ISSN 2277 -3061

DOMAIN ORINTED ONTOLOGY BASED SEMANTIC WEB SERACH METHODOLOGY USING SPARQL QUERY

P. Nandhakumar[1]
Research Scholar, Karpagam University
Coimbatore, Tamil Nadu, India
nandhap@gmail.com[1]
M. Hemalatha[2]
Dept. of Computer Science, Karpagam University
Coimbatore, Tamil Nadu, India
csresearchhema@gmail.com[2]

Abstract

Semantic web facilitates the use of automated processing of descriptions on the web and exchange and representation of information is done in a meaningful way. But a conventional search engine, the context and semantics of the user query is not analyzed fully and the data is not well structured so it does not provide the relevant content needed by the user. Hence to overcome this problem, semantic web search has become an essential part in today's world. In this proposed method the user given query is analyzed semantically and the web data is stored in an ontology which is well structured and conventional search is performed using SPARQL query viewer plug-in. Finally, ranking algorithm is used to rank the extracted links for the given query. The results obtained are accurate enough to satisfy the request made by the user. The level of accuracy is enhanced since the ontology is made consistent and query is analyzed semantically to retrieve the correct result. The domain specific evaluation time obtained shows promising results.

Keywords: Data Mining, Semantic web, Domain oriented Ontology, Search methodology

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 12, No. 9

editor@cirworld.com

www.cirworld.com, www.ijctonline.com

1. Introduction

The main objective of the semantic web is to make the meaning of information available in the web precise using semantic mark-up, which helps in added valuable access to knowledge available in the web. Semantic search performs a prominent role in fulfilling this objective by producing accurate answers to the user given queries. Recently large number of search engines and its related tools has been developed and the available the overview of state-of-art semantic search tools has been surveyed in our previous paper which reveals that, naïve users cannot use this tool but this enhances the performance when compared to the conventional search technologies. Hence, a search engine or tool has to be designed to facilitate semantic web inference through text search. In order to design a tool the drawbacks faced in the current web search techniques has been analyzed and can be listed as follows:

- The traditional search mechanism does not adapt to indexing and retrieval of semantic mark up.
- Search engines commonly use words and its variants for indexing the database.
- The search engines simply ignore the markup.
- Semantic markup is not used by the current search technique for text retrieval perfection.
- Current search engines make use of simple statistical terms to search the relevant document to the query.
- Some techniques like blind relevance feedback and thesaurus expansion can be integrated to the retrieval process and can be compared to the semantic markup.

From the drawbacks analysed, it is clear that current search engines faces the problem due to unstructured data and lack of using semantic mark up in the search retrieval process. Hence, the main purpose of this work is to improve the ability of both people and software agents to find information, documents and relevant answers to queries on the Web. Here the focus is made on combining the retrieval with ontology browsing using SPARQL query language. The Proposed technique provides the relevant and reliable result to the user. As ranking algorithm is used the user need not sift through the results obtained the relevant results will be in higher rank which can be easily retrieved. The relevancy of results provided by the proposed techinique is accurate and up to the mark.

The paper is organised as follows: section2 disscusses about the existing techniques in the form of literature review, section 3 discusses in detail about the proposed technique, section 4 deals with the results and disscussion and section 5 illustrates the conclusion and future work.

2. Literature Review

Kandogan et al. have developed a novel semantic search engine-Avatar, which combines the techniques used in traditional text search engine along with the use of ontology annotations [8]. The technique adopted in this search engine consists of two main components to achieve the above mentioned functions like: semantic optimizer and user interaction engine.

Bhagwat and Polyzotis have introduced a Semantic-based file system search engine- called Eureka, which makes use of an inference model to create the links between files and a File Rank metric to rank the files based on their semantic importance [2].

Wang et al. Proposed a semantic search methodology to retrieve information from normal database tables, and adapts to three main steps like: identifying semantic relationships between table cells; converting tables into data in the form of database; retrieving objective data from query languages [10].

Cohen et. al have presented a semantic search engine- XSEarch for searching the contents from XML based database [4]. The Simple query language (SQL) is used for retrieving data. The usage of SQL for quering helps the naive user to submit their queries efficiently. The ranking technique is adopted to retrieve the semantically related document and convince the user by returning the appropriate query answers.

Corby O et.al[5] have proposed a novel ontology-based search engine “Corese” for the semantic web: This search engine is specifically dedicated to the retrieval of web resources annotated using resource description framework and stored using a OWL query language. The ontology representation for this corese is built using RDF which enables ontology representation by providing concept hierarchy and a relation hierarchy. This search engine has proved to be a good example for RDF-based querying language supported search engines.

Georges Gardarin et al. have introduced an ontology-based Web information system SEWISE [11] which supports Web information description and retrieval. In this technique domain ontology technique is adopted to map text information from various Web sources into one uniform XML structure. This technique also displays some hidden semantic available in the text accessible by the program.

SemSearch proposed by Yuangui Lei, Victoria Uren and Enrico Motta, takes the focus of user queries into consideration when generating formal queries, thus being able to produce precise results that on the one hand satisfy user queries and on the other hand are self-explanatory and understandable by end users. Thus, SemSearch makes it possible for ordinary end users to harvest the benefits of semantic search and other semantic web technologies without having to know the underlying semantic data or to learn a SQL-like query language.

The SHOE search engine introduced by Heflin et.al [7] is the first form-based semantic search engines which returned refined web pages for the user given queries. The web pages returned by this search engine were suitable only to users who had well defined knowledge on ontologies and knowledge bases.

Swoogle proposed by Ding et.al [6], is a crawler based search engine whose index currently contains information on over a million RDF documents. Swoogle implements a hybrid approach by including several components like Google meta-search engine, a focused HTML crawler and an RDF crawler.

Staab et.al [9] has introduced a new technique which can work on the Ontotext RDF crawler. This technique can build the knowledge base using the data collected from RDF based internet downloaded fragments. At each phase of crawling it maintains the URI filtering conditions and the host of URIs to be retrieved.

Cinamon P. [3] has proposed new technique ORAKEL to search the web. This technique is ontology-based question answering search engines which makes use of natural language processing technologies to reformulate natural language queries into ontological triples or into specific query languages.

3. Proposed Semantic web Methodology

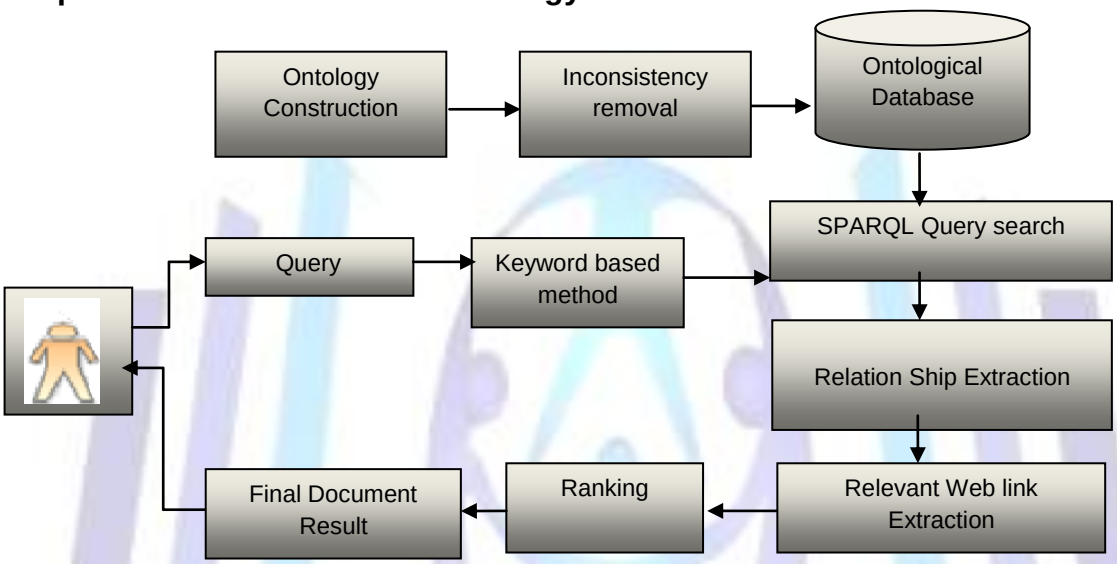

Figure 1: Proposed Semantic Web Methodology

3.1 Ontology Construction

Ontology is a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they can or cannot be related to each other, and how they are related to each other. Hence, Ontology provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary.

The use of ontology to represent the vocabulary helps the users in many ways

- It enables more number of machines to distribute and share their knowledge.
- It enables a machine to use the knowledge in various applications.

In this work Resource Description Framework or RDF is used to represent knowledge which describes a subject in terms of its classes and their relationships.

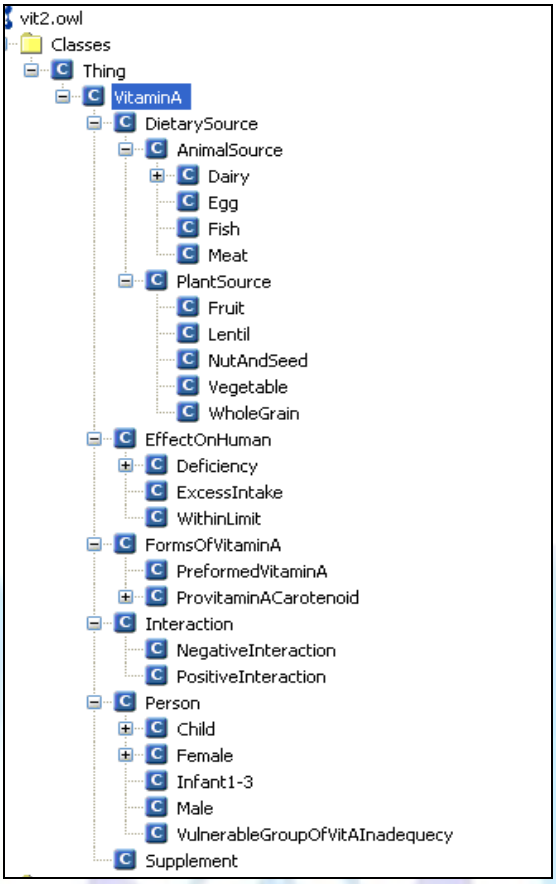

Figure 2: Vitamin A Ontology

In this work vitaminA ontology is chosen as the domain, the designed vitamin A ontology is shown in Figure 2 which depicts the available classes, instances, and relations among them in the specific domain.

3.1 Ontology Inconsistency Removal

The ontology being constructed is not structured and hence when search query is made it increase the recall rate. Hence to overcome this problem, the constructed ontology has been checked to remove the inconsistencies. The figure 3a below shows the initial ontology which is inconsistent and figure 3b shows the procedure materializes being adopted to remove the inconsistencies in the current ontology.

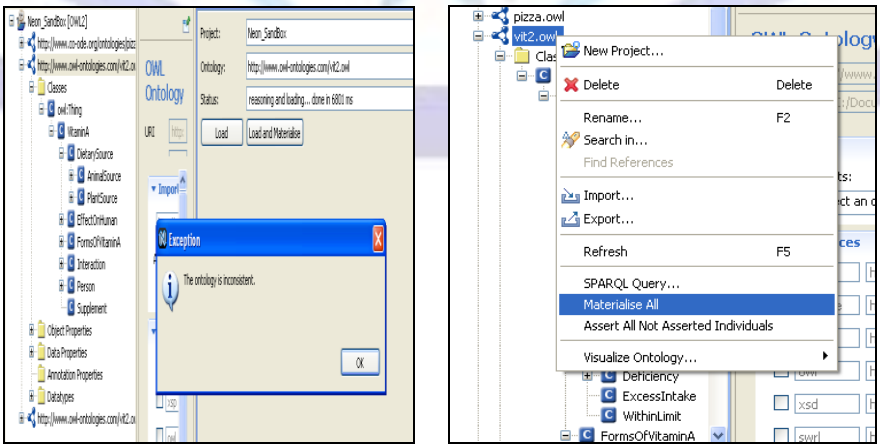

Figure 3: a. inconsistent ontology b.Ontology inconsistency removal using Materialise

There are three ways to translate an inclusion: internal, material and strong. For every ontology that is going to be translated, it would be asked that if you want to use always the same way or decide for every axiom in the knowledge base independently after the translation process the refreshment of the ontology project is to be done. Depending on the original ontology and the chosen inclusion type some new classes may appear. Now reasoning tasks can be done using the new ontology that can return possible results.

3.2 Query Construction

3.2.1. Resource Description Framework

Resource Description Framework (RDF) framework is used for describing and interchanging metadata and provides machine understandable semantics for metadata. Using this RDF frame work leads to better precision in resource discovery than full text search, it assists applications as schemas evolve, it increases the interoperability among metadata.

In this RDF framework everything is described by an RDF expression called resource and each resource is provided with a URI that may link to an entire Web page or a part of a Web page. “A property is a specific aspect, characteristic, attribute, or relation used to Describe a resource” – W3C, Resource Description Framework (RDF) Model and Syntax Specification [12]. A statement combines a resource, a property and a value. These three individual parts are known as the “subject,” “predicate” and “object”.

For example, “Strawberry has VitaminA in IU” is an RDF statement having the following parts:

Subject (Resource) Strawberry

Predicate (Property) has vitamin A in IU

Object (value) 12

This can be represented as an RDF graph as shown in Figure below.

```
<?xml version="1.0"?>
<Fruit rdf:ID="Strawberry">
<HasVitaminAinIU rdf:datatype="&xsd;int">12</HasVitaminAinIU>
<HasName rdf:datatype="&xsd;string ">Strawberry</HasName>
<HasRetinolActivityEquivalent rdf:datatype="&xsd;int">1</HasRetinolActivityEquivalent>
<HasBetaCarotene rdf:datatype="&xsd;int">7</HasBetaCarotene>
<HasLuteinZeaxanthin rdf:datatype="&xsd;int ">26</HasLuteinZeaxanthin>
</Fruit>
```

This RDF code describes the resource “strawberry” which in an instance of the “fruit” class. “26” is the value of the property “HasLuteinZeaxanthin.” Similar documents have been created for all the classes and instances of each class as specified in the ontology. A separate subclass has been described for RDF code and its relationships.

RDF Schema is a simple data-typing model for RDF [12] which can be used to describe groups of related resources and the relationships among these resources [12]. For example, we can say “strawberry” is a type of “fruit” and “fruit” is a subclass of “Plant source”. The use of RDF Schema helps in easy inference of data, and it also enhances the search results. The figure 4 below shows the ontology being visualized using ontology visualizer.

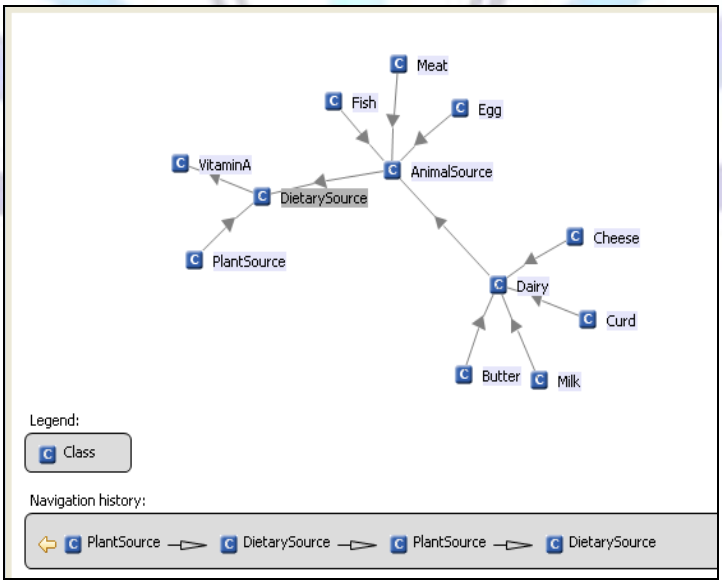

Figure 4: Ontology visualizer for vitamin A database

3.2.2 SPARQL Query Construction using RDF

The query is being constructed using SPARQL language and the sample query is shown below which retrieves the results for the query which has vitamin A in IU and also has Retinol activity equivalent.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ub: <http:// xmlns="http://www.owl-ontologies.com/vit2.owl#>
SELECT ?X, ?Y, ?Z
WHERE
{?X rdf:type ub: HasName
  ?Y rdf:type ub: HasVitaminAinIU
  ?Z rdf:type ub: HasRetinolActivityEquivalent
  ?X ub: HasName ?Z .
  ?Z ub: HasVitaminAinIU ?Y .
  ?X ub: HasRetinolActivityEquivalent ?Y}
```

3.2 Query Execution

The Ontology being constructed is used to obtain high relevancy factor when search for the documents. For this a domain knowledge specific to a particular scope is collected and organized. In this work three genuine ontologies has been used and compared to identify the performance of the proposed technique. It receives OWL ontology and a query expressed using SPARQL syntax as inputs and outputs the answers in a table. This reasoning task is performed using the inference engine in KAON2. The figure 5 below shows the execution of the query and results obtained for the relevant query.

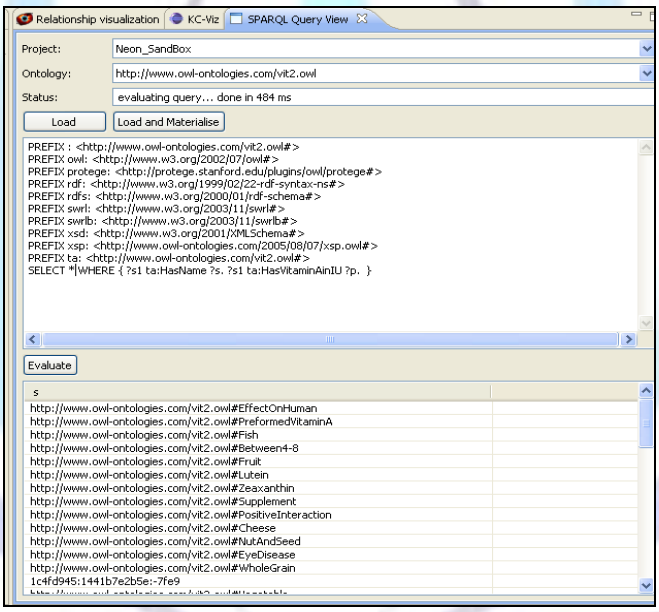

Figure 5: Query execution interface

3.3 Relationship Extraction

In this phase, the hierarchies among relationships that exist among ontology classes are obtained. These concepts, facts and relationships are implemented using Neon as shown in figure 6. This helps to define the classes existing in this specific domain ontology, the relationships that exist among the classes, i.e. either parent child relationship or sibling relationship. The disjoint subclasses which don't share any common instance can also be identified and defined as disjoint.

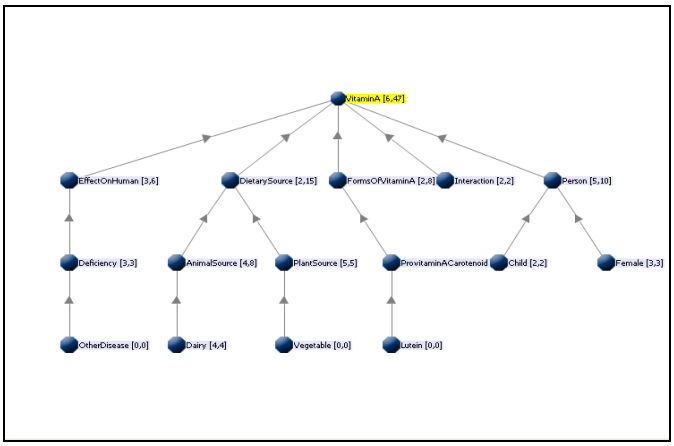

Figure 6: Vitamin Ontology Diagram in Neon

The list of terms available at VIT2 ontology is retrieved and the relationship among them is obtained using the relationship visualization Plugin. This helps to represent information as 'subject', 'predicate' and 'object' analogous to English language grammar structure. Subject and object is an instance of classes while predicate defines the relationship or property between them. With respect to mapping, the set of subjects are domain that is mapped to set of objects as range through relationships. This relationship graph is shown in figure 7.

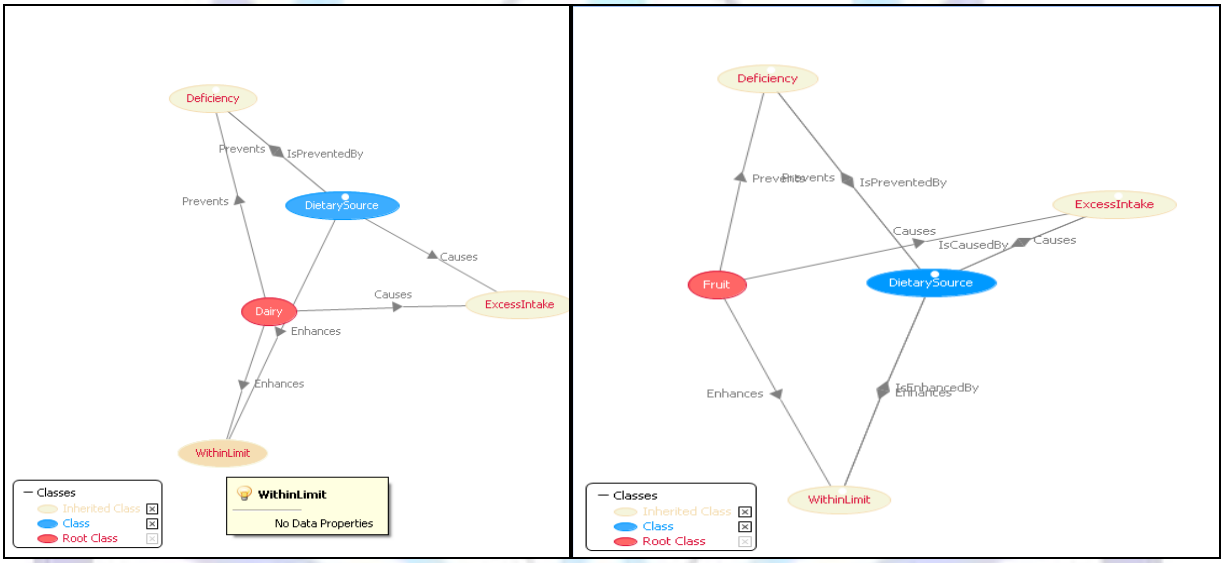

Figure 7: Domain range relationship among entities

3.4 Web Links Extraction

The query is analyzed semantically using word net and semantically analyzed words are obtained and these words are matched to the concepts stored in the specific ontology to get the related keywords to the query. These words are then used to retrieve the appropriate documents from the web. The process of extracting the relevant links to query is a critical part and it has be done using KAON2 which extracts the links relevant to the criteria specified in the query. The sample query and the links extracted for the query is shown in figure 8.

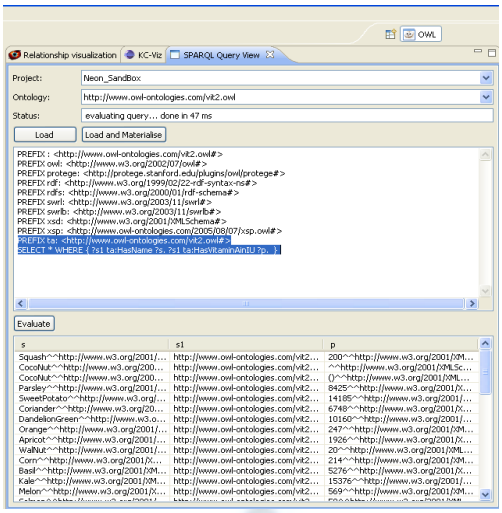

Figure 8: web link extraction

3.5 Ranking Mechanism

So in order to rank the sources, the ontology being chosen has some pre specified criteria and on this basis ranking is done on various sources of interest**.** This phase returns the set of links which are semantically related to the user query. The ranking is done based on the semantic relevance which is attained by means of the extracted domain keywords from the specific ontology. Permutation is applied on the retrieved links to rank the links which are more relevant and it is re-ranked based on its relatedness. The result of ranking is shown in the figure 9 below for the specific query.

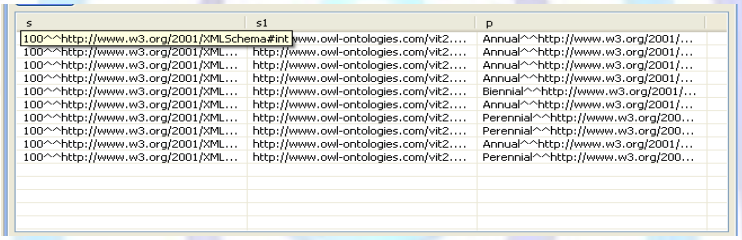

Figure 9: Web Links extracted after ranking

4. Experiment and Results

4.1 Sample SPARQL Queries

Among 100 queries some sample queries and the evaluation detail has been listed below to evaluate the performance of the proposed method. Figure 10 below shows the web links being extracted for a specific sample query. **Sample Query1:** SELECT * WHERE { ?s1 ta:HasName ?s. ?s1 ta:HasVitaminAinIU ?p. }

Sample query2: SELECT * WHERE { ?s1 ta:HasQuantity ?s. ?s1 ta:HasPlantType ?p. }

Sample query3: SELECT * WHERE { ?s1 ta:HasVegtype ?s. ?s1 ta:HasVitaminAinIU ?p. }

Sample query4: SELECT * WHERE { ?s1 ta:HasTolerableUpperIntakeLevel ?s. ?s1 ta:HasRecommendedDietaryAllowance ?p. }

Sample query5: SELECT * WHERE { ?s1 ta:SufferingFrom ?s. ?s1 ta:HasBetaCarotene ?p. }

Sample query6: SELECT * WHERE { ?s1 ta:HasRetinol ?s. ?s1 ta:HasVegType ?p. }

Sample query7: SELECT * WHERE { ?s1 ta:HasRecommendedDietaryAllowance ?s. ?s1 ta:HasAlphaCarotene ?p. }

Sample query8: SELECT ?X WHERE {?X rdf:type ta:GraduateStudent . ?X ta:takesCourse?p.}

Sample query9: SELECT ?X WHERE{?X rdf:type ta:Publication . ?X ta:publicationAuthor ?p.}

Sample query10:SELECT ?X WHERE {?X rdf:type ta:Student}

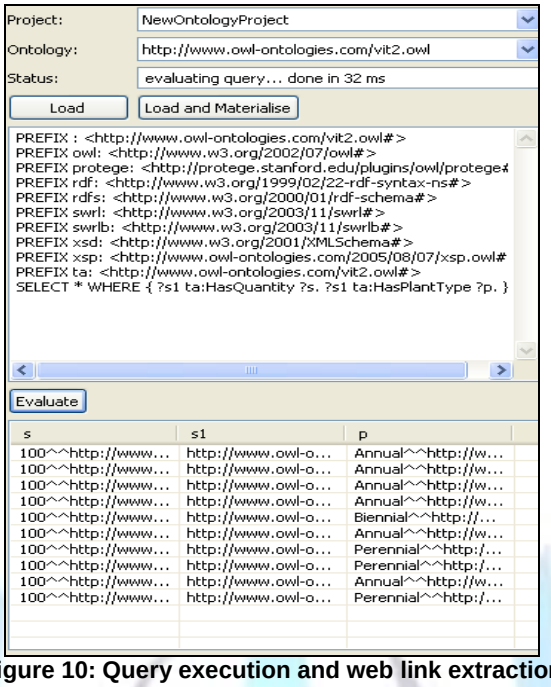

Figure 10: Query execution and web link extraction

The Proposed technique has been evaluated to know its performance. The Time taken for the execution of the query using the proposed technique and the common web are collected. The number of links extracted for each query is also analyzed. The results obtained for the sample of ten queries are listed in table 1.

Table 1: Sample query results based on time

Queries	Time taken by proposed method	Time taken by common web based method	No of links extracted by proposed method	No of links extracted by Common web based method
Query 1	47ms	1562 ms	25	112
Query2	32 ms	922 ms	10	89
Query3	16 ms	688 ms	14	110
Query4	15 ms	687 ms	36	105
Query 5	16 ms	625 ms	22	86
Query6	15 ms	906 ms	45	96
Query7	16 ms	859 ms	51	118
Query8	21ms	458 ms	32	154
Query9	19ms	328 ms	21	115
Query10	19 ms	678 ms	15	78

4.1 Evaluation Criteria

The proposed system is being evaluated using various criteria like accuracy, precision and recall.

Accuracy: It is defined as the measure of how much of the information the system returns is correct is calculated as accuracy. The accuracy of the proposed technique has been evaluated against the result set generated by running the query using “strawberry,” “HasName,” “HasBetaCarotene,” “HasVitaminAinIU,” “HasRetinolActivityEquivalent” etc. From the Web pages returned, it can be observed how there exist possibly out-of-scope pages that have been ranked as very relevant, while potentially interesting pages are positioned at the end of the list. This can evaluated using Precision and recall rates.

Precision Rate: Precision is the fraction of retrieved documents that are relevant to the search.

$$\text{Precision} = \frac{\#\text{of relevant links given by thesystem}}{\text{Total}\#\text{of links retrieved}}$$

Recall Rate: Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Re}call = \frac{\#\text{ of relevant links given by the system}}{\text{Total}\#\text{ of relevant links in web and proposed system}}$$

To evaluate the performance of the proposed method 3 genuine ontology has been selected and tested. The results obtained for the 3 ontological databases when executed using the proposed technique has been compared with the web results. Similarly 100 queries have been executed and the results have been obtained. The results obtained are listed in table 2.

Table 2: Comparison using various ontological databases

Evaluation measures	Proposed Method			Common Web		
	Vitamin A Database	Pizza Database	Education Database	Vitamin A Database	Pizza Database	Education Database
Precision	0.9371	0.9241	0.9394	0.4663	0.5231	0.5762
Recall	0.599	0.524	0.612	0.213	0.3126	0.2873
Accuracy	94	93	94	47	52	58
Time taken	22.25ms	23.301ms	25.932ms	838.37 ms	873.142	912.20

The results obtained clearly shows that precision, recall and accuracy rate of the proposed technique is higher than the common web. The overall precision, recall, time taken and accuracy of the proposed technique has been evaluated and compared with the common web. The results obtained are listed in table 3. The results obtained clearly show that the proposed method has higher performance rate than existing web. Following charts in figure 11 shows the results for precision rate, recall rate for this proposed approach as compared to common search engines.

Table 3: Comparison using various evaluation measures

Method	Avg. Precision	Avg. Recall	Total Time taken	Accuracy
Proposed method	0.9335	0.5783	23.8276	93.66
Common web	0.5219	0.2709	874.5706	52.33

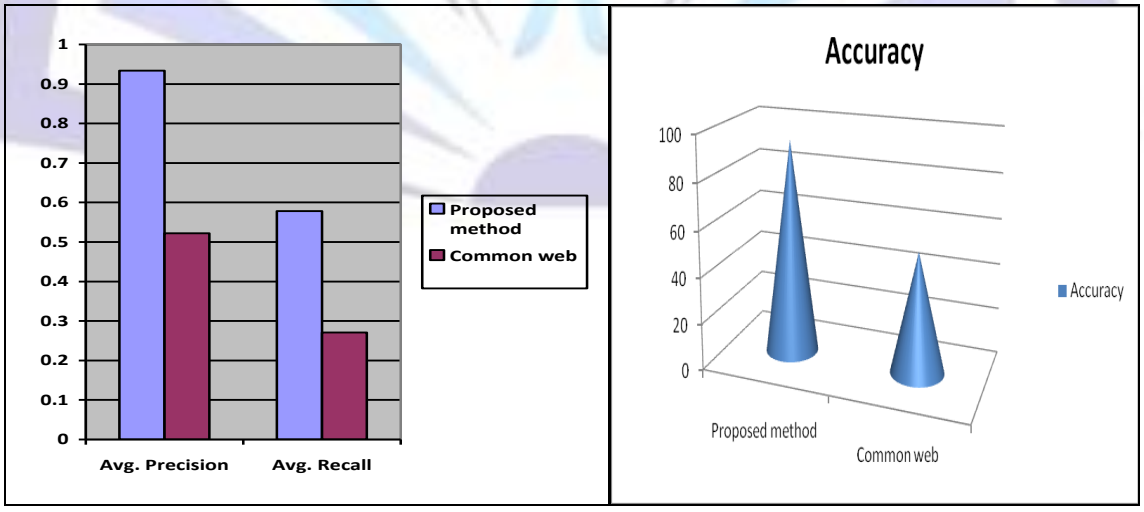

Figure 11: Chart showing precision, recall and accuracy obtained for the proposed technique.

From the results obtained it can be inferred that the higher value of average precision and recall for the proposed technique when compared to common web depicts that the proposed technique has a better performance and accuracy in retrieving the results than the generic search engines. The recall values depict the coverage of the system. The average relative recall value of the proposed technique is also higher compared to that of common web. The higher value denotes the best coverage of the proposed system compared to other common generic search engines.

5. Conclusion and Future Work

The data available in the web today consists of billions of web pages and are represented in HTML which does not aid in exact retrieval correct web page for the user given query. This inefficiency is due to unstructured data which does not make use of semantic markup for the retrieval of information. This paper focuses on the ways to enhance the search process by making use of structured ontological data which represented by means of semantic mark up.

The proposed technique first removes the inconsistencies in the ontological data and then it obtains the semantically related words from the user query and then matches it with the semantic markup to retrieve the exact link for the query. As the query is analyzed and information is extracted from RDF store using SPARQL query that is being executed using SPARQL plug-in. Finally the extracted links are ranked to obtain the most relevant link to the query.

Experimental results obtained prove that the proposed technique has higher performance results than the other common web search engines. This work can be further enhanced to search the web using various data mining related search algorithms to achieve higher accuracy.

References

1. Berners-Lee, T., Hendler, J. and Lassila, O., 2001 “The Semantic Web”, Scientific American.
2. Bhagwat D. and N. Polyzotis, 2005 "Searching a file system using inferred semantic links," in

   Proceedings of HYPERTEXT '05 Salzburg: 85-87.
3. Cimiano P.,2004. ORAKEL: A Natural Language Interface to an F-Logic Knowledge Base. In Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems: 401–406.
4. Cohen, S. Mamou, J. Kanza, Y. Sagiv, Y,2011. “XSEarch: A Semantic Search Engine for XML” proceedings of the international conference on very large databases.1(1):45-56.
5. Corby O., R. Dieng-Kuntz, and C. Faron-Zucker,2004. Querying the Semantic web with Corese Search Engine. In Proceedings of 15th ECAI/PAIS, Valencia (ES)
6. Ding L., T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs, “Swoogle: A Search and Metadata Engine for the Semantic Web,” Proc. 13th ACM Int’l Conf. Information and Knowledge Management (CIKM ’04), pp. 652-659, 2004.
7. Heflin J. and J. Hendler, 2000. Searching the Web with SHOE. In Proceedings of the AAAI Workshop on AI for Web Search, pages 35 – 40. AAAI Press.
8. Kandogan E., R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu, 2006"Avatar"Avatar semantic search: a database approach to information retrieval," in Proceedings of SIGMOD '06 Chicago, PP: 790-792.
9. Staab S., K. Apsitis, S. Handschuh, H. Oppermann, (2004) Specification of an RDF Crawler.
10. Wang H. L., S. H. Wu, I. C. Wang, C. L. Sung, W. L. Hsu, and W. K. Shih, 2008 "Semantic Web mining” Research, Reflections and Innovations in Integrating ICT in education. of CIKM '00 McLean, PP:243-249.
11. www.georges.gardarin.free.fr/Articles /Sewise_NLDB2003.pdf.
12. World Wide Web Consortium (W3C).http://www.W3C.org

Authors

P. Nandhakumar (nandhap@gmail.com) completed M.C.A., M.Phil. and currently pursuing Ph.D in computer science at Karpagam University, Coimbatore under the guidance of Dr.M.Hemalatha, Professor and Head, Dept. of Software System, Karpagam University, Coimbatore. He is working as Senior Software Engineer in Easy Design Systems Private Limited, Coimbatore.

Dr. M. Hemalatha (csresearchhema@gamil.com) completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. She received best researcher award in the year 2012 from Karpagam University. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering, Bioinformatics and Neural Network. She is a reviewer in several National and International Journals.