JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-4 E-ISSN: 2347-2693

Keyword Based Efficient Web Crawler for Next Generation Semantic Web

Sheetal[1*] and Megha Bansal[2]

[1*,2] *Department of Computer Science, MDU , India*

www.ijcseonline.org

Received: 03/03/2014 Revised: 19/03/2014 Accepted: 15/04/201x4 Published: 30/04/2014

***Abstract*—** This paper will be implement Ontology Based Search Using Semantic Web. The method of web crawling with filter is used. This approach is query based approach using Jena API. The proposed approach solves the problem of revisiting web pages by crawler. The Semantic Web is an extension of the current Web that allows the meaning of information to be precisely described in terms of well-defined vocabularies that are understood by people and computers.

***Index Term*—**Sementics Web, Metadata, An Ontology,RDF,Search Engine

I. INTRODUCTION

The number of web pages available on the internet is growing day by day and so, in the case searching the relevant information in the internet is very hard. But usually, searching the relevant information in the internet is very hard. But usually, searching the information on www (World Wide Web) can be done by finding the so many lists of links. The crawler is an program which downloads the information or data from the www for search engines is called web crawler and this process is called web crawling. To fulfill the requirements of uniquely identifying a web page ,each web page is assigned an URL (Uniform resource locater) that effectively serves as the pages world wide name.Url having the three parts :the protocol(called as a scheme),the DNS name of machine on which the page is located ,and a local name indicating the specific page.web crawler searches the web for new information.In this paper we design and develop a semantic web architecture that can relieve users from overburden of doing lot of keywords based search before getting the appropriate result or information. in this proposed system the topic based web interface has to be develop for accomplishing the goal of semantic browsing in a semantic web environment. The proposed system exhibits the refinement with higher accuracy and in automatic way.The proposed system architecture has to be divided in two different modules:First module takes the query in the form of topic description and the second module provides a mechanism for the user to enhance the search results for the image filtering tools.In this semantic web,ontolgy RDF,OWL is used.

II. RELATED WORKS

1.1 AN ONTOLOGY: Ontology is a model of the world, represented as a tangled tree of linked **concepts**. Concepts are language-independent abstract entities, not words. They are expressed in this ontology using English words and phrases only as a simplifying convention.

1.2 Jena ontology API: Jena is a programming toolkit, using the Java programming language. Through the Ontology API, Jena aims to provide a consistent programming interface for ontology application development, independent of which ontology language you are using in your programs.

The Jena Ontology API is language-neutral: the Java class names are not specific to the underlying language. For example, the `OntClass` Java class can represent an OWL class or RDFS class.

1.3 RDF:RDF uses (XML) as a common Select the required for the exchange and processing of metadata. Semantic Web :It is an idea of having data on the web defined and linked in a way that it can be used by machines not just for displaying purpose.The main intent of semantic web is to machines much better access to information resources . For example, a user query is “Give me the name of users that works in ABC University and are age below 60”. These queries are not built using natural langue (such as phrases), but with an easy to use user interface that help users to build the queries they want. Different person can give this query in the different forms.

SPARQL: SPARQL allows users to write queries against data that can loosely be called "key-value" data, more specifically it is data that follows the RDF specification of the W3C. The entire database is thus a set of "subject-predicate-object" triples.

A. *Example:*

Another SPARQL query example that models the question "What are all the country capitals in Africa?":

III. METHODOLOGY

for Input Output Architecture Proposed System

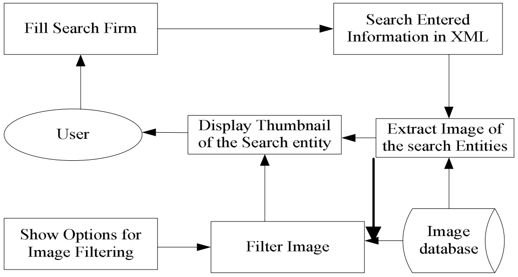

Fill Search Firm
Search Entered Information in XML
User
Display Thumbnail of the Search entity
Extract Image of the search Entities
Show Options for Image Filtering
Filter Image
Image database

Figure 1. Input/ Output Architecture

Corresponding Author: *Sheetal*
Department of Computer Science, MDU , India

The above mentioned tentative figure represents the input output architecture for proposed search engine.

A. Training the System

The data of various users is collected and organized around ontology of users. The system has a large database of images belonging to various categories. These images are passed into an algorithm which extracts various metadata of image such as file type, file size, file dimension, date created on etc. An algorithm “Nearest neighbor interpolation” method was used to calculate the average color of the image by resampling the image to a 1 x 1 dimension. All the details along with the URL of image file and it s category is stored in a database. The category of an image is identified manually and it can be anything like age, place where he works, location etc.

B. Structure of Topic based search form

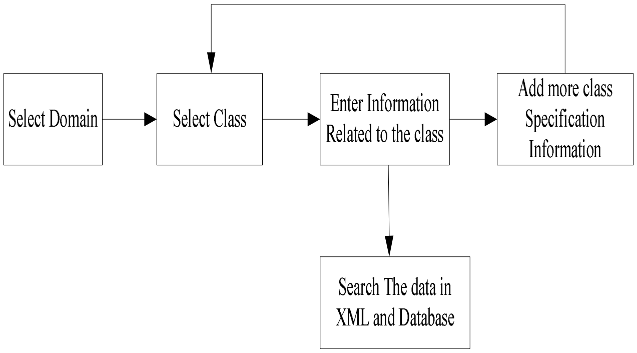

Select Domain
Select Class
Enter Information Related to the class
Add more class Specification Information
Search The data in XML and Database

Figure 2 Structure of Topic based search form

As user enters his natural language query by selecting category and topics and then entering the details he has. He can enter multiple queries similarly. Once the user has built his query, it has passed through a search mechanism where the data is first checked in the xml file then images are retrieved from the database

C. User Interface for image display and filtering

User interface is the program that user can see and use. For a particular domain, user enters relevant search keywords. These keywords are then searched in the database using SQL query. Figure 5 depicts the metadata extraction from the image.

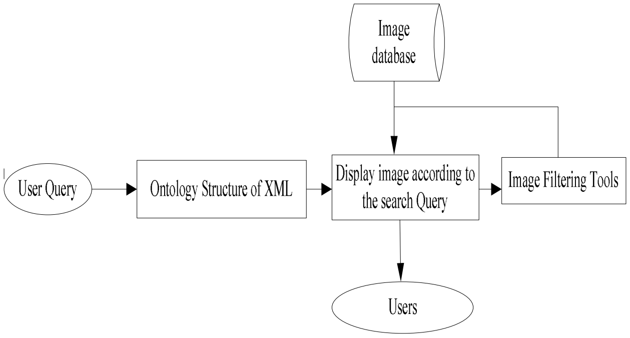
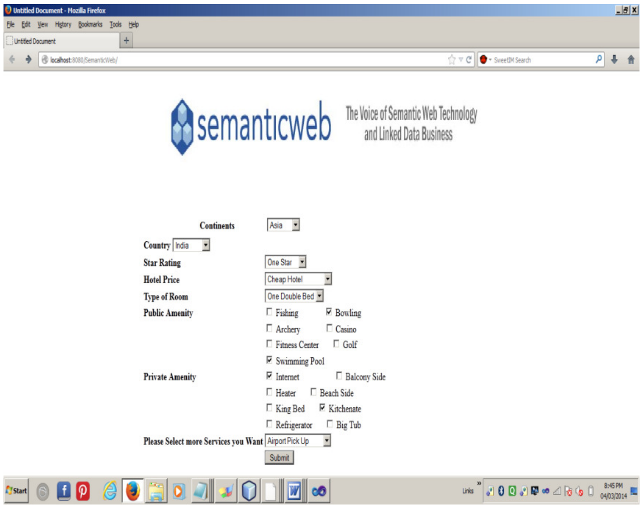


Image database
User Query
Ontology Structure of XML
Display image according to the search Query
Image Filtering Tools
Users

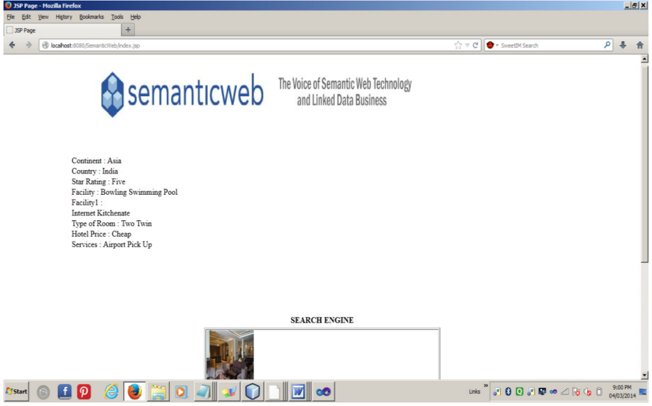
Figure 3. Extracting metadata from image

IV. Result & Discussion


Untitled Document - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8080/SemanticWeb/
semanticweb The Voice of Semantic Web Technology and Linked Data Business

Continents Asia
Country India
Star Rating One Star
Hotel Price Cheap Hotel
Type of Room One Double Bed
Public Amenity ☐ Fishing ☑ Bowling
☐ Archery ☐ Casino
☐ Fitness Center ☐ Golf
☑ Swimming Pool
Private Amenity ☑ Internet ☐ Balcony Side
☐ Heater ☐ Beach Side
☐ King Bed ☑ Kitchenate
☐ Refrigerator ☐ Big Tub
Please Select more Services you Want Airport Pick Up
Submit

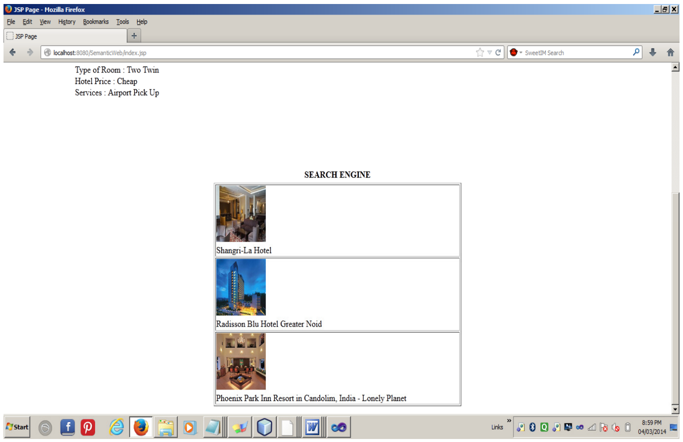
<a> Select the requirements


JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8080/SemanticWeb/index.jsp
semanticweb The Voice of Semantic Web Technology and Linked Data Business

Continent : Asia
Country : India
Star Rating : Five
Facility : Bowling Swimming Pool
Facility1 :
Internet Kitchenate
Type of Room : Two Twin
Hotel Price : Cheap
Services : Airport Pick Up

SEARCH ENGINE

In the figure:4 the user will give the query to search the hotels in countries of rating one star and two star and so on...user will search for the facilities regarding to his requirements. Search engine will search the list of all that hotels having one star in goa according to the users requirements with images of that hotels with ranking because when the next time the user will search for website hotel ,then the all images will be stored in the data base.


JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8080/SemanticWeb/index.jsp

Type of Room : Two Twin
Hotel Price : Cheap
Services : Airport Pick Up

SEARCH ENGINE

Shangri-La Hotel

Radisson Blu Hotel Greater Noid

Phoenix Park Inn Resort in Candolim, India - Lonely Planet

V. Conclusion & Future Work

We have presented a framework for understanding ontology applications using Jena API, and used it to highlight the many similarities between work being done in different areas.Phase I is dedicated to the study of the research work published by different view and guideline of my work to follow to develop the agent for searching.In Phase II the steps are developed for the creations of the agent which will access the web and RDF Files and created the database for the personalized access. The last Phase is the most important phase it is with the implementation of the agent. Phase III is concerned about developing system for taking inputs from the users and comparing them and showing them in useful forms.

Content to be stored in Ontology:

- Unique Identification Number used for fetching Hotel Image from Database
- Hotel Website URL
- Location (same as above 4 things)
- Ratings (same as above)
- Type of Room= List of types of room available in that hotel......if the option 3. selected by user is there in this list then this hotel image will be displayed.

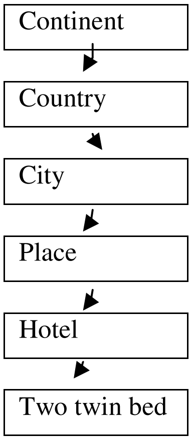


FIG: A Simple Concept hierarchy of Hotel

Acknowledgment

The research presented in this paper would not have been possible without the support of my college. I wish to express my gratitude to people who helped turn the World-Wide Web into the useful and popular distributed hypertext it is. I also wish to thank the anonymous reviewers for their valuable suggestions.

References

[1] Raman Kumar Goyal1, Vikas Gupta2, Vipul Sharma3, Pardeep Mittal4, —Ontology Based Web Retrieval, 1Lecturer (Information Technology), International Journal of Computer Sciences and Engineering RIEIT, Railmajra, 2AP (CSE), RIEIT, Railmajra,3Student, UIET, Panjab University, Chandigarh, 4AP (CSE), BFCET, Bathinda.

[2] Felix Van de Maele, —Ontology-Based Crawler for the Semantic Web, Faculty of Science, Department of Applied Computer Science, Vrije Universiteit Brussel, May 2006.

[3] Subhendu Kumar Pani,Deepak Mohapatra,Bikram Keshari Ratha," Integration of Web Mining Web Crawler Relevance &State of Art",Vol.02,No.3,2010,772-776.

[4] Jan Paralic, Ivan Kostial, —Ontology Based Information Retrieval‖, Department of Cybernetics and AI, Technical University of Kosice, Letna 9, 040 11 Kosice, Slovakia.

[5] Sriram Raghavan,Hector GarciaMolina, —Crawling the Hidden Web‖, Computer Science Department, Stanford University, USA.

[6] Chang Su, Yang Gao, Jianmei Yang, Bin Luo —An Efficient Adaptive Focussed Crawler Based on Ontology Learning, Proceedings of the Fifth International Conference on Hybrid Intelligence Systems- 2005 IEEE.

[7] Debajyoti, Arup Biswas, Sukanta —A New Approach to Design Domain Specific Ontology Based Web Crawler, 10th InternationalConference on Information Technology – 2007 IEEE. Ringe et. al.

[8] Ganesh S, Jayaraj M, Aghila G —Ontology Based Web Crawler‖ Information Technology;Coding & Computing, 2004 volume 2, 2004 page (s) -337-341-IEEE.

[9] Yuan X, H Macgregor and J. Harms, "An efficient scheme to remove crawler traffic from the internet." Proceedings of the 11th International Conference on Computer Communications and Networks, Oct 2002. 14-16, IEEE CS Press, (pp: 90-95).

[10] Kai Song, Yonghong Tian, Tiejun Huang, Wen Gao, "Diversifying theImage Retrieval Results".