Jr. of Industrial Pollution Control 33(2)(2017) pp 1534-1541
www.icontrolpollution.com
Research Article

HPL ALGORITHM FOR SEMANTIC INFORMATION RETRIEVAL WITH RDF AND SPARQL

PRATHYUSHA KANAKAM[1*], S. MAHABOOB HUSSAIN[2] AND D. SURYANARAYANA[3]

[1]Assistant Professor, Department of Computer Science and Engineering, MVGR College of Engineering, Vizianagaram, Andhra Pradesh, India

[2]Assistant Professor, Department of Computer Science and Engineering Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India

[3]Professor, Department of Computer Science and Engineering Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India

(Received 07 August, 2017; accepted 20 September, 2017)

Key words: Semantic web, RDF, Ontology, Information retrieval, Search engines, LODQA, SPARQL

ABSTRACT

As the web composed with lots of unstructured data, retrieving the accurate information for the user's posed query from it is a critical issue. Most of the search engines are fails to achieve the accurate outcomes for the users. In order to overcome this and to obtain efficient results, SPARQL query language is used to convert users' posed natural language queries to machine understandable format. Semantic web technology based on SPARQL is used to acquire useful information from the RDF knowledge base, which gives beneficial information to the user. In this paper, HPL algorithm is effectively utilized to aware of querying the semantic web. It also makes use of Linked Open Data Quality Assessment (LODQA) system to perform a semantic search that converts normal user defined queries into machine understandable formal logic. By combining both these systems, the interpretation of a natural language query into SPARQL queries is made easy that grabs knowledge from ontological database that are stored in Resource Description Framework. Accordingly, technologies and data storage possibilities are analyzed and evaluated to retrieve accurate results and a test case for the career opportunities for students.

INTRODUCTION

At the initial stage of computer processing (in the past 1980's) obtaining knowledge from the web is a challenging issue (Mowery, *et al.*, 2002). During this era, user followed a well-defined set of rules to obtain useful information by logging into remote system and search for the content or file. If the file found, user has to download it from that remote system to local system, which is bit complicated and time-consuming task (Hussain, *et al.*, 2016). This process involves various pitfalls like expert knowledge required for information access to learning about commands, special languages, and syntaxes. To overcome these obstacles, in 1990's Internet has changed as the World Wide Web. In this, the user is provided with graphical user interface to find the details of the document by typing URL, which contains hyperlinks which connects other documents and by clicking on those links, information is accessed so that the information is navigated from a web server which does not require expert's knowledge (Hölscher, et al., 2000). As the web growing day by day, there is a need for information retrieval techniques along with semantic web technology to obtain accurate results in an efficient manner.

When a system is integrated with both semantic techniques and knowledge retrieval techniques, it provides better results. Semantic Web designed that makes uses of these techniques will be in a machine understandable format. This format designing is

***Corresponding authors email: prathyusha.kanakam@gmail.com; mahaboobhussain.smh@gmail.com; suryanaray-anadasika@gmail.com**

the prime responsibility for the question-answer processing system in order to acquire triplets from users' query. It is a type of Information Retrieval system that defines keywords from the question for searching exact content in the documents. The obtained answer gives the information about each entity whether it is person/time/location. Some systems map the relevant information to the question posed by the user that depends mostly on the keywords in the query (Fig. 1).

Natural language processing systems interprets user's posed queries in a machine understandable format to obtain useful knowledge from web. These systems process each and every word in the textual query of the users. Over the past years of 1960's, both Information Retrieval systems and knowledge based systems are served for question answering purpose. Later the web gradually increases in its size due to large amounts of information that makes the search engine a challenging issue to answer users' queries. Google and Bing type frameworks mainly challenges the quality of the large information in complex format that are stored in various datacentres which are both interconnected and as well as scattered over different locations in the world. Due to this growing data, everything is obtained from single platform called web that constitutes several documents related to specific organisations like national statistical agencies, e-commercial websites provided with web forms for searching their databases. Search engines are familiar to retrieve the results but not effective to retrieve semantic information from the Web and thus, it is called as the deep Web (Michael, 2001). Every Search engine is automated in such a way that maintaining of HTML query forms to acquire high valued results in an efficient manner from the web with low cost. Using an HTML and crawler it is an effortless task to build a search page and a search engine to search in the huge database but, retrieving the precise and predicted information for the users is complicated factor. Therefore semantic Web provides absolute, essential information and can easily synchronise computers and individual to achieve results in collaboration. Ontology is a logical information repository which consists of triplets i.e., resource description framework correlated with facts and practices as shown the below architecture in (Fig. 1).

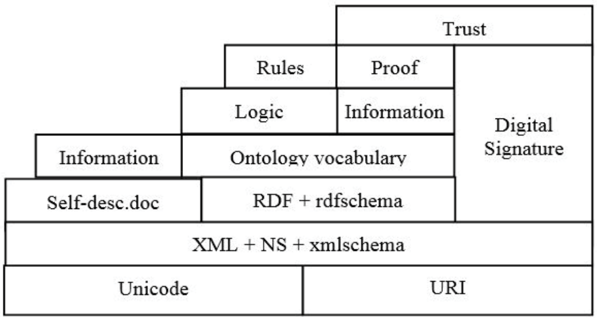

Fig. 1 DWT decomposition model.

Semantic Web is knowledge of connections and the Web of data is an upgrade of the Web of documents. It consists of a decentralised database which is accessible by the semantic search engine machine. Here, every document should be in resource description framework will be considered for query processing on this ontological database and it is a conceptual model of the domain of career pathfinder. Suggestions will be offered for the post education of tenth, inter and degree accomplished students. The proposed approach is a translation of a natural language query with Natural Language Processing (NLP) to SPARQL from the ontology knowledge base (Shaik, *et al*., 2016). In this current work, career ontology is used to explain the semantic Web and its retrieving process as shown in (Fig. 2). Keywords are mapped to concepts in the ontology, and the graph then is used to retrieve available relations between these concepts. These relations are most likely the gaps in the query, which are not specified by the user. The user knows the complexity, e.g. searching for a post-graduation course and duration of engineering, the obvious relation would be "post-graduation course". RDF is used in the description of triples such as {Subject, Predicate and Object} to bypass inaccurate results in search time (Fig. 2).

PRELIMINARIES

Semantic search furnishes accurate results by providing answers to user posed queries based on meanings rather than keyword search. The pitfall lies in query languages to search a particular content. For that, a framework is required to describe resource for storing the information of the web and a special language to query on that ontological repository. To retrieve exact results from the semantic Web documents which are in the form of RDF format, a unique query language is used called SPARQL. From a Resource Description Framework (RDF) format data, SPARQL can easily retrieve and handle the information and therefore it is called as an RDF and a semantic language for querying ontological databases. This specification defines the semantics as well as syntax of the SPARQL to RDF. Finally, the outcome of the queries in SPARQL syntax will be in triplet or in graphical representation called RDF graphs. Mostly, the syntax of the query of SPARQL represents conjunctions, disjunctions and some optional patterns. Therefore, the entire semantic web documents are in <Subject, Object, Predicate> triples.

Resource Description Framework (RDF)

Knowledge can be represented more expressively using resource description schemas (RDFS) which are an RDF vocabulary description language used to illustrate vocabularies for RDF, means the models that are dealing. In RDF schema everything discusses resources and it define hierarchical relationships like sub-classes, super-classes and also sub-properties and super properties [3]. Besides classes, it defines the properties of classes, properties connect with classes either with literals. It can define a property via rdf:Property. On property, one can define restrictions on domain and range according to type via rdfs:domain and rdfs:range. Ontologies are developed as a further semantic Web standard as a reason that the RDF will not accomplish more complicated constraints concerning properties of the classes and the resources.

Web Ontology Language (OWL)

The knowledge through ontology is represented using Web Ontology Language and are used to describe relationships among classes and classifications (McGuinness and Harmelen, 2006).

SPARQL

SPARQL (a recursive acronym for Simple Protocol and RDF Query Language and pronounced as "sparkle") is an RDF query language, i.e., a semantic query language for retrieving from RDF databases (Prud, *et al*., 2006). SPARQL is a language for retrieving the knowledge from RDF and it consists of some set of standard rules followed to process the SPARQL queries on RDF data to retrieve most accurate results.

SPARQL endpoint is RDF triple database on the server usually, which is available on web and top of web transfer protocol there is an SPARQL protocol layer means via HTTP SPARQL query transfers to server and server gives its results to the client. It is like SQL but works on RDF graphs, not on tables. RDF is a connected graph representation with some patterns of variables. These patterns combined to get different patterns of more complex results. The association between SPARQL and RDF Graphs is shown in (Fig. 3).

Querying on the ontology generated with the RDF is a critical task which employs syntax based formal query language. Here in this paper, SPARQL query language is used to retrieve the information from these ontologies. RDF subgraphs include URIs, blank nodes, typed and untyped literals with aggregate functions, subqueries, complex joins, property paths to extract the data. Efficient retrieval of information

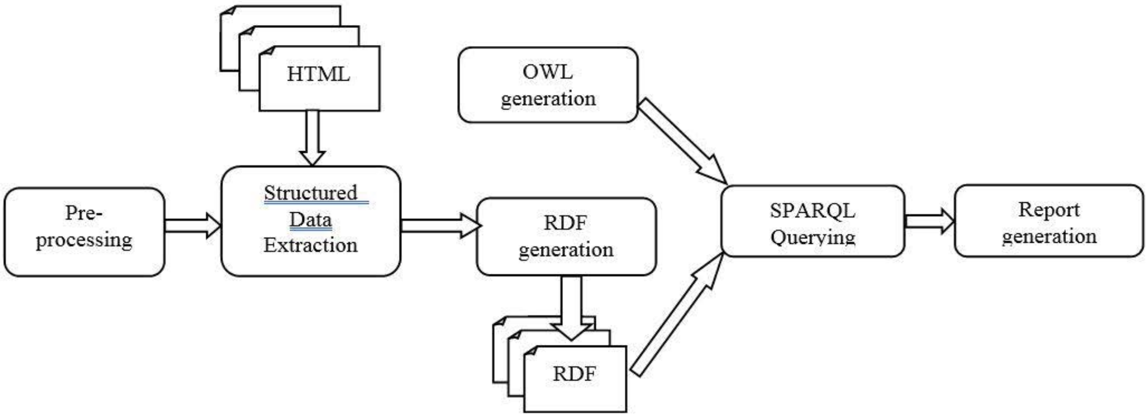

Fig. 2 Process of semantic web.

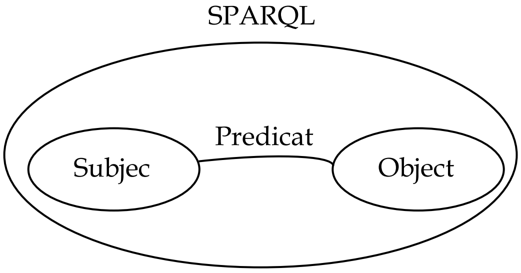

Fig. 3 Association between SPARQL and RDF graphs.

is performed using queries on unknown relations. Rather than full data manipulation, and constructing of new RDF graphs, updating is carried out on repository using query languages. RDF subgraphs also transform data from one vocabulary to another vocabulary. Thus RDF, RDF Schema, ontology Web Language are adopted to obtain the logical and accurate results. Federated queries are distributed over different SPARQL endpoints (Zviedris and Guntis, 2011). This framework is suitable for implementing in multiple programming languages (Don, *et al.*, 2003). A machine understandable specification of a shared abstract model is used to define meanings of all concepts called Ontology. It is easy to retrieve the information from the ontology-based database by using an SPARQL query language, but the application which takes a query in the form of natural language can be converted into the formal query (SPARQL). This paper works on the process of converting the natural language into the formal query which can post on the ontology directly to retrieve the information.

HPL Algorithm

In the previous work, the authors designed a High-Performance Linguistics scheme which is a cognition-applied machine to learn how to infer the content of the natural language sentence (Suryanarayana, *et al.*, 2016). A question-answering system concentrates on Linguistic models which makes the meaning of each stage of words in a user's natural language query. The path to give output from the input is a trivial task to done. For that, a systematic procedure will give a clue to a machine to interpret natural language sentence. Applying cognition to a machine to comprehend various categories of the text and mapping of text to document is a complicated task. Text Lemmatization process is involved in this question-answering system, to find lemmas from the natural language sentence as well to assign some categories to those particular lemmas (Suryanarayana, 2016). It gives the best solution to solve the problem of grasping enormous amounts of data and handle it more efficiently. (Fig. 4) shows the entire collection of ontology and it is termed as semantic knowledge bases that enable exciting applications such as question answering on open domain collections. This system automatically learns ontology from texts in a given domain. The domain specific ontology that results from these knowledge acquisition methods incorporated into Lexico-Semantic database that various natural language processing systems

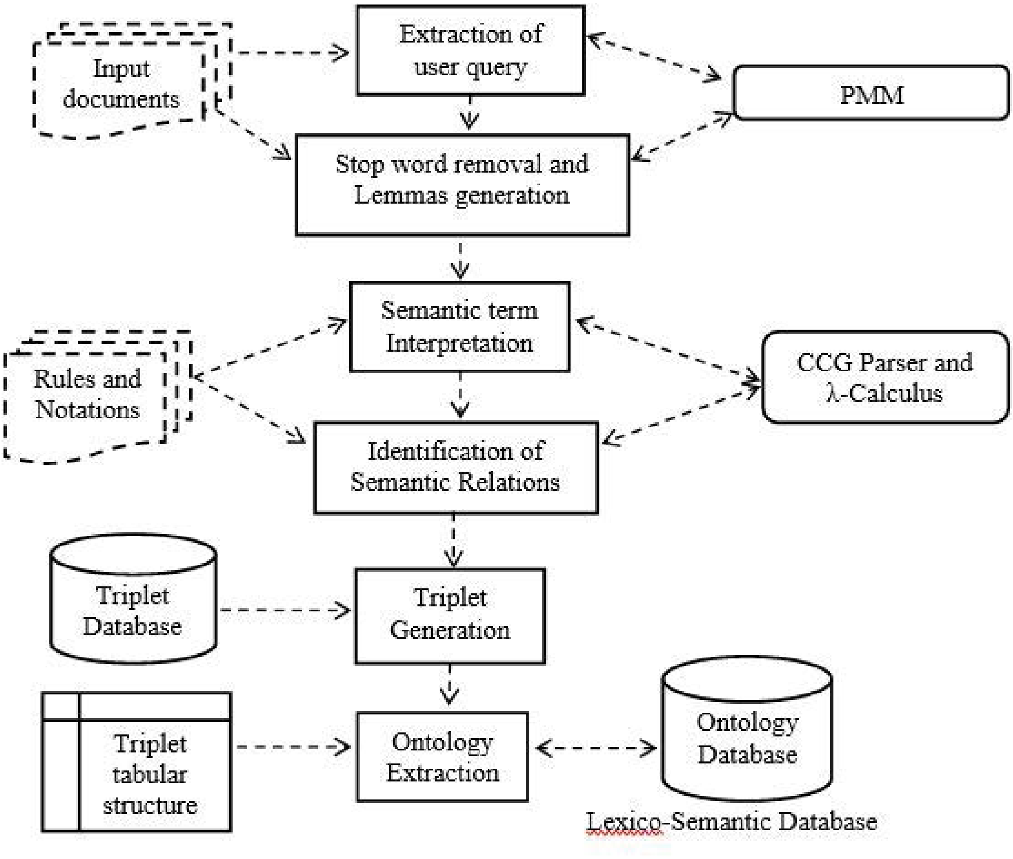

Fig. 4 Association between SPARQL and RDF graphs.

may employ. This system helps to extract specific knowledge and for searching that knowledge from unstructured text on the web. It uses ontology-based domain knowledge base known as Lexico-Semantic database (Suryanarayana, 2015) (Fig. 4).

CAREER PATH FINDER WITH HPL: AN EXPERIMENTAL ANALYSIS

In this paper, HPL's step-by-step procedure is imitated to convert a sentence into SPARQL as well as applied this SPARQL on ontological databases to retrieve the accurate information according to the user posed natural language query. As in (Fig. 5), the system takes natural language query as input and finally produces SPARQL queries as output, together with the answers to them certain SPARQL endpoint. It has 3 modules in LODQA- Graphicator, TermFinder, and GraphFinder (Fig. 5).

Linked Open Data Quality Assessment (LODQA) system

Graphicator

It is the prime module of the LODQA system, which resembles the lemmatization process in HPL to remove stop words from the Natural Language query (NLQ) and to find root lemmas in that NLQ. Firstly, it parses NLQ, and then produces a graph representation for that, which is called pseudograph pattern (PGP). A PGP contains nodes and relations. According to Jin-Dong Kim and Shigeru Nakajima (Bretonnel and Jin-Dong, 2013) the nodes are typically correspond to the basic noun phrases (BNPs) in the natural language query, and the relations to the dependency paths between the BNPs as expressed in the natural language query. For example, consider an NLQ as "What is the Post Graduation course after engineering?" (Fig. 6).

The graphicator produces pseudograph pattern as shown in (Fig. 6) to obtain the lemmas of that NLQ. From the above, "Post Graduation course" is subject, "engineering" is the object and both are related on predicate "after". Thus PGP gives the graphical representation to obtain the root words in NLQ to provide accurate information to the user.

Term Finder

Using PGP produced by graphicator for a given NL query, the TermFinder is responsible for finding the URIs and values of the nodes in the PGP where the URIs and the values have to be actually present in the target dataset. Other than that, there is no chance for the PGP to be matched with any part of the dataset. After normalisation, each node of the PGP is connected to a URI in the dataset then the PGP is converted as an anchored PGP (APGP). A natural language term may be normalised to more than one RDF terms due to ambiguity. Therefore, more than one APGPs may be produced from one PGP through normalisation. For example, consider the below query.

NLQ: "What is the Post Graduation course after engineering?"

Termfinder: {"Post Graduation", "engineering", "after"}

Graph Finder

After APGP is obtained, an NLQ enters into graph finder module, which is the last module of LODQA system. The GraphFinder module is responsible for searching the target dataset for corresponding parts, considering possible variations, which may occur in the dataset (Unger, *et al.*, 2014; Gole's, 2017). GraphFinder attempts to generate SPARQL queries for all the possible structural variations to absorb structural discrepancy between APGP and actual structure in the target dataset. The SPARQL queries are then applied on ontological databases to provide accurate information to the users' posed queries.

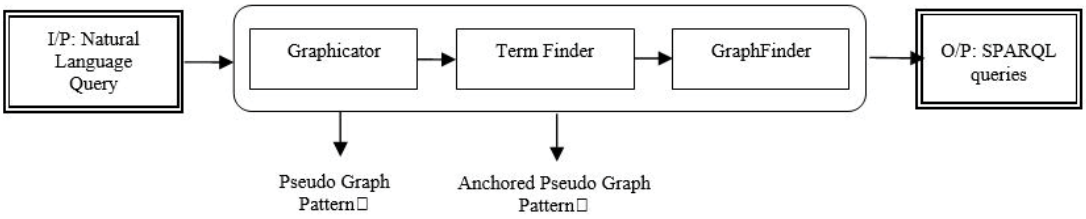

Fig. 5 Association between SPARQL and RDF graphs.

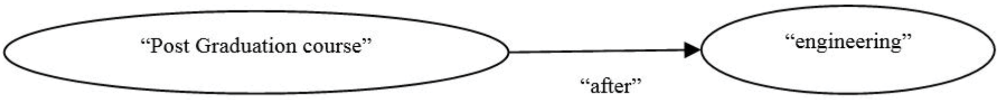

Fig. 6 Pseudo graph pattern.

For the above-mentioned query, the output of graph finder is shown in (Fig. 7). In this, stop words can be categorised as arguments of the second level (ARG2) whereas the remaining root words are as arguments of the first level (ARG1). Therefore, the triplets "Subject, Predicate and Object" are generated from the the lemmas of natural language query which are its first level arguments (Fig. 7).

The graph finder converts arguments and relations into SPARQL query. There are 425 lexical categories like noun, pronoun, determiner etc. For semantic representation, directionalities (forward/backward) are applied and there by combinatory rules are generated. Finally from XML to RDF documents are used to create respective ontology. The graph finder having Parts of Speech(POS) grammar notations are POS tagging and their descriptions as shown in Table 1.

Ontology conceptualizes a domain into a machine-readable format. Mostly information on web represented as natural language documents. Extraction of the knowledge from a large document and indexing of the will be done easily whenever the whole data undergoes into a tabular structure, i.e., answering to user's queries. HPL system mainly depends on triplet generated and domain ontology mapping to that triplet for extracting exact content or semantics from natural language queries posed by the user. Sample ontology for career pathfinder is shown in (Fig. 8).

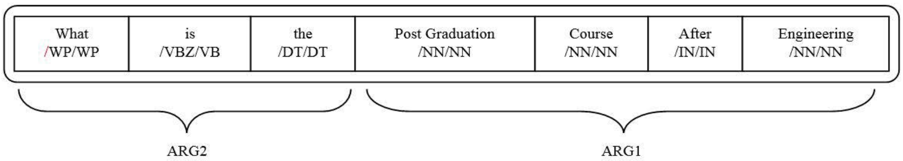

Fig. 7 NLQ in form of arguments and their relations.

Table 1. POS Tags (Source from [15])

Tag	Description
DT	Determiner
IN	Preposition subordinating conjunction
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
WP	Wh -pronoun
VBZ	Verb, 3rd person singular present
VB	Verb, base form

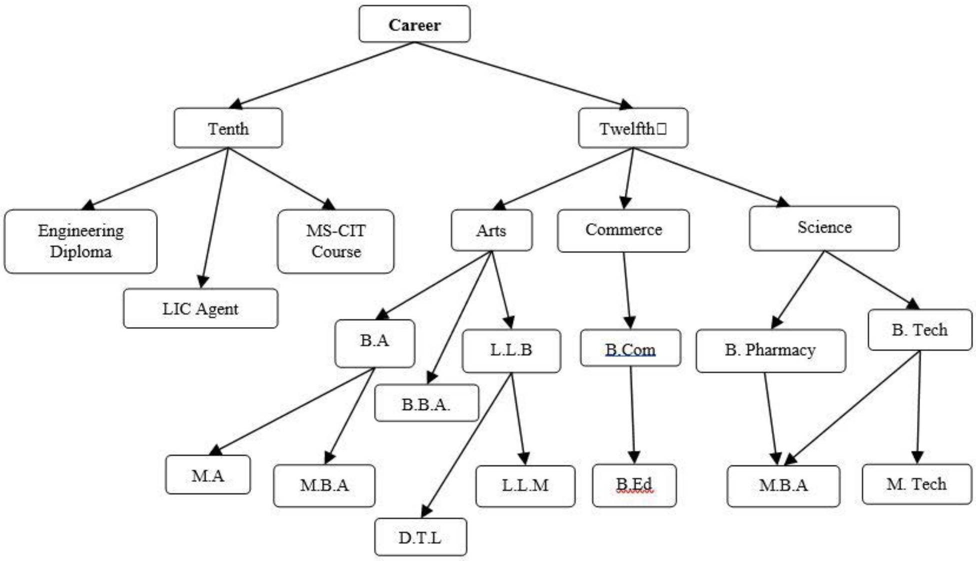

Fig. 8 Ontology for career path finder.

RESULTS AND DISCUSSION

Representing the information of the web and processing on that information with machines is the key idea of the semantic Web. This question-answering system combines both formal logics and linguistics to represent the knowledge in an efficient manner. Knowledge can be represented more expressively using RDF schema which is an RDF vocabulary description language used to describe Vocabularies for RDF. RDF is a logic-based knowledge representation framework to store the resources related to domain carrier pathfinder and describing their relationships.

:Career rdf:type rdfs:Class.

:Tenth rdf:type :Career.

: Twelfth rdf:type :Career

The entire properties, domains and range are related to a single resource and all the subclasses, superclasses, sub properties, super properties are defined as hierarchical relationships among themselves (Hornsby, *et al*., 2010).

In the above-mentioned example,

Career → rdf type class

{Tenth, Twelfth} → resources (Career) $\wedge$ member class (Career)

property of the class → literals

property definition → rdf: Property

type → rdfs:domain $\wedge$ rdfs: range

For the natural language query “What is the Post Graduation course after engineering?” the query formulation will be done as shown as below.

select

?thing ?post-graduationcourse ?duration ?entrance exam ?job

where {

?thing rdf:type dbo:engineering.

After POS tagging, arguments of the first level for NLQ “What is the Post Graduation course after engineering?” are taken into consideration to form triplet for the NLQ. Then using the triplet, SPARQL query is generated. SPARQL doesn't follow the syntax conventions because it is not based on any markup languages like XML. Names beginning with a ? or a $ are variables. As per the syntax, the variables that returned as result are the variables that declared after the after SELECT keyword. The natural language query translated into SPARQL query for retrieving of accurate results from RDF knowledge base. To extract data from ontology, the SELECT statement is used in SPARQL and the results can be displayed as tabular form. Table 2 shows the result of a SPARQL query on carrier pathfinder ontology.

Therefore, for any query given, results will be displayed accurately and the keywords were given priority. To make the search on the database effective the redundant and useless words should be discarded. The essential intention of this proposed method is to provide more accurate information to get knowledge on their career using the intelligence of the machine which can understand the perception of the given query.

CONCLSION

Semantic Web represents the meaningful information after which it is processed by the machines. This approach uses Linked Open Data Quality Assessment (LODQA) system and High-Performance Linguistics question-answering system combines both formal logics and linguistics to represent the knowledge in an efficient manner. Knowledge in an ontological database of carrier pathfinder can be represented more expressively using RDF schema which is an RDF vocabulary description language used to describe vocabularies for RDF. Given natural language query will be processed to convert into Subject, Object, Predicate called as triplets. A SPARQL query will be formed from these triplets to process on the ontology knowledge base. Therefore, any natural query language will be easily retrieving the semantic information with this scheme. SPARQL is used to express and formulate query on databases of RDF. Thus, this method reduces the complexity in retrieving accurate information from any NLQ. Hence, the practical approach with different types of NLQ are checked with this system and observed that it retrieved accurate and semantic results. Here, authors worked on a query and the results show the semantic information retrieval. In this paper, the entire approach follows the HPL algorithmic process to demonstrate the proposed work.

Table 2. Experiment result for an SPARQL query

Thing	Post-Graduation Course	Duration	Entrance Exam	Job
B. Tech	{M.B.A, M. Tech}	4.0	EAMCET	Software job

REFERENCES

Bretonnel, C.K. and Jin-Dong, Kim. (2013). Evaluation of SPARQL query generation from natural language questions. Joint Workshop on NLP&LOD and SWAIE: Semantic Web, Linked Open Data and Information Extraction. 3.

Don, C., Daniela, F., Jonathan, R., Jérôme, S. and Mugur, S. (2003). XQuery: A query language for XML. *SIGMOD Conference*. 682.

Gole's, S. (2017). Part-of-speech tagging using OpenNLP - Sagar Gole's Blog. [online] Available at: http://blog.thedigitalgroup.com/sagarg/2015/06/18/part-of-speech-tagging-using-opennlp/

Hölscher., Christoph. and Gerhard, S. (2000). Web search behavior of Internet experts and newbies. *Computer Networks*. 33(1) : 337-346.

Hussain, M.S., Prathyusha, K., Suryanarayana, D., Swathi, G. and Sharmela, S. (2016). Semantic Information Retrieval: An Ontology and RDF-based Model. *International Journal of Computer Applications* 156(9) : 34-38.

Hornsby, Kathleen, S. and Kripa, J. (2010). Combining ontologies to automatically generate temporal perspectives of geospatial domains. *Geoinformatica*. 14(4) : 481-505.

Michael K.B. (2001). White paper: the deep web: Surfacing hidden value. *Journal of Electronic Publishing*. 7(1).

Mowery., David, C. and Simcoe, T. (2002). Is the Internet a US invention? —an economic and technological history of computer networking. *Research Policy*. 31(8) : 1369-1387.

McGuinness, D.L. and Harmelen, F.V. (2004). OWL web ontology language overview. *W3C Recommendation*. 10.

Prud., Eric. and Andy, S. (2006). SPARQL query language for RDF.

Shaik, S., Prathyusha, K., Hussain, S.M., Suryanarayana, D. (2016). Transforming Natural Language Query to SPARQL for Semantic Information Retrieval. *International Journal of Engineering Trends and Technology*. 41 : 347-350.

Suryanarayana, D. Prathyusha, K., Hussain, S.M. and Sumit, G. (2016). High Performance Linguistics Scheme for Cognitive Information Processing. 4th International Conference on Advanced Computing. Networking and Informatics.

Suryanarayana, D. (2016). Cognitive analytic task based on based on search query logs for semantic of semantic identification. *IJCTA*. 9 : 273-280.

Suryanarayana, D., Hussain, S.M. Prathyusha, K. (2015). Stepping towards a semantic web search engine for accurate outcomes in favor of user queries: Using RDF and ontology technologies. Computational Intelligence and Computing Research (ICCIC). IEEE International Conference on. IEEE.

Unger, C., Corina, F., Vanessa, L., Axel-Cyrille, N.N., Elena, C., Philipp, C. and Sebastian, W. (2014). Question answering over linked data (QALD-4). Working Notes for CLEF 2014 Conference.

Zviedris., Martins. and Guntis, B. (2011). ViziQuer: A tool to explore and query SPARQL endpoints. *The Semanic Web: Research and Applications*. 441-445.