Comparing SPARQL Tools for Logical Inferencing in Semantic Web

Aman Jolly, Shailender Kumar

Abstract:--- In today's modern era, where search engines are still using traditional information retrieval system which usually involves string matching and comparison of text and character. Semantic web technologies are considered an only silver bullet that has the potential to lead the transition from keyword-based search to context-based search. Semantic web technology can provide a system where the web of the document can be made machine-understandable. First, this paper projects how a semantic web's linked data can lead us to more robust and machine understandable system. Second, this research shows how a logical inference can be made in linked data by using SPARQL queries on different SPARQL tools. Third, this paper shows the analysis of different SPARQL tools such as Twinkle 2.0, Jena ARQ 3.5, and Protégé 5.0. A comparative tabular analysis has been evaluated in order to compare the features of these SPARQL tools and describe shortcomings in their present version.

Index Terms: Semantic web, Linked Data, SPARQL tools, Protégé 5.0, Jena ARQ 3.5, Twinkle 2.0, Logical inference .

I. INTRODUCTION

The semantic web is one of the most prominent and budding fields in IT-oriented research. However, research on adding semantics in the web was initialized as soon as the web was started but a robust semantic web that is the large repository distributed metadata that was integrated into human readable document emerged only over fourteen years. The Semantic Web (Web 3.0), a web of data, methods and Technology permit machine to comprehend the meaning of the information stored on World Wide Web (WWW) [15]. It provides a standardized, powerful, worldwide, omnipresent communication mechanism whose benefits are not viable to ignore [9].In the history of web there was the time when ARPANET [2], the father of the first generation of web was connected with only fewer academic institutions in late 1960's and networking were highly based computer-centric processing and way of retrieving the information from the source was connecting to a remote system via terminal, one had to browse through their file system data and retrieve the desired file by downloading it to base workstation. This type of approach requires expert knowledge as one has to remotely login via advanced commands and information retrieval was more costly in terms of hardware cost. In this first generation of web networks, only the experts of academic institutions were connected that made the flow of information limited to some. This restriction of information gave rise to the second generation of the web also called web 2.0 which marks the iconic birth of World Wide Web in 1990's. The networking in the second generation of the web is based on document centred processing unlike web 1.0 and the mode of information retrieval is making a query in the search engines like Google via web browsers [2]. The search engines display the list of relevant hyperlinks that are meant to satisfy the required user query. Thus by clicking on the hyperlink, it provides access to the document that is being hosted on other servers across the internet. This state of the art is still being used actively. About 44% of the world population that is 3.236 billion people in 2016 had an access to internet according to World Bank's world development report. Web 2.0 is much more convenient and don't require any expert knowledge that can be easily used by normal users. Information retrieval became much easier due to web crawlers and web search engines as they maintain web index that made the search for required information easily. The statistics that are shown by World Bank clearly reveal that the web is very big and it is growing continuously at an astonishing level.

A. Need for Semantic Web

After the arrival of "Internet of Things", a network of different devices such as household and network devices is needed to establish connections among them [10]. All these devices are having different types of sensor that gather different type heterogeneous data. Because all these devises are gathering data from internet and environment, they can put data on the web or on any social media platform so that devices can communicate with each other and share their information. The issue lies in the enormity of data and incapability of machines to make sense of this data. Humans try to solve the problem with contextual knowledge, world knowledge and experience if the content is written in the language that is understood by them. Exploiting machines to their full potential have become a necessity, so as to provide better filtering and searching capabilities in this vast cyberspace. The web documents that are currently based on hypertext markup language are meant for human consumption [1]. The web pages are designed for providing only the descriptions about the presentation of information in the web browsers and the linkage of information from one document to another. But it does not show the semantics of data. Thus the widely used web standards are human based, rather than machine-based. Also, there is an ambiguity in the natural language. For e.g.: The word 'father' in WorldNet lexical database contains different concepts. Father can be a male parent or can be used to address male priests in some churches (Padre).

Revised Manuscript Received on April 05, 2019.

Aman Jolly, Assistant Professor, KIET Group of Institutions, Ghaziabad, Uttar Pradesh, India.

Dr. Shailender Kumar, Assosciate Professor, Delhi Technological University, Delhi, India.

It can be someone who is having an important or distinguished position in some organization. It can also be used interchangeably for representing God, founder, beginner, founding father and don (the head of an organized crime family). So in the natural language, the interpretation is complex because same words are actively used in different context and different words are used to specify the same concept that is depicted in Figure 1. Due to this, traditional information retrieval system that usually involves string matching and comparison of text and character show multiple irrelevant results.

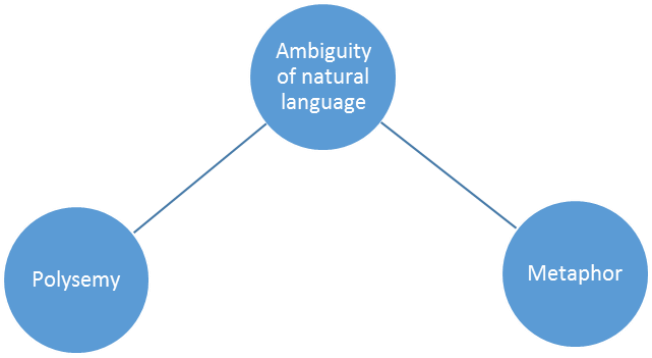

Fig. 1. Ambiguity of natural language

For successful communication in natural language, the syntax and meaning (semantics) of the information must be interpreted correctly. The understanding (correct interpretation) depends upon the context of sender and receiver and pragmatics of the sender. The context depends on the experience of both sender and receiver. In the traditional web, there is no provision for extracting implicit semantics which means that the information is there but it has hidden meaning, this makes it difficult to understand the correct meaning. Semantic web (web 3.0) [1] is the extension of web 2.0 (the web of documents) is known as the web of data. Web 3.0 is a web as a huge decentralized knowledgebase of machine accessible data. According to T B Lee, there is a huge potential in the web when human-readable text is mixed with machine-readable data [2]. To make the documents on the web read and interpret correctly by the machine, web content is explicitly annotated with semantic metadata [9]. This semantic metadata encodes the meaning of contents of the web that can be read and interpreted correctly by machines [2].

II. LITERATURE SURVEY

In this section, the prominent work achieved in the field of semantic web technologies and its applications in various domains have been addressed. In this survey, tools and related key technologies that have been applied to achieve their proposed objectives.

Rafael S. Gonçalve [5] used Semantic web linked data for data acquisition. They collected the data from the input gathered from these form as semantically aware ontologies. They also performed data acquisition from form ontology to perform inference for investigating the eligibility for disability benefits using SPARQL query.

Sergio Cerón-Figueroa [6] implemented pattern classification for matching ontology instances. They have referred data sets from ADRIADNE and MERLOT, in Learning Objects Metadata format. They also proposed a new pattern classification model to match instances from different ontologies which are related to e-learning material which is used in the same context of the knowledge society. Their model showed high accuracy result when compared with some best existing method of ontology matching [6].

M. English [7] has proposed his work based on trust layer of semantic web layer cake that can be implemented in blockchain technologies used in bitcoin where it was described that how a blockchain technology can have the potential that can be applied in making semantic web architecture more robust and resilient by fulfilling three key requirements in the Uniform resource identifiers that are security, human readability and decentralization. They also presented how data can be stored in a blockchain and the concept of distributed trust can contribute to the linked data system in the semantic web. In addition to that, they formalized a path where semantic web linked data approach can contribute to blockchain technique. They proposed a model of representing the transaction in RDF so that the semantic representation can ensure trust at the human understandable level, unlike present system where trust is established only at technical level [7].

C. Fluit [8] has presented his work on Ontology Data analysis, querying and data exploring for inferencing and described a cluster map technique for visualizing ontologies, representing classes and their hierarchies, and also described three key applications of Clustered mapping technique that are as follows:-

- Dope browser
- Xarop/SWAP: Peer-to-Peer Knowledge Management
- Aduna AutoFocus

III. TECHNIQUES USED

This section, various techniques that have been used in making the linked data and extracting implicit knowledge from them have been addressed.

A. Resource Description Framework

Resource description framework [3] can be used to represent simple facts of knowledge. In the semantic web, the information exchange is done in the standardized format. The Resource description framework is the basic building block of information exchange in the web of data that is used in the representation of knowledge in a standard schema. It is introduced because XML schema has multiple ways to represent a fact which are standardized using RDF [4].

- Resource: indicates an abstract concept, identified via URI.
- Description: depicts the properties and relationship among the Resources describing as a graph.
- Framework represents the amalgamation of Web protocols such as HTTP, XML, URI, etc.

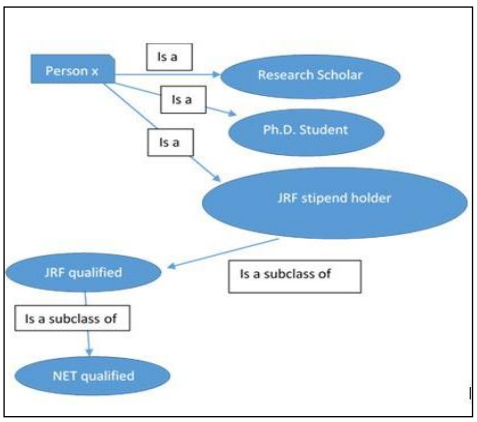

Fig. 2. Defining Semantic Explicit Knowledge.

Now suppose there is an individual "Person X" who explicitly belongs to a class of "Research scholar", "Ph.D. student" and several other classes such as "JRF stipend holder" as depicted in Figure 2. The "JRF stipend holder" class is a subclass of the "JRF qualified" then it can deduce the hidden knowledge that Person x is "JRF qualified" and "NET qualified" from the given explicit information as shown in Figure 3.

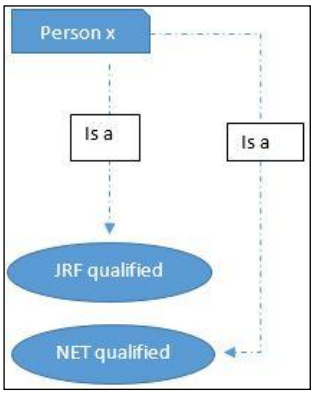

Fig. 3. Logical Inference of Implicit Knowledge.

The meaning of information is made explicit by formal (structured) and standardized knowledge representations (ontologies). Thus it is possible to relate and integrate heterogeneous data, process the information automatically and deduce implicit (non-evident) information in an automated way [14]. The semantic web is a global database that contains a semantic universal network of semantic prepositions [10].

B. SPARQL (Querying RDF)

SPARQL query is used for making a logical inference in semantic web linked data. The linked open data which is programmed in resource description framework, where data is denoted in a triple-based data model which is a linked data representation, a query language is needed to acquire the linked data [10]. SPARQL is a W3C standard inspired by SQL, which has three components, i.e. a query language for RDF graph traversal second is a protocol layer for using SPARQL via HTTP and third is the XML output format specification for results. SPARQL queries fetch linked data and represent it as tables for user view. The basic query components include a Prefix Set and a Variable set. Prefix set is the endpoint locators and common Prefix is to be used to have concise queries whereas Variable Set includes variables with name lead by a question mark like ?V1, ?V2 etc. The SELECT-WHERE statements – comprises of two components namely a set of question words and a question pattern. The SELECT statement chooses which data is to be displayed whereas The WHERE keyword indicates the selection pattern, i.e. the data specified to pull out. Another component is the result set which is used to show the results in tabular structure or in XML or Text formats. Due to ease of typing XML based linked data formats, any text editor can be used to program linked data. Sublime Text is an open source text editor that is used in our research. The installation package was directly downloaded from their official website. The linked data is programmed in the turtle syntax notation. There are different tools which are used to execute SPARQL queries. Twinkle 2.0, Jena ARQ 3.5 and Protégé 5.0 have been used in this research. Twinkle [11] is the tool to implement SPARQL Query. For installing Twinkle tool, JDK 1.5 or higher should be installed and configured. Jena framework is developed in java which is used to build semantic web application whereas Jena ARQ [13] is the SPARQL query processor for Jena. Protégé [12] is a free, open-source, GUI based tool for semantic web development.

IV. EXPERIMENT AND RESULTS

In this section, we are making linked data and infer new knowledge that was explained in Figure 2 and Figure 3 by using SPARQL queries in different tools. The experiment is demonstrated by a flowchart depicted in Figure 4.

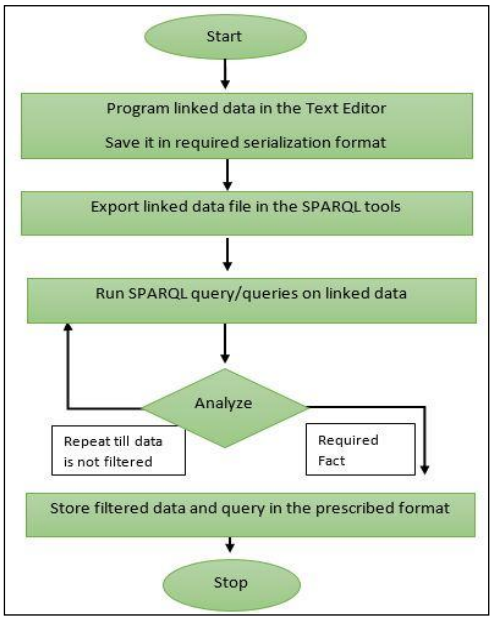

Fig. 4. Steps Involved in the Experiment for Extracting Semantic Inference.

A. Results

These results show that how the data can be extracted and filtered to such extent that logical inference can be calculated by using SPARQL query. The set of queries have been performed in order to deduce that named individual "Person_X" belongs to "JRF_qualified" and "NET_qualified" class. The data was exported and the SPARQL query was used in Twinkle 2.0, Jena ARQ 3.5 and Protégé 5.0 to infer the following result in different tools.

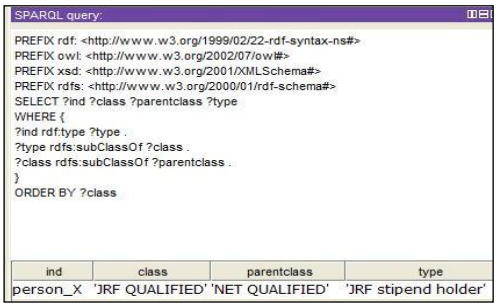

Fig. 5. SPARQL results in PROTÉGÉ 5.0

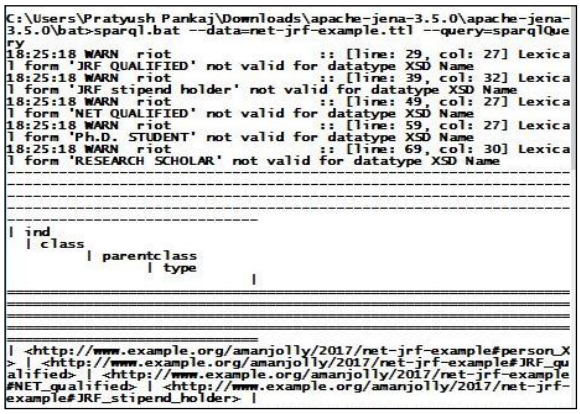

Fig.6. SPARQL results in JENA ARQ 3.5

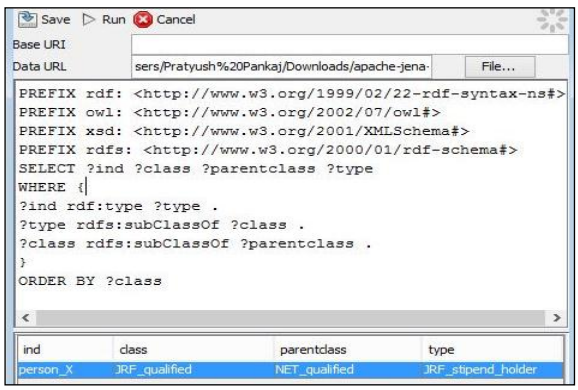

Fig.7. SPARQL results in TWINKLE 2.0.

. Figure 5 shows the result that was extracted in Protégé 5.0 whereas Figure 6 and Figure 7 shows the result that was extracted in Jena ARQ 3.5 and Twinkle 2.0 respectively by importing linked data explained in Figure 2 and Figure 3. The result that is shown in Protégé 5.0 is represented in the tabular form as depicted in Figure 5. Whereas results that are extracted by Jena ARQ 3.5 are shown in a textual format as shown in Figure 6. However, Twinkle 2.0 gives the flexibility to the user to visualize data in both textual as well as tabular format. In figure 7 the SPARQL results of Twinkle 2.0 are shown in tabular format.

After checking the SPARQL query result in Twinkle 2.0, Jena ARQ 3.5, Protégé 5.0, the serialization format of linked open data that was imported in the these tools in turtle syntax notation was changed to RDF/XML, OWL/XML, LATEX, N3 (.n3), N-Triples (.nt), Manchester format, KRSS2 & OBO formats in order to check the support of these formats in the SPARQL tools. Based on their support of different formats and features a comparative analysis is evaluated in Table 1.

Table 1 Comparison of Twinkle 2.0, Jena ARQ 3.5 & Protege 5.0.

Serialization & Features supported	Twinkle 2.0	Jena ARQ 3.5	Protégé 5.0
1) RDF/XML support	✓	✓	✓
2) OWL/XML support	×	×	✓
3) TURTLE(.ttl) support	✓	✓	✓
4) LATEX support	×	×	×
5) N3 (.n3) support	✓	✓	✓
6) N-Triples(.nt) support	✓	✓	✓
7) Manchester support	×	×	✓
8) KRSS2 support	×	×	✓
9) OBO support	×	×	✓
10) Query saving feature	✓	×	×
11) Query loading from file	✓	✓	×
12) Format Independent Parsing	×	×	✓
13) Graphical user interface	✓	×	✓
14) Long running SPARQL queries cancellation	✓	×	✓

A. Observations

From above analysis, it is clear that parser(s) of Twinkle 2.0 and Jena ARQ 3.5 are file extension dependent which means that it parses the linked data file based on its extension. For example, if the extension of the linked data file which is programmed in turtle notation and saved in '.ttl' file extension format. It scans the data with the turtle parser but if same data is saved in '.n3' file extension format, then the parser fails to parse the data because the n3 parser is trying to parse turtle notation. Whereas protégé 5.0 doesn't scan linked data file extension before parsing but it checks the linked data file content with all its available parsers. The one who correctly interprets the linked data is correctly selected as a valid parser.

V. CONCLUSION AND FUTURE WORK

The paper offered an empirical study of how logical inference can be extracted from semantic web linked data. A Practical approach of extracting implicit data has also been provided by filtering the data set in Twinkle 2.0, Jena ARQ 3.5, Protégé 5.0 using SPARQL query in aggregation with a generated linked dataset that was programmed in Turtle syntax in a triple format using a Text editor. The visualized data graph has been provided for the same. The comparative analysis of these tools based on their features and support on various serialization formats have been evaluated in the tabular format. File extension dependent feature in SPARQL tools limits its functionality of parser support but it has its own advantage of execution speed than those who employ file extension independent feature. There are several extensions which can be clearly done to this work. For future work, this research may be extended to develop a hybrid framework that employs both file extension dependent and independent feature for SPARQL tools that can support various new serialization formats. Semantic web technology and their applications in various domains have proven their worth in giving an intellectual way of resolving various dynamic and real-time problems by exploring uncharted areas of the web.

REFERENCES

1 Berners-Lee, T., Hendler, J. and Lassila, O. "The semantic web." Scientific American (2001).
2 Berners-Lee, Tim., Semantic web roadmap, 1998, http://ww w.w3.org/DesignIssues/Semantic.html.
3 https://www.w3.org/2001/sw/wiki/RDF.
4 https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.
5 Gonçalves, R S.," An ontology-driven tool for structured data acquisition using Web forms."(2017). In Journal of Biomedical Semantics (Vol. 8, p. 26). Journal of Biomedical Semantics.
6 Figueroa, S C., "Instance-based ontology matching for e-learning material using an associative pattern classifier." (2017). In Computers in Human Behavior (Vol. 69, p. e53). Elsevier Ltd.
7 English, M., "Block Chain Technologies & the Semantic Web : A Framework for Symbiotic Development." (2016). In Computer Science Conference for University of Bonn Students (p. 47–61).
8 Fluit, C., "Ontology-Based Information Visualization: Toward Semantic Web Applications." (2006). In Visualizing the Semantic Web (pp. 45–58).
9 Hitzler, P., Krotzsch, M., & Rudolph, S. "Foundations of semantic web technologies." (2009). CRC press.
10 DuCharme, B., "Learning SPARQL: querying and updating with SPARQL 1.1." (2013). O'Reilly Media, Inc. p. 1-16.
11 Twinkle 2.0 Download Link -
12 http://www.ldodds.com/projects/twinkle/twinkle-2.0-src.zip. Protégé 5.0 Download Link -
13 https://github.com/protegeproject/protegedistribution/releases/tag/protege-5.0.0#downloads.
14 Jena ARQ 3.5 Download Link -http://www-eu.apache.org/dist/jena/binaries/apache-Jena-3.5.0.zip.
15 Malik, S. K., & Rizvi, S. (2012)." A framework for SPARQL query processing, optimization and execution with illustrations." International Journal of Computer Information Systems and Industrial Management Applications
16 Kumar, S., & Kumar, S. "Semantic Web attacks and countermeasures." 2014 International Conference on Advances in Engineering and Technology Research, ICAETR 2014.

AUTHORS PROFILE

Aman Jolly is currently working as an Assistant Professor in KIET group of institutions. He has done his B.Tech in computer science from GGSIPU and M.Tech in Information Security from GGSIPU.

Dr. Shailender Kumar is currently working as an Associate Professor at Delhi Technological University. He is having specialization in Database Systems, Data mining, Big Data Analytics, Machine Learning and Information Security