Ontology based Information Retrieval: A Case Study for Sports and Eminent Personalities Domain

R. Lakshmi Tulasi
QIS College of Engineering & Technology
Ongole
India.

M. Srinivasa Rao
Professor
School of Information Technology
JNTU, Hyderabad

G. Rayana Gouda
Sr. Assistant Professor
Department of Networking and Telecommunication,
IIIT, Pune

ABSTRACT
The objective of this paper is to develop software framework to improve information creation, maintenance and retrieval by introducing semantic technologies. This paper analyzes the drawbacks of traditional keyword based search engines and proposes the need for semantic based intelligent information retrieval systems. It also analyzes technologies specified by W3C, procedure for the creation of ontology from scratch, the evolution of ontology. This paper presents Ontology Based Information Retrieval (IR) for Sports and Eminent Personalities Domain which is developed using Protégé tool.

General Terms:
Computer Science, Information Retrieval

Keywords
Semantic Web, Ontology, OWL, Protégé tool, IR.

1. INTRODUCTION
The World Wide Web has become a vast resource of information. It is growing rapidly from last few decades. The problem is that finding the information that the individual desires is often quite difficult, because of complexity in organization and quantity of the information stored. So there is excessive demands for tools and techniques that can handle semantically.

1.1 Problems with the current Search Engines:
In conventional search engines Information Retrieval (IR) is based on keyword, set of keywords or with a natural language query which is entered by the users. It can not understand what really is the user needed, so it retrieves large number of documents in the ranked order which have poor semantic relationships among the documents. It is very difficult and time consuming for the user to navigate[1] and select the appropriate document.

This keyword based approach results *poor precision* - List of retrieved documents contains a high percentage of irrelevant documents, and *poor recall*- List of relevant retrieved among possible relevant .

This happens because of following reasons [2][3]

- A huge amount of information is currently available in information systems worldwide in the form of unstructured text.
- Documents hold the value and vocabulary of their own.
- Due to inability to handle synonyms and polysemy.
- Most search engines consult databases of the most frequently used words in documents, such as words drawn from documents title and first few sentences. Hence they won't retrieve documents in which the keywords for which you are searching for and are buried somewhere within the document.

1.2 How it can be rectified:
Semantic web**:** The semantic web is an evolving development of the WWW in which the meaning of information and services on the web is defined making it possible for the web to understand and satisfy the requests of people and machines to use the web content. The intelligent extraction system searches the information on the concept but not on the matched words. It can give the answers to users quickly and precisely. Ontologies can do this.

1.3. Technologies Involved:
The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it. The collection, structuring and recovery of linked data are enabled by technologies that provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies are specified as W3C standards and include [4] :

- Resource Description Framework (RDF), a general method for describing information
- RDF Schema (RDFS)
- Simple Knowledge Organization System (SKOS)
- SPARQL, an RDF query language
- Notation3(N3), designed with human-readability in mind
- N-Triples, a format for storing and transmitting data
- Turtle (Terse RDF Triple Language)
- Web Ontology Language (OWL), a family of knowledge representation languages.

The Semantic Web Stack illustrates the architecture of the Semantic Web [5] is shown in Fig 1.

The functions and relationships of the components can be summarized as follows:

- XML provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML is not at present a necessary component of Semantic Web technologies in most cases, as alternative syntaxes exist, such as Turtle is a de facto standard, but has not been through a formal standardization process.

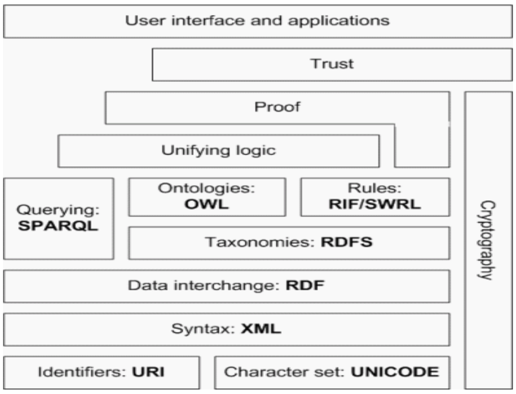

Fig 1: The Semantic Web Stack

- XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.
- RDF is a simple language for expressing data models, which refer to objects and their relationships. An RDF-based model can be represented in a variety of syntaxes, e.g., RDF/XML, N3, Turtle, and RDFs. RDF is a fundamental standard of the Semantic Web.
- RDF Schema extends RDF and it is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.
- SPARQL is a protocol and query language for semantic web data sources.
- OWL can be used explicitly represent the meaning of the terms in vocabularies and relationships between those terms. OWL has more ability to represent machine interpretable content on the Web than XML, RDF, RDF-S.OWL is a revision of DAML+OIL. OWL has three increasing expressive sub languages: OWL Lite, OWL DL, and OWL Full. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjoint ness), cardinality (e.g. "exactly one"), equality, richer typing of properties and characteristics of properties (e.g. functional), and enumerated classes.

The rest of the paper is organized as follows: A brief detail of ontology process creation is given in section 2. In section 3 we give the case study of sports and eminent personalities domain is given.

2. ONTOLOGY FRAMEWORK

2.1 What is Ontology? [6]

Ontology is the structural framework for organizing information .It formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain.

2.1 Ontology building from scratch [7]:

The methodology proposes the following stages:

2.2.1 Capture motivating scenarios: Elaborate stories or examples that describe the motivation for the proposed ontology in terms of its intended applications. At this stage, one identifies possible applications and intended solutions.

2.2.2 Formulate informal competency questions: Formulate the questions that the ontology must be able to answer based on the motivating scenarios. These questions can be stratified, that is, the answer to a question can be used to answer more general questions. At this stage, one identifies queries (Questions and corresponding answers) to be handled by the ontology.

2.2.3 Specify the terminology in a formal language: At this stage, one chooses constants, functions and predicates to be used because the methodology explicitly uses classical first-order logic (FOL).

2.2.4 Formulate formal competency questions in FOL: The identified questions are represented using the terminology defined in the previous stage.

2.2.5 Specify axioms and definitions for the terms in the formal language: Axioms restrict the possible interpretations for the formal terminology. They are necessary and sufficient conditions to express competency questions and characterize their solutions.

2.2.6 Evaluate competency and completeness: One demonstrates the competency of the ontology with respect to the set of questions that arise from the applications that use the ontology and define the conditions under which the solutions to the questions are complete.

2.2 Process of ontology creation [8]

Now methods of creating ontology have two ways: one way is that ontology is reconstructed from thesaurus and class, another way is that ontology is created by domain expert. The second way is popular now. Till now several methods are available for ontology construction. These are TOVE method, METHONTOLOGY method, frame work method, KACTUS method, SENSUS method and DEF5 method. Ontology creating tools are Ontolingua, OntoSaurus, Protégé, WebODE, and OntoEdit and so on.

Ontology can be created using protégé tool. Protégé-OWL editor is an extension of Protégé that supports the Web Ontology Language (OWL). It is a free open source platform that provides a growing user community with a suite of tools to construct domain models and knowledge based applications with ontologies. It is the most recent development in standard ontology languages, endorsed by the World Wide Web Consortium (W3C) to promote the Semantic Web vision. OWL ontology may include descriptions of classes, properties and their instances. Given such ontology, the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics. To use the OWL Tools we must ensure that the OWL Tools are installed and configured in Protégé. Process of ontology creation involves following steps [9]:

2.3.1 Create a new OWL Ontology: Start protégé .When welcome protégé dialog box appears, press ‘Create New OWL ontology’. Enter our ontology name in the place of default URI and save it in our PC.

2.3.2 Add a comment to ontology: To add a comment which describes our ontology ensure that the 'Active Ontology Tab' is selected in the 'Ontology Annotations' view, double click to the right of the comment property name. An editing window will appear in the table. Enter a comment and press *CTRL+ENTER.*

2.3.3 Create named classes as subclasses of 'Thing': The main building blocks of an OWL are the classes. The empty ontology contains a class called 'Thing'. OWL classes are interpreted as sets of individuals (or sets of objects). The class 'Thing' is the class that represents the set containing all individuals. Because of this all classes are subclasses of Thing. Subclasses of Thing are created with 'Add subclass' button of 'Classes' tab.

2.3.4 Add OWL Properties: OWL properties represent relationships. There are two main types of properties, Object properties and Data type properties. Object properties are relationships between two individuals. These link an individual to an individual. Object properties may be created using the 'Object Properties' tab. Use the 'Add Object Property' button to create a new Object property and name the property using the 'Property Name Dialog'.

2.3.5 Add Inverse Properties: Each object property may have a corresponding inverse property. If some property links individual a to individual b then its inverse property will link individual b to individual a. Inverse properties can be created/specified by using 'Add Inverse property' button in the inverse property view.

2.3.6 Create OWL Object Property Characteristics: Enrich the properties of OWL with Functional, Inverse Functional, Transitive, Symmetric, Anti Symmetric, or Reflexive characteristics in the 'Property Characteristics View'.

2.3.7 Specify Property Domain and Range: Properties may have a domain and a range specified. Properties link individuals from the domain to individuals from the range. Range can be created by using 'add' button on the 'Range view'.

2.3.8 Specify Property restrictions: Based on need add and edit property restrictions like Quantifier Restrictions, Cardinality Restrictions, Has Value Restrictions using 'Class Description View' of 'Classes' tab. Quantifier Restrictions are categorized into *E*xistential restrictions and *U*niversal restrictions.

- *Existential* restrictions describe classes of individuals that participate in *at least one* relationship along a specified property to individuals that are members of a specified class.
- *Universal* restrictions describe classes of individuals that for a given property 'only' have relationships along this property to individuals that are members of a specified class.

2.3.9 Invoke a Reasoner: Invoke the Reasoner to compute the *inferred* ontology class hierarchy and to perform *consistency* checking. This can be invoked via the 'Start Reasoner' option in the Reasoner drop down menu.

2.3.10 Set Data type Properties: Describe relationships between an individual and data values. These can be created using 'Add Data type Property' button of the 'Data type Properties' tab.

3. CASE STUDY

3.1 Ontology for Sports and Eminent Personalities Domain

The main objective of this ontology is to create a knowledge base for sports and eminent personalities. It provides relevant results based on domain specific knowledge and improves both the precision and recall.

3.1.1 Create new Ontology for Sports and Eminent Personalities Domain: Start protégé .When welcome protégé dialog box appears, press 'Create New OWL ontology'. Enter 'Sport Ontology. Owl' name in the place of default URI and save it in our PC.

3.1.2: Add a comment to ontology: Using 'Active Ontology' Tab in the 'Ontology Annotations' view adds a comment, "This is the ontology that explains the sports and eminent personalities".

3.1.3: Create subclasses of 'Thing': With 'Add subclass' button in the 'Classes' tab creates Continent, Field, Persons, and Sport as subclasses of 'Thing'. Repeat same process to create Africa, America, Asia, Europe, and Oceania as the subclasses of Continent. North_America, South_America as subclasses of America. Artists_and_Entertainers, Leader_and_Revolutionaries, Scientist_and_Thinkers, Sports_Personalities as subclasses of Persons as shown in Fig 2.

3.1.4: Add individuals to a class: Add individuals to a class using *'Add individual'* button in 'individual members list view'. Ex: Cameroon, Egypt, Ivory_coast, Malawi, Nigeria, Rwanda, South_Africa, and Swaziland are members of Africa as shown in Fig3.

3.1.5: Add OWL Properties: With 'Add Object Property' button of the 'Object Properties' tab create Object Properties like died_in, grew_up_in, is_led_by, is_the_birth_country_for, is_the_death_country_for etc. as shown in Fig4.

3.1.6: Add OWL Object property characteristics: Add functional characteristic to died_in, grew_up_in, is_the_father_of, lived-in, plays, and was_born_in object properties. Add Inverse functional characteristic to is_the_birth_country_for, is_the_death_country_for, is_well_played_by, was_a_living_country_for, and was_developed_by etc.

3.1.7: Invoke a Reasoner: Invoke Reasoner to check asserted class hierarchy and the inferred class hierarchy. The asserted class hierarchy matches with the inferred hierarchy, and no inconsistencies so nothing is displayed in the 'Class hierarchy' view as shown in Fig5.

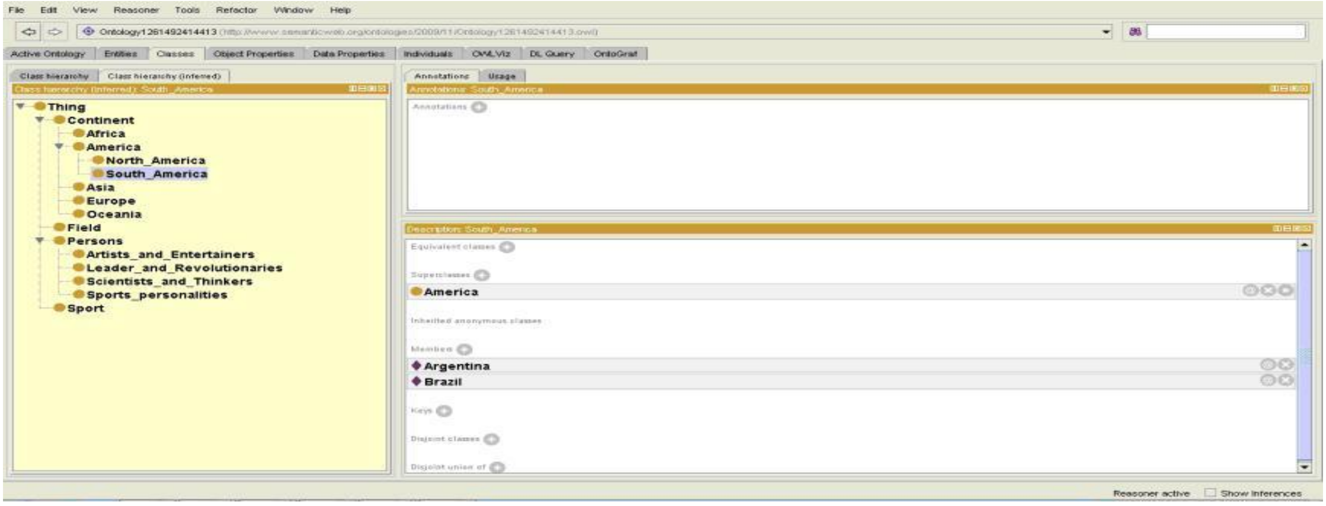

Fig 2: Named subclasses of 'Thing'

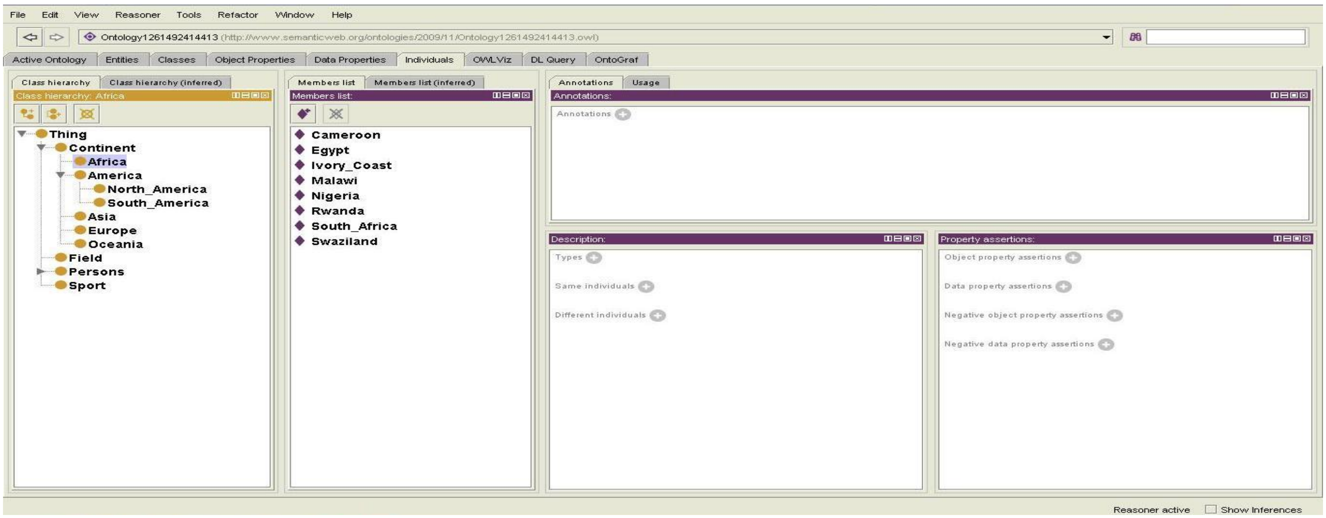

Fig 3: Individuals to the class Africa

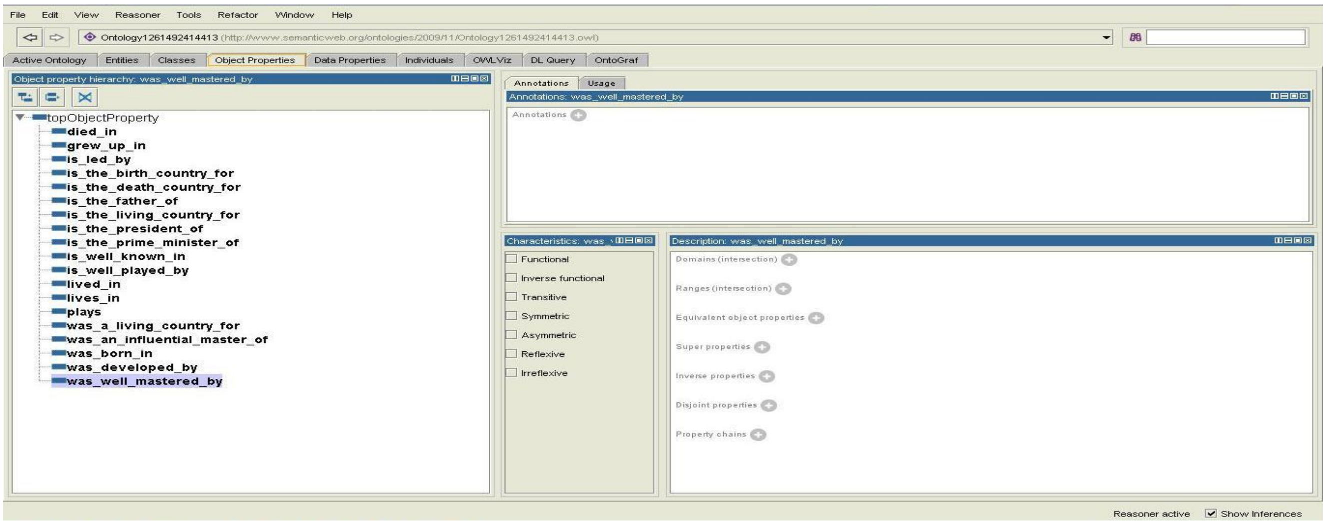

Fig 4: OWL Object Properties of classes

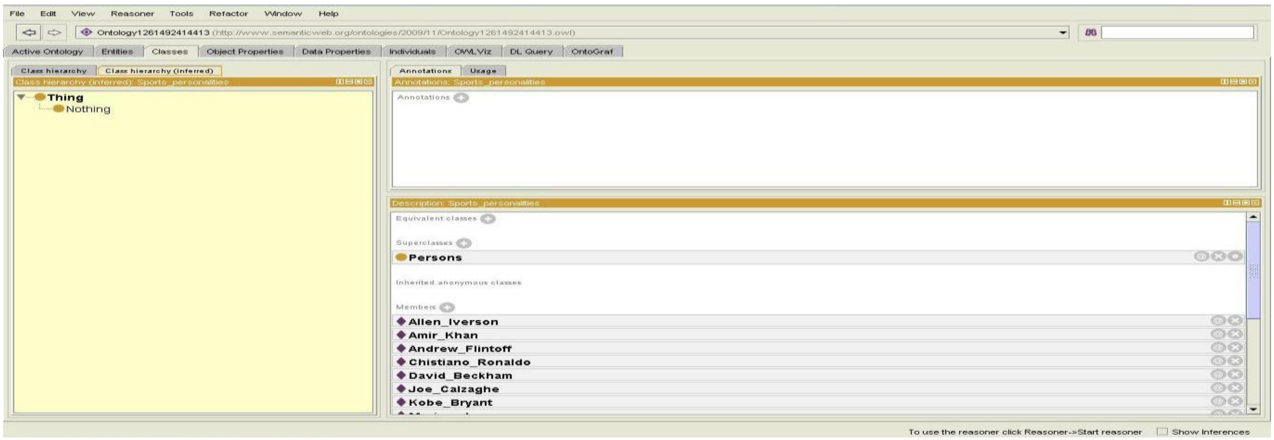

Fig 5: Class hierarchy view –Invoke Reasoner

3.1.8. Executing DL Query[10]*:* Ontology can be tested in the query search engine of the Protégé tool for the given query.
Query1: Scientists_and_Thinkers and lived_in only Europe

Result: Search engine displays Scientists_and_Thinkers lived_in Europe. Results are shown in Fig 6.

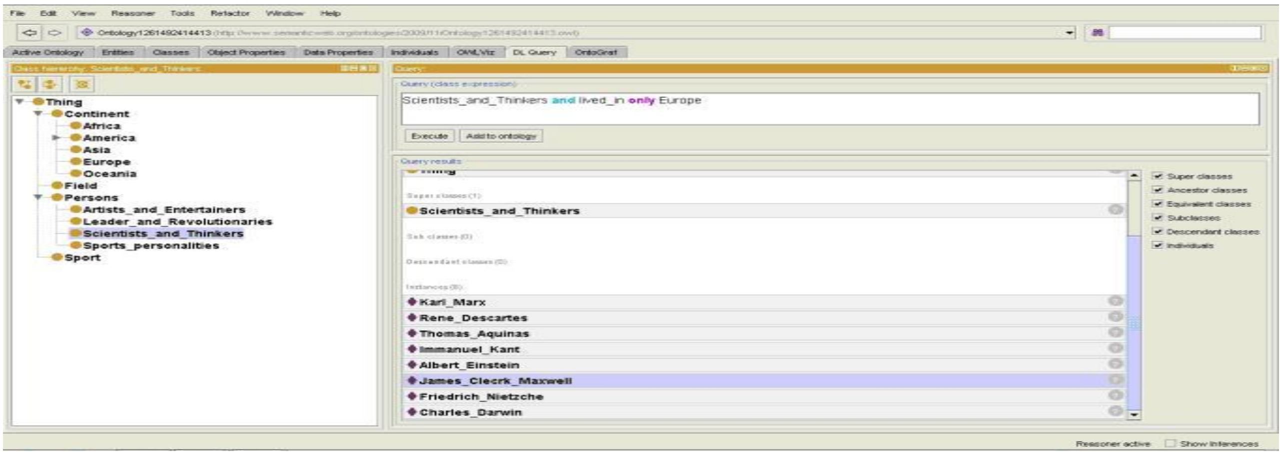

Fig 6: Query result for "Scientists_and_Thinkers and lived_in only Europe"

3.1.9. Represent Ontology Graph which shows semantic relationships between classes and instances of Sports and Eminent Personalities Domain using 'Onto Graf' tab. It is shown in Fig7.

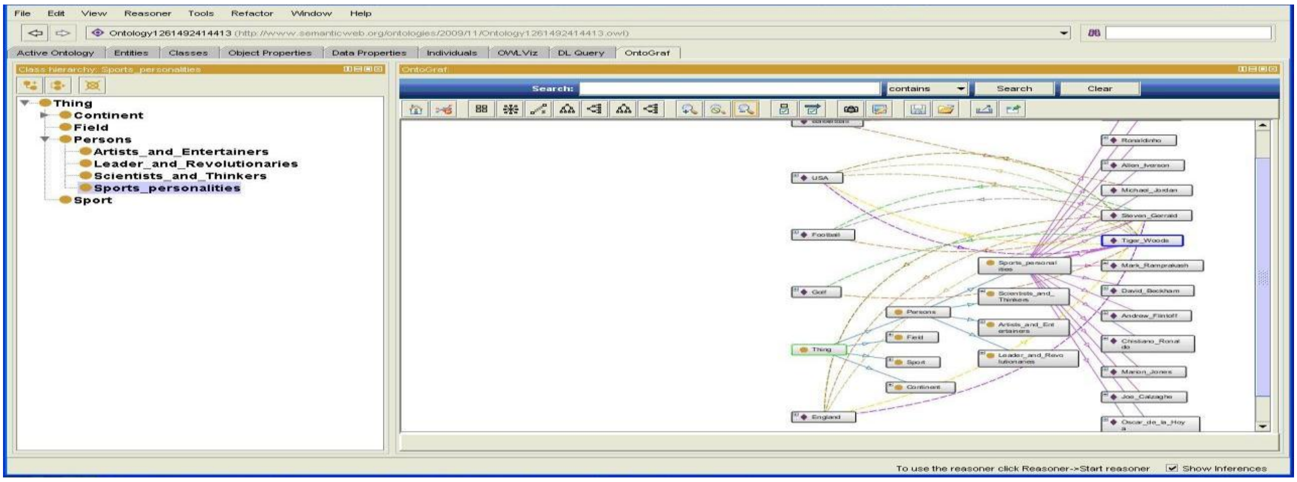

Fig 7: Ontology graph for Sports and Eminent Personalities domain

3.2 A singular activity: Evaluation [11]

Ontologies have been initially introduced as one of the solutions to the problems that impair knowledge sharing and reuse. Therefore, one important feature is the fact that they are shared by a group. All methodologies for building ontologies from scratch recognize the importance of evaluation because one needs to guarantee the quality of the resulting ontology to its users.In OE, one distinguishes between two kinds of evaluation:

1) Technical evaluation: Judge the ontology and documentation against a framework. There are two activities involved in evaluation: (1) verification, which guarantees its correctness according to the accepted understanding about the domain of specialized knowledge sources, and (2) validation which guarantees that it corresponds to what it is supposed to, according to the specification requirements document.

2) User assessment: Judge from the user point of view the usability and usefulness of the ontology and its documentation when (re)used or shared in applications

4. CONCLUSIONS

The drawbacks of the conventional search engines can be rectified with semantic web technologies. Among the technologies involved in this we have chosen OWL ontology to develop efficient retrieval of documents. The proposed system is an effort to retrieve relevant documents in a Sports and Eminent Personalities domain which is represented in the form of OWL ontology.

5. REFERENCES

[1] Ralf Bergmann, “Experience Management Foundations, development Methodology and Internet based applications”, 1st Ed, Springer, 2002, ch3, pp50-52.

[2] Gerald Kowalski and Mark T May bury , *Information Storage and Retrieval Systems*: *theory and implementation*, 2nd ed., Kluwer Academic Publishers, Springer,1997, ch.1, pp 6-10.

[3] W.B Frakes and R.A.Baeza-Yates, eds., *Information Retrieval: Data structures and Algorithms*, Prentice-Hall, 1992.

[4] Van Dam, A.,”Hypertext’87 Keynote Address”, Communications of the ACM, Vol 31, No.7, July 1988,pages 887-895.

[5] Aurora Gerber, Alta Vander Merwe, and Andris Barnard,” A functional software Architecture” European Conference”, June 2008,pp 273-287

[6] Lee W Lacy, *OWL:representing information using the Web Ontology Language*, 2nd ed., Trafford Publishers, 2005, ch3,pp 90-103.

[7] C. Debruyne, P. De Leenheer, and R. Meersman, “Fact Type Reuse in the DOGMA ontology Framework,” in *Proc. On the Move to Meaningful Internet Systems: OTM 2009:Confederated International Conferences*, Vilamoura ,Portugal,2009, pp 1148-1163 .

[8] Ellis C. Montero Hernandes, Deysiane sande , and Sandra Fabbari, “Using visualization and a collaborative Glossary to support Ontology Conceptualization”, presented at the 12th Enterprise Information Systems, Portugal, June 8-12, 2010.

[9] protégé owl tutorial http:// owl.cs.manchester.ac.uk /tutorials/

[10] EduardoMena, and Alatza Illaramendi “Ontology Based Query Processing for global information Systems”, 2ed, Kluwer Academic publishers, 2001, Ch 4,107-108.

[11] Oliver Boyd Barrett, and Pamela and Pamela OMalley “*Educationreform in democratic spain*”, Taylor & Francis, 2003, Ch 5, pp 60-63.