SEMANTIC KNOWLEDGE ACQUISITION OF INFORMATION FOR SYNTACTIC WEB

G.Nagarajan[1] and K.K.Thyagharajan[2]
[1]Research Scholar, Sathyabama University, Chennai,India
nagarajanme@yahoo.co.in
[2]Professor,Dept. of Information & Technology, RMK College of Engineering & Technology, Chennai, Tamil Nadu ,India

ABSTRACT

Information retrieval is one of the most common web service used. Information is knowledge. In earlier days one has to find a resource person or resource library to acquire knowledge. But today just by typing a keyword on a search engine all kind of resources are available to us. Due to this mere advancement there are trillions of information available on net. So, in this era we are in need of search engine which also search with us by understanding the semantics of given query by the user. One such design is only possible only if we provide semantic to our ordinary HTML web page. In this paper we have explained the concept of converting an HTML page to RDFS/OWL page. This technique is incorporated along with natural language technology as we have to provide the Hyponym and Meronym of the given HTML pages. Through this automatic conversion the concept of intelligent information retrieval is framed.

KEYWORDS

Ontology, OWL, RDFS, Name entity recognition , machine learing,Probability Reasoner;

1. INTRODUCTION

Information is the main source of intelligent. Information is poured all over the internet but when we search for particular, the result would be again trillion of informative and non informative information; again we need a refine search manually. This can be overcome by Semantic approach. Research in information retrieval (IR) community has developed different techniques to help the people locate relevant information in large document repositories. The variety of techniques is besides classical IR models (i.e., Vector Space and Probabilistic Model) [1], extended models such as Latent Semantic Indexing [2],Machine Learning based models (i.e., Neural Network, Symbolic Learning, and Genetic Algorithm based models) [3] and Probabilistic Latent Semantic Analysis (PLSA) [4] has been devised with hope to improve semantic information retrieval process.

In this paper we proposed ontology based information retrieval system named as Intelligent Semantic Information Retrieval System. The primary goal of this paper is to design a Intelligent search engine which has to provide only the needed relevant information regarding the given query.

Semantic search engine [5] is the only key answer for this kind of search. As said in here both the machine and the user tries to search some information on web. There are many research papers regarding the design of a new semantic search engine. In [6] even listed top five to ten Semantic Search Engine. The main drawback of Semantic Search Engine is that the available Semantic Web page[7] on web is very few. As the concept of Semantic Web had started on around 2000, we have very few Semantic Web page. As the creation and design of Syntactic Web page is ease of work also we have lot of in-built software for it still people are interested in creating simple

DOI : 10.5121/ijwest.2012.3409

Syntactic web page with general XHTML,XML,PHP,ect. coding instead of construct ontology for it.

Thus to build a Intelligent Semantic Information Retrieval System the domain related syntactic web page is converted to its corresponding Semantic web page using this collection of Semantic web page we can build a Intelligent Semantic Information Retrieval System. The specific domain taken for our research is sports domain where the events Cricket, Croquet, Tennis and Volleyball in short we called as CCTV Sports Conceptual model.

The paper is organized in such a way that first the concept of syntactic to semantic conversion steps is explained through which the concept of Intelligent semantic information retrieval system framework is drafted.

2. RELATED WORK

In [8] discuss the way of converting the HTML to OWL using table. They consider the TABLE tag of HTML page and tried to convert to OWL. This won't produce any semantic to the ontology. In [9] they tried to convert the HTML to OWL using the FRAME set tags they also tried to incorporate UML to identify the class and subclasses. In [10] the conversion is done by first annotating the web page. The annotation they consider is the semantic annotation thus they tried to provide semantic of the page. They use the tool called GATE to analyze the semantic through natural language processing. In [11] the conversion is take place using the tag and they used GRDDL tool for conversion.

3. LINGUISTIC CONVERSION OF SYNTACTIC TO SEMANTIC WEB PAGE

As we search for intelligent information we need to design an intelligent system for this Syntactic to Semantic conversion. Figure1 shows the proposed framework, where the collection of Syntactic web pages are collected via a Web crawler as the output of an web crawler is list of URL we required a genius system to filter out the unwanted URL. Then with the available list of URL of input the XML is created with that entity concern XML a conversion of XML to OWL is implemented and the crated ontology is collected in repository which can be used by an of the Semantic Web search services. The conversion of HTML to XML and the XML to OWL is explained in detail in the forth coming session. The concept of Web Crawler is already explained in [12] we are not specifying in this paper.

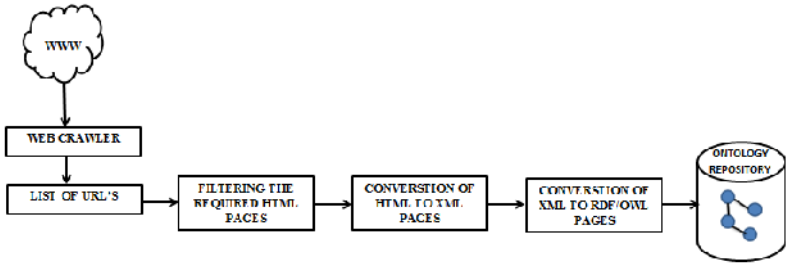

Figure 1 Web Intelligent Framework

3.1. HTML to XML Conversion

The first phase of this Web Intelligent Framework is the conversation of all the web page collected from a web crawler to a standard XML files with name entity as the main entity. Name Entity Recognition is a concept of Natural Language Processing. In short it is called as NER.

The main technology used here are patterns and Lexicons. For the given Corpus text NER classifies the entity as Person Name, Organization Name, Location and Miscellaneous (Date, Time, Number, Percentage, Monetary expression, Number expression and Measurement expression.)

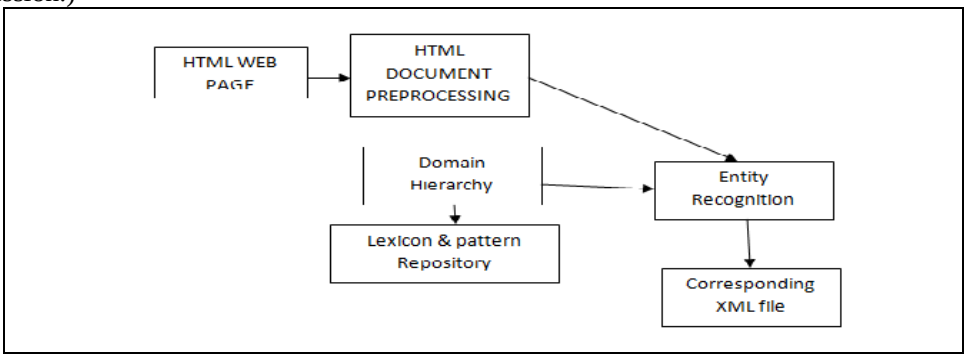

Figure.2. HTML to XML Conversion

Figure 2 shows the general framework for converting HTML document to XML using Name Entity concept this technique is derived from [13].

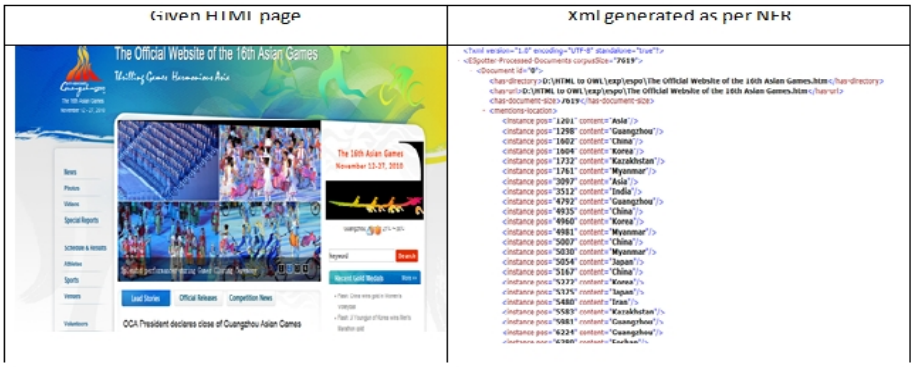

Figure 3. Output of the conversion

Figure 3 shown the output of XML creation of the given website which is relevant to Asian games. The concern entity relation XML for Organiztion is given below:

```
<mentions-organization>
    <instance content="Guangzhou Online News Centre" pos="5954" />
    <instance content="Guangzhou Asian Games Organising Committee"
pos="7257" />
    <instance content="Spectator Services" pos="393" />
    <instance content="Media Services" pos="412" />
    <instance content="Olympic Council" pos="1198" />
    <instance content="Press Conferences" pos="3112" />
    <instance content="The Radio Management" pos="5807" />
    <instance content="News Coverage Tour" pos="5977" />
    <instance content="Media Friends'" pos="5983" />
  </mentions-organization>
```

3.2. XML to RDFS/OWL Conversion

The next phase of the searching technique is the XML to OWL/RDFS conversion. Thus through this conversion we provide semantic to the web page. As the conversation is take over

automatically we need same format of well formatted XML file, that's the reason we use the generalized NER technique for XML conversion which provide same entity name tag. With this how we can convert is the main focus of this work.

The semantic of the web page can be given by defining the RDFS which provide the rules of the web page and also defining OWL which define the conceptual ontology of the web page. In our work we have done this through two main techniques one is via Syntactic Analysis and another technique is via Semantic Analysis

3.2.1 Syntactic Analysis

Here we generally map the XSD element [14] and convert to OWL element for the mapping the strategy shown in Table 1 is used

Table 1: XSD to OWL

SN	XSD	OWL
1	Xsd:elements,containing other elements or having at least one attribute	Owl:class,coupled with owl:ObjectProperties
2	Xsd:elements,with neither sub-elements nor attributes	Owl:DatatypeProperties
3	Named xsd:complexType	Owl:class
4	Named xsd:SimpleType	Owl:DatatypeProperties
5	Xsd:minOccurs,xsd:maxOccurs	Owl:minCardinality,owl:maxCardinality
6	Xsd:sequence,xsd:all	Owl:intersectionOf
7	Xsd:choice	Combination of owl:intersectionOf, owl:unionOf and owl:complementOF
8	xsd:simpleType	owl:Datatype
9	xsd:simpleType with xsd:enumeration	Becomes an owl:Class as a subclass of EnumeratedValue. Instances are created for every enumerated value. An instance of Enumeration, referring to all the instances, is created as well as the owl:oneOf union over the instances.
10	xsd:complexType over xsd:complexContent	owl:Class
11	xsd:complexType over xsd:simpleContent	owl:Class
12	xsd:element (global) with complex type	owl:Class and subclass of the class generated from the referenced complex type
13	xsd:element (global) with simple type	owl:Datatype
14	xsd:element (local to a type)	owl:DatatypeProperty or owl:ObjectProperty depending on the element type. OWL Restrictions are built for the occurrence.
15	xsd:group	owl:Class and subclass of A_AbstractElementGroup
16	xsd:attributeGroup	owl:Class and subclass of A_AbstractAttributeGroup
17	xsd:minOccurs and xsd:maxOccurs	Cardinality specified in minimum cardinality, maximum cardinality and universal (allValuesFrom) OWL restrictions.

18	Anonymous Complex Type	As for Complex Type except a URI is constructed from the parent element and the nested element reference. Also, the class is defined as a subclass of A_Anon.
19	Anonymous Simple Type	As for Simple Type except a URI is constructed from the parent element and the nested element reference.
20	xsd:default on an attribute	Uses dtype:defaultValue to attach a value to the OWL restriction representing the associated property.
21	Substitution Groups	Subclass statements are generated for the members. Instance files resolve their types by consulting the OWL model at import-time.
22	Annotation attributes on elements	OWL Annotation properties are created and placed directly on the relevant class.
23	Annotations using xsd:annotation	Become, based on user selection, dc:description, rdfs:comment and/or skos:definition OWL annotations.
24	xsi:type on an XML element	Overrides the schema type with the specified type.

Ontologies main elements are the owl classes, Object property, Data Property and all those constrain and cardinality element. Here as shown in Table 1 the XML element is converted. The main drawback of this approach is that some time an irrelevant data element would be tagged in OWL may produce irrelevant output. Figure 4 shows a pictorial representation of conversion of XML to OWL conversion

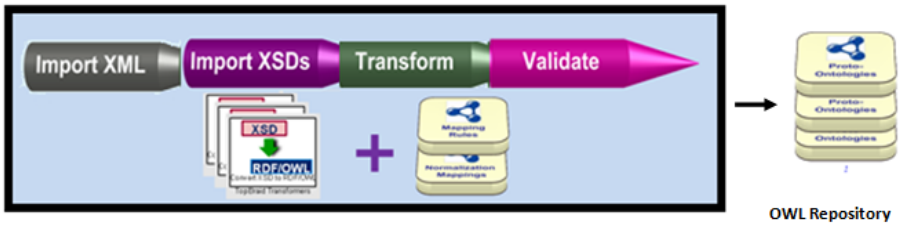

Figure 4. Syntactic Analysis

3.2.2 Semantic Analysis

In this analysis the RDFS/OWL is generated using Natural Language Processing techniques [15]. For a Semantic Web page we have to create both RDFS and OWL. RDFS which Resource Descriptor Framework is a kind a rules and logic regarding the content on the page. A human identifies and analysis any intelligent information only via logical reasoning. Likewise RDF produces logic to the web page. OWL, Web ontology language used to produce the ontology of the web page with this only we will have the whole conceptual idea of any general concept we can give accurate results.

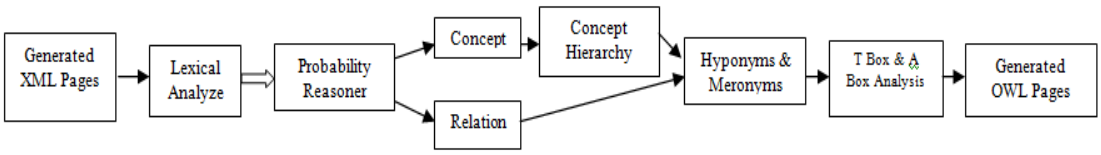

Figure 5. Semantic Analysis

Figure 5 show a general framework for generated OWL from the generated XML. To analyze the content of XML we uses Lexical Analyzer which analyze each entity XML tag and represent whether they are noun or verb or any other verbal notation. Once analyzed we use the concept of Probability Reasoner to determine the concept and relationship between them as ontology is of considering the concept and their relationship between them. We use Probability as they used to handle uncertainty with the help of deductive logic i.e using a set of hypothesis for reasoning. The primary relationship between the concept is to be identified are "is-a" and "part-of" relation. To identify this kind of relationship [16] to create a full structured ontology we have to determine the Hyponym and Meronym of the identified tag. There is an automatic way to determine and extract the Meronym with the following linguistic pattern

```
1) Such NP as NP,*(or|and) NP
2) NP, NP* or other NP
3) NP, including NP, or|and NP
```

Probability Deductive Reasoning

Probability reasoning is used to handle uncertainty by using the basics of Mathematical induction concepts. Where it provide knowledge for a given conceptual model via some logical reasoning technique. There are so many reasoning techniques available such as Deductive reasoning, Inductive reasoning, Abductive reasoning, Analogical reasoning and Fallacious reasoning. Among them to handle uncertainty we go for deductive reasoning logic. In deductive reasoning, knowledge acquisitions can be done by using one or more domain conceptual model. In our work we are using the visual and textual domain conceptual model as the input to the probability deductive reasoner.

Through the Reasoner the terms, concepts and the relationship between the concepts are determined from the visual and textual domain conceptual model. The concept such as cricket, ball, bat etc., and there object properties like has_ball, has_bat etc., is been extracted from the domain model. From these the hyponyms and meronyms are analyzed. Where hyponyms of a concepts are, semantically related concepts. Thus it provide a semantic relationship "is-a" between two related concepts or terms. For example has_cricket_batting_ball and has_cricket_bowling_ball are semantically related . Metonyms provides the "part-of" relationship between the concepts or terms for example has_croqent_hoop is a part of has_croquent_mallent. Likewise the "part-of" and "is-a" relationship between all the classes and object properties is been analyzed over here thus to provide an semantic knowledge to the intelligent semantic search system.

T box and A box

One of the main components of Semantic web architecture layer is Logic and rules (7). The mere idea of Logic introduced in semantic web is that to provide a logical agent vise decision making when comes to semantic oriented approach. So, in our work we need some logical representation

to provide an efficient information retrieval system. These logical can be derived from the descriptive logical notation which has the mere similarity with ontology.

The concept in Description logic is knows as classes in ontology, likewise Role as property and individual as object. With this we can introduce Rule-based ontology reasoning A-box and T-box. These are the facts associate with the visual and textual domain concept, relation and object to the knowledge base conceptual model. A T-box provides the associate classes and property whereas and A-box provides the instance of those classes.

The semantics of this domain can be defined by the interpreting concept i.e. represent one conceptual model agent with the another conceptual model thus the semantic between visual and textual agent is describe using interpreting description logic.

Where interpreting I,

I = (domain ,{classes, property and object})

Thus for an T-box

$$I \models C \subseteq D$$

Where C = classes

D = domain

For A-box

$$I \models a : C$$

Where a is the class instant of C and

$$I \models (a, b) : R$$

Where a is class instance of R related to b an instance

Using these reasoning techniques the CCTV ontology can be pruned and refined using the multi agent domain conceptual model.

4. INTELLIGENT SEMANTIC INFORMATION RETRIEVAL SYSTEM

In any information retrieval system the search would be effective only if the search content is organized in some way and only the efficiency depends on how the search take place. A human can acquire an knowledge only if he knows what he is searching for likewise a system can provide intelligent information only it knows the semantic of the content it searching for. Thus in all of our work we try to provide semantic information visually and textually to our intelligent system.

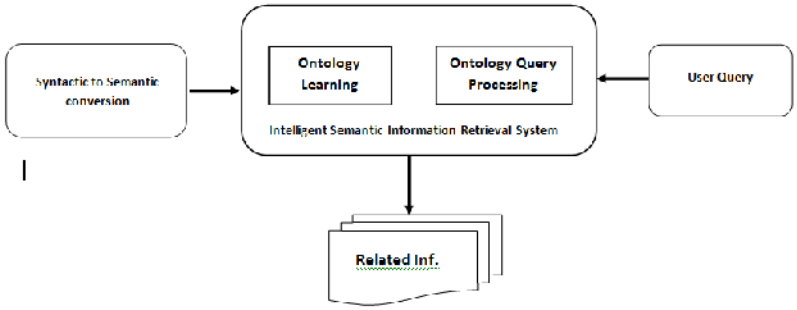

Figure 6. Intelligent Semantic Information Retrieval System

Figure.6 shows the overall framework designed for information retrieval using ontology concepts. Searching the ontology is done by SPRQL or OWL-QL language (18) . as seen in former session logical descriptive language is one of the strong foundation of ontology. These query language is basically used to search an ontology for relevant result of each classes in the ontology is represented by a specific URI.

For textual keywords or for an sentence the probability reasoning concept is used to analyze the query to extract the main concepts and relation words from which the related pages is displayed.

5. CONCLUSIONS

If we have knowledge about what we are searching for, we can easily retrieve the desire information. The main drawback in information retrieval procedure in web technology is that the technology doesn't know the semantic and syntax of what the user searching for. This gives birth to the Semantic Web Technology. In this paper we deal with the reusability technique of using converting the available HTML pages to an Ontology enriched Semantic web page.

REFERENCES

[1] R A Baeza-Yates and B A Ribeiro-Neto(1999)," Modern Information Retrieval", ACM Press / Addison-Wesley

[2 S C Deerwester, S T Dumais, T K Landauer, G W Furnas and R A Harshman,(1990) "Indexing by latent semantic analysis. "JASIS, Vol. 41, No. 6, pp. 391–407.

[3] H Chen,(1995) "Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms". JASIS, Vol. 46, No. 3, pp. 194–216.

[4] T Hofmann,(1999)" Probabilistic latent semantic analysis". UAI , pp. 289–296.

[5] Wang Wei, Payam M. Barnaghi, Andrzej Bargiela,(2008)," Search with Meanings: An Overview of Semantic Search Systems"

[6] Kyumars Sheykh Esmaili, Hassan Abolhassani (2005). " A Categorization scheme for semantic web search engines".

[7 Berners-Lee T, Hendler J, Lassila O.(2001)" The Semantic Web". Scientific American 284(5):35–43.

[8] Yuri A Tijerino et al (2004) "Towards ontology generation from tables" Kluwer academic Publishers (2004)

[9] Sidi Benslimane et al (2006) "Towards ontology extration from data intensive web sites: An html forms based reverse engineering approach" International arab journal of information technology.

[10] Debajyoti Mukhopadhyay et al (2007)" A New semantic web services to translate HTML pages to RDF" Int, Conference of IT .

[11] Hoon Hwangbo et al (2008) " Reusing of information constructed in HTML document : a conversion of HTML to OWL" Int. conference on control, automation and systems .

[12] Hsien-Tsung Chang,"Web Image retrival systems with automatic web image annotaing techniques":WSEAS trancaction on Information science and application.

[13] Zhu,J.,Uren,V.&Motta,E.(2005)"Espotter:Adaptive named entity recognition for web browsing". In Intelligent IT tools for knowledge Management System,KMTOOLS pp.518 – 529

[14] Hannes Bohring and Soren Aure (2004) "Mapping XML to OWL ontologies"

[15] Alessandro Lenci et al (2006) "NLP based ontology learning from legal texts. A case study" .

[16] Saminda Wishwajith Abeyruwan (2010) "Prontolearn: unsupervised lexico semantic ontology generation using probabilistic methods" Workshop Uncertainty Reasoning for the Semantic Web. These of universtiy of Miami.

[18] P Cimiano, (2006) "Ontology Learning and Population from Text: Algorithms, Evaluation and Applications", Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Authors

G.Nagarajan has received his Diploma in Electronic & Communication Engineering from Directorate Of Technical Education 1997. He has received his BE degree in Electrical & Electronic Engineering from Manonmaniam Sundaranar University 2000. He received his ME degree in Applied Electronic Engineering from Anna University 2005. He also received his ME degree in Computer Science Engineering from Sathyabama University 2007. He is at present a PhD Scholar in Computer Science Engineering from Sathyabama University. His research areas are web Image Mining , Artificial Intelligent, Ontology Learning, Machine Leaning, NLP and Semantic Web.

Dr. K.K. Thyagharajan has received his B.E., degree in Electrical and Electronics Engineering from PSG College of Technology (Madras University). He received his M.E., degree in Applied Electronics from Coimbatore Institute of Technology and Post Graduate Diploma in Computer Applications from Bharathiar University. He has received his Ph.D., (Multimedia Streaming) degree in Information and Communication Engineering from College of Engineering Guindy, Anna University. He has written 5 books in Computing. His book "Flash MX 2004" published by McGraw Hill (INDIA) has been recommended as text / reference book by many universities. He has published more than 30 papers in National and International Journals and Conferences. He is a grant recipient of Tamil Nadu State Council for Science and Technology. His biography has been published in the 25th Anniversary Edition of Marquis Who's Who in the World Directory. He has been invited as chairperson and delivered special lectures in many National and International conferences and workshops. He is reviewer for many International Journals and Conferences. His current interests are Multimedia Networks, Mobile Computing, Web services, Data Mining, e-learning, Image Processing, Microprocessors and Microcontrollers. He has guided 10 M.E. projects and now 9 students are doing Ph.D. under him in the area of Multimedia, Image Processing and Data Mining. He is a life member of Computer Society of India and Chairman of the ISTE chapter of RMK College of Engineering and Technology.