Challenges in Information Mining from Semantic Web Data

R. Gomathi, S. Logeswari

Abstract - The quantity of web pages is growing in our day to day life and quantity of semantic web data on the web is also growing quickly. With the raise in information, the semantic web data also gets increased. This becomes a policy for the researchers to use data mining algorithms to manipulate the semantic web data. Although Semantic web has a wide and extensive selection of applications, there is a less amount of investigation in applying the data mining techniques in semantic web. This manuscript explains the research behaviour in the semantic web applications and data mining. It also addresses the challenges faced in this area of investigation.

Keywords: Semantic Web Data, Internet, Data Mining, RDF Data

I. INTRODUCTION

In the current scenario of information expertise, data in the web is not stored in a single processor. Instead data is stored in different computers and so the access of the data becomes hard. Semantic web data is stored in different formats which includes the Resource Description Format (RDF), RDF/XML, N3, Turtle, N-triples and OWL (Web Ontology Language). There are a wide range of functions for the semantic web. Since data is in stored in different formats, the semantic web data administration in a mind-numbing task. Also mining information from the semantic web data to extract knowledge becomes a challenging issue. Web mining refers to the method of applying data mining algorithms to web logs[1], the contents and the structures. This research concentrates on the concepts of semantic web and the role of data mining in extracting knowledge from the semantic web information. The problem of applying the data mining procedures into the semantic web helps in incorporating the domain knowledge [2].

II. SEMANTIC WEB

Internet consists of a group of web pages and the meaning of the web page is stored using the semantic web technology. A review of in progress study in the area of semantic web mining is performed in literature [3].

The mechanism of semantic web consists of the following elements:

- SPARQL
- XML (Extensible Markup Language)
- XML Schema
- RDF
- RDF Schema
- OWL(Web Ontology Language)

The commonly used structure to articulate information regarding the Semantic Web facts [4]. RDF intends to connect data resource on the web. It is basically a data copy that makes applications to easily process and control the web data. It concentrates on the exchange of data and interoperability.

The results of applying data mining techniques on the learning objects meta data was performed in research [5]. In RDF,a resource refers to a a web page or a collection of many web pages. In the context of RDF, a resource is indicated by a unique URI(Uniform Resource Identifier). The basic component of RDF model is the triple. An RDF file is a collection of triples. There are three different representations of RDF data. They are

- Triples
- RDF graph
- RDF/Extensible Markup Language(XML) format

RDF is a data model designed for providing metadata [6] to describe web resources. A triple consists of subject, predicate and object. The other names of subject, predicate, object are resource, property, literal respectively. The predicate is a property which describes the association between subject and object. The subject and predicate of a triple is URI's and the object is a URI or may be a literal.

For example, consider the following sentence,

Ram is the owner of the resource http://www.xx3.org/home/Ram.

The above statement consists of the following parts,

Subject part : http://www.xx3.org/home/Ram
Predicate part: owner
Object part : Ram

The above three parts constitute a triple.

A example set of RDF triples are shown in Table 1.

Table 1 A Sample Group of RDF Triples

Subject	Predicate	Object
_:N8f8fa34ea91949b9be72008e6cf3f8e4	<http://www.xx3.org/2007/vcard-rdf/3.0#Given>	"masah"
_:N8f4fa67ea91949b9be72008e6cf3f8e3	<http://www.xx3.org/2007/vcard-rdf/3.0#Family>	"Janet"
<https://some/RebeccaSmith/>	<http://www.xx3.org/2007/vcard-rdf/3.0#FN>	"Becky Smith"
<http://someplace/RebeccaSmith/>	<http://www.xx3.org/2007/vcard-rdf/3.0#N>	_:Ncbc191cb7d68458b80ad05e8b45eff86
<www://somewhere/SarahJones/>	<http://www.xx3.org/2007/vcard-rdf/3.0#FN>	"masah Janet"

The design of the RDF graph is revealed in Figure 1.

Revised Version Manuscript Received on 25 November, 2018.

Dr. R. Gomathi, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam-638401, India.
Dr. S. Logeswari, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam-638401, India.

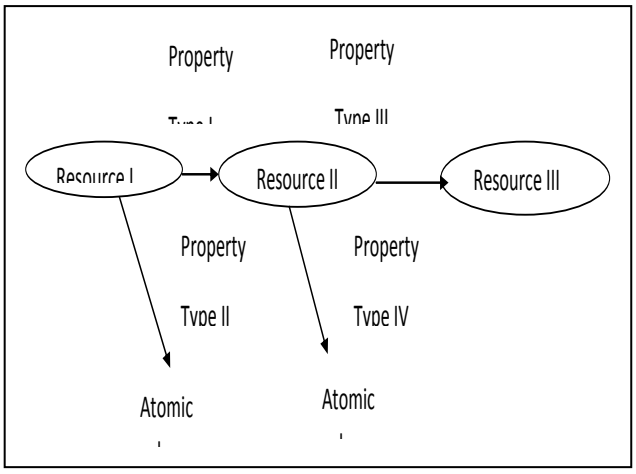

Figure 1 Design of an RDF graph

The RDF statements are also pictorially represented as RDF graphs with nodes and arcs. An RDF graph refers to a directed graph with subjects and objects of triples in the nodes and predicates labeled at the edges of the graph.

A third type of representing RDF data is the RDF/XML format. An example RDF/XML format for the sentence "Ram is the owner of the resource http://www.xx3.org/home/Ram." is written as

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:rdf="http://www.xx3.org/1999/02/22-rdf-syntax-ns#" xmlns:s="http://description.org/schema/">
 <rdf:Description about="http://www.xx3.org/home/Ram">
  <s:owner>Ram</s:owner>
 </rdf:Description>
</rdf:RDF>
```

Different query languages exist for querying the RDF data. The most popular of them is the SPARQL Protocol and RDF Query Language (SPARQL). It's alike to Structured Query Language (SQL) and is used for querying the RDF data [7].

SPARQL is the defacto RDF query language and it is the recommendation of World Wide Web Consortium (W3C).

The syntax of a SPARQL query consists of the following elements,

a) Prefixes and their declaration
b) Definition of Dataset
c) Result
d) Query pattern
e) Query modifiers

An example of the syntax is given below,

```
# Declaration of Prefix foo:http://yyy.com/resources/>
.....
# dataset definition

FROM.....
# result clause
SELECT .....
# query pattern
WHERE {
.....
}
# query modifiers
ORDER BY .....
```

The following is a simple example for a SPARQL query,

```
SELECT ?x
WHERE { ?x <http://www.xx3.org/2001/vcard-rdf/3.0#FN > "John Jones" }
```

The above sample query is executed on the RDF dataset in Table1.1 and it lists out the URI for firstname="John Jones". The sample output is

```
| x                                  |
| <http://somewhere/JohnSmith/> |
```

The SPARQL query language supports four types of queries. They are

- SELECT
- CONSTRUCT
- ASK
- DESCRIBE

The query appearance takes a WHERE block to confine the query with some conditions. In the case of the DESCRIBE query the WHERE block is non-compulsory. This research focuses on simple SELECT queries with varying number of predicates.

The following are the key areas in which semantic web is applied

- Knowledge management
- Ontologies
- E-commerce services
- Web services

III. DATA MINING

A part of research which is applied in many fields like statistics, machine learning and databases, communications, storage technologies and a large collection of scientific and commercial data is the data mining.

The purpose of data mining is to predict or describe information. Prediction focuses on using few variables in the data to calculate unidentified or future values of other variables. Description aims at finding human interruptible patterns to describe data. Data mining depends on three technologies which includes [8]

- Mass data collection
- Dominant computers
- Data mining algorithms

Data mining uses the following techniques in common [9],

- Artificial neural networks
- Decision trees
- Classification
- Association
- Regression
- Machine Learning
- Genetic algorithms
- Nearest neighbor method
- Rule induction

Semantic web mining technology is a new area of research which combines semantic web and web mining.

Study in this area is carried out in the present era and the research faces a lot of issues and challenges. It is possible to bring new knowledge based on learning matters stored in databases. Semantic associations can also be set up among the attributes that describe learning objects.

IV. SEMANTIC WEB MINING CHALLENGES

When applying data mining techniques to semantic web data, more than a few probable challenges appears. The barrier faced by the researchers and solutions to agreement with some complex issues were also discussed.The following are the issues and challenges [10]

Challenges in audio/video data mining - There is no standard support available to extract valuable information from audio and video files available in the web.

Quality in searching the semantic web data based on keywords – The keyword based searches suffers from problems like search returning many answers, returning low quality results and missing of many pages related to our search.

Querying the semantic web – The execution of a query on the semantic web data takes more time because of constraints like availability of large amount of data, data in different formats.

Lack of global standards in semantic web mining – The number of international standards are available on Semantic Web Mining is very less. Non-availability of broad, rough and internationally recognized set of principles addressing a mix of semantic web, web mining and unstructured data mining is a challenge being faced by researchers from corner to corner in the world.

Multi-dimensional data analysis and mining – Multi dimensional data analysis is apprehensive with a variety of different tools and methods that have been developed to question existing data, discover exceptions, and verify hypotheses from multiple sources. Gathering data from multiple dimensions and then performing analysis by combining them is a tedious task and suffers from issues like execution time, missing of some data while retrieval and so on.

Semantic search engines use one or more of the following methodologies

- Contextual investigation
- Logic engines
- Natural language understanding
- Ontology search

V. CONCLUSION

Semantic Web Mining is a pioneering and fast-growing investigation field integrating data mining and Semantic Web. In this document a detailed modern examination of on-going examination in the field of Semantic Web Mining has been described. These readings analyze the challenges and issues that researchers face in this area of research.

REFERENCE

1. O. Mustapaşa, A. Karahoca, D. Karahoca and H. Uzunboylu, “Hello World, Web Mining for E-Learning,” Procedia Computer Science, Vol. 3, No. 2, 2011, pp. 1381- 1387. doi:10.1016/j.procs.2011.01.019
2. H. Liu, “Towards Semantic Data Mining,” Proceedings of the 9th International Semantic Web Conference, Shanghai, 7-11 November 2010, pp. 1-8.
3. D. Jeon and W. Kim, “Development of Semantic Decision Tree,” Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, Macau, 24-26 October 2011, pp. 28- 34.
4. A. Jain, I. Khan and B. Verma, “Secure and Intelligent Decision Making in Semantic Web Mining,” International Journal of Computer Applications, Vol. 15, No. 7, 2011, pp. 14-18. doi:10.5120/1962-2625
5. A. Segura, C. Vidal-Castro, V. Menéndez-Domínguez, P. G. Campos and M. Prieto, “Using Data Mining Techniques for Exploring Learning Object Repositories,” The Electronic Library, Vol. 29, No. 2, 2011, pp. 162-180. doi:10.1108/02640471111125140
6. Hitzler, P, Krotzsch, M & Rudolph, S 2011, ‘Foundations of Semantic Web technologies’, CRC Press.
7. DuCharme, B 2013, ‘Learning Sparql’, O'Reilly Media.
8. http://www.thearling.com/text/dmwhite/dmwhite.htm
9. Shah Neha K,” Introduction of Data mining and an Analysis of Data mining Techniques”, Indian Journal of Applied Research,vol.3,No.5,2013.
10. T. Sunil kumar,, Dr. K. Suvarchala,” A Study: Web Data Mining Challenges and Application for Information Extraction”, IOSR Journal of Computer Engineering,vol.7,No.3,pp.24-29,2012.
11. Menemencioglu, Oguzhan, and Ilhami M. Orak. "A review on semantic web and recent trends in its applications." In Semantic Computing (ICSC), 2014 IEEE International Conference on, pp. 297-303. IEEE, 2014.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication