International Journal of Database Theory and Application
Vol.10, No.7 (2017), pp.77-88
http://dx.doi.org/10.14257/ijdta.2017.10.7.07

A Novel Framework for Data Extraction from Multiple Repositories and Generation of Ontologies using Inverted Indexing Technique

Sudeepthi Govathoti[1], M. Surendra Prasad Babu[2]

Anurag Group of Institutions, Hyderabad, India
Andhra University, Visakhapatnam, India
[1]sudeepthicse@cvsr.ac.in, [2]profmspbabu@gmail.com

Abstract

Recent years have observed the tremendous growth of information through the large number of domains available in the web. Social media (LinkedIn, Twitter etc.) concentrate on handling massive data obtaining from various sources. It is a fact that information retrieval and data extraction are difficult tasks in handling the large collection of web documents. Semantic web is a new technology used to handle the massive raw data to transform it into knowledgeable representation. Traditional search engines use page ranking algorithms to find data from a large data sources. The proposed work is aimed at designing a user interface for data extraction from multiple repositories using Uniform Resource Identifiers (URIs) and applying inverted indexing techniques for generation of Ontologies. These methods may be used to develop efficient semantic web knowledge based systems for retrieving relevant information from the web.

Keywords*: Semantic Web; Ontology; Knowledge base.*

1. Introduction

The WWW (World Wide Web) contains a large volume of interlinked documents that allows the people to share the information from large database repositories globally. The structure of the web has taken many evolutions from web of documents to web of applications [1]. The Semantic framework provides strong architectural foundation for next generation of web. It allows data to be shared and reused across applications [2]. Resource Description Framework describes data information in graphical form in terms of things and their interrelationships [3]. It provides framework to describe resources where each component of RDF statement is represented in form of triples. [4]. SPARQL (Simple Protocol and RDF Query Language) is a fundamental RDF querying language for the semantic web [5]. Ontologies are the key technologies used in the semantic web. Ontologies are usually built on concepts which have finite set of terms and their relationships, properties, constraints and possibly axioms [6]. It is a Conceptual network linked to each other by semantic taxonomic relationships and enables the communication with a unique standardized vocabulary.

Ontologies provide standard terminology for describing objects and their relationships. The Web ontology language (OWL) has taxonomy and richer set of relationships and set of inference rule [7]. OWL concepts are used to express rules in Semantic Web Rule Language (SWRL [8]. The present paper is arranged as follows: In Section 2 related work is presented. The proposed work is plotted in Section 3.Algorithm & flowchart for proposed work is illustrated in section 4. Experimental results are described in Section 5; conclusions are appended in Section 6.

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2017 SERSC

2. Related Work

Nupur Choudhury *et al.* [9] discussed the evolution of web by encouraging participation, collaboration, and information sharing. Web 1.0 refers to the first generation innovation World Wide Web capable of sharing the web of documents. Web 2.0 is popularly known as web of people facilitating interaction between web users and websites. Web 3.0, Web of data, refers to semantic framework aimed at providing stronger architectural foundations for interactive environment to the people and machines. Chandrasekaran John R *et al.* [10] proposed a conceptual introduction to ontologies and its role in information systems, natural language understanding, and knowledge based systems and AI. Opined that agricultural knowledge is heterogeneous and distributed across the institutions and people all over the world. . Even if it is digitalized using the existing web technologies, data sharing in this domain is very meagre and is unable to produce the desired results. Kaushik Deb *et al.* [11] proposed Fruit Ontology system for providing fruit related data and information. They used ontology based semantic web architecture for faster retrieval. Wang Shuyan *et al.*, [12] proposed an online agricultural ontology service system. It mainly concerns with efficiency and evolution of ontology in semantic search applications but it is does not concentrate on integrating a knowledge from multiple repositories.

J.A.Gokhale *et al.* [13] proposed a Global Information System (GIS) incorporating geospatial data for providing various aspects such as yield measurement, crop estimation, crop predication, forecasting, zoning issues . Muhammad Shoaib *et al.* [14] proposed a Integrated Agriculture Information Framework (IAIF) framework for extracting knowledge from multiple domains. It is developed and designed to enable knowledge extraction from multiple repositories. IAIF has been designed and developed with the core aim of meaningfully representing; combining, merging and aggregating the data present in existing knowledge repositories through the use of metadata and domain ontologies.

Liu Xuninget *et al.* [15], proposed a convenient means of searching information using a searching tool that is developed using neural network technology. The extracted knowledge is classified automatically using neural network. Maryam Tayefeh Mahmoudi *et al.* [16] proposed a System for retrieving and organizing the educational materials using the semantic based approach, a frame based approach for representing conceptual entities for handling large text representations is suggested.

3. Proposed Work

A novel framework is designed for data extraction from multiple repositories. Ontologies are generated by applying inverted indexing technique. The framework is explained in six phases namely data collection, data extraction, data validation, XML conversion, indexing and Ontology generation phase.

In data collection phase data fusion method is adopted to collect larger set of data. In data extraction phase combination of HTML Parsing, Text Pattern matching along with DOM parsing is adopted to extract the web page structure. In data validation unwanted data and duplicate content is verified and removed by handling only the necessary part of URLs. Data is converted into XML notation by adding namespaces to each field of the database. Ontologies are generated by applying inverted indexing technique. The generated ontologies are stored in the database by excluding the stop words along with their frequency. The methodology for the proposed work is explained section 4.

 Copyright ⓒ 2017 SERSC

4. Methodology

4.1 Data Collection

The user interface is designed to collect the data from multiple repositories like Google, Bing, Yahoo etc., and maintained as a relational database. URI stands for "Uniform Resource Identifier" is a string of characters that uniquely identifies a resource in a uniform format. The scraped results are stored in database in the form of Keyword, Title, URL, and Description. The data collected is in unstructured format, which is converted into structured and Semi-Structured formats. The general approaches used to handle this bulk data representation are as follows:

- Collection fusion
- Data Fusion

Collection Fusion: It focus on ranking of the data collected from different sources and provides re-ranking after storing the data collected from multiple repositories [17][18].

Data Fusion: In Data Fusion data from different sources is merged and final ranking is performed on the merged data. [19] In our proposed work, we adapted data fusion to collect the large set of data.

4.1.1. Stepwise Analysis of Data Collection Phase:

1. Input test query.
2. Query is processed by presenting to multiple repositories.
3. The data is retrieved as a search results obtained from different search engines.
4. Repeat steps 1-3 until required data is obtained.

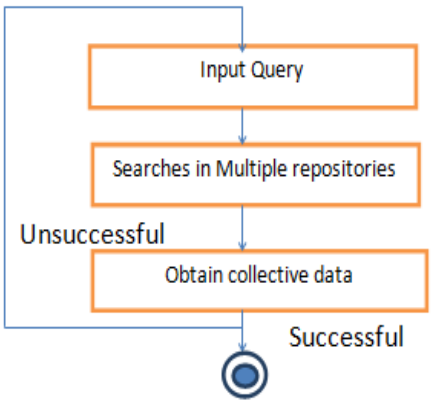

Figure 1. Workflow Diagram of Data Collection Phase

4.2 Data Extraction

In this phase data extracted from multiple repositories is maintained in a unique database. In order to handle data extraction with respect to the user query many methods can be adopted from the web as follows:

Human copy-and-paste: This technique is a manual approach used by humans if the search engines limit the results.

Text pattern matching: This method applies regular expression based approach to parse the web page

Copyright ⓒ 2017 SERSC

HTTP programming: It Uses Socket programming for handling the static and dynamic web page retrievals by posting HTTP requests.

HTML parsing: Web page content is parsed using HTML notations and converted into relational database.

DOM parsing*:* The data retrieved on a client machine is referred as a client-side scripts and these webpage's are parsed as a DOM tree, based on which programs can retrieve parts of the data.

Semantic annotation recognizing*:* This can be represented as a special case of DOM Parsing, which allows parsing data schema presented in the web pages.

Among these techniques, combination of HTML Parsing, Text Pattern matching along with DOM parsing is adopted to extract the web page structure. The software resource used in this phase is CURL, which is a command line library for obtaining and transferring data with URLs supporting all the http methods like GET, POST, cookies, etc. in PHP.

4.2.1. Stepwise Analysis of Data Extraction Phase:

1. Obtain Search results from data collection phase.
2. Perform data extraction using HTML parsing & Regular expression parsing.
3. Store it in the data base

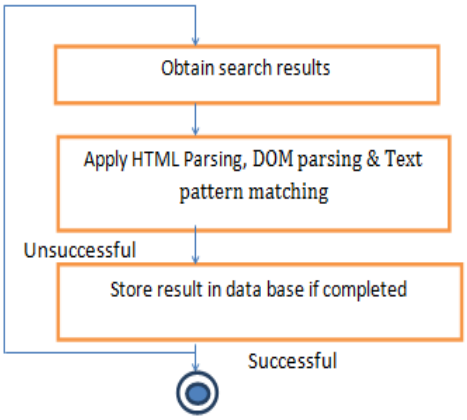

Figure 2. Workflow Diagram of Data Extraction Phase

4.3. Data Validation

In this phase, the data collected from the previous phase, unwanted data and duplicate content is verified and removed by handling only the necessary part of URLs [20].If the Bing Search results page is considered, the required part is present only in one section and the data present at the left and right ends of the page may not considered by the user so, these sections are removed and the concepts of Title, URL Description parts of the result is displayed by considering the document section of the web page using the tags of html like *<a href> * and *<h3></h3>* and ** as shown in *Figure:4.*

 Copyright ⓒ 2017 SERSC

Figure 3. Search Result Page with Various Sections

4.4. XML Conversion Phase

In this phase, the data stored in database is converted into XML notation by adding namespaces and rules of XML to each field of the database, i.e. the each field from the database is represented as <title> </title>,<URL></URL> and <description> </description> tags.

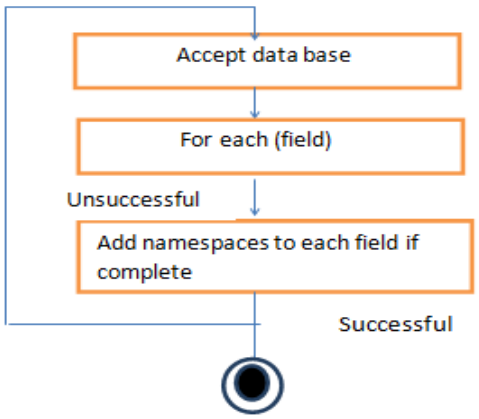

Figure 4. Workflow Diagram of XML Conversion Phase

The XML conversion phase includes the steps below:

1. Access the database.
2. For each (field) add namespaces of its field i.e. <*title*> welcome movie </*title*>.
3. Repeat step 2 until all records are connected into XML format.

4.5. Indexing Phase

The data collected from multiple repositories is validated and converted into XML formats and maintained in a structured format in database. The keywords are extracted and indexing process is performed on the keywords obtained by removing the stop words by maintaining each words frequency count. The inverted indexing mechanism is used in this phase provides the flexibility of the retrieval of ontology by removing repeated and unnecessary data that was collected from various sources.

Copyright ⓒ 2017 SERSC

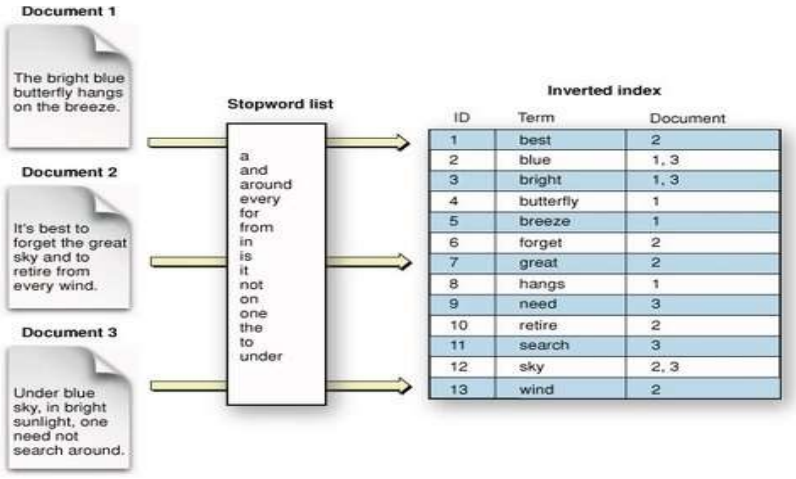

Figure 5. Inverted Indexing Architecture

The inverted indexing phase stores the list of occurrences of words for each description and maintains its frequencies in the form of hash table. It is a data structure which holds the list of terms obtained from the document along with their frequencies of occurrence. Figure 5 represents the inverted indexing architecture, where the documents are considered and list of words are extracted by removing the stop words like "is", "the" etc. as it may reduce the performance of search.

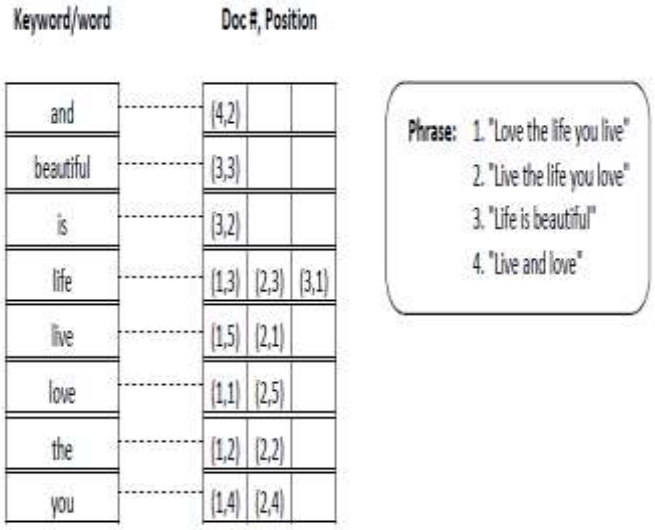

Figure 6. Inverted Indexing Data Structure

4.6. Ontology Generation Phase

Ontology generation phase plays crucial role in generating OWL rules for the semantic web and creating axioms which supports SWRL. In this phase, the Ontology database is created by excluding the stop words from the description text of the relational database and maintains the keywords separated as axioms along with their frequency.

The Step wise analysis of ontology generation phase:

1. Access the description field data from the database.
2. For each (description text) remove stop words.
3. List the elements with description keywords and their frequencies (i.e.) count of

 Copyright ⓒ 2017 SERSC

the corresponding word in the list of description phrases.

4. Repeat step 3 & 4 until all phrases from description test is completed.

4.7 Algorithm & Flowchart of Proposed System:

The algorithm for the proposed work is stated as follows:

Step 1: Initiate a search query to search engines.
Step 2: Extract results from multiple repositories like Google, Bing, yahoo etc.
Step 3: Store the Structured data if obtained, otherwise go to *Step 4*(check the data scraped).
Step-4: Perform data validation for each (data) check if data (unstructured) convert into structured format. Else Semi-Structured, convert into XML format.
Step5: Apply inverted indexing on the data in the database.
*Step 6***:** Generate ontology for the data indexed, repeat steps 3 to 6 until the tracing of the data is performed.

Flowchart of Proposed System:

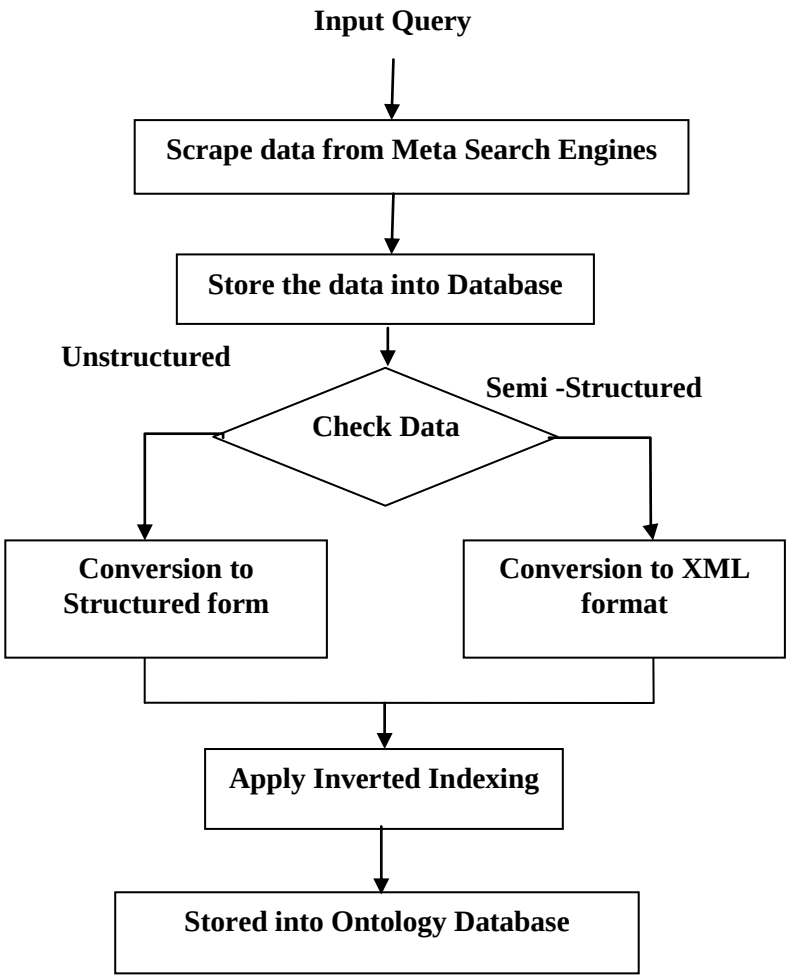

Figure 7. Flowchart for the Proposed System

Copyright ⓒ 2017 SERSC

5. Experimental Results

Figure 8. User Interface

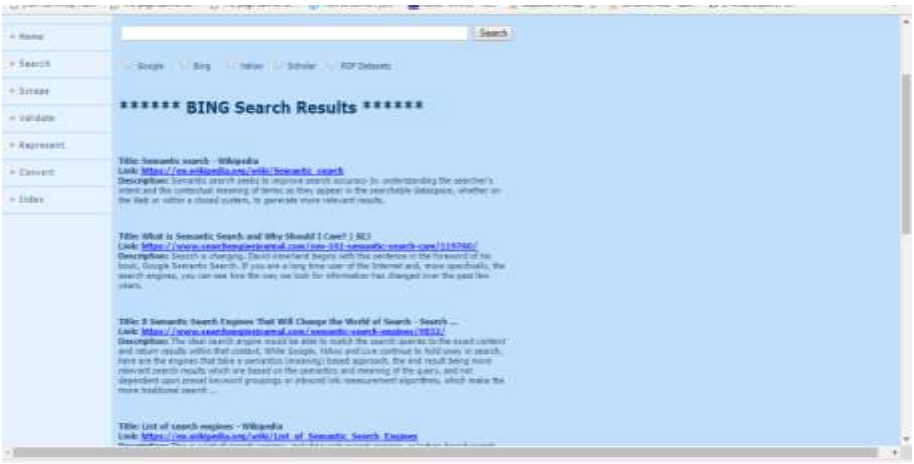

Figure 9. Data Extraction

Figure 10. Data Validation by Removing unwanted URLs

`The Scraped result contains unwanted data like advertisement links to images etc. The data is validated by considering the title, URL and description parts of the results.

 Copyright ⓒ 2017 SERSC

Figure 11. Conversion into XML Format

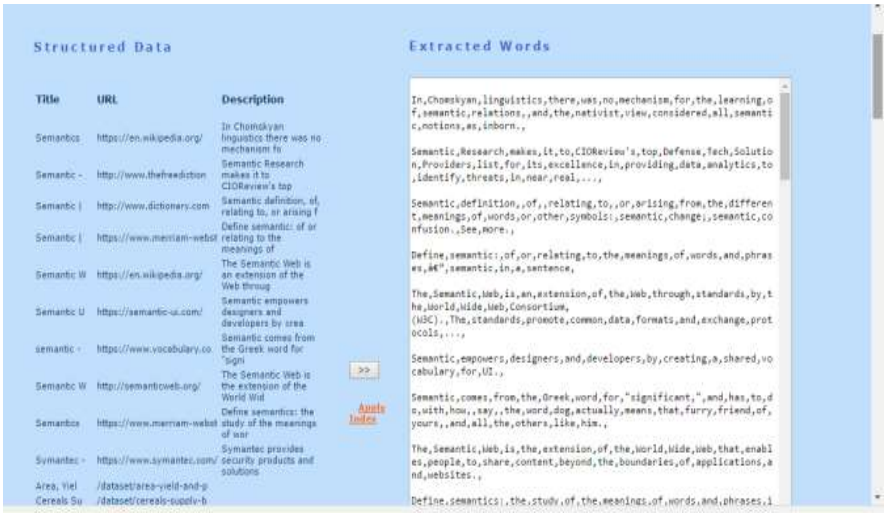

Figure 12. Structured Format Representation

The fig-12 presents the data collected from collection phase is validated and represented in structured format. For representing the ontological notation, we need to extract the data from the description field to store various keywords or terms.

The extracted words can be seen on the other side (right) in the above fig. In general the extracted words consists of various stop words , by removing all those index as we focus on ontological representations, which can be viewed in the next page and gets indexed using inverted indexing mechanism in the next page by removing stop words from the list of words extracted.

The fig-13 presents the ontology database with the keywords along with the count of frequencies in the descriptions obtained from the results of multiple repositories.

Copyright ⓒ 2017 SERSC

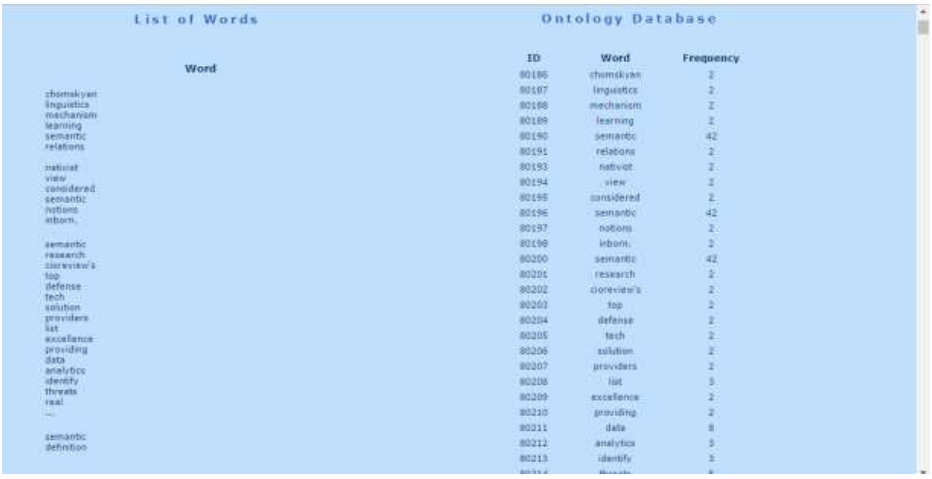

Figure 13. Ontologies generated after applying indexing.

6. Conclusion

Several techniques are adopted to handle semantic web, towards latest research trends in web development. In our work, we proposed a framework for extracting data from multiple repositories using an interface, which collects data from multiple repositories and validates them by considering only the necessary sections in the result. We propose inverted indexing mechanism which removes the stop words creates the index. It also helps the user to process the user query with fast and more relevant results. The main objective of this indexing method is to facilitate finding valuable and relevant information in Web. The words that are indexed are maintained in ontology database further moves onto ontology generation which helps to frame the rules. The scope of this work is extended towards the implementation of automated RDF text creation from the ontology generated which helps to frame out the rules for the design of an expert system.

References

[1] S. Aghaei, M. A. Nematbakhsh, H. K. Farsani, "Evolution of the World Wide Web: From Web 1.0 to Web 4.0", International Journal of Web & Semantic Technology (IJWEST), vol.3, no.1, **(2012)**.
[2] G. Kuck, "Tim Berners-Lee's Semantic Web", South African journal of information management, vol. 6, no.1, **(2004)**.
[3] A. Altar, Samuel *et al.*, "RDF & RDF Query Languages – Building Blocks for the Semantic Web", Journal of Information Systems & Operations Management.
[4] RDF. Resource Description Framework (RDF). May 9, 2007, http://www.w3.org/ RDF/. SPARQL 1.1. **(2013)**.
[5] SPARQL 1.1 Overview. October 29, 2013, http://www.w3.rg/TR/ Sparql11-Overview/
[6] D. Dou, H.ao Wang, H. Liu "Semantic Data Mining: A Survey of Ontology-based Approaches" IEEE 9th International Conference on Semantic Computing, **(2015)**.
[7] I. Horrocks, P. F. Patel-Schneider, "A Proposal for an OWL Rules Language",. Proceedings of the Thirteenth International World Wide Web Conference, **(2004)**, pp. 723–731
[8] Y.-J. Hu, and C.-L. Yeh, "Challenges For Rule Systems On The Web", Ruleml-2.0, October 25-26, 2007,Orlando, Florida Horrocks, I., Patel-Schneider, P., And Van Harmelen, F, " From SHIQ And RDF To OWL: The Making Of A Web Ontology Language", Journal Of Web Semantics, **(2003)**, pp.7–26.
[9] N. Choudhury, "World Wide Web and Its Journey from Web 1.0 to Web 4.0",(IJCSIT) International Journal of Computer Science and Information Technologies, vol. 5, no. 6, **(2014)**, pp. 8096-8100.
[10] B. Chandrasekaran, J. R. Josephson, V. R. Benjamins, "What Are Ontologies, and Why Do We Need Them?", IEEE Intelligent Systems, **(1999)**.
[11] S. Chowdhury, M. M. Kaysar, K. Deb, "Designing a Semantic Web Ontology of Agricultural Domain", **(2013)**.
[12] X. Jinhui, S.Yang, Y. Yong, Y. Zhifeng, W. Shuyan, "An Online System for Agricultural Ontology Service", Third International Conference on Intelligent Networks and Intelligent Systems, **(2010)**.

 Copyright ⓒ 2017 SERSC

[13] N. Agarwal, M. Rao, S.S. Mantha, J.A.Gokhale, "Annotation of Geospatial Data Based on Fabrizio Lamberti, Member, IEEE, Andrea Sanna, and Claudio Demartini, Semantics for Agriculture: Case Study for India, **(2011)**.
[14] M. Shoaib, A. Basharat "Semantic Web based Integrated Agriculture Information Framework", Second International Conference on Computer Research and Development, **(2010)**.
[15] D. Liying, L. Hongjuan, H. Dongbin, W. Zhe, L. Xuning, "Research on Intelligent Searching of Agricultural Information Based on Ontology", International Conference on Computer Science and Service System, **(2012)**.
[16] M. T. Mahmoudi, F. Taghiyareh and K. Badie, "A Semantic Rule-based Framework for Efficient Retrieval of Educational Materials", The Electronic Journal of e-Learning, vol. 11, no. 3, **(2013)**,pp. 182-192.
[17] E. M. Voorhees, N. K. Gupta, "Learning CollectionFusion", International ACM SIGIR Conference on Research and Development in Information Retrieval, USA, Special Issue, **(1995)**.
[18] G. Towell, E. M. Voorhees, N. K. Gupta, B. Johnson-Laird, "Learning Collection Fusion Strategies for Information Retrieval", Twelfth Annual Machine Learning Conference, Lake Tahoe, **(1995)**.
[19] G. Paltoglou, M. Salampasis, and M. Satratzem, "Simple Adaptations of Data Fusion Algorithms for Source Selection", Advances in Information Retrieval, 31th European Conference on IR Research, **(2009)**.
[20] I.P Keerthana, T. Aby Abahai, "An Intelligent Meta Search Engine for Efficient Web Document Retrieval", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 17, no. 2, **(2015)**, pp. 45-54.

Copyright ⓒ 2017 SERSC

Copyright ⓒ 2017 SERSC