Estimating Domain Models from Metadata Instances to Improve Usability of LOD Datasets

Poster

Ryota Kinjo
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
S1721665@s.tsukuba.ac.jp

Mitsuharu Nagamori
Faculty of Library, Information and Media Science, University of Tsukuba, Japan
nagamori@slis.tsukuba.ac.jp

Shigeo Sugimoto
Faculty of Library, Information and Media Science, University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp

Keywords: metadata; metadata schema; domain model; schema extraction; application profile;

Introduction

Linked Open Data(LOD), which is one of the efforts to help realize semantic web, has gradually become popular. Many Linked Open Data datasets, however are not well utilized. There are multiple reasons for this, such as limited recognition of LOD, limited usability of LOD datasets and so on. In attempting to solve these issues, we focused on a metadata schema that describes the structure about metadata instances in each LOD dataset. As information about metadata schema are not typically released, it is difficult to use LOD datasets. Therefore, in this research we extract the domain model (Dublin Core) as directed graph, which is one piece of information about a metadata schema, from metadata instances. For example, FIG1 illustrates a domain model extracted from a metadata instance.

Domain models are suitable for understanding the rough structure of a metadata instances in an early stage. We developed an estimation method to generalize a process of understanding metadata schema when people who are not familiar to the datasets handle them. We then apply the estimation method to existing datasets.

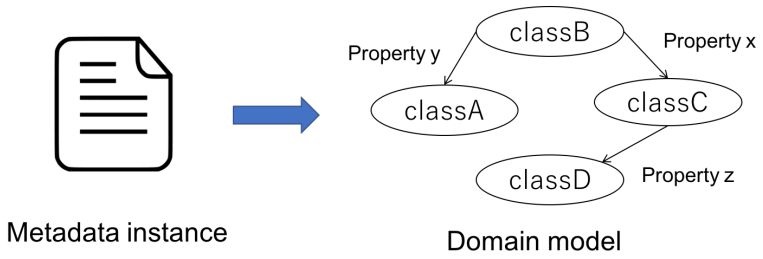

FIG. 1. Estimate domain model from metadata instances.

Method for Estimating Domain Models

People generally try to grasp the rough structure of datasets by executing SPARQL queries and seeing text formatted metadata. When people understand things and relation among the things in metadata, people can better understand metadata schema. Therefore, the purpose of a domain model that we estimate is to help people understand main class and properties which belong to those main classes. The method for estimating domain model is divided into 3 steps.

First is the execution of a series of SPARQL queries, as shown in FIG.2, to get statistics that are needed to estimate the domain model. Second is to determine which information is put into the domain model. This is the most important step in estimating the domain model. We implemented this through the comparing of each number of class as a subject with number of class as an object.

"Number of class as a subject" means number of times that a class behaves as a subject as shown in FIG.3. For example, if a resource which belongs to class A appears three times as a subject in metadata instances, the number of class A as a subject is three.

```
SELECT DISTINCT ?type
WHERE{
  ?s rdf:type ?type .
}
```

(a) A query to get all classes

```
SELECT DISTINCT ?v
WHERE{
  ?s rdf:type <given_class> ;
  ?v ?o.
  FILTER(?v NOT IN (rdf:type))
}
```

(b) A query to get all properties of *given class*

```
SELECT (GROUP_CONCAT (DISTINCT ?range ; SEPARATOR = ",") AS ?gc)
WHERE{
  ?s <given_property> ?o ;
  rdf:type <given_property>.
  ?o rdf:type ?range .
}order by (?o)
```

(c) A query to get range of *given* combination of *class and property*

FIG. 2. A series of SPARQL queries.

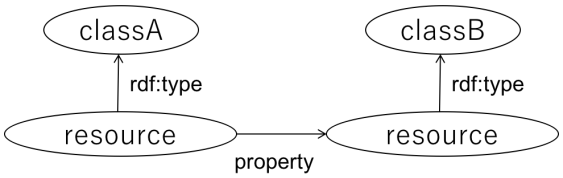

FIG. 3. Class A as subject and class B as object.

"Number of class as an object" is obtained in a similar way. Then if number of the class as a subject is larger than number of the class as an object, we define the class as main classes. Main classes and properties which are relevant to main classes are put in domain model. Third is to describe domain model as directed graph.

Experiment

We conducted an experiment to verify the validity of domain model estimated by our method. We prepared 5 LOD datasets and manually determined their correct domain model. We then estimated 5 domain models of 5 datasets using our method, followed by a comparison of the domain models made through both methods. Correct domain model is determined on the basis of published information about the metadata schema, such as text description or using RDF metadata found in the website. Correct domain model was confirmed by members of our lab. We must add that we use existing domain model, which is in graph format, if it exists. we calculated precision and recall by converting the directed graph into RDF triples. We show datasets used in experiment TABLE 1. We used 0.1% of all data found in Europeana. The purpose of this was to verify whether our method is useful when applied to a portion of metadata instances.

Results and Discussion

Table 2 shows results of the experiment. As can be seen in TABLE 2, the results from Europeana and Kyoto Kokusai Manga Museum (KKM) were below standard. In Europeana, our conclusion is

that there was an insufficient number of metadata instances. To address this, we plan to prepare a sufficient quantity of metadata instances, or establish a random sampling method for a small quantity of metadata instances. A cause of bad precision and recall in KKM is that there are some unused classes and properties in metadata instances, while still being described in published information about metadata schema. The problem of terms, which are not used in metadata instances nevertheless being described in published information, is common among many datasets. This problem needs to be discussed in the future.

TABLE 1: used datasets

Datasets name	Correct domain model	Memo
Aozorabunko LOD	Made by hand	
CiNii	Made by hand	
Europeana	Existing model	1 / 1000 of Overall
Kyoto Kokusai manga museum	Existing model	
NDLSH	Made by hand	

TABLE 2: Results of experiment.

Datasets name	precision	Recall
Aozorabunko LOD	0.85	1
CiNii	0.83	0.83
Europeana	0.07	0
Kyoto Kokusai Manga Museum	0.23	0.2
NDLSH	0.63	0.63

Relevant study

ELLIS(Gottron) is most similar study. While ELLIS provides exhaustive schema information by use of an interactive interface, our approach provides rough limited schema information by use of static image.

Conclusion

In this research, we proposed a method for estimating a domain model and conducted an experiment to verify validation of the method. We concluded that our method needs a greater number of metadata instances in order for the experiment to produce better results. A primary problem is an evaluation for validity of our method. We can't say that the purpose of domain model estimated by our method is same as purpose of existing domain model. The purpose of our domain model is to let users understand the dataset. As existing domain models are typically determined at the stage of design, the domain model often contains classes or properties which are not used in metadata instances. Therefor to verify whether coherence of LOD datasets is improved, comparing estimated domain model with existing one is not suitable. In the future, we need to verify whether the coherence of LOD datasets is improved when utilizing our estimated domain model.

References

Dublin Core singapore-framework. Retrieved February 19, 2017, from http://dublincore.org/documents/singapore-framework/ .

Aozorabunko LOD. Retrieved February 19, 2017, from http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/ .

CiNii. Retrieved February 19, 2017, from https://support.niii.ac.jp/ja/CiNii/api/api_outline .

Europeana. Retrieved February 19, 2017, from http://pro.europeana.cu/ .

Kyoto kokusai manga museum. Retrieved February 19, 2017, from http://mdlab.slis.tsukuba.ac.jp/lodc2012/kmm/ .

Web NDL Authorities. Retrieved February 19, 2017, from http://id.ndl.go.jp/information/download/ .

Tsunagu Honma, Mitsuharu Nagamori, and Shigeo Sugimoto. (2014). Extracting Description Set Profile from RDF Datasets using Metadata Instances and SPARQL Queries. Graduate School of Library Information and Media Studies University of Tsukuba. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2014, 109-118.

Thomas Gottron, Malte Knauf (2016). ELLIS: Interactive Exploration of Linked Data on the Level of Induced Schema Patterns. ESWC