DEVELOPING A SEMANTIC WEB BASED SEARCH USING ONTOLOGY CLASSIFICATION


*Assistant Professor, School of Computer Science and Engineering, VIT University, Vellore.
**Associate Professor, School of Computer Science and Engineering, VIT University, Vellore.
***Associate Professor, School of Computer Science and Engineering, VIT University, Vellore.
+++Student, School of Computer Science and Engineering, VIT University, Vellore.

Email: virtual.viji@gmail.com

Received on 06-08-2016

Abstract:
Adding Semantic meaning to our knowledge helps to branch across domain of knowledge. Without the use of semantic, it might be tough to find meaningful connection among them. Our graph based model primarily consist data in the form of RDF. OWL (Web Ontology Language) is language that aims to know the web information into the machine-readable content with semantics. Stanford Medical Informatics has developed Protégé-OWL which is an open-source ontology.

W3C has suggested SPARQL to represent the Resource Description Framework graph—which contains subject and predicate. In our case we use RDF data for VIT University website. This has been implemented in a tool referred to as a Protégé that is freely available in the market and an open source tool that give platform to construct totally different models & knowledge-base.

Protégé has been used to build ontology for VIT University domain. Ontology is the characterization of ideas within the domain of knowledge base that are represented as classes and subclasses. Further these concepts or the classes can have attributes and restriction outlined on them. In addition to that, if we want to build a very big ontology, we can amalgamate many existing Ontology to depict a bigger domain. We can also reuse general ontology and expand it to describe our domain of interest. In this paper for executing query we have use the SPARQL query language which is in RDF format which is used to retrieve the data which is being stored.

Keywords: SPARQL, PROTÉGÉ -OWL, Semantic Web, Resource Description Framework (RDF), Web Ontology Language (OWL).
Introduction: The internet of today’s world comprises of an enormous repository of web pages, resources and web services. Because of such huge amount of information, it is difficult for an internet user to find out the desired information. Search engines like Google solves the problem to some extent by indexing the key words of the web pages and then search for user information based on these keywords. However, the basic problem with internet is that there is no semantic information associated with contents over the web. Due to its inherent limitation, machines can only perform syntactic processing of the web contents.

To address this Tim Berners Lee has introduced the term semantic web, which has an aim to add the capabilities of internet by attaching semantic information to internet contents\(^1\). The objective of the semantic web is to provide a better platform for the knowledge representation of Linked data to allow machine processing on a global scale by adding logic, inference and rule systems to the web which allows data to be shared across different applications and boundaries\(^2\). Semantic web is an extension of the existing web, which descript shared conceptual model of a given domain by ontology and provide the semantic description for the data\(^3\). Semantic Web is a combination of a number of elements that includes a number of formal and informal knowledge representation languages, editing and publishing tools and a set of methodologies and standards for semantic web applications\(^2\). It is proposed that knowledge acquisition in the semantic web, has become a hot research.

Ontology supports information exchange based on semantics rather than just syntax\(^7\). Ontology has proven to be useful for knowledge representation, information sharing and system modeling\(^9\). It is widely used in knowledge-based systems, semantic web, natural language processing, and other applications\(^15\). Ontology represents set of precisely defined terms about a specific domain and accepted by this domain’s community. Ontology is an explicit specification of a conceptualization\(^3\). Ontology by combining with a large set of individual occurrences forms a knowledge base\(^2\). In this paper, for working with the ontology the Protégé software has been used\(^5\). Protégé is used as the development tool which is an open-source ontology editor, developed by Stanford Medical Informatics and coded by JAVA\(^12\), and OWL is used as the standard ontology language, combined with the general guidelines and methodology for constructing ontology\(^11,4\).

Related work:

Search engine retrieves information from the web and is very crucial complex operation. Various researchers proposed many ranking models like hyperlink based model, statistical model, Boolean model, conceptual model and many more which are widely used\(^2\). The natural language processing techniques such as relaxation algorithm and language model
are used by some of them. Use of the natural language techniques for these models helps to consider semantic, syntactic structure and Morphological form of terms. Some other document ranking models such as Fuzzy Sets, Neural Networks, and Relevance Feedback Models can be used to increase the performance of the ranking models effectively. Lample R gave the Ontology-based design Information extraction as well as retrieval. Nowadays many ontologies are being formed as formal underpinnings for RDF-based data. They used the domain-specific ontology and the natural language processing so that from the unstructured documents structured and semantic based representation can be created easily. Linguistic patterns recognized the relationship between the concepts. Moreover, the Relationship and concepts is represented in the form of the conceptual graph. Finally, the integration of this conceptual graph builds the domain specific ontology gets built by integrating that conceptual graph and that can be compared to any other ontology to find the semantic similarity between the two documents. Focus of the various researchers in all the contribution is on to introduce the semantics either by taking Relationship or ontology that exists between the concepts. It is necessary to compare the complete semantic similarity between the documents and the query given by the user to make ranking of the documents by the search engine having only the relevant pages in the result-set.

Methodology:

Design of Traditional Web

For simplicity we are considering two different websites. These websites must be initialized independently from each other. From the two one called “all_India_ Universities” maintain and hosts information or database about all universities in India and the other website will maintain or hosts information about a particular college in our case VIT University. Both of the above website shown in Fig. 1. contains interrelated information in their websites database. From the two one fronting an MS SQL database for all universities in India, and the other one fronting a MYSQL database of VIT University. The URIs of the website are respectively at URI 1:
V. Vijayarajan* et al. International Journal of Pharmacy & Technology

The URI1 contains information about all the universities in India and contains synopsis about each of those universities in which state or region it resides. It also includes for each of universities the number of colleges in that university and also weather the university is approved by AICTE or has UGC approved their rankings, their ratings and so on. Moreover, it contains per university how many numbers of students are studying and the all India rank of that University. The URI2 contains details of a particular college in a state of India which belongs to one of the universities stored in “all_India_Universities” website. It contains all the academic details of the college including number of student, number of faculties, each student information, each faculty information, head of the department, vice chancellor of the college, director of the college, also it contains information about how many courses the college is offering, the fees structure, past years placement details, and rank among the other colleges under the same university to which VIT belongs and where the university head office is and strength of whole college and also part courses.

Design of Semantic Web

Fig. 2. Domains in Semantic Web.

If it is not providing any benefits, then it is meaningless to add semantics to your data. We can branch our data across the domain of knowledge and that is the key advantage of adding semantic meaning to data which is shown in Fig. 2. Now let’s start with what is domain of knowledge. If we want to retrieve related information containing details from both the website, then information sharing must be done between these two websites. We will show both way first sharing of information without use of semantics and after then we will see the information sharing with use of semantics and advantages over the former method. We want to share information between these two sites. We can do it easily with the use of RDF. Now we will see how it is become easier with the effective use of RDF and semantic web using these it will be done automatically and not manually. Following definition Introduces some of the key terms of semantic web.
Vocabulary - A collection of many terms will be given a well-defined meaning which will remain consistent across context.

Ontology - Allows you to define *contextual relationships* following a defined vocabulary. It is the keystone of defining a knowledge domain. A basic syntax for defining ontology is OWL (Web Ontology Language) which is an extension to RDFS (RDF Schema). Now we are going to see using semantic modeling how can we model the two site scenario? First requirement is that the both of the sites must have common, standard vocabulary for describing their data which is contextually consistent across context. For example, the term university head office should mean the same thing for both the websites. This can be done either by using same ontology or by using a common vocabulary, for defining or expressing the data they expose, and publish the same data on query able endpoint, so that the two sites can effectively communicate and can share information with each other across the web. Advantage of using standard, consistent vocabulary in place: Both of the sites can traverse each other by querying each other using the same terms predefined in vocabulary. The all_India_Universities website can now query the particular college website here VIT for gaining more internal details about the college. The VIT college website can now query the all_India_Universities website for gaining the more information about university it belongs to. With use of the contextual information defined in the formal web ontology, the university website can query the college website for number of students, rank among all its other college, number of courses it offers, faculty strength, placement details, the events happening in college, all these things have been found via the linked terminology without the client even thinking that information initially existed. Same goes for the college website (VIT). It can query the all_India_Universities for extracting different information which it cannot get otherwise like establishment year of that university, head office of the university, rank of it among other university, approval (UGC, AICTE), total strength of university and many more. These all can be happened without any need of information transformation, mapping between two websites, or contract setup between them. It happens just because we have used semantic web. This occur without any necessity for alteration, joining, or agreement being set up between the two sites. It all happens through semantics.

**VIT University Design Model:** OWL ontology consists of Properties, Individuals and Classes, which are roughly correspond to the Protégé Instances, classes and slots. The same is shown in Fig. 3.

![Fig. 3. Asserted Model.](image-url)
Class Elements:

Initially, the `owl: Thing` is the only class which is contained by an empty ontology. As mentioned previously, the class `owl: Thing` is the class that represents the set containing all individuals because all these classes are subclasses of `owl: Thing`. For example, an Academics class and its subclasses can be defined as:

```xml
<owl:Class rdf:ID="Academics">
  <rdfs:subClassOf rdf:resource="#VIT_UNIVERSITY"/>
  <owl:disjointWith rdf:resource="#Campus_Life"/>
  <owl:disjointWith rdf:resource="#Faculty"/>
  <owl:disjointWith rdf:resource="#Placement"/>
</owl:Class>

<owl:Class rdf:about="professor"/>
```

![Fig. 4. VIT University domain class hierarchy tree.](image)

Property elements:

Properties of the OWL represent relationship among any two individuals. Two types of properties are there. One Object properties and the other are Data type properties. Object properties link an entity with other entity. Data type properties links the individual with a XML Schema Data type value. The following example shows the “Designation” data type property which links Faculty class as its domain and allows any string of characters as its range

```xml
<owl:DatatypeProperty rdf:ID="Designation">
  <rdfs:domain rdf:resource="#Faculty"/>
  <rdfs:range rdf:resource="&xsd:string;"/>
</owl:DatatypeProperty>
```

The following example shows the object property has DMC faculty, which relates the Digital Media Computing class as its domain and their associated faculties.
Instances:

Instances represent individuals in ontology and are defined as the actual entities. The following example shows Prof. Vijayarajan is an instance of faculty class having attributes as Designation, Faculty name and Intercom Number assigned values as Associate Professor, Prof. Vijayarajan and 2717.

```xml
<Faculty rdf:ID="Prof_Vijayarajan">
  <Designation rdf:datatype="&xsd;string">Associate Professor</Designation>
  <Faculty_Name rdf:datatype="&xsd;string">Prof. Vijayarajan</Faculty_Name>
  <Intercom_No rdf:datatype="&xsd;int">2717</Intercom_No>
</Faculty>
```

Result and analysis:

Resource Description Framework (RDF) is being used by the semantic web as basic data format, which represents information about the resources. The W3C has developed SPARQL as a query language for RDF data. SPARQL query language is a graph pattern query language, it queries RDF data. The query shown below retrieves the list of Computer Architecture Faculty. In the query subject named “Subject” is associated the object named “Faculty”. These two classes are interconnected by the attribute named “hasCAFaculty”. Fig. 5. shows the output of the query with subject and its associated faculty.

```
SELECT ?Subject ?Faculty WHERE
{ ?Subject :hasCAfaculty ?Faculty }
```

![Fig. 5. List of Computer Architecture Faculty.](image-url)
In the similar fashion described, the following query retrieves the list of Parallel and Distributed Systems and with subject as “Subject” and object as “Faculty” associated with the property hasPDCfaculty using the below query and the result is shown in Fig. 6.

```
SELECT ?Subject ?Faculty WHERE { ?Subject: hasPDCfaculty ?Faculty}
```

![Fig. 6. List of Parallel and Distributed Systems.](image)

The following Query retrieves the list of all the subjects and their correlated faculties in the VIT University Domain. In the SPARQL, the new UNION operator is being introduced which merges the results of all subjects and faculties in the domain and is shown in Fig. 7.

```
SELECT ?Subject ?Faculty WHERE { {?Subject:hasPDCfaculty ?Faculty } UNION {?Subject :hasCAfaculty ?Faculty} UNION {?Subject :hasDMCfaculty ?Faculty} UNION {?Subject :hasTCSfaculty ?Faculty}} ORDER BY ?Subject
```

![Fig. 7. List of all Faculties.](image)

Next, we retrieve list of all faculty details order by faculty name. SPARQL query works in the similar manner to SQL query selecting the faculty name, designation intercom number of each faculty. The order by clause of SPARQL query sorts that faculty according to their name alphabetically and is shown in Fig. 8.

```
```
Fig. 8. List of all Faculty Details.

Performance Analysis:

Fig. 9. Performance graph of search engine by semantic web.

Here by this performance analysis we can find the number of search need for successful processing of query. In a regular search engine to find a single query it takes ‘n’ no. of search where n is number of pages. Fig. 9, we have exposed the total number of search required for different pages. Also it seems apparent for a domain which contains 1000 pages; search required for search engine of semantic web query is about 138. While for general search engine we must look for each and every 1000 pages as shown in Fig. 10. From above graph it seems apparent that for a small query the search time required for both the case is almost equal. But when we add more pages, the performance of search engine by semantic web is better. So, for a large domain search like WWW, performance search engine by semantic web is far better than a regular search engine.

Fig. 10. Performance graph of search engine by regular keyword.
Conclusion:

In this paper we have used the **PROTÉGÉ-OWL** editor to form ontology for our University domain. OWL (Web Ontology Language) is a web language which intends to communicate the web information into the machine-readable content with semantics. This OWL ontology includes classes, properties & its relationship with each other. In this paper for executing query we have use the SPARQL query language which is in RDF format which is used to retrieve the data which is being stored. Our graph based model consist data in the form of RDF. In our case we use RDF data for VIT University website. Protégé has been used to build ontology for VIT University domain. We wanted to share information between these two sites and we have done it easily with the use of RDF. Now we have shown how it is become easier with the effective use of RDF and semantic web.

Using these retrieval and querying information of each other done automatically which was previously done manually. We have used the tool such that Both of the sites can traverse each other by querying each other using the same terms predefined in vocabulary, the university website can query the college website via the linked standard terminology these all can found without the user can imagining that information was initially existed. Also VIT website can query the all_India_University for extracting different information which it cannot get otherwise like establishment year of that university, head office of the university, rank of it among other university and many more. These all happens without any need of information transformation, mapping between two websites, or contract setup between them. It happens just because we have used semantic web.

References:

6. E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF-W3C working draft” 4 October 2006.

**Corresponding Author:**

V. Vijayarajan*

Email: virtual.viji@gmail.com