ONTOLOGY BASED MEANINGFUL SEARCH USING SEMANTIC WEB AND NATURAL LANGUAGE PROCESSING TECHNIQUES

K. Palaniammal[1] and S. Vijayalakshmi[2]

Department of Computer Applications, Thiagarajar College of Engineering, India

E-mail: [1]sudharoentgen19@gmail.com, [2]svlcse@tce.edu

Abstract:

The semantic web extends the current World Wide Web by adding facilities for the machine understood description of meaning. The ontology based search model is used to enhance efficiency and accuracy of information retrieval. Ontology is the core technology for the semantic web and this mechanism for representing formal and shared domain descriptions. In this paper, we proposed ontology based meaningful search using semantic web and Natural Language Processing (NLP) techniques in the educational domain. First we build the educational ontology then we present the semantic search system. The search model consisting three parts which are embedding spell-check, finding synonyms using WordNet API and querying ontology using SPARQL language. The results are both sensitive to spell check and synonymous context. This paper provides more accurate results and the complete details for the selected field in a single page.

Keywords:

Ontology, Semantic Web, Information Retrieval, Natural Language Processing Techniques

1. INTRODUCTION

Meaning based search has traditionally been an interesting research area within the semantic web and Natural Language Processing (NLP) field [1]-[2]. The reason is that meaningful search is a key aspect for human conversations. In fact, the fast development of the semantic web has led researchers to focus on the development of techniques based on synonym recognition to improve the discovery of resources on the WWW [3]. The internet information resources increased day by day. The generic search engines such as Yahoo, Google use traditional search forecasts, not satisfied user requirements to catch high grade web information resources. Hence they want to search related reliable and latest information more precisely and efficiently [4]. On account of in this paper, to use semantic web and Natural Language Processing technique to develop meaning based search mechanism. Semantic web was envisioned by Tim Berners-Lee as the next generation of the Web. It is necessary in the first place for mark-up data on the web semantically, so that they can be understood and processed by agents autonomously [5]. The Semantic Web vision is based on structuring the knowledge that is present in the current web, so that it is understandable by machines without human intervention. Semantic web aims to provide a new framework that can enable knowledge sharing and reusing. Semantic Web uses agent technology, ontology, and a number of standard markup languages, such as Resource Description Framework (RDF), Ontology Web Language (OWL) to formally model information represented in web resources. Ontology removes the difficulties to find, present, access, or maintain available electronic information on the web and it provides the method for a data representation to enable software products (agents) to provide intelligent access to heterogeneous and distributed information. This mechanism is capable of improving the traditional problem of the keyword search and enables the user to perform a semantic-based query and search for the required information, thereby improving the search information. Ontology is an agreed vocabulary that provides a set of well-founded constructs to build meaningful higher level knowledge for specifying the semantics of terminology systems in a well defined and unambiguous manner. For a particular domain, ontology represents a richer language for providing complex constraints on the types of resources and their properties. Compared to taxonomy, ontology's enhance the semantics by providing richer relationships between the terms of a vocabulary. Ontology's are usually expressed in a logic-based language, so that detailed and meaningful distinctions can be made among the classes, properties, and relations. Ontology can be used to increase communication both between humans and computers. The three major uses of ontology are: first, to assist in communication between humans; second, to achieve interoperability and communication among software systems and third is to improve the design and the quality of software systems. Currently, the most prominent ontology language is the OWL used in this paper. OWL is a vocabulary extension of RDF and is derived from the DAML + OIL language, with the objective of facilitating a better machine interpretability of Web content than the one supported by XML and RDF [6].

This paper is organized as follows. Section 2 focuses on the related work; in section 3 presents proposed approach; in section 4 are focused on the system implementations; section 5 presents experimental results and evaluation and section 6 ends with the conclusion.

2. RELATED WORK

A very recent system called PowerAqua [7] is an ontology-based Natural Language Information (NLI) system which surpasses traditional systems by managing multiple ontology sources and high scalability. Since it is NL processing module remains the same as in the previous AquaLog system [8], AquaLog is a portable NLIKB system which handles user queries in a natural language (English) and returns answers inferred from a knowledge base. The system uses GATE1 libraries (namely the tokenizer, the sentence splitter, the POS tagger, and the VP chunker). ORAKEL [9] is an ontology-based NLI system. It accepts English factoid questions and translates them into first-order logic forms. This conversion uses full syntax parsing and a compositional semantics approach.

ORAKEL can be ported into another domain but such porting requires a domain expert to create a domain-dependent lexicon. The lexicon is used for an exact mapping from natural language constructs to ontology entities. A possible drawback of ORAKEL's approach is that the system can neither handle ungrammatical questions nor deal with unknown words. Wang et al. [10] tossed a semantic seek methodology to collect knowledge from normal tables, which has the three main steps: identifying semantic relationships between table cells; converting tables into data in the form of a database; and retrieving objective data by query languages. This work demonstrates how intelligent agents can extract the tabular information for answering queries. With the assistance of ontological knowledge, the intelligent agents can distinguish concepts and instances in each table cell. Sara Cohen Jonathan Mamou et al., Presented a semantic search engine for XML (XSEarch) [11]. It has a simple query language, suitable for a naïve user. It returns semantically related document fragments that satisfy the user's query. However those search methods still suffer from complex query syntax. F. Shaikh et al., [12] proposed the semantic web based search engine named (SWISE). The XML meta-tags deployed on the web pages to searching queried information. The XML page will be consisted of built-in and user defined tags. The metadata information on the pages is extracted from this XML into RDF. The RDF graphs are populated by inputting through forms.

3. ONTOLOGY BASED MEANINGFUL SEARCH USING SEMANTIC WEB AND NLP TECHNIQUES

This paper is to develop a reliable and an efficient search engine to retrieve the accurate results for the user's query. It also aims at retrieving the same result for synonymous words which prevents the appearance of irrelevant search results. It provides the complete details for the query about the education domain with the correct URL and metadata in which to search for, which consumes more time in the syntactic search engine. The details are generated with the help of ontology and relations among classes, entities, individuals are also created. Hence now the user can query upon the information stored within the ontology. The querying of the ontology is supported towards the properties, classes, individuals and entities created in the ontology.

4. SYSTEM IMPLEMENTATIONS

The prototype implementation can be divided into two tasks, which are creation of ontology knowledge base and search module. The search module consist the following sub-process such as embedding Spell check, finding Synonyms Using WordNet API and Querying Ontology using SPARQL.

4.1 CREATION OF ONTOLOGY KNOWLEDGE BASE

Ontology is created using Protégé. Protégé is an open-source tool developed at Stanford medical informatics. It has a community of thousands of users. Although the development of Protégé has historically been mainly driven by biomedical applications the system is domain-independent and has been successfully used for many other application areas as well. Like most other modeling tools, the architecture of Protégé is cleanly separated into a "model" part and a "view" part. Protégé model is the internal representation mechanism for ontology and knowledge bases. Protégé view components provide a user interface to display and manipulate the underlying model. Protégé model is based on a simple yet flexible Meta model, which is comparable to object-oriented and frame-based systems. It basically can represent ontology consisting of classes, properties (slots), property characteristics (facets and constraints), and instances. Protégé provides an open Java API to query and manipulate models. An important strength of Protégé is that the Protégé meta model itself is Protégé ontology, with classes that represent classes, properties, and so on.

This paper made use of education ontology for querying upon the desired event, the required components to build up the ontology such as classes, instances and relationships are being created. The classes created are college, college type, school, university, district, etc., The subclasses created within college are engineering, arts and science, law college, medical, polytechnic, institute etc. with regard to the metadata, currently no such ontology available. Therefore, we collect and develop an education system ontology resource from the websites mentioned in protégé OWL [13]. The properties of the classes are created.

Properties are of two types:

- Data type Property: It is being used to set properties enhancing the existence of an individual
- Object Property: It is used to create a relationship between two different class individuals.

The data type properties created is about college, contact, location, URL. The object properties created are type of college, present in the location, the name of institution etc.

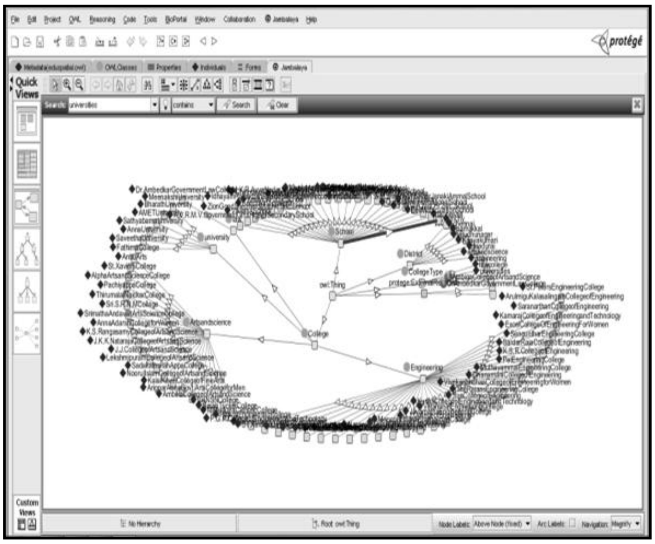


Fig.1. Education Ontology in Radial Layout

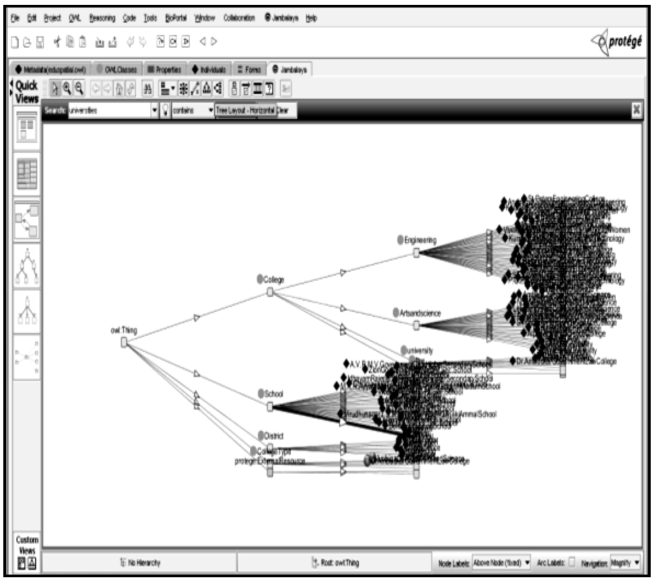


Fig.2. Education Ontology in Tree Layout

The individuals for each OWL classes are created for Engineering Institutions, Arts & Science Colleges etc. The Fig.1 and Fig.2 shows the sample ontology screen shot [14].

4.2 SEARCH MODULE

The search module split into the following sub-parts such as embedding spell check, finding synonyms using WordNet API and querying ontology using SPARQL language.

4.2.1 Embedding Spell Check:

This paper used embedding spell check module for proficient web search. Owing to the user enter word in incorrect or spelling mistake the (spell check module) Java program, providing suggestions for unknown (misspell) words based on custom dictionary and system administrator can create a list of preferred words and assign higher weight to the list. As a basic implementation Suggester can serve as a spell checker. In this case all words have the same weight. The basic implementation includes high speed suggestion engine, based on fast edit-distance calculation algorithm enhanced with Lawrence Philips Metaphone algorithm and private fuzzy-matching algorithm. Basic Suggester (free) uses one dictionary, where all words have the same weight. The Suggester Spell Check uses Basic Suggester. Suggester can be used as a Spellchecker, Search engine suggestions, based on your custom word list, Misspelt word suggestions in any other fields, which require custom dictionaries.

The BasicSuggester uses ConFigureuration and Dictionary objects. To use the suggester, load English dictionary from jar file and basic suggester configuration from file. An instance for BasicSuggester is created as Suggester, based on configuration and dictionary is being attached to the Suggester. An array list is being created and the return type of getSuggestion method which retrieves the suggested words for the word given is stored in the array list. The number of suggestions made is dependent on argument passed to getSuggestion within the loop each of the suggested word in array list is retrieved as a string and used.

4.2.2 Finding Synonyms Using WordNet API:

The WordNet API employed for find relevant meaningful search. In WordNet, words and their relationships to each other are organized in a hierarchical manner similar to the taxonomies which may be found in the natural sciences. Words which are closely related to each other may be found in the same branch of the hierarchy's tree. Each word belongs to a set of synonyms, also known as a synset. These synsets are the foundation upon which the WordNet database is constructed. Formally, a synset is a set of one or more synonymous words that may be substituted for each other in context without changing the overall meaning of the sentence in which they are contained. Words which have multiple meanings or “word senses” appear in more than one synset. WordNet provides a polysemy count for each word which is used to track the number of synsets which contain the word.

Since different word types follow different grammatical rules, WordNet makes the distinction between four of the primary word types in the English language, which include nouns, verbs, adjectives, and adverbs. The nouns category contains words which refer to entities, qualities, states, actions, or concepts, and can serve as the subject of a verb. Words classified as verbs may serve as the predicate of a sentence and describe an action, occurrence, or state of existence. Adjectives are words that may modify nouns. The final word classification stored in WordNet, the adverb, is similar to the adjective and contains words which modify word types other than nouns.

The WordNetDatabase class provides access to the information stored in the WordNet database and must be instantiated before use. A method, getFileInstance, returns an implementation of the class that works with the local WordNet database and may be used when creating a new instance of the WordNetDatabase class. Other than WordNetDatabase, another critical component of the JAWS API is the Synset interface. This interface represents WordNet's collections of related words, or Synsets. These synsets are stored as an array of word forms. Several overloaded methods of the WordNetDatabase class known collectively as getSynsets can be used to retrieve synsets from the WordNet database by providing a starting word in the form of a string when the getSynsets is called. When instantiating a Synset, the getSynsets method is used to populate the new instance of the Synset interface with WordNet information.

The getWordForms method may be used to retrieve the individual groups of word forms for each Synset stored as an element of this array, which may themselves be stored as arrays of strings containing all words similar to the original word.

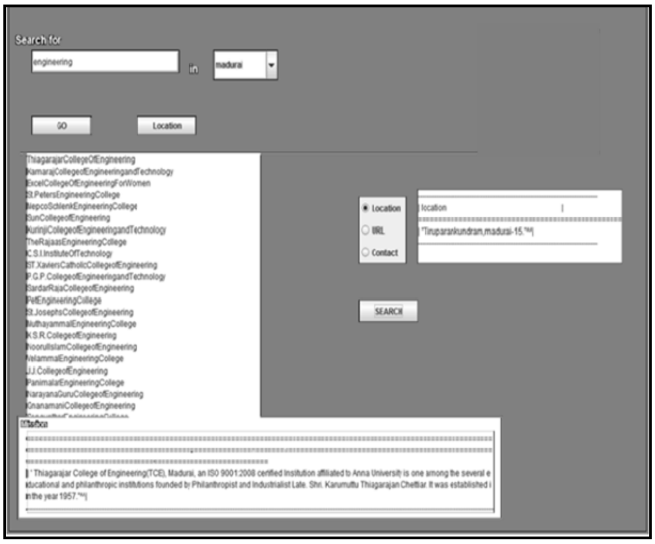
4.2.3 Querying Ontology Using SPARQL:

In this paper, SPARQL query is being used to retrieve relevant information from ontology. SPARQL (Simple Protocol and RDF Query Language) is the same as in SQL and used to access more reliable and accurate results. The SPARQL language applied for specific search module, this module for the service requester who has domain knowledge relevant their service queries, in order to help them to quickly retrieve results from the ontology knowledge base [13].

5. EXPERIMENTAL RESULTS AND EVALUATIONS

5.1 RESULTS

In this section, we experimentally evaluate the search model. We have run it with different inputs such as universities, schools, college etc.



engineering

madurai

ThiagarajarCollegeOfEngineering
KamarajCollegeofEngineeringandTechnology
ExcelCollegeOfEngineeringForWomen
St.PetersEngineeringCollege
MepcoSchlenkEngineeringCollege
SunCollegeofEngineering
KurinjiCollegeofEngineeringandTechnology
TheRajaasEngineeringCollege
C.S.I.InstituteOfTechnology
ST.XavierisCatholicCollegeofEngineering
P.G.P.CollegeofEngineeringandTechnology
SardarRajaCollegeofEngineering
PetEngineeringCollege
St.JosephsCollegeofEngineering
MuthayammalEngineeringCollege
K.S.R.CollegeofEngineering
NoorulIslamCollegeofEngineering
VelammaEngineeringCollege
J.J.CollegeofEngineering
PanimalarEngineeringCollege
NarayanaGuruCollegeofEngineering
GnanamaniCollegeofEngineering

location

"Tiruparankundram,madurai-15."

"Thiagarajar College of Engineering(TCE), Madurai, an ISO 9001:2008 certified Institution affiliated to Anna University is one among the several educational and philanthropic institutions founded by Philanthropist and Industrialist Late. Shri. Karumuttu Thiagarajan Chettiar. It was established in the year 1957."

Fig.3. Screen shot of Search System

The Fig.3 shows the designed search engine with input as engineering, retrieves the list of engineering college and when selected a specific college from the list gives its corresponding location, URL and contact followed by the mission of the institution. When the required district to which the query term relevant to which the query term, it displays results corresponding to the district. The user enters the word with the wrong spelling; it retrieves the output with the correct word from ontology. The result is being retrieved the same for synonymous words which results in the reduction in storing the same ontology again. The location button for when clicked with the required location of the college, outputs its route along with directions to reach it in a map.

5.2 EVALUATIONS

To evaluate our model, precision and recall widely used performance measures from the information retrieval system, are adopted in the following experiment, a proper threshold value needs to be decided to filter the irrelevant concepts for metadata.

$$\text{Precision P} = \frac{\text{number of retrieved relevant data}}{\text{number of retrieved data}} \qquad (1)$$

Precision is used to measure the preciseness of a search system [15]. In this experiment, Precision P is defined as the number of retrieved relevant data among the retrieved data.

$$\text{Recall R} = \frac{\text{number of retrieved relevant data}}{\text{number of retrieved data}} \qquad (2)$$

Recall is used to measure the effectiveness of a search system [15]. In this experiment, Recall R is defined as the number of retrieved relevant data to total number of relevant data in the knowledge base.

To evaluate, the performance of our approach from the perspective of information retrieval field, the mechanism and algorithm concerning the model referred from [15]. Different queries are made to compare the performance of the system. All the parameter results are averaged by 100.

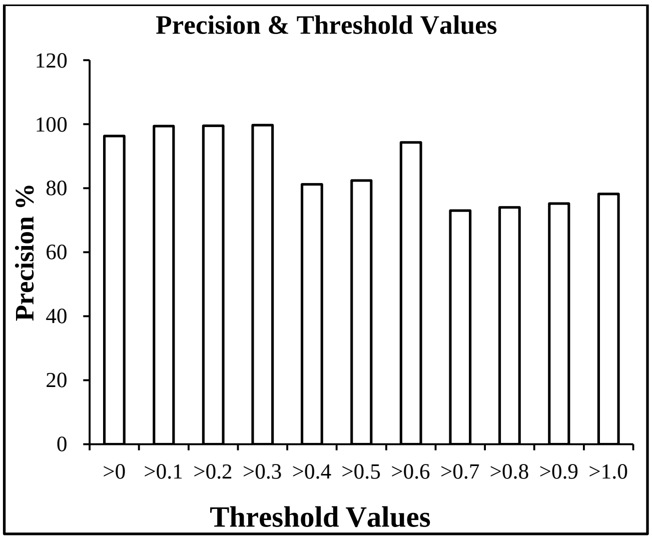


Fig.4. Precision at different Threshold Values

The Fig.4 shows the performance of our model on precision with a variation of threshold values from 0 to 1.0 at the interval of 0.1. The precisions values are gained higher values at 0.3.

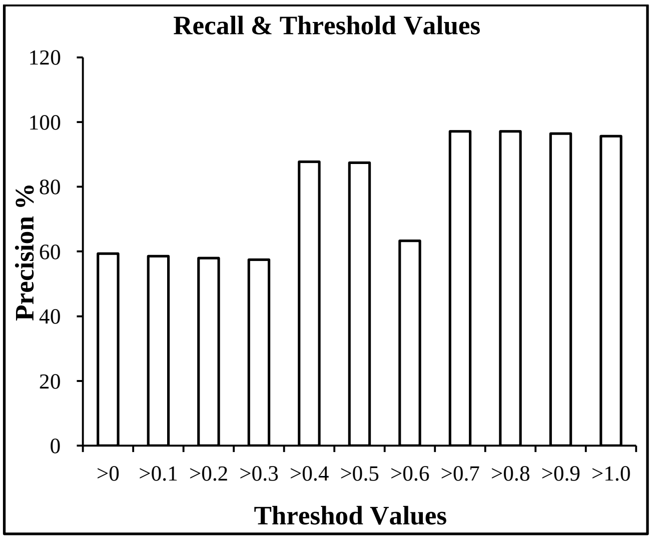


Fig.5. Recall at different Threshold Values

The Fig.5 shows the performance of our model on recall with a variation of threshold values from 0 to 1.0 at the interval of 0.1. The recall values are gained higher values at 0.7 and 0.8.

6. CONCLUSION

We have implemented a reliable and an efficient system, which suggests the user all the effective details to know about an educational domain. It too filters the query based on the user's spatial interest and it displays the location of the selected institution along with the listed direction to visit the institution. It is reliable because though it is being inputted with synonymous words and misspell, it retrieves the similar result and does not provide an irrelevant results. All the details can be retrieved in a single page, so it saves the user's inconvenience to move on to more pages to search for the right result. The system can be further refined of with more words in the search interface which can yield more filtration of the query result. The system can be better used with more performance indicators which can better model user requirements.

REFERENCES

[1] S. Nirenburg, M. McShane, T. W. Finin, J. English and A. Joshi, "Using a natural language understanding system to generate semantic web content", *International Journal on Semantic Web and Information Systems* Vol. 3 No. 4, pp. 50-74, 2007.

[2] C. D. Manning and H. Schuetze, "*Foundations of Statistical Natural Language Processing*", First Edition, MIT Press, 1999.

[3] E. Kaufmann and A. Bernstein, "Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases", *Journal of Web Semantics*, Vol. 8, No. 3, pp. 377–393, 2010.

[4] K. Palaniammal, M. Indra Devi and S. Vijayalakshmi, "An Unfangled approach to semantic search for e-tourism domain", *Proceedings of International Conference on Recent Trends in Information Technology*, pp. 130-135, 2012.

[5] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", 2001, http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html.

[6] C. Ming-Yen, C. Hui-Chuan and C. Yuh-Min, "Developing a semantic-enable information retrieval mechanism", *Expert Systems with Applications*, Vol. 37, No. 1, pp. 322-340, 2010.

[7] V. Lopez, M. Fernandez, E. Motta and N. Stieler, "PowerAqua: Supporting users in querying and exploring the semantic web", *Semantic Web*, Vol. 3, No. 3, pp. 249–265, 2012.

[8] V. Lopez, V. Uren, E. Motta and M. Pasin, "AquaLog: An ontology-driven question answering system for organizational semantic intranets", *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 5, No. 2, pp. 72–105, 2007.

[9] P. Cimiano, P. Haase, J. Heizmann, M. Mantel and R. Studer, "Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system", *Data and Knowledge Engineering*, Vol. 65, No. 2, pp. 325–354, 2008.

[10] H. L. Wang, S. H. Wu, I. C. Sung, C. L. Sung, W. L. Hsu and W. K. Shih, "Semantic Search on Internet Tabular Information Extraction for Answering Queries", *Proceedings of the ninth International Conference on Information and Knowledge Management*, pp. 243- 249, 2000.

[11] S. Cohen, J. Mamou, Y. Kanza and Y. Sagiv, "XSEarch: A Semantic Search Engine for XML", *Proceedings of the International Conference on very large data bases*, Vol. 29, pp. 45-56, 2003.

[12] F. Shaikh, U. A. Siddiqui, I. Shahzadi, S. I. Jami and Z. A. Shaikh. "SWISE: Semantic Web based intelligent search engine", *Proceedings of International Conference on Information and Emerging Technologies*, pp. 1-5, 2010.

[13] H. Dong and F. K. Hussain, "Focused Crawling for Automatic Service Discovery, Annotation, and Classification in Industrial Digital Ecosystems", *IEEE Transactions on Industrial Electronics*, Vol. 58, No. 6, pp. 2106-2116, 2011.

[14] K. Palaniammal and S. Vijayalakshmi, "Semantic Web Technique Apply for E-Museum Domain Using the Protégé Tool for Ontology Based Semantic Search", *Proceedings of the Second International Conference on Computer Application*, Vol. 4, pp. 195-199, 2012.

[15] R. Baeza-Yates and B. Ribeiro-Neto, "*Modern Information Retrieval*", First Edition, Addison Wesley, 1999.

[16] http://www.semanticweb.org/

[17] http://www.w3.org/

[18] http://www.w3.org/TR/rdf-sparql-query/

[19] http://www.w3.org/TR/owl-features/

[20] http://jena.sourceforge.net/ARQ/Tutorial