A Simple Semantic Web Crawler for Intelligent Information Retrieval from Academic Websites

V.Kiran Kumar, Ramya

Abstract**: *In various applications data is shared and reused through a common framework like Semantic Web. In essence, in the ever-expanding sphere of the web, huge quantities of web content is created and made available for men and machines for their interpretation and application. In the present paper, an attempt has been made to develop a "scutter", other wisely known as semantic crawler that essentially collects and stores information in a centrally-located database by crawling through the semantic content. The projected 'scutter' is based on Jena 3.0 Framework, which is a freely downloadable language software available on https://jena.apache.org/download/. In this process, a RDF file is taken as the seed input file after which the 'scutter' accesses other RDF documents by crawling through 'rdfs:seeAlso' property, thereby designated as an automatic extraction of semantic information emanating from various websites. Also, certain privacy related issues, especially in FOAF metadata is discussed in this paper.***

Index Terms**: *Semantic Web, RDF, RDFS, Semantic.***

I. INTRODUCTION

The vast number of tools for preparing the web content has enabled users to contribute instant web information, thus leading to accelerated growth in Web sphere.

The immense success the web enjoys can be attributed to the freedom it endows its users. However, the created host of web content has a predefined purpose of serving the intellectual and information needs of human beings, not fit for any sort of processing by machines.

For instance, by using a search engine, we retrieve web content, without human connect, it is not possible to derive any effective result.

In the advanced future generation of web applications, the attention is to generate web content that can be used by humans, which is also understood and interpreted by machines.

Such a futuristic version of the web is popularly referred to as future generation web or Semantic Web.

As defined by W3C, "The Semantic Web, with its a common framework, allows data to be shared and reused across application, enterprise, and community boundaries."8 Tim Berners-Lee coined the term semantic web to refer to a web of data that machines find fit to process9.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author

Dr V.Kiran Kumar*, Associate Professor, Department of Computer Science, Dravidian University, Kuppam, Andhra Pradesh, India. kirankumar.v@rediffmail.com
Mrs Ramya, Research Scholar, Department of Computer Science, Dravidian University, Kuppam. Andhra Pradesh, India.

In the remaining parts of this paper, the various sections are organized as hereunder: in section 2, an introduction to Semantic Web is presented; Section 3 presents in detail Unified Resource Identifiers as well as their significance. A brief introduction to RDF (Resource Description Framework) is provided in Section 4 with a detailed description about Scutter Algorithm and the linkages between RDF files through the rdfs:seeAlso property. In section 5, "Scutter" algorithm and its process of linking rdf files together, along with extraction process and storage in central database is discussed.

A. Introduction to Semantic Web

Berners-Lee, Tim; James Hendler and Ora Lassila (2008) considered the Semantic Web as the technology of the future with potential for transforming 'documents' on the World Wide Web (WWW) as Knowledge that is fit for machine-based processing. The W3C with its leader in web inventor Tim Berners-Lee defined the standards as guidelines for making assessment of website quality through the presentation of the web content. Through web mining, various aspects of websites are investigated. The World Wide Web Consortium (W3C) defined the Semantic Web as "a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". The Tim Berners-Lee invented term referred to a web of data fit for processing by machines. In figure 1, the Semantic Layer alongside various technologies useful in the design of Semantic Web applications is exhibited. Such languages primarily represent information that machines understand which help in maintaining interoperability among applications on the existing web. When proper meaning is added to the existing website, semantic web applications can be designed for use. All resources on the web having a unique name is represented by Uniform Resource Identifier (URI). RDF, RDFS and OWL are some of the key technologies that fall under this domain. RDF is used to represent simple facts. RDFS gives little bit semantics comparing with RDF.

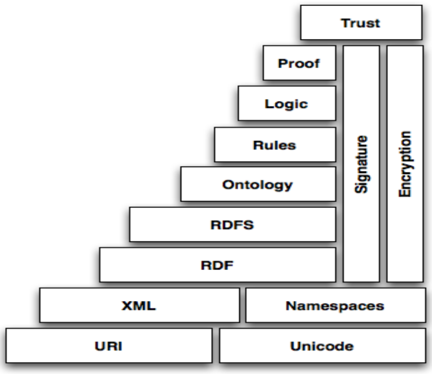

Figure 1: Semantic Layer Cake.

II. URI REF AND NAMESPACES

A Uniform Resource Identifier (URI) represents a character string used to identify an abstract (or) physical available resource on the web. A URI with an integrated optional fragment identifier is referred to as the URI reference (URI ref), with character "#" as the precursor. For instance, in the given string a URI ref is [], while "http://www.dru.ac.in" is the designated URI. "All_Genders" as a fragment identifier is used only after "#". *http://www.dru.ac.in#All_Genders*

A collection of names is referred to as name space which is represented by URI ref. The qualified names (Qnames) represent the names from namespaces that include 'p:u', wherein 'p' refers to the namespace prefix, while the local part is represented by 'u'. For instance, 'rdf: description' is a qualified name in which 'rdf' represents namespace prefix, while the local part is 'description'.

III. TRIPLES

The statements in RDF formats are composed of a subject, a predicate (property) as well as an object.

A triple refers to the link between the

subject->Predicate->Object,

wherein an RDF URI reference or a blank node can be designated as subject. A RDF URI reference can be designated as the predicate, while an RDF URI reference or a literal or a blank node can be selected as an object. This triple format may be used to represent any statement, for instance,

"Tim Berners-Lee invented World Wide Web" ,

wherein

◦ Tim Berners-Lee is the subject
◦ invented is the predicate
◦ World Wide Web represents an object..

IV. RDF (RESOURCE DESCRIPTION FRAMEWORK)

As the World Wide Web (WWW) had the original purpose to create content for the use of human beings with all its content machine-readable. However, no data is machine-understandable except few meta data available on the web. The use of metadata (data about data) provides the required solution that is fit to be processed by machines. The metadata about web resources is largely supported by Resource Description Framework (RDF) with automatic processing by applications, more than as shown by people. The RDF as W3C recommendation was released by the World Wide Web Consortium in In February 2004. Trastour, David, Claudio Bartolini, and Chris Preist (2002) opined that as a general method RDF is effective in decomposing knowledge into minuscule parts while justified on the basis of some rules of semantics. Figure 2 presents RDF graph for the above statement in the form of a graph model as shown below in Figure 2.

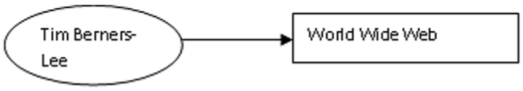

Figure 2: RDF Graph.

V. INTRODUCTION TO SCUTTER

In the example given below, the creation of RDF file having properties such as vice-chancellor, registrar and rector has been explained:

```
<?xml version="1.0"?>
<rdf:RDF
          xmlns:rdf="http://www.w3.org/1999/02/22-
                     rdf-syntax-ns#"
           xmlns:pers="http://dravidianuniversity.ac.in#"
           xml:base="http://dravidianuniversity.ac.in#">

<rdf:Description rdf:about="#DravidianUniversity">
     <pers:vice-chcancellor>G.Lokanatha
                Reddy</pers:vice-chcancellor>
     <pers:registrar>T.Anuradha</pers:registrar>
     <pers:rector>Nil</pers:rector>
</rdf:Description>

</rdf:RDF>
```

The example explains about Dravidian University with "G.Lokanatha Reddy" as Vice-Chancellor, "T.Anuradha" as Registrar and "Nil" as Rector.

A. Publishing RDF files into HTML

In this section, the linking of RDF files with the web is enumerated in detail. The linkage between a document and external source is defined by the <link> tag, wherein an RDF file is considered as an external source.

```
<html>
<head>
….
<link rel=”meta” type=”application/rdf+xml”
title=”Univ-info”
href=”http://www.dravidianunivesity.ac.in/du.rdf”/>
….
</head>
<body>
….
</body>
</html>
```

As seen above, our RDF document is indicated in our homepage through href, while the crawler takes some time to reach the homepage to locate the RDF document. The crawler is referred to as scutter in the parlance of RDF terminology with the primary task of visiting the homepage for discovery of RDF files. As soon as the RDF document is located, the document is parsed, before being stored into its centralized data system as the triples for usage later.

B. Linking one RDF file to another RDF file

<?xml version="1.0"?> The key property for linking two RDF files can be formalized through rdfs:seeAlso Property. In the instance given below,

the process of interlinking of two RDF files viz., du.rdf and sv.rdf is presented;

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
     syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-
     schema#"
   xmlns:pers="http://dravidianuniversity.ac.in/adm#"
          xml:base="http://dravidianuniversity.ac.in#">
<rdf:Description rdf:about="#DravidianUniversity">
      <pers:vice-chancellor>G.Lokanatha
                  Reddy</pers:vice-chcancellor>
      <pers:registrar>T.Anuradha</pers:registrar>
      <pers:rector>Nil</pers:rector>
</rdf:Description>
<rdfs:seeAlso
rdf:resource="http://svuniversity.ac.in/sv.rdf"/>

</rdf:RDF>
```

rdfs:seeAlso property works similar to hyperlinks connecting more than one document in HTML. In similar ways, such a property is used for connecting more than one RDF documents. In the instance given earlier, the key RDF documents are joined in hyperlink with RDF document like “sv.rdf”. After noticing such a property, the scutter moves onto rest of the RDF documents.

C. How to get into the Circle?

As shown in the previous sections, scutter resembles a crawler while visiting the webpage for RDF documents, before storing them in a centralized database system. However, the problem lies in the fact that till the visit, a scutter makes to our web page, the waiting continues, which may require more time in case of designing more web pages more often. Alternatively, one may approach such a problem by introducing rdf:seeAlso property to be made part of each RDF file, after which the scutter crawls the page to decipher the last updates. Finally, it connects us also to become indexed.

D. Pulling out of RDF metadata on Web

In this section, the RDF scutter or crawler used to extract RDF properties on web is discussed, which is identical to ‘spider’ in the parlance of web technology. As a software, the ‘scutter’ follows ‘rdfs:seeAlso’ property to ascertain peculiarities in the relationship between two individuals. By applying ‘rdfs:seeAlso’ property, one can connect to one more person in the circle of knowledge. Such a relation is not to be confused with one implying friendship, approval or even a face-to-face meeting having occured4. Semantic web applications use such a property for updating the known circles about others activities. The RDF files are connected together by such a property along with ‘rdfs:seeAlso’ property, which is identical to anchor element in HTML. RDF database can be built by writing applications with such a property. The ‘rdfs:seeAlso’ property is used to write the proposed ‘scutter’ for collecting data on seed link, prior to its storage in a local file or database. The initial point of the link at wherein the process begins is referred to as the seed link.

The projected ‘scutter’ is based on Jena 3.0 Framework, which is a freely downloadable language software available on https://jena.apache.org/download/. In this process, a RDF file is taken as the seed input file after which the ‘scutter’ accesses other RDF documents by crawling through ‘rdfs:seeAlso’ property, thereby designated as an automatic extraction of semantic information emanating from various websites.

```
Procedure scutter (start_file)
(uri_vector : a vector to store RDF files)
Begin
1. Insert the first rdf file < start_file > to the vector <uri_vector> to start the program.
2. Remove the first element from <uri_vector> and then start converting it into a Jena Model
3. Extract the contents of statements from Jena Model one by one.
4. if the statement contains a property called ‘rdfs:seeAlso’, then add the value of the property to <uri_vector>.
Else
   Extract the property and its value , then store the contents into a local database.
5. Continue to execute the steps from 2 to 5 until <uri_vector> gets empty.
End
```

The above program is largely dependent upon a certain limit. Based on the limit provided in the program, the output file will be created. The depth is also considered in this program to limit the execution time. Here depth of the program is nothing but the number of links to be traversed. In case an unspecified limit, the scutter run continues without end or at least, till uri_limit reaches a stage of complete emptiness. The presented ‘scutter’ makes use of “http://drav.ac.in/du.rdf” as well as the uri_limit to 40 and depth_limit to 50. After the scutter was made to crawl based on ‘rdfs:seeAlso’ property under the specified uri_limit and depth the RDF classes were generated.

VI. CONCLUSION AND FUTURE WORK

As an uncomplicated Java program, our proposed ‘Scutter’ makes sure that RDF files are crawled from a variety of websites. In such a process, a seed RDF file is selected to initiate the program.

However, the proposed Scutter program is found to be encumbered by a limitation in which each website must have an RDF file with a mandatory linkage in html through rdfs:seeAlso property.

fact, the experimentation of the presented scutter algorithm was performed only on a handful of websites, just to understand the basic process in which data collection from selected university websites occur. After collecting the credentials of the Vice-Chancellor, Registrar and rector from each university, this algorithm dumps it in a centralized database, after which applications are designed. More significantly, in future works, it might become possible to generate more information about other vital details of a university than just the limited information consisting of the names of Vice-Chancellor, Registrar and Rector.

REFERENCES

1. Berners-Lee, T. (2000). Weaving the Web – The Past, Present and Future of the World Wide Web by its Inventor. Texere.
2. Berners-Lee, T., Hendler, J., & Lassila, O. (2001).
3. The Semantic Web. Scientific American Magazine; Retrieved March 26, 2008.
4. Breitman, K., Casanova, M., & Truszkowski, W. (2006). Semantic Web. Concepts. Breitman, K. K., Truszkowski, & Felicissiomo. (2006).
5. The automatic semantic desktop; Helping users copy with information system and complexity. In Proceedings of IEEE International workshop, (pp. 156-162). Brickley, D., & Guha, R. V. (2003).
6. Resource Description Framework (RDF) Schema specification 1.0: RDFSchema. W3C Working Draft. Decker, S., & Frank, M. (2004).
7. The social semantic desktop. In Proceeding of the WWW 2004 Workshop Application Design, Development and Implementation Issues in the Semantic Web. Fikes, R; Horrocks, I. (2003).
8. OWL-QL – A Language for Deductive Query Answering on the Semantic Web. KSL. Halpin, H; Tuffield, M. (2010). A Standards-based, Open and Privacy-aware Social Web.
9. W3C Social Web Incubator Group Report. W3C Incubator Group Report.
10. Kumar, K. (2013). Towards Web 3.0: An Application oriented approach. IOSR Journal of Computer Engineering, 15(5), 50-53. Kumar, K., & Rao, R. (2009a).
11. Semantic Extension of Syntactic table data. International Journal of Systems and Technologies. Kumar, K., & Rao, R. (2009b).
12. TBL2RDF: Html Table To RDF Translator. International Journal of Web Applications.
13. Martinez, O., & Botella, F. (n.d.). Building E-Commerce Web Applications: Agent and Ontology-based Interface Adaptivity. Operations Research Center, University Miguel Hernández of Elche, Avda. Universidad.
14. Sandahl, Z., & Sandahl, K. (2003). Potential advantages of Semantic Web for Internet commerce. In Proceedings of the International Conference on Enterprise Information Systems (ICEIS).
15. Sauermann, L., Sebastian, T., & Linux, M. (2008). Case Study: KDE 4.0 Semantic Desktop Search and Tagging. Academic Press. Sreedhar, G. (2016).
16. Identifying and Evaluating Web Metrics for Assuring the Quality of Web Designing. In Design Solutions for Improving Website Quality and Effectiveness (pp. 1-23).
17. Hershey, PA: Information Science Publishing (an imprint of IGI Global). Trastour, D., Bartolini, C., & Preist, C. (2002).
18. Semantic web support for the business-to-business e-commerce lifecycle. In Proceedings of the 11th international conference on World Wide Web, (pp. 89-98). ACM. doi:10.1145/511446.511458 Williams, H., Li, F., & Whalley, J. (2000).
19. Interoperability and electronic commerce: A new policy framework for evaluating strategic options. Journal of Computer-Mediated Communication, 5(3).

AUTHORS PROFILE

Dr. V. Kiran Kumar, Working as Associate Professor in the Department of Computer Science, Dravidian University, Kuppam, Chittoor Dt., Andhra Pradesh, India. Completed his Ph.D from Acharya Nagarjuna University, Guntur, Andhra Pradesh, and M.Sc., (Computer Science) from Acharya Nagarjuna University, His Major Areas of Academic Interest are Semantic Web, Web Technologies and Programming Languages.