An Outline Implementation for Habitual Production of Ontology from RDF to Medical Data

Yan Zhigang
Inner Mongolia Normal University, China
3762533081@qq.com

Abstract

Information on the web is extremely growing in current years with the quicker velocity. This enormous or capacious data has driven intricate troubles for information recovery and information organization. As the data resides in a network with numerous forms, the Knowledge management in the web is a challenging task. Here the novel ‘Semantic Web’ concept may be used for understanding the web contents by the machine to offer intelligent services in an efficient way with a meaningful knowledge representation. The information recovery in the conventional web source is centered on ‘page positioning’ strategies, though in the semantic web the information recovery forms depend on the ‘idea-based learning’. The proposed work is gone for the improvement of another system for programmed age of cosmology and RDF to some continuous system information, removed from numerous storehouses by following their URL’s and Text Documents. Enhanced altered ordering method is connected for cosmology age and turtle documentation is utilized for RDF documentation. A program is composed for approving the extricated information from different archives by expelling undesirable information and considering just the report segment of the website page.

Keywords: Semantic web, Resource description framework, Ontology, Improved inverted indexing, Knowledge management

1. Introduction

Internet (WWW) is considered as a worldwide data storehouse that recognizes archives and other web assets by Uniform Resource Locators, interlinked by hypertext joins. Web indexes are utilized to recover the data from the web. Information overburden is the most concerning issue in nowadays for the current framework. Advancement of web incorporates the web renditions of web 1.0, 2.0 and so on. In this arrangement, the web rendition 3.0 is alluded to as semantic web [1] is advanced as an information administration bolster over the globe. Web indexes ought to be improved with semantic web abilities that break down site page content and give more important outcomes relating to the client question. Semantic web guidelines incorporate Resource Description Framework (RDF), web metaphysics, RDF Schema and Rule Interchange Format (RIF) for taking care of information. RDF gives a calculated depiction of data for speaking to the web assets like Turtle punctuation, N-Triples and so on. RDF portrays
information on the Web in chart shape [2]. Ontologies comprise of the limited arrangement of terms, connections, limitations and sayings [3]. Ontologies have ended up being valuable for compelling learning displaying and data recovery. The remaining paper is arranged as follows: In Section 2 the related work is presented. The proposed work and its methodology are discussed in Section 3 & 4. The results are presented in Section 5. Comparative analysis is explained in Section 6 and Conclusions are given in Section 7.

2. Literature Review


Amel Grissa Touzi et.al [6] suggested the Fuzzy Ontology of Data Mining (FODM) for processing automated generation of ontologies in the domain of data mining.

Amira Aloui et.al [7] implemented a plugin named “FO-FQ Tab plug-in,” which can be integrated with protégé editor for building the fuzzy ontologies from large databases. To overcome the drawbacks of the existing system for accessing the related science information,


Tahani Alsubait et.al [9] developed the e-learning suite, with the set of questions designed using ontological representation.

A.H.M.Rupasingha et.al [10] suggested that the performance of the ontology generation is always dependent on the specificity of the terms.


3. Proposed Methodology

Semantic web capacities like RDF and metaphysics are connected to advance the learning. The proposed work is a usage of the system proposed by the creators [8]. The system is planned with reference to the semantic web Stack. It is completed in two stages, specifically Data extraction stage and Data portrayal stage. Web scratching is performed utilizing HTML parsing procedure in information extraction stage by giving example seek inquiry as a contribution to numerous stores. DOM parsing and HTML parsing procedures are connected to approve the information recovered from different storehouses by considering just the record area of the site page. Extensible Markup Language (XML) is the base for the semantic web portrayals; the approved data is changed over into semi-organized documentations by utilizing XSD statement from DOM tree and go as a contribution for the following layers of the proposed structure. XML documentation is given as a contribution to information portrayal stage. RDF documentation is produced and spoken to in graphical frame utilizing Graphviz device. A printed portrayal of RDF diagram is given utilizing Turtle, the Terse RDF Triple Language. Enhanced Inverted Indexing system is connected for ontological portrayal of words by barring the stop words.
4. Methodology

Implementation of the framework proposed in Section III will be carried out in two phases namely data extraction and data representation phases. The details are given below.

Phase 1: Data extraction
Data extraction phase is sub-divided into three steps namely web scraping, data validation, XML Conversion.

Step 1: Web Scraping
Web scraping, also be referred as screen scraping or Web harvesting, is used to fetch and extract the data from a web document using HTML parsing techniques. Here Web pages are crawled and the content of the Web page is extracted, reformatted and stored in a database. HTML parsing technique is used for scraping data from the web documents.

Step 2: Data Validation
In the Data Validation Step, the data collected from step 1 is validated using HTML and DOM parsing techniques. Here unwanted data is removed and the necessary portion of URLs is retained. The validated data is stored in a database by considering Document Object Model (DOM). The Data validation process is explained in [Figure 2].

Step 3: RDF-ization
RDF is the basic building block in semantic web, promoting conceptual modeling of web data [13]. It identifies things using Uniform Resource Identifiers (URIs) and describes resources with properties and values. Each tuple in a relational database is decomposed as RDF triples.
An Outline Implementation for Habitual Production of Ontology from RDF To Medical Data

Figure 2. RDF triple

Decomposition of tuple creates a new blank node, corresponding to the row, and a new triple set is obtained. RDF can be viewed as a directed graph with labeled nodes and arcs. The triple is represented as a <subject, predicate, object> format by exploring the relationship among the nodes [14]. A node can be a URI reference, literal or the blank node. The predefined data type used in RDF is “RDF: XML Literal,” for embedding XML in RDF.

5. Result Analysis

A sample search query is given as an input and web scraping results are shown in [Figure 3] as follows,

Figure 3. Web scraping results

Web scraping performance is evaluated by considering the following parameters like database size, number of URL’s extracted, number of keywords, response time, latency time. Scraped data from multiple repositories is given as an input to data validation step. The validated data is obtained as an output to data validation process by applying HTML and DOM parsing technique. Data validation considers only the document section of a web page. The validated data stored in database is converted into the XML notations by applying XSD declaration from DOM tree.
RDF is a recommended standard of World Wide Web Consortium (W3C). RDF generation for sample relation named “testrdf” which has an attribute as <name, description, freq> is considered. The “testrdf” represents the relation’s name is considered as a class in an RDF graph and has set of three nodes that are connecting the testrdf in depth wise manner represents the tuple of a relation. Ontology generation for the data obtained from multiple repositories as well as the text file.
Improved inverted indexing technique is applied for extracting the words with their frequencies discarding the stop words, in the order of highest precedence. The highest frequency word is considered as a frequent search term for the purpose of rule framing using description logic. The rule mapping is done for the efficient retrieval operation which will be future work. RDF is a recommended standard of World Wide Web Consortium (W3C). RDF representation of data in turtle form is shown in [Figure 5].

6. Conclusion

The advancement of information portrayal on the web is taking numerous varieties step by step which prompts the tremendous significance of the data portrayal and information stockpiling to help top of the line data recovery frameworks, Hence the exploration on the semantic web is stimulated and numerous types of research introduced an approach to speak to information utilizing the semantic web. The current framework gathers the information from web crawlers is showing normal execution in recovery and in this manner the proposed system enhances the execution by applying semantic web abilities RDF-ization and metaphysics age which gives quicker recovery. Metaphysics age utilizing devices like Protege, Ontorat, RDF2OWL and so on. is a manual handling of making classes and subclasses. Enhanced rearranged ordering strategy is adjusted for cosmology age. Surmising guideline encircling can be connected to any sort of question given by client and database can be secured by applying proficient cryptography calculation which is done in our future work.

References


This page is empty by intention.