Data Mining in Semantic Web Data

K. Chomboon, N. Kaoungku, K. Kerdprasop, and N. Kerdprasop

***Abstract*—This research aims at studying the data mining role in semantic web data. Semantic web is popular in a variety of different applications, but research in data mining in semantic web data, appears less. As open source software for data mining in semantic web open source is minimal, and data model of the semantic web requires RDF or OWL format. These specific formats cannot be used directly in most data mining tools. We thus propose a methodology to mine data that appear in an RDF format. The mining process has been demonstrated through the use of R packages.**

***Index Terms*—Data mining, semantic web, R language.**

I. Introduction

Current data is not stored on a single computer, because the current is the era of information technology and social media, data can be stored in many computers on the internet, is difficult for them to access data quickly and easily. The researchers presented the technology to help manage these data called semantic web. The data in the format or the same specification as RDF/XML, N3, Turtle, N-Triples and OWL.

Semantic web [1], [2] has been used in various fields such as Information Systems, Search Engine etc. Large data technology to handle with this is data mining, because the large data analyzed find patterns or relationships of data is an advantage of data mining. Research in the field of data mining in semantic web data is not yet widely, since there is a management tool for data mining of semantic web is less, and data from the semantic web is stored in a format that cannot be used directly in data mining. The research in data mining has appeared very little.

Research in the field of data mining in semantic web data applied to various algorithms of data mining, such as data classification, association rule mining etc. Most research using the licensed software such as Microsoft Data Mining Extension (DMX) which is Microsoft SQL Server.

From the above it can be seen that the present data are not stored on a single computer always, is difficult to put that information in the internet is analyzed find patterns or relationships with the data mining. This research has proposed methods for data mining in semantic web data.

II. Background

A. Semantic Web

Semantic web, have been developed since the storage is only human to understand the meaning, but the machine cannot understand it, because data without structure. Semantic web has been developed to provide useful data on the Internet that can be analyzed and applied to various tasks. The language used for defining the data structure is RDF [3]-[5] (Resource Description Framework). Which is written in the form of sentences consists of the subject, predicate and objects show in Fig. 1.

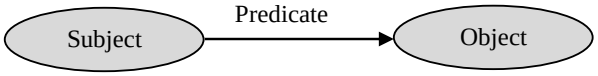

Fig. 1. RDF triples.

The standardization for semantic web in the context of web 3.0 shows in Fig. 2.

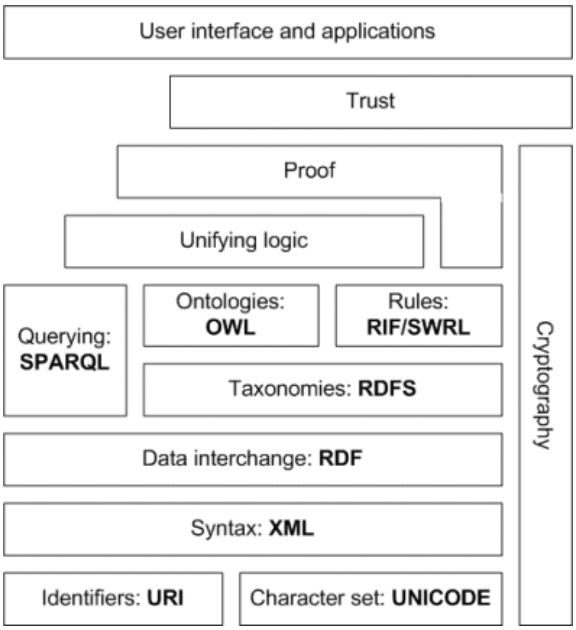

Fig. 2. Standard of semantic web in the context of web 3.0.

The components of semantic web are as follows:

- XML stands for Extensible Markup Language: XML is a markup language much like HTML, but XML was designed to transport and store data, not to display data. XML provides an elemental syntax for content structure within documents.
- XML Schema: XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.
- RDF stands for Resource Description Framework: RDF is a language for expressing data models. RDF was designed to provide a common way to describe information so it can be read and understood by computer applications.
- RDF Schema: RDF Schema extends RDF and is a

Manuscript received December 13, 2013; revised March 14, 2014. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

The authors are with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: chomboon.k@gmail.com).

vocabulary for describing properties and classes of RDF-based resources.
- OWL stands for Web Ontology Language: OWL was designed to provide a common way to process the content of web information. OWL and RDF are much of the same thing, but OWL is a stronger language with greater machine interpretability than RDF. OWL comes with a larger vocabulary and stronger syntax than RDF.
- SPARQL: SPARQL is a protocol and query language for semantic web data sources.

B. XML

XML stands for Extensible Markup Language. XML was designed to carry data, not to display data. Tags are not predefined. The components of XML are as follows:

1) Tree structure

XML documents must contain a root element. This element is "the parent" of all other elements. All elements can have sub elements. The example shows as Fig. 3.

```
<root>
<child>
<subchild>…….</subchild>
</child>
</root>
```

Fig. 3. Show tree structure of XML documents.

2) Syntax

The syntax rules of XML are very simple are shows as follows:

- All XML elements must have a closing tag:
   This is a paragraph
- XML tags are case sensitive:
  <Note>This is incorrect<note>
  <note>This is correct<note>
- XML elements must be properly nested:
  This text is bold and italic
- XML documents must have a root element:
  <book>
  <title>XML Manual</title>
  <price>25.00</price
  </book>
  The root element is <book>.
- XML attribute values must be quoted:
  <book category="MANUAL">
  <title>XML Manual</title>
  <price>25.00</price
  </book>
- White-space is Preserved in XML:
  Input = Hello World
  HTML output: Hello World
  With XML, the white-space in a document is not truncated.
- Entity References.
  There are 5 predefined entity references in XML.
  < <less than
  > >greater than
  & &ampersand
  ' 'apostrophe
  "" quotation mark

3) Elements

An XML document contains XML elements. An XML element is everything from (including) the element's start tag to (including) the element's end tag. Example shows as Fig. 4.

```
<bookstore>
<book category="CHILDREN">
<title>Harry Potter</title>
<author>J K. Rowling</author>
<year>2005</year>
<price>29.99</price>
</book>
<book category="WEB">
<title>Learning XML</title>
<author>Erik T. Ray</author>
<year>2003</year>
<price>39.95</price>
</book>
</bookstore>
```

Fig. 4. Show elements of XML documents.

In the Fig. 4. <bookstore> and <book> have element contents, because they contain other elements. <book> also has an attribute (category="CHILDREN"). <title>, <author>, <year>, and <price> have text content because they contain text.

XML elements must follow these naming rules:

- Names can contain letters, numbers, and other characters.
- Names cannot start with a number or punctuation character.
- Names cannot start with the letters xml (or XML, or Xml, etc).
- Names cannot contain spaces.

4) Namespace

XML Namespaces provide a method to avoid element name conflicts. Example show as follows:

- This XML carries book properties:
  <book>
  <weight>150</weight>
  <length>29.7</length>
  <width>21</width>
  </book>
- This XML carries book information:
  <book>
  <title>Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
  </book>

If these XML fragments were added together, there would be a name conflict. Both contain a <book> element, but the elements have different content and meaning. An XML parser will not know how to handle these differences. Name conflicts in XML can easily be avoided using a name prefix show as follows:

<p:book>
<p:weight>150</p:weight>
<p:length>29.7</p:length>
<p:width>21</p:width>

```
</p:book>
<i:book>
< i:title>Harry Potter</ i:title>
< i:author>J K. Rowling</ i:author>
< i:year>2005</ i:year>
< i:price>29.99</ i:price>
</ i:book>
```

There will be no conflict because the two <book> elements have different names. When using prefixes in XML, a so-called namespace for the prefix must be defined. The namespace is defined by the xmlns attribute in the start tag of an element show as follows:

```
<root>
<p:book xmlns:p="http://www.ex.com/properties">
<p:weight>150</p:weight>
<p:length>29.7</p:length>
<p:width>21</p:width>
</p:book>
<i:book xmlns:i="http://www.ex.com/information">
< i:title>Harry Potter</ i:title>
< i:author>J K. Rowling</ i:author>
< i:year>2005</ i:year>
< i:price>29.99</ i:price>
</ i:book>
</root>
```

C. SPARQL Language

SPARQL [6] is a query language for the semantic web, which is format in RDF / XML or OWL. SPARQL language to access data through a Triple (Basic Graph Pattern) consists subject predicate and object. The main structure consists of a "SELECT" to define a variable to store the results of the query and "WHERE" as a condition for the query. Table I show example data and query with SPARQL language.

TABLE I: EXAMPLE DATA AND QUERY WITH SPARQL LANGUAGE

Data	Query

@prefix foaf: <http://xmlns.com/foaf/0.1/> .
_:a foaf:nameb "Alice" .
_:a foaf:mbox <mailto:Alice@example.com> .
_:b foaf:name "Bob" .
_:b foaf:mbox <mailto: Bob@example.org> .
_:c foaf:name "Peter" .
	PREFIX foaf: <http://xmlns.com/foaf/0.1/>
	SELECT ?name
	WHERE {
	?x foaf:name ?name .
	}

Table I shows example data and query with sparql language are as follows:

- PREFIX is defines a resource that is defined in the head.
- SELECT is set the variable to store the results of query, by define variable need the "?" before variable name.
- WHERE is the conditions used for the query (e.g. ?x foaf:name ?name etc.)

D. R Language

R language [7], [8] is a functional and object-oriented language that was developed to replace the S language for statistical, developed in 1995 from the Department of Statistics, University of Auckland, New Zealand. R language has been applied to various fields, the data mining applied R language in the research, Because of strength of the R language for data mining is to analyze large data and open-source software. In this research study the data mining in semantic web data. The command in R language for data mining in semantic web data as follows.

Rrdf is a package on R language, handling triples on ontology within "RDF/XML" format. Now we use rrdf package to transformation data on "RDF/XML" to data frame format on R. Then we easily to use data on data frame format to mining.

- load.rdf(filename, format = "RDF/XML") is command to load data from file format RDF / XML.
- sparql.rdf(model, sparql) is command to query data with SPARQL language.

III. METHODOLOGY

In this section we present the process of data mining on semantic web dataset. We use R language for implementing our method. The overview our techniques show as Fig. 5.

Step 1: use library rrdf on R language to import dataset from RDF file as data frame

Step 2: use data on R language to classification.

Step 3: use model to predict data

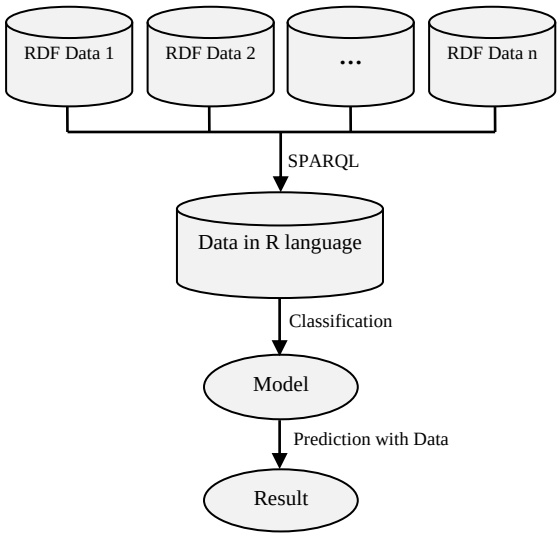

Fig. 5. The overview of techniques to mining dataset.

IV. EXPERIMENT RESULT

This research experimentation used Iris dataset from UCI Machine Learning Repository. Iris dataset has 5 attributes and 150 data instances.

Step 1: use library rrdf on R language to import iris dataset from RDF file as data frame. We can use this command to load data from RDF file.

"RdfData <- load.rdf("iris.rdf")".

Then convert data to data frame can use this command.

```
"Data <- data.frame(sparql.rdf(RdfData,
"PREFIX ir: <http://127.0.0.1/iris#>
SELECT ?SepalLength ?SepalWidth
?PetalLength ?PetalWidth
?Species { ?x ir:SepalLength ?SepalLength.
?x ir:SepalWidth ?SepalWidth.
?x ir:PetalLength ?PetalLength.
?x ir:PetalWidth ?PetalWidth.
?x ir:Species ?Species.}))"
```

Convert iris data to data frame on R language (show in Fig. 6).

```
Console ~/R/
> Data <- data.frame(sparql.rdf(RdfData,"PREFIX ir: <http://127.0.0.1/iris#> SELECT ?
SepalLength ?Sepalwidth ?PetalLength ?PetalWidth ?Species{ ?x ir:SepalLength ?
SepalLength. ?x ir:Sepalwidth ?Sepalwidth. ?x ir:PetalLength ?PetalLength. ?x
ir:Petalwidth ?Petalwidth. ?x ir:Species ?Species. }"))
> Data
    SepalLength SepalWidth PetalLength PetalWidth    Species
1           6.1        2.6         5.6        1.4  virginica
2           5.1        2.5         3.0        1.1 versicolor
3           5.4        3.7         1.5        0.2     setosa
4           7.9        3.8         6.4        2.0  virginica
5           6.7        3.1         4.7        1.5 versicolor
6           5.4        3.4         1.5        0.4     setosa
7           7.3        2.9         6.3        1.8  virginica
8           4.5        2.3         1.3        0.3     setosa
9           6.9        3.2         5.7        2.3  virginica
10          4.9        3.6         1.4        0.1     setosa
11          6.0        2.7         5.1        1.6 versicolor
12          6.7        3.0         5.0        1.7 versicolor
13          5.0        3.5         1.3        0.3     setosa
14          4.4        3.0         1.3        0.2     setosa
15          6.0        2.2         4.0        1.0 versicolor
```

Fig. 6. Convert iris data to data frame on R language.

Step 2: then use data on R language to classification. First we can generate model form data with this command.

"model <- ctree(Species ~ ., data = Data)", use column Species to decision show in Fig. 7.

```
Console ~/R/
> model <- ctree(Species ~ ., data = Data)
> model

  Conditional inference tree with 3 terminal nodes

Response:  Species
Inputs:  SepalLength, SepalWidth, PetalLength, PetalWidth
Number of observations:  148

1) PetalLength == {1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.9}; criterion = 1, statistic
= 266.364
  2)*  weights = 48
1) PetalLength == {3.0, 3.3, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6,
4.7, 4.8, 4.9, 5.0, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 6.0, 6.1, 6.3, 6.4, 6.6,
6.7, 6.9}
  3) PetalWidth == {1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7}; criterion = 1, statistic =
80.388
    4)*  weights = 54
  3) PetalWidth == {1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5}
    5)*  weights = 46
```

Fig. 7. Generate model from data use Species to decision.

Step 3: we use model to predict data and view the result of model with this command "table (predict(model, data = Data, Data[,5])", show in Fig. 8.

```
Console ~/R/
> table(predict(model,data = Data), Data[,5])

             setosa versicolor virginica
  setosa         48          0         0
  versicolor      0         49         5
  virginica       0          1        45
```

Fig. 8. Result of the model.

V. Conclusion

This research aims to study how to use dataset from semantic web in format RDF/XML and apply to mining with R language. Because R language is open source program, we can use library is already in R and easy to create function for mining data. We can use semantic web dataset or open linked dataset to improve performance of data mining.

References

[1] P. Hitzler, M. Krötzsch, and S. Rudolph, "Foundations of semantic web technologies," *Textbooks in Computing*, Chapman and Hall/CRC Press, 2009.

[2] V. Nebot and R. Berlanga, "Finding association rules in semantic web data," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 51-62, 2012.

[3] E. Willighagen. (2013). RRDF - support for the resource description framework. [Online]. Available: http://cran.r-project.org/web/packages/rrdf/rrdf.pdf

[4] W3C. (2004). Resource Description Framework (RDF): Concepts and abstract syntax. [Online]. Available: http://www.w3.org/TR/rdf-concepts/

[5] W3C. (2008). SPARQL query language for RDF. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[6] C. Kiefer, A. Bernstein, and A. Locher, "Adding data mining support to SPARQL via statistical relational learning methods," *The Semantic Web: Research and Applications*, vol. 5021, pp. 478-492, 2008.

[7] K. Kerdprasop. (2012). Data mining with R. [Online]. Available: https://sites.google.com/site/kittisakthailand55/home/datamining2-55

[8] E. Paradis, J. Claude, and K. Strimmer. "APE: Analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289-290, 2004.

Kittipong Chomboon is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology, Thailand in 2012, and master degree in computer engineering from Suranaree University of Technology, Thailand in 2013. His current research includes ontology and classification.

Nuntawut Kaoungku is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology, Thailand in 2012, and master degree in computer engineering from Suranaree University of Technology, Thailand in 2013. His current research includes semantic web and association.

Nittaya Kerdprasop is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in Computer Science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes knowledge discovery in databases, artificial intelligence, logic programming, and intelligent databases.

Kittisak Kerdprasop is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, U.S.A. in 1999. His current research includes data mining, artificial intelligence, functional and logic programming languages, computational statistics.