, and Ora Lassila. The Semantic Web. Scientific American, 2001, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C7084A9809EC588EF21&pageNumber=2&catID=2

[2] Semantic web Activity, W3C, http://www.w3.org/2001/

[3] T. R. Gruber, A Translation Approach to Portable Ontology Specifications[J], Knowledge Acquisition, vol. 5, no.2, 1993, pp.199~220.

[4] Tim Berners-Lee and Eric Miller, 'The Semantic Web lifts off'. ERCIM News No. 51, October 2002

[5] Tim Berners-Lee, Semantic Web - XML2000, talk in 2000, http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview-2.html

[6] L. Miller, A. Seaborne, and A. Reggiori, "Three Implementations ofSquishQL, a Simple RDF Query Language," Proc. Int'l Semantic WebConf. (ISWC 02), LNCS 2342, Springer, 2002, pp. 423–435

[7] Frank Manola and Eric Miller, RDF Primer, W3C, http://www.w3.org/TR/2004/REC-rdfprimer- 20040210/ , February 2004

[8] Taboada, D.Martinez, J.Mira; Experiences in reusing knowledge sources using Protege and PROMPT; Int.J.Human-Computer Studies 62(2005) 597-618.

[9] J. Broekstra, A. Kampman, and F. van Harmelen, "Sesame: A GenericArchitecture for Storing and Querying RDF and RDF Schema," Proc.Int'l Semantic Web Conf. (ISWC 02), Springer, 2002, pp. 54–68Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[10] P. Martin and P. Eklund, "Knowledge Retrieval and the World Wide Web," IEEE Intelligent Systems, vol. 15, no. 3, 2000, pp. 18–25

[11] N. Guarino, C. Masolo, and G. Vetere, "OntoSeek: Content-Based Access to the Web," IEEE Intelligent Systems, vol. 14, no. 3, 1999, pp. 70–80

[12] M. Holi and E. HyvÃ¶nen, "A Method for Modeling Uncertainty in Semantic Web Ontologies," Proc. 13th Int'l Conf. World Wide Web (WWW 04), ACM Press, 2004, pp. 296–297.

[13] N.F. Noy and M.A. Musen, "Ontology Versioning in an Ontology Management Framework," IEEE Intelligent Systems, vol. 19, no. 4, 2004, pp. 6–13

[14] Neches R, Fikes R E, Gurber T R, etal. Enabling Technology for Knowledge Sharing[J]. AI Magazine,1999,12（3）:36~56

[15] Gruninger, M. and Fox, M.S. (1995). Methodology for the Design and Evaluation of Ontologies. In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95,Montreal.

[16] Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, Methods and Applications[M]. Knowledge Engineering Review 11(2).

[17] Rosch, E. (1978). Principles of Categorization[J]. Cognition and Categorization. R. E. and B.B. Lloyd, editors. Hillside, NJ,Lawrence Erlbaum Publishers: 27~48.

[18] Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F. and Lorensen, W. (1991). Object-oriented modeling and design[M]. Englewood Cliffs, New Jersey: Prentice Hall.

[19] Protégé , http://protege.stanford.edu/index.html

[20] Protégé plug-ins, http://protege.stanford.edu/download/plugins.html

[21] Using The Protege-OWL Reasoning API , http://protege.stanford.edu/plugins/owl/api/ReasonerAPI Examples.html

[22] Ontoviz, http://protege.cim3.net/cgi-bin/wiki.pl?OntoViz

[23] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen, Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema, In The Semantic Web ISWC 2002.