SPARQL Endopoint を利用したメタデータインスタンスに基づくドメインモデル推定

A Method for Domain Model Estimation from Mata Data Instance by using SPARQL Endpoint

金城良大 [1] ,三原鉄也 [2] , 永森光晴 [2,3],杉本重雄 [2]

Ryota Kinjo [1], Mihara Tetsuya[2], Mitsuharu Nagamori [2,3], Shigeo Sugimoto[2]

[1] 筑波大学情報メディア創成学類
[1]College of Media Arts, Science and Technology
[2] 筑波大学図書館情報メディア系
[2] Faculty of Library, Information and Media Science, University of Tsukuba.
[3] 筑波大学知的コミュニティ基盤研究センター
[3] Reserch Center for Knowledge Communities, University of Tsukuba.

Abstract: Linked Open Data has become popular gradually. But almost LOD datasets are not utilized still. In this research we considerd difficulties to understand metadata schema as cause that datasets are not utilized . In other to solve it we developed a method for estimating a domain model from metadata instances. The domain model is one of the metadata schemas. It expresses relation among things in the metadata. Then we evaluated our method to compare correct domain model generated manually with one generated by our method. we understood that estimating domain model from metadata instance needs sufficient amount of instances .

1. はじめに

セマンティック Web の取り組みの一つである Linked Open Data (LOD)の普及が進んでいる。一方で多くの LOD データセットは利活用されていないという問題がある。この理由としては世間一般にまで LOD の認知が及んでいないことや、個々のデータセットの有用性の不足など様々な要因が考えられる。本研究ではその中でもデータセットと共にあるメタデータスキーマ情報の公開が行われていないことに着目した。LOD データセットの利活用にはその LOD のメタデータスキーマの理解が必要になる。しかし現状では多くの LOD データセットではメタデータスキーマの情報が公開されておらず、その理解の妨げとなっている。

そこで本研究ではメタデータスキーマ情報の一つであるドメインモデルをメタデーインスタンス（LOD の実データ）から推定する。ドメインモデルとはメタデータ記述対象とその記述対象間の関係を表すものである[10]。ドメインモデルはデータセット利活用の初期段階に必要な、メタデータスキーマのおおまかな理解に適している。

本研究では、利用者がメタデータスキーマを持たないデータセットを利活用するときに行う一連のデータセット把握の流れを機械的に再現することによる推定手法を考案した。そして実際の LOD データセットについてその手法を適用する実験を行った。

2. データセット構造理解のためのドメインモデルの提案

2.1. インスタンスからのスキーマ理解の困難さ

データセットの利活用にはそのメタデータスキーマを理解する必要がある。一方、現状では多くの LOD データセットが十分なメタデータスキーマ情

報を公開していない[3]。そのためデータセット利用者はインスタンスからメタデータスキーマを理解しなければならない。どこまで詳細にメタデータスキーマを理解するのかは利用者の目的に応じて変化するが、まずそのデータセットの主要な記述対象（事物）と大まかな構造が理解できることが望まれる。

LOD 利用者が未知のメタデータスキーマを把握する方法として、メタデータインスタンスのテキストデータを直接参照しながら,SPARQL クエリのやり取りを行う方法がある。この方法は専門性の必要な作業であり、体系化されていない。データセットの規模や内容に依存するが、非常に手間のかかる作業である。

本研究ではこのインスタンスからスキーマを推定する一般的なプロセスを下記のように想定する。まず最初にそのデータセット内で出てくる事物を把握するために、全クラスの列挙する SPARQL クエリ処理を行う(図 1)。一般的に LOD データセットでは事物にクラスをつける。したがってこのクエリで事物を確認することができる。

次に各事物の属性を知るために、各クラスに所属するリソース集合ごとのプロパティを列挙するクエリ（図２）を各クラスごとに実行する。属性の数が多いクラスほどそのデータセット内で主要な役割を担っていることが予想できる。

次に行うのは、主要な事物間の関係を探るために列挙されたプロパティの内いくつかのレンジを探る作業である。図３にクラスプロパティを指名した時のレンジを出力するクエリを示す。主要な事物に関係すると思われるプロパティの数だけ図 3 のクエリを繰り返していく。そのうちクラスとクラスをつなげるプロパティを見つけることができ、データセット内の事物間の関係を把握することができる。この作業でのプロパティへの目星のつけ方は作業者の経験や恣意的な判断に依存すると考えられる。

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

図 1：全クラスを列挙するクエリ

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

図 2：各クラスごとのプロパティを列挙するクエリ

```
SELECT　(GROUP_CONCAT(DISTINCT ?range ;
SEPARATOR = ",") AS ?gc)
WHERE{
  ?s <プロパティ名> ?o ;
     rdf:type <クラス名名>.
  ?o rdf:type ?range .
}order by (?o)
```

図 3：レンジを出力するクエリ

2.2. ドメインモデルを利用した利活用支援

前節では十分なメタデータスキーマが公開されていない場合のスキーマ理解の困難さを述べた。このときにドメインモデルがメタデータスキーマ情報として公開されているとメタデータスキーマの理解に役立つ。図 4 はドメインモデルの例である。利用者は図 4 から *aozora:Bibresouce* クラスが *aozora:Person*,*foaf:Person* クラスに *dc:creator*,*bibo:translator* の関係を持つ事がわかる。また *rdf:seeAlso*,*dc:subject* の二つのプロパティで外部リンクを持つ事や、空白ノードを介して *aozora:fileData* プロパティで青空文庫のウェブサイトとリンクを持つ事などのメタデータの構造が理解できる。この例から適切なドメインモデルがあれば主要な事物間の関係が明らかになり、メタデータスキーマ理解に役立つと言える。

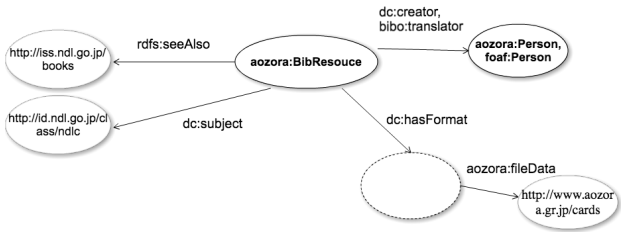

図 4：青空文庫 LOD ドメインモデル[4]

2.3. メタデータスキーマ理解のための ドメインモデル推定

本研究では LOD データセットのメタデータスキーマの構造理解支援のためにメタデータインスタンスからドメインモデルを推定する。本節では LOD 利用者に必要なメタデータスキーマ理解のための要求要件を定め、それを満たすドメインモデルの定義をする。詳しいドメインモデル推定手法については 3 章で後述する。

研究室の知見から明らかになった LOD 利用者に必要なメタデータスキーマ理解のための要求要件は、1)主要なクラスがわかること,2)主要なクラスを中心にクラス間をつなぐプロパティがわかることである。

これらの要求要件を満たすためのドメインモデルの仕様を下記に示す。

- １つの有向グラフから構成される
- ノードはクラス,外部リソース,空白ノードのいずれかを表し、それぞれを区別可能にする
- エッジはプロパティを表す

また、推定に用いるメタデータインスタンスにはSPARQL エンドポイントを介してアクセスし、出力は人が解釈し易いように画像ファイルとする。図５は本研究で出力するドメインモデルの例である。図５のようにメインクラスを始点に有向グラフを作成する。

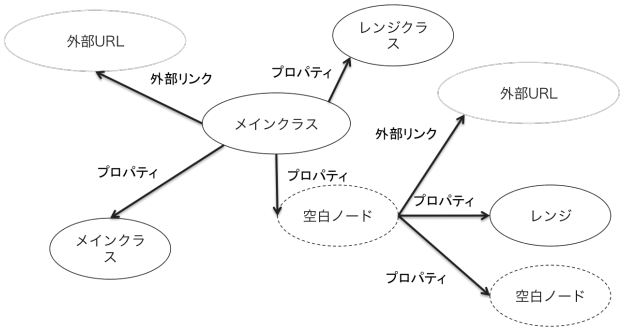

図 5：推定するドメインモデルの例

2.4. 関連研究

関連研究にはメタデータスキーマを抽出する研究とデータセットの可視化に関する研究があげられる。

メタデータ抽出の研究については本間らの先行研究[1]がある。文献[1]はメタデータインスタンスからメタデータスキーマを推定している点で本研究と類似している。しかし本研究がメタデータのクラス間関係に注目しているのに対して、[1]ではプロパティの記述規則に注目している点で異なる。

次にデータセット可視化には Florenzano らの研究[2]がある。文献[2]はデータセットを可視化するシステムを構築している。インタラクティブなインタフェースをもつアプリケーションでユーザの入力に応じて,情報を提供している。文献[2]のシステムがデータセットの構造的情報を網羅的に全て提供しているのに対し、本研究は限られた情報を提供する点で異なる。

3. ドメインモデル推定手法

3.1. 推定手法の概要

本章ではドメインモデルの推定手法について述べる。推定手法の方針は、利用者がメタデータスキーマを持たないデータセットを利活用するときに行う一連のデータセット把握の流れを機械的に再現することである。それによって 2.3 節で述べた要求を満たすドメインモデルを推定する。下記の３つのステップでドメインモデルの推定を行う。

1. 統計情報の取得
2. 統計情報の解析
3. 有向グラフの描画

ステップ１の統計情報の取得は SPARQL クエリによるエンドポイントへの問い合わせである。またステップ２の統計情報の解析は SPARQL クエリの結果から、どの情報をドメインモデルに組み込むべきか、決定するステップである。ステップ２はドメインモデルに組み込むノードとエッジを決定する最も重要なステップである。本研究は非常に単純な手法でステップ２を実現した。また３のステップは有向グラフをどのように描画するかを決めるステップである。以降それぞれのステップについて説明する。

3.2. 統計情報の取得

本ステップの目的は次の統計情報の解析に必要な全ての情報を抽出することである。次の統計情報の解析では主にクラス間関係に注目した解析を行う。したがってクラスの情報を中心に情報を抽出する必要がある。本ステップでは SPARQL エンドポイントに対して一連の推定クエリを使って問い合わせることで、統計情報を取得する。下記に取得する主な統計情報の一部を示す。

- 全クラスの情報
  - 使用回数, 主語として使用される回数,目的語として利用される回数等
- 各クラスがもつプロパティの情報とそのレンジの情報
  - 目的語はリソースなのかリテラルなのか、レンジは複数あるのか、そのクラスに属するリソースは必ずそのプロパティをもつのか等

3.3. 統計情報の解析

本ステップは抽出した統計情報を使って有向グラフに組み込むノードとエッジを決める。本研究が目指す有向グラフは 2.3 節の要求要件を満たす必要がある。その手段として、本ステップでは有向グラフに採用されるトリプル（以下採用トリプル）を任意の数選出し、結合することで、有向グラフを作成する。採用トリプルとは図６に示すノード、エッジ、ノードの組み合わせである。

図６の採用トリプルの選出方法について述べる。

図 6：採用トリプル

エッジの出て行くノードではデータセットでの主要なクラスを選ぶ。これをメインクラスと呼ぶ。採用トリプルではメインクラスから出るプロパティとそのレンジを組み込む。レンジには,クラス,外部リンク,空白のノードが含まれる。したがってリテラルが目的語のプロパティは採用トリプルには含まれない。なおレンジの中でも空白ノードが目的語のプロパティは、その空白ノードからレンジにつながるプロパティがある場合のみ組み込む。

本ステップでは適切なメインクラスを推定することが、要求要件を満たすことに繋がる。したがってメインクラスの決定は非常に重要である。本手法では所属するリソース集合が、主語として機能している回数と目的語としての機能している回数を比較し、主語としての回数が上回るものをメインクラスとしている。所属するリソース集合とはそのクラスに対して *rdf:type* プロパティをもつ全てのリソースを指す言葉である。図 7 に *man* クラスが主語として使われている RDF の例を示す。図 7 では *man* クラスが主語として,*woman* クラスが目的語として機能している。本手法では主語としての役割の強いクラスをメインクラスとして判定するとも言える。

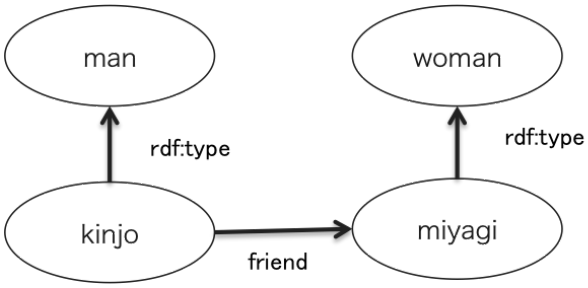

図 7：man クラスが主語として使われる例

3.4. 有向グラフの描画

本ステップでは全ステップで決まった採用トリプルを結合し有向グラフとして出力する。有向グラフの描画には Gviz[5]を利用している。

4. 実験

本章では前章で述べた手法により推定したドメインモデルの妥当性を確かめるための検証実験を行う。本実験では５つの LOD データセットを用意し,ドメインモデルを推定する。その後それぞれの正解のドメインモデルと比較する。正解のドメインモデルは公開されているメタデータスキーマ情報をもとに作成した。なお正解セットは長年メタデータスキーマについて研究している著者らの研究室で確認を行った。またドメインモデルが公開されているものはそれを正解セットとして利用する。表 1 に実験で利用するデータセットを示す。Europeana のデータセットでは全体の 1/1000 のインスタンス量で実行した。その理由としては Europeana のようなデータサイズの大きいデータセットに対して、インスタンスの一部に本手法を適用した場合、適切なドメインモデルが推定されるのか、検証するためである。

表 1：実験対象のデータセット

データセット名	正解セット	メモ
青空文庫 LOD	作成	
Cinii[6]	作成	
Europeana[7]	既存	全体の 1/1000
京都国際漫画ミュージアム[8]	既存	
Ndlsh[9]	作成	

推定ドメインモデルと正解セットの比較は,一度有向グラフを RDF トリプルのテキスト形式にして比較する。比較では全トリプル、全クラス、外部リンクの適合率、再現率を評価する。

5. 結果・考察

前章で述べた実験の結果を表 2,3,4 に示す。表２から Europeana と京都国際漫画ミュージアムの２つのデータセットで適合率、再現率が 0.24 以下になっている。それに対して他のデータセットは 0.63 以上となっており、適合率再現率に差が開いている。

この原因は適合率、再現率の低い２つのデータセットにあった。Europeana では十分なインスタンスが用意できなかったことが原因だと考えられる。これを解決するには全てのデータを用意するかランダムサンプリングの手法を考える必要がある。また京都国際漫画ミュージアムでは公開されているメタデータスキーマで定義されているクラスがインスタンスで使用されていないことが原因であった。表３の京都国際漫画ミュージアムの全クラスの評価において、適合率に対して再現率が低いという結果も同様の理由から起きている。メタデータスキーマでは定義されているタームがインスタンスで利用されない事例への対処は、今後の課題である。

さらに表 4 を見ると Europeana と京都国際漫画ミュージアムの既存ドメインモデルから作成した正解セットには、外部リンクが含まれていないことがわかる。外部リンクは他の情報資源とのリンクを指す。本研究では外部リンクをデータセットを理解する上で重要なものとしてドメインモデルに組み込んだ。しかし既存のドメインモデルではそれらは含まれていなかった。実験データセット数が少ないため断定できないが、今後実験するデータセット数を増やし、外部リンクをドメインモデルに組み込むべきか検討していく必要がある。

表 2：全トリプルの適合率と再現率

データセット名	全トリプル	
	適合率	再現率
青空文庫 LOD	0.85	1
Cinii	0.83	0.83
Europeana	0.07	0
京都国際漫画ミュージアム	0.23	0.2
Ndlsh	0.63	0.63

表 3：全クラスの適合率と再現率

データセット名	全クラス	
	適合率	再現率
青空文庫 LOD	1	1
Cinii	1	1
Europeana	0.33	0.4
京都国際漫画ミュージアム	0.9	0.53
Ndlsh	1	1

表 4：外部リンクの適合率と再現率

データセット名	外部リンク	
	適合率	再現率
青空文庫 LOD	1	1
Cinii	0.75	1
Europeana	0	
京都国際漫画ミュージアム	0	
Ndlsh	0.33	0.33

6. おわりに

本研究ではメタデータインスタンスからドメインモデルを推定するための手法を提案し、検証実験を行った。実験結果から適切なドメインモデル推定には十分な量のインスタンスが必要なことがわかった。本手法のドメインモデル推定手法は要求要件を定義しそれを満たすようなドメインモデルを目指した。しかし既存のドメインモデルが同様の要求要件を満たす、メタデータスキーマ理解のために妥当なものとは限らない。したがって推定ドメインモデルと既存ドメインモデルの比較による推定手法の評価は適切な評価とは言えない。今後は推定したドメインモデルの有無で LOD データセットの利活用性が向上するのか、検証する必要がある。

謝辞

本研究は JSPS 科研費 15K00444 の助成を受けたものである。

参考文献

[1] Tsunagu Honma ,Mitsuharu Nagamori, Shigeo Sugimoto : “Extracting Description Set Profile from RDF Datasets using Metadata Instances and SPARQL Queries”, Graduate School of Library Information and Media Studies University of Tsukuba(2014)

[2] Fernando Florenzano , Denis Parra , Juan L.Reutter:“A visual Aid for Understanding Endpoint Data” . ISWC (2016)

[3] Yoritsugu Nishide , Tsunagu Honma,Mitsuharu Nagamori, “日本の Open Data 活用を目的としたデータセットのスキーマ分析とリンク関係の調査” , IPSJ SIG Technical Report (2013)

[4] 青空文庫 Linked Open Data . http://mdlab.slis.tsukuba.ac.jp/lodc2012/aozoralod/ ,（参照 2017-2-17）

[5] Gviz . https://github.com/melborne/Gviz ,（参照 2017-2-19）

[6] Cinii 全般 – メタデータ・API . https://support.nii.ac.jp/ja/cinii/api/api_outline ,（参照 2017-2-19）

[7] Europeana . http://pro.europeana.eu/ ,（参照 2017-2-19）

[8] 京都国際漫画ミュージアム . http://mdlab.slis.tsukuba.ac.jp/lodc2012/kmm/ ,（参照 2017-2-19）

[9] Web NDL Authorities . http://id.ndl.go.jp/information/download/ ,（参照 2017-2-19）

[10] Dublin Core singapore-framework. http://dublincore.org/documents/singapore-framework/ ,（参照 2017-2-19）