International Journal of Advances in Computer Science & Its Applications – IJCSIA
Copyright © Institute of Research Engineers and Doctors
Volume 9 : Issue 1- [ISSN : 2250-3765] - Publication Date: 10 May, 2019

Semantic Framework for Big Data in Chemoinformatics

Jungkee Kim

Abstract— **In recent years, the new data sources for chemical compounds, proteins, and genes have been widely spread. However, the sources are commonly isolated and cannot be connected each other. The relationships between the data sources and connections produce a semantic integration. The integration between chemical data resources and life science information provides a better chance to understand some aspect of small molecules in biological systems. In this paper, we review the existing researches on Resource Description Framework technologies in life sciences and propose to apply a big data framework to process big data for such technologies.**

Keywords— **Big Data, RDF, Chemoinformatics, Semantic Network**

I. Introduction

Big data is produced from the explosion in the current scientific data growth as a result of the developing technologies. Scientists wish to find new and exclusive patterns in data sets that explain scientific phenomena not yet known. The detection of new patterns showing the causes and effects of various scientific phenomena is often beyond the compass of single datasets. For example, systems biologists study the complex biological systems that integrate microarray datasets to biological pathways, using a large number of other data sets to provide evidence for the links [1].

Another notable example of the link is drug discovery, in which a new distinct chemical entity is designed or discovered based on the descriptions of its required chemical properties. This process requires the effective link of many scientific datasets that are both spread and growing independently each other[2].

Chemical Entities of Biological Interest (ChEBI) [3] is an ontology of molecular entities based on molecular chemical compounds as well as conventional structural databases. The ontology is part of Open Biomedical Ontologies (OBO) [4] that is developed as an effort to create the semantic connections among different biological and medical domains. The molecular entity is an unique atom, molecule, ion, radical, and complex, etc., which is identifiable as separately distinguishable entity. Ontology for Biomedical Investigations (OBI) [5] is another part of OBO effort and the BOI is an integrated ontology for the description of life science and clinical investigations.

ChEMBL[6] is a chemical database of bioactive molecules with drug chemical compound properties. It contains descriptions for chemical movements involving over a million chemical entries. This provides an exclusive resource for drug researcher. It is updated on a properly frequent basis as the existing data is well-maintained and new data is added. The ChEMBL dataset is available for downloading and can be browsed through a web. The ChEMBL is also partly mapped to RDF by European Bioinformatics Institute [7] and other groups [8, 9].

This paper introduces RDF and proposes to apply a big data frame work for RDF in life sciences.

II. Resource Description Framework

The Resource Description Framework(RDF)[10] is a W3C recommendation for a standard representation of metadata. This framework is described in XML format. RDF has an innate function for machine-oriented data exchanging between applications because of its XML features. XML and RDF provide semantic interoperability in the current Web domain, but XML only describes the document structure. RDF emphasizes semantic meaning on the Web resources by adding a capability as a data model for knowledge representation.

The basic block of RDF consists of three object types - resources, properties, and statements. A resource is anything that can be written as a Uniform Resource Identifier (URI) in the RDF expression. It can be not only a Web page but also an XML element. Anything written in URI could be a resource. A property is a specific characteristic, attribute, or relation of the resource - for example, "owner." Each property has a specific meaning, which can be classified by a schema related to the name of the property. A statement is a combination of a resource, a property, and a value. Each part of a statement is also known as the subject, the predicate, and the object. The object can be another resource or a literal, which might be any string or XML. Figure 1 presents an example of RDF graph for the Web page of the University.

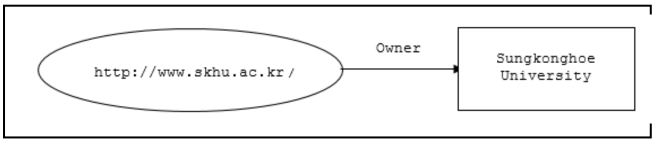

Figure 1. An Example of RDF graph

Jungkee Kim
Sungkonghoe University
Korea

In the figure, the oval shape node denotes a subject, the arc denotes a named property, and the rectangular shape is a node, which represents a literal. The graph represents the following statement:

"Sungkonghoe University is the owner of the resource http://www.skhu.ac.kr/."

The statement can be written in the XML format:

```
<?xml version="1.0">
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1992/02/22-rdf-syntax-ns#"
 xmlns:rdf="http://descriptio.org/schema/">
<rdf:Description about="http://www.skhu.ac.kr/">
  <s:owner>Sungkonghoe University</s:owner>
</rdf:Description>
</rdf:RDF>
```

The basic block of RDF consists of three object types - resources, properties, and statements. A resource is anything that can be written as a Uniform Resource Identifier (URI) in the RDF expression. It can be not only a Web page but also an XML element. Anything written in URI could be a resource. A property is a specific characteristic, attribute, or relation of the resource - for example, "owner." Each property has a specific meaning, which can be classified by a schema related to the name of the property. A statement is a combination of a resource, a property, and a value. Each part of a statement is also known as the subject, the predicate, and the object. The object can be another resource or a literal, which might be any string or XML. Figure 1 presents an example of RDF graph for the Web page of the University.

The statement can be written in the XML format. The RDF XML syntax has a root element, <RDF>, but this element is optional when the description is known to be RDF from the application content. In RDF element, the namespace attributes designate the location of the declarations of the RDF elements with the prefix "rdf:", and the location of the schema declaration associating with the prefix "s:". The namespace declaration can alternatively appear in a specific description element, or even in property elements. The description element has the subject; the child elements describe the properties and the objects. For example, "<s:owner>" and "</s:owner>" - a pair of tags – can show the property. The object is "Sungkonghoe University."

As in XML Schema for XML document, RDF Schema provides a vocabulary constraint facility for RDF document. In RDF Schema, the classes of the resources are defined. The classes have the same role as in the object-oriented programming models. The classes have hierarchical structures and they are extended with subclass refinement. The terms such as "Class," "subPropertyOf," and "subClassOf" are used for the basic type system for RDF to define such classes. By using class concepts, the reusability of metadata can be increased because sharing schemas and adding subclasses to the existing schemas will produce sufficient mechanisms in many schema specifications.

Ontologies play a crucial role in the "Semantic Web" – a machine-understandable Web with intelligent services. They provide shared and precisely defined terms in a particular domain for communications between human users and application systems. DAML+OIL – combined terminology from old versions, DARPA Agent Markup Language (DAML) and Ontology Inference Layer (OIL) - is an ontology language submitted to W3C in 2002 as a semantic markup language for Web resources [11]. It extends RDF and RDF schema. The object-oriented structure of domains in DAML+OIL consists of the terms "Class" and "Property." The usage of the term class is similar to that of RDF schema, but the classes of DAML+OIL are less restricted. For example, "subClassOf" class elements of DAML+OIL allow cyclic subclass-relations. DAML+OIL was superseded by Web Ontology Language (OWL) [12]. The W3C created OWL working group as part of their Semantic Web Activity in 2001. In 2004, the OWL became a formal W3C recommendation.

III. Data Repositories for Large Chemical Set

Public databases like PubChem[13], BindingDB[14], and ChEMBL represent some examples of large public domain repositories of compound activity data. ChEMBL and BindingDB contain manually extracted data from many articles. PubChem was originated from a central repository of High Throughput (HTS) screening experiments for the National Institute of Health's Molecular Libraries Program, but also includes data from other repositories. Scientific and Technical Information Network (STN) is another public database for Chemical Abstracts Service (CAS). Commercial databases, such as SciFinder and Reaxys have collected a large amount of data extracted from publications of articles and patents. Similarly to public and commercially available repositories, industry has produced large private collections. The data quality in databases usually depends on data source, data acquisition processes and care efforts. Accumulated chemical patents represent another rich resource for chemical information. Large-scale text mining has been done on patent corpus to extract useful information. IBM has contributed chemical structures from 2000 patents in PubChem. SureChEMBL database was launched in 2014 providing the wealth of knowledge hidden in patent documents and currently contains 17 million compounds extracted from 14 million patent documents.

IV. Map Reduce Framework

A small number of processes presented a temporal extension of the OWL for expressing time dependent information. Urbani et al. [15] proposed that a distributed reasoning method for computing the closure of an RDF graph based on MapReduce. They implemented it on top of Hadoop not with a straightforward way but with a parallel and distributed method for more expressive reasoning. When the volume of data increases and the ontology is modified, the dependent solutions should calculate the entire RDF again with new data attainment, and it is a time consuming job. The reasoning methods are required to avoid such time consuming processes. The MapReduce divides input data into many distinct independent pieces for map functions, and the reduce functions join the map results back into a final production.

A classical workflow for collecting related data and inserted into a local database management system (DBMS) before generating chemical RDF triples. Figure 2 summarizes the workflow for the collection, intermediate storage, and generation of information in RDF format.

The collections are usually obtained from querying on typical chemical or medical databases such as PubChem and ChEMBL. Then the collected data is saved in relational tables in a DBMS. A set of SQL queries are used by scripts and the scripts generate RDF triples from the result of the queries.

We propose an architecture in which the Hadoop Distributed File System (HDFS) replaces the local database for a temporary storage. Apache Hadoop is a promising system to store the extracted huge datasets from chemical databases. Each job in the MapReduce divided into two phases – a map and a reduce. The map phase divides the input data by relating each element with a key. The reduce phase handles each split independently, and all data is processed based on key-value pairs. The map function processes a certain key-value pair and produces a certain number of new key-value pairs. The reduce processes all intermediate values grouped by the same key into another set of key value pairs as output. As mentioned before, we can utilize the distributed reasoning method for computing the closer of RDF graphs in the chemical compound data in life sciences, too.

V. Conclusion

We briefly describe the RDF usage in presenting chemical structures. Big data in chemistry is largely focused from industry and academic organizations. To develop application systems for this area needs new infrastructure of map reduce and distributed systems. This is the suggestion for the framework and we expect further progress in this view.

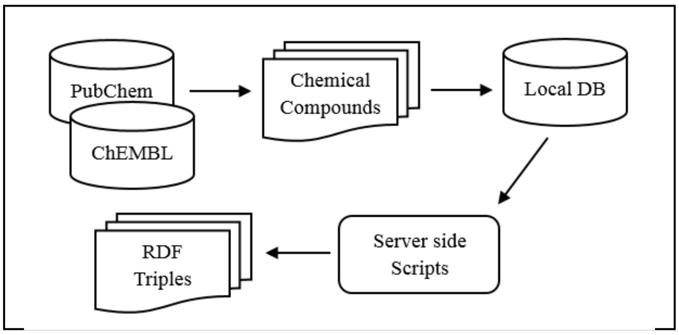

Figure 2. Workflow for Collecting and Saving Chemical Compounds

References

[1] Y. Saeys, I. Inza, and P. Larranaga. "A review of feature selection techniques in bioinformatics," *bioinformatics* vol. 23, pp.2507-2517, 2007.

[2] F. Belleau, M. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, pp. 706-716, 2008.

[3] K. Degtyarenko, et al., "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, pp.344-350, 2007.

[4] A. Bandrowski, et al., "The ontology for biomedical investigations," J. of *PloS one*, vol. 11.4, 2016..

[5] OBO Thecnical Working Group, "The OBO Foundry," WWW, http://www.obofoundry.org/.

[6] A. Gaulton, et al., "ChEMBL: a large-scale bioactivity database for drug discovery," Nucleic acids research, vol. 40, D1100-D1107, 2011.

[7] S. Jupp, et al., "The EBI RDF platform: linked open data for the life sciences," *Bioinformatics*, vol. 30, pp. 1338-1339, 2014.

[8] B. Chen, et al, "Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data," BMC bioinformatics, vol. 11:255, 2010.

[9] E. Willighagen, et al., "Linking the resource description framework to cheminformatics and proteochemometrics," *J. of Biomedical Semantics*, vol. 2:S6, 2011.

[10] O. Lassila, R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," WWW, http://www.w3.org/TR/REC-rdf-syntax/, February 1999.

[11] DAML+OIL "Web Ontology Language," *Submission request to W3C*, http://www.w3.org/Submission/2001/12/, December 2001.

[12] D. McGuinness and F. Van Harmelen, "OWL web ontology language overview," *W3C recommendation*, 2003.

[13] E. Bolton and others, "PubChem: integrated platform of small molecules and biological activities," *Annual reports in computational chemistry*, Vol. 4. Elsevier, pp.217-241, 2008.

[14] T. Liu, and others. "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research* vol.35, pp. 198-201, 2006.

[15] J. Urbani, S. Kotoulas, E. Oren, and F. Harmelen, "Scalable distributed reasoning using mapreduce," *International Semantic Web Conference*, pp. 634-649, 2009.