Volume 2, Issue 6 June 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

IRJIF IMPACT FACTOR: 3.01

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi. All research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.01

SPECIFIC RANKING USING SEMANTIC WEB FOR INCREASING THE EFFICIENCY OF THE WEB CRAWLER: REVIEW

MOHIUDDIN ANSARI & ASST. PROF. MR. YATIN CHOPRA

CBS Group of Institutions, Maharshi Dayanand University, Haryana

ABSTRACT:

A Web Crawler is an internet bot that downloads data from World Wide Web for search engine and Indexing. Web information is constantly changing and is updated without prior notice. The Web Crawler checks the World Wide Web for the updated information. People visiting the website frequently are actually the familiar websites and this makes it high ranked website. In this paper the network traffic solution is used to get desired information. This paper will be actualize Ontology Based Topic Specific Search Using Semantic Web. The strategy for web crawling with filters is utilized. It is a query based approach with Jena API. The proposed approach takes care of the issue of re-visiting web pages by crawler. The Semantic Web is an extended version of the present Web that permits the meaning of data to be decisively described regarding all around characterized vocabularies that are comprehended by individuals and computers. As Topic based pursuit is an inquiry interface paradigm taking into account a long running library tradition of faceted classification. Furthermore, effective search systems frameworks have proven that they are both capable and instinctive for end – users, especially in drafting complex queries. In this way, topic – based search shows a promising path for semantic searching interface design only if it gets successfully combined with Semantic Web Technologies. Topic based web search-engines is different from other search engines like Google, MSN/BING, Yahoo! as it only integrates information, indexes it and answers the queries of user.

Keywords: ***Web Crawler, Semantic Web, search engine, ontology.***

1. INTRODUCTION

Keyword or Topic-Based-Search is helpful particularly to a client who realizes what keywords are utilized to list the image and in this way can do the formulation easily and quickly. This methodology is problematic, in any case, when the client does not have a clear objective at the top of the priority list, does not comprehend what there

is in the database, and what sort of semantic concepts are included in the area. The target is to make a Topic-Based-Semantic web search tool which is highly user-friendly that helps in providing advanced search options with topics. A client shouldn't know about the concepts supporting the semantic web to utilize it. The user must feel that whatever they are searching on this semantic web right now is similar to what they do normally with search engines daily.

2. OBJECTIVES

The objective of the work is about providing information about the hotel domain. The objectives are as follows:

- Using SPARQL providing hotel relevant information.
- Hotel Ranking
- Representation of hotels using Knowledge-Base

SPARQL provides better performance than SQL. Hence SPARQL is used. Here Hotel Ontology is developed using RDF since RDF is a scheme language as it has pointer at the top of the document in which RDF scheme is being used. Anyone can easily make a new scheme document. General representation is for Knowledge Representation as in RDF, a document assumes that specific things have properties that has values.

3. LITERATURE REVIEW

Gateway of Semantic Web is to provide access to ontologies and Sematic Data. Hence it has 3 main roles:

- Collection of Semantic Content from web that are available.
- Extract metadata and index which can be useful
- Implementing the query access to the data.

Raman Kumar, Goyal Vikas Gupta, Vipul Sharma, Pradeep Mittal in June 2015 described **"Tourism Ontology and Semantic Management System: State of the Arts Analysis"** stating the global importance of steadily rising tourism creating a new opportunity in many countries. The solution for information management for complex tasks of tourism are still at an early stage from a semantic point of view. This paper aims applying, evaluating and concretizing semantic web technologies such as ontologies, semantic search to information rich tourism domain and semantic annotation of contents. Identification of seven tourism ontologies are suitable as a base for creating problem specific ontologies.

P. Jourlin, R. Deveaud, E. Sanjuan-Ibekwe, in the year 2012 has described a web crawler based on GNU/Linux and Postgre SQL which is a novel, focusable, scalable and distributed web crawler. They have released a GNU public license. The report shows the use case related to the analysis of Twitter's stream about the presidential election of French in 2012 in its URL.

SA Patel and JM Patel in the year 2012 has described web crawler as an intelligent agent in which pages available in internet are growing tremendously day by day and in this case searching the required information in the internet becomes very hard work. Generally, the searching of information in WWW can be done by searching the list of the link.

SS Vishwakarma, A Jain, in the year 2012, described a **web crawler algorithm with query based approach with increasing efficiency**. In his paper the network traffic solution is proposed in which approximately 40% of the web traffic is by web crawler and the web crawling filter is used which is query based approach. The approach actually solves the problem of revisiting web pages by a web crawler.

Marc Najork has described Web Crawler Architecture in which a web crawler, being an important component of search engines, is given one or more seed URLs, download the web pages associated with that URLs. It also extracts the hyperlinks and recursively continues to download the pages by that identified hyperlinks. They are also used in many applications like data mining, shopping engines, indexing of web pages within search engines and so on. Their two main data structures, one which is set to be crawled URLs and second the content providers without overloading any particular web server.

A. Agarwal, D. Singh, A. Kedia, A. Pandey, V Goelin the year 2012 described **a design of a parallels migrating web crawler** that poses certain drawbacks such as generations of large amount of redundant data and wastage of bandwidth of the network due to transmission of such unwanted data. In order to overcome these drawbacks with traditional crawler techniques, they have proposed a parallel migrating web crawler and presented with detailed requirements along with the Crawler Architecture.

N. Singhal, A. Dixit, RP Aggarwal in the year 2012 described**" Regulating Frequency of Migrating Web Crawler based on Users Internet"** shows that due to the lack of efficient techniques, the crawlers add up unwanted traffic to already overloaded Internet. Optimization of frequency of visiting sites can be done by calculating the refresh time. This helps in optimizing the effectiveness of the crawling system by managing the

revisiting frequency. An alternate approach for the optimization of frequency of visiting websites can be done based on user's interest.

RK Rana and N. Tyagi in the year 2012 described **"A Novel Architecture of Ontology based Semantic Web Crawler"**. They presented that searching meaningful information from billions of resources of data on web is difficult task by looking at the growing popularity of internet. The future of WWW is semantic web where ontologism are used to give a valid meaning to the web content. On Web Semantic the data will be linked to different ontologies and processing of information becomes very hard without proper knowledge of sematic mappings between different ontologies. The Architecture can exploit the semantic Meta data to discover and take out information from semantic web.

D. Khurana and S. Kumar described **"An Improved Approach for Captain Based Image Web Crawler"** in which the WWW is a global, read write information space. All the information in the form of text documents, images, multimedia and other sort of information are targeted to be resources by short and unique identifiers knows as Uniform Resource Identifiers so that it can be easily found and accessed in the simplest way. It becomes a very big reservoir of information providing unrestricted access to a Hugh unending reservoir of information present in the form of hypertext markup language. These documents contains hyperlinks to each other documents.

Web Crawler Design Issues: A Review present that the big and dynamic nature of web increases the requirement for updating for web based information for retrieving desired information from the system. Crawlers encourage the procedure by taking after the hyperlinks in Web pages to automatically download a partial preview of the Web. While a few systems depend on crawlers that comprehensively crawl the Web, others concentrate on subject particular accumulations. In present paper the different sorts of crawlers are talked about. The paper likewise talks about a few web crawler design issues alongside their answers.

4. ONTOLOGY

An ontology is a model of the global world, displayed as a tangled tree of associated concepts. Concepts are language-independent abstract entities, not just words. These are portrayed in this ontology using English content only as a simplifying convention. Whereas machines wouldn't care if concepts were referenced by, say, numbers - to make sure they are language independent - such a naming convention would make the ontology completely opaque to the individuals who have to create and use it. So we use English names for ideas and, both in the

ontology and in every writings of the ontology, use capital letters/ words to distinguish ideas, like DOG, from words in confirmed language, like English "dog" or French "chien". The main purpose of Semantic Ontology is to make use of automated text-processing by using knowledge based representation of abstracts of the world. Ontology displays how concepts are related like DOG and CAT are somewhat closely related as they are MAMALs and properties of each has TAIL, FUR, but CAT could be the AGENT of HISSING where DOG could be AGENT of BARKING. We can take an example of TABLE referring a horizontal flat surface with 4 legs and being made of WOOD/METAL and is located at BUILDING/ROOM and so on.

5. MATERIALS AND METHODS

Jena is a programming tool which is used in this proposed system with the help of Java programming language. With Ontology API, Jena aims to give a consistent interface for programming for ontology application programming that is independent of ontology language used in our programs. Jena API for Ontology is language neutral. The java class names used are actually not specific to the language used. For example, the java class name 'OntClass' represents 'OWL" class or RDFS class. We need to have a required language to design ontologies. In this proposed system of RDFS, the Jena language is used which is supported by ontology language. RDFS helps the oncologists to make simper hierarchy concepts and similar properties.

SPARQL is pronounced as "sparkle" is an acronym for SPARQL protocol and RDF query language, is also used in this proposed system. RDF is query languages that are used for databases. SPARQL query consists of triple pattern - conjunction, disjunction and optional pattern. It helps in retrieving and manipulating data stored in Resource Description Framework (RDF) format. This format was made standard format by Data Access Working Group (DAWG) of WWW Consortium and is also known as one of the important key technology of semantic web. SPARQL 1.0 became official W3C Recommendation in January 2008 and SPARQL 1.1 became in March 2013.

6. FUTURE SCOPE

This Hotel Ontology Structure is developed using RDF and rdf scheme. RDF provides framework for development of logical languages for coordinated effort in semantic web. It is an XML based language that represents exchanging of information. It is giving data on the meaning of data. In RDF a document assume that a specific thing have properties that has a value. In this way, it is a mechanism for knowledge representation.

Semantic Web approach can be used in:

- Resource Recovery for Search Engine Efficiency:
- For describing Intellectual Property Rights of Websites.
- For Rating Content as per clicks.
- By Intelligent Software's to provide Knowledge Sharing.
- As an inventory for content, contents for specific web pages or digital library.
- In gathering similar web pages as a single logical document.
- In providing privacy policies for users and web pages.

7. CONCLUSION

This work executes the Semantic Web. It is an extended version of the present web in which data is given well defined significance, making computers and individuals to work in collaboration. The Semantic Web gave a typical system that permits information to be shared and reused crosswise over application, enterprise, and community limits. Semantic Web methods will help structuring more and more information and will integrate it at Webpage level. This will help avoiding reinvention of data description. We need to stop worrying about how to address 25 percentage of problem when 75 percentage is already addressed with present Semantic Web Standards. Committing more new standards, 90 percentage of it will be addressed soon. Semantic Web Methodology helps in describing different types of information. When the pieces of information are defined, system will describe even more. Initially it will help us describing ourselves which is critical to personalize. Further it will help in determining places and things that are linked in different.

8. REFERENCES

[1] "Ontology Based Web Retrieval", Raman Kumar Goyal, Vikas Gupta, Vipul Sharma, Pardeep Mittal, UIET, Panjab University, Chandigarh, 4AP (CSE), BFCET, Bathinda, June 2015.

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

[3] "Ontology Focused Crawling of Web Documents"--Marc Ehrig, Alexander Maedche

[4] "Ontology Based Information Retrieval"--Department of Cybernetics and AI, Technical University of Kosice, --by Jan Paralic, Ivan Kostial, Slovakia.

[5] "Crawling the Hidden Web"-by Sriram Raghavan, Hector Garcia Molina, --Computer Science Department, Stanford University, USA.

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

[7] "A New Approach to Design Domain Specific Ontology Based Web Crawler", --By Debajyoti, Arup Biswas, Sukanta, 10th International Conference on Information Technology, -- 2007 IEEE. Ringeet. al.

[8] "Ontology Based Web Crawler"--By Ganesh S, Jayaraj M., Aghila G., --Information Technology-Coding & Computing, 2004 volume 2, 337-341-IEEE Sept 2012.

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

[10] “The Ethicality of Web Crawlers”-By Sun. Y, Council G. Isaac and Giles C. Lee, in the proceedings of 2010 "IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology"-Toronto Canada Aug-2010.(pp: 668-675)

[11] 'Web Crawler' From "Wikipedia"- Link: http://en.wikipedia.org/wiki/Web_crawler

[12] 'World Wide Web' From "Wikipedia"- Link: http://en.wikipedia.org/wiki/World_Wide_Web

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com