International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 4, Issue 7 (December 2014) PP: 15-19

Efficient Mining Technique for High Rank Web Pages

Neha Singh[1], Bharat Bhushan Agarwal[2]
Computer Science Department, IFTM University, Moradabad

ABSTRACT: *The Web crawler is a computer program which downloads the data or information from World Wide Web for the search engine. Web information is updated or changed rapidly without any information or notice. The Web crawler searches web for updated or new information. Approximate 40 % of web traffic is by the web crawler. In this paper the web or network traffic solution has been proposed to get the relevant information. This paper will be implement Ontology Based Topic Specific 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.*

I. Introduction

The Web crawler is a computer program which downloads the data or information from World Wide Web for the search engine. Web information is updated or changed rapidly without any information or notice.The Web crawler searches the web for updated or new information. Approximate 40 % of web traffic is by the web crawler. In this paper the web or network traffic solution has been proposed to get the relevant information. This paper will be implement Ontology Based Topic Specific 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. IAs Topic based search is a search interface paradigm based on a long running library tradition of faceted the classification and efficient search systems have proved the paradigm both powerful and intuitive for end – users, particularly in drafting complex queries. Thus, topic – based search presents a promising direction for the semantic search interface design, if this can be successfully combined with Semantic Web Technologies. Topic based search engines differs from traditional search engines such as the Google, Yahoo! Or MSN, only in information this aggregates, the index and use to answer the users queries, Instead of using the human readable documents such as HTML, PDF or DOC, Topic – based search engines will use semantic web documents (RDF, XML, OWL).

University and the age below 60". These queries are not built using natural langue, but with an easy to use a user interface that help users to build the queries they want. Different person can give this query in their own wordings

1.2 AN ONTOGOLY

In a widely-quoted definition, an ontology is a specification of the conceptualization [Gruber T. 1993], Let's unpack that brief characterization a bit. An ontology allows the programmer to specify in an open, meaningful, way, the concepts and the relationships that collectively characterize some domain of the interest. An ontology is a model of the world, represented as the tangled tree of a linked concepts. Concepts are the language-independent abstract entities, not words. They are expressed in this ontology using English words and phrases only as a simplifying convention. That is, whereas machines wouldn't care if concepts were referenced by, say, numbers – to make them look really language independent – such a naming convention would make the ontology completely opaque to the people who have to build and work with it. So we use (quasi-) English names for concepts and, both in the ontology and in all our writing about the ontology, use capital letters to distinguish concepts, like DOG, from words in a given language, like English "dog" or French "chien".

The purpose of the Ontological Semantic ontology is to improve automated text processing by providing language-independent, meaning-based representations of concepts in the world. The ontology shows how concepts are related (e.g., DOG and CAT are closely related, both being MAMMALs) and what properties each has (e.g., both CAT and DOG have FUR and a TAIL, but CAT can be the AGENT of HISS, whereas DOG can be the AGENT of BARK).

Unlike words in a language, each ontological concept is unambiguous: i.e., it has exactly one meaning. For example, the concept TABLE refers to a flat horizontal surface with legs and has properties including what it's typically made of (WOOD, METAL, etc.), where it's typically located (in a BUILDING or ROOM), etc.

1.3 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 the ontology application development, independent of that 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 the OWL class or the RDFS class. To represent the differences between various representations each of ontology languages has profile, that lists the permitted constructs and the names of the classes and properties.

1.3.1 RDFS

RDFS is the weakest ontology language supported by the Jena. RDFS allows the ontologist to build a simple hierarchy of the concepts, and the hierarchy of properties. Consider the following trivial characterization

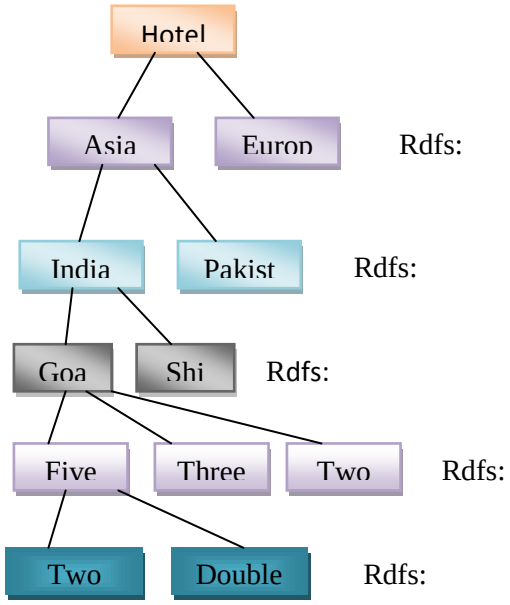


Table 1: A simple concept hierarchy

Using RDFS, we can say that my ontology has five classes, and that Asia is a sub-class of Hotel and so on. So every animal is also the organism. The good way to think of these classes is as describing sets of individuals: Hotel is intended to describe a set of Hotels, some of that are in Continent Asia (i.e. a sub-set of the set of Hotels is the set of continents then each continent has countries.), and some hotels are in Europe, Australia etc.

We can describe this simple hierarchy with RDFS. In RDFS, we can only name classes, we can not construct expressions to describe the interesting classes,However, for many applications this is the sufficient to state the basic vocabulary, and RDFS is perfectly well suited to this.

Note also that we can both describe the classes, in the general terms, and we can describe the particular instances of those classes. So there may be a particular hotel in Asia (i.e. has rdfs: continent “Asia”), and has hotels in India ,Pakistan etc. For countries we are using subclass rdfs: country (i.e. has rdfs: country “India” or rdfs: country “Pakistan”) further each country has states in rdfs: state i.e rdfs: state “Goa”.

1.4 Proposed System

Here in this proposed system we design and develop a semantic web architecture that can relieve the users from the overburden of doing a lot of keyword based search before getting the desired result. This system takes the natural language user query in the form of topics in a user friendly environment. In this proposed system the topic based web interface has to be developed for accomplishing the goal of semantic browsing in a highly heterogeneous semantic web environment. The proposed system exhibits the refinement with higher accuracy and in automatic way. Unlike the existing systems in this research work in order to deal with ontology a common interface has to be designed.

The proposed system architecture has to be divided into 2 different modules: First module takes the user query in the form of topic description and the second module provides a mechanism for the user to enhance search results by providing various image filtering tools. But for the system to understand user query and give best result, it should be trained first. RDF file structure used in the work :

II. Ontology Structure

2.1 RDF in Proposed System

Here in this proposed system we design and develop a semantic web architecture that can relieve the users from the overburden of doing a lot of keyword based search before getting the desired result. This system takes the natural language user query in the form of topics in a user friendly environment. In this proposed system the topic based web interface has to be developed for accomplishing the goal of semantic browsing in a highly heterogeneous semantic web environment. The proposed system exhibits the refinement with higher accuracy and in automatic way. Unlike the existing systems in this research work in order to deal with ontology a common interface has to be designed.

The proposed system architecture has to be divided into 2 different modules: First module takes the user query in the form of topic description and the second module provides a mechanism for the user to enhance search results by providing various image filtering tools. But for the system to understand user query and give best result, it should be trained first. The SPARQL query used in the work:

2.2 Input Output Architecture For Proposed System

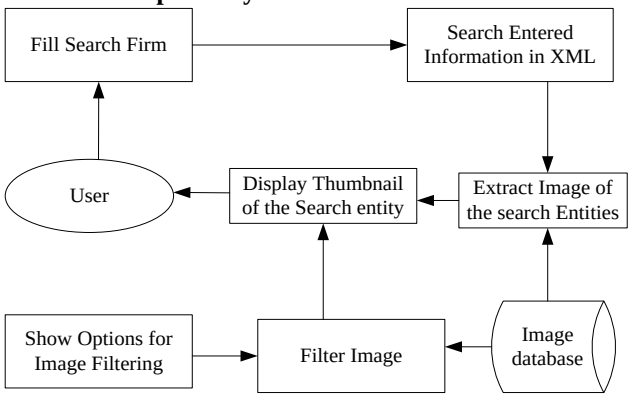


Figure 2.2.1 Input/ Output Architecture

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 the various categories. These images are passed into the algorithm which extracts various metadata of image such as the 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 the database. The category of the image is identified manually and this can be anything like age, place where he works, location etc.

B. Structure of Topic based search form

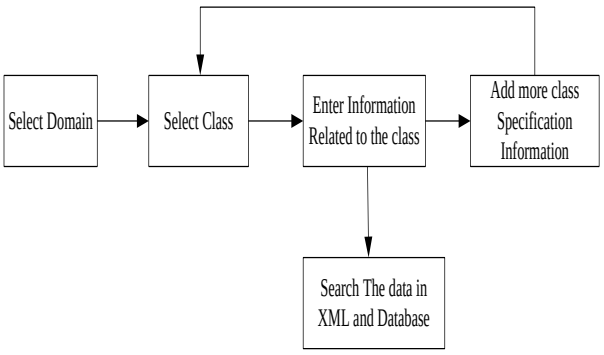
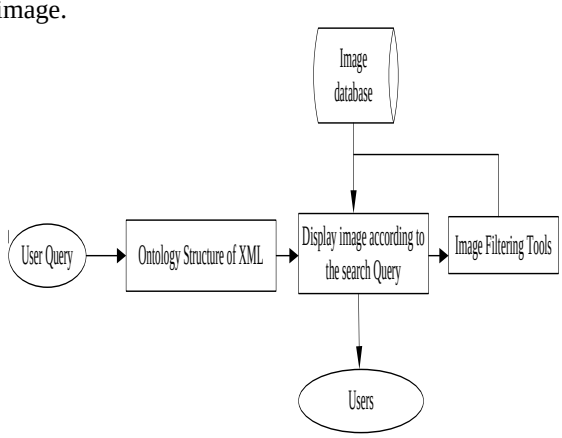


Figure 2.2.2 Structure of Topic based search form

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 below depicts the metadata extraction from the image.



III. Results & Discussion

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 authors and creating the 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

- The 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.
- Facilities it offers = List of facilities like SPA, SPORTS, INTERNET, TV, AC, NON-AC.....if ANY of the options selected by user are there in this list then display the hotel image.

Contents stored in Database:-

- Unique Identification Number (Primary Key)
- Hotel Image or its path.
- Rank (You can decide if you want to keep this factor in Ontology, fact is that later you may need to update it).

IV. Future Scope

HOTEL ONTOLOGY structure is developed using RDF.RDF offers a framework for the construction of logical languages for collaboration in the semantic web. It is an XML based language representing exchange of data. It is providing information on the meaning of the information. In the RDF a documents makes assumptions that specific things have properties which have values. So, in a certain way it is a general mechanism for knowledge representation. The definition of the mechanism is domain-neutral: the semantics of the specific domains is not fixed, but a mechanism is usable for the description of the information from any domain.

RDF is a scheme language. The RDF document has a pointer to its RDF the scheme at the top. This is the list of the terms of data that are used in document. Anybody can make the new scheme document.

Semantic Web approach can be used in a variety of the application domains. this can be used:

- In resource discovery to provide better search engine capabilities.
- In cataloguing for describing the content and content available at the particular web site, page or the digital library.
- by intelligent software agents to facilitate knowledge sharing and exchange.
- In content rating as used in the work ror rating the hotels on clicks.
- In describing collections of pages that represent a single logical 'document'.
- For describing the intellectual property rights of web pages.
- For expressing the privacy preferences of a user as well as the privacy policies of a web site.

Ontologies are the building blocks for the semantic web. Ontologies can play an important part on the web as they allow the processing, sharing and re-use of knowledge between programmes. An ontology is a classification system for concepts and their underlying connections within a specific domain of knowledge as we have used Hotel domain in this work. It is a kind of proto-theory, indicating which elements exist within a specific domain and how these elements can be related to each other. They support the integration of heterogeneous and distributed information resources

REFRENCES

[1]. "XML and Semantic Web W3C the Standards Timeline". 2012-02-04.
[2]. "W3C Semantic Web Activity". World Wide Web Consortium (W3C). November 7, 2011. Retrieved November 26, 2011.
[3]. Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). "The Semantic Web".The Scientific American Magazine. Retrieved March 26, 2008.
[4]. Lee Feigenbaum (May 1, 2007). "The Semantic Web in Action". Scientific American. Retrieved February 24, 2010.
[5]. Berners-Lee, Tim (May 1, 2001). "The Semantic Web". Scientific American. Retrieved March 13, 2008.
[6]. Nigel Shadbolt, Wendy Hall, Tim Berners-Lee (2006). "The Semantic Web Revisited". IEEE Intelligent Systems. Retrieved April 13, 2007.
[7]. Allan M. Collins; M. R. Quillian (1969). "Retrieval time from semantic memory". Journal of verbal learning and verbal behavior 8 (2): 240–247. doi:10.1016/S0022-5371(69)80069-1.
[8]. Allan M. Collins, A; M. Ross Quillian (1970). "Does category size affect categorization time?". Journal of verbal learning and verbal behavior 9 (4): 432–438. doi:10.1016/S0022-5371(70)80084-6.
[9]. Allan M. Collins, Allan M.; Elizabeth F. Loftus (1975). "A spreading-activation theory of semantic processing". Psychological Review 82 (6): 407–428. doi:10.1037/0033-295X.82.6.407.
[10]. Quillian, MR (1967). "Word concepts. A theory and simulation of some basic semantic capabilities". Behavioral Science 12 (5): 41430. doi:10.1002/bs.3830120511.PMID 6059773.
[11]. Semantic memory |book:Marvin Minsky (editor): Semantic information processing, MIT Press, Cambridge, Mass. 1988.
[12]. Berners-Lee, Tim; Fischetti, Mark (1999). Weaving the Web. HarperSanFrancisco. chapter 12. ISBN 978-0-06-251587-2.
[13]. Gerber, AJ, Barnard, A & Van der Merwe, Alta (2006), "A Semantic Web Status Model, Integrated Design & Process Technology", Special Issue: IDPT 2006
[14]. Gerber, Aurona; Van der Merwe, Alta; Barnard, Andries; (2008), "A Functional Semantic Web architecture", European Semantic Web Conference 2008, ESWC'08, Tenerife, June 2008.
[15]. Artem Chebotko and Shiyong Lu, "Querying the Semantic Web: An Efficient Approach Using Relational Databases", LAP Lambert Academic Publishing, ISBN 978-3-8383-0264-5, 2009.
[16]. Victoria Shannon (June 26, 2006). "A 'more revolutionary' Web". International Herald Tribune. Retrieved May 24, 2006.
[17]. Introducing The Concept of Web 3.0
[18]. Lukasiewicz, Thomas; Umberto Straccia. "Managing uncertainty and vagueness in description logics for the Semantic Web".
[19]. Semantic Web Standards published by the W3C
[20]. "OWL Web Ontology Language Overview". World Wide Web Consortium (W3C). February 10, 2004. Retrieved November 26, 2011.
[21]. "RDF tutorial". Dr. Leslie Sikos. Retrieved 2011-07-05.
[22]. "Resource Description Framework (RDF)". World Wide Web Consortium.
[23]. "Standard websites". Dr. Leslie Sikos. Retrieved 2011-07-05.
[24]. Allemang, D., Hendler, J. (2011). "RDF –The basis of the Semantic Web. In: Semantic Web for the Working Ontologist (2nd Ed.)". Morgan Kaufmann. doi:10.1016/B978-0-12-385965-5.10003-2.
[25]. Raman Kumar Goyal1, Vikas Gupta2, Vipul Sharma3, Pardeep Mittal4, ―Ontology Based Web Retrieval, 1Lecturer (Information Technology), RIEIT, Railmajra, 2AP (CSE),