人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-01

サンプリングに基づくLODの構造推定に関する基礎的検討

Investigation on LOD Structure Estimation Based on Sampling

矢部彩佳 *　高間康史
Ayaka Yabe , Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: 近年 LOD によるデータ公開が進められており，これらを活用したサービス開発なども期待されている．しかし，他者が公開したデータを利用する場合，データ構造が不明な場合があり，活用を阻害する一要因となっている．本稿では LOD を探索的にブラウズする作業を支援するシステムの実現を目的として，その要素技術となる LOD の構造推定に着目する．SPQRL クエリによるサンプリングに基づく推定方法に関する基礎的な検討を行った結果について報告する．

1　はじめに

本稿では，RDF (Resource Description Framework) で記述された LOD を探索的にブラウズする作業を支援するシステムの要素技術として，SPARQL クエリによるサンプリングに基づく LOD 構造の推定手法に関する基礎的な検定を行った結果について報告する．

近年，計算機で処理しやすい形式でデータを公開･共有する仕組みとして LOD (Linked Open Data) が注目されている．LOD は自分の手元にない外部リソースを扱えることが利点だが，他者が公開したデータを使用する場合，データ構造が不明という問題点がある．このため，探索的に LOD をブラウズし，その構造を把握する必要があると考える．

探索作業を支援するために，探索の起点として有効なリソースの抽出･提示を行う．起点として有効なノードを発見するためには，LOD のデータ構造を分析する必要があるが，全データを取得して分析を行うのでは，外部リソースの活用という LOD の利点が生かせないと考える．そこで本稿では，LOD データを SPARQL クエリを用いてサンプリングし，LOD の構造推定を試みる．

RDF とは，リソースの関係を主語 (subject)・述語 (predicate)・目的語 (object) の 3 つの要素 (トリプル) を用いて表現するデータモデルであり，データセットは，主語と目的語をノード，述語をエッジとするグラフ構造で表現される．本稿ではこの構造を用いてサンプリングを行う．

現在日本で公開されている LOD データを調査したところ，Excel 等のテーブルデータを RDF データに変換したものが多く発見された．そこで本稿では LOD を表構造を持つもの（テーブル型）とそれ以外に分類し，サンプリングにより両者を区別可能であるかを検証し，その結果に基づきサンプリングによる構造推定の可能性について考察する．

*連絡先：首都大学東京大学院 システムデザイン研究科
〒 191-0065 東京都日野市旭ヶ丘 6-6
E-mail: yabe-ayaka@ed.tmu.ac.jp

2　関連研究

2.1　表データの RDF データ化ツール

現在日本で公開されている LOD データには表形式のものが多くある．その理由として，すでに表形式で管理していたデータを公開する場合が多いことと，表形式データの RDF データ変換ツール・サービスが整備されていることが挙げられる．前者は，総務省や市町村が公開しているデータが該当する．後者に関しては，オープンデータ活用支援プラットホーム LinkData.org[1] などが存在する．

LinkData.org は，データ・アプリ・アイディアの作成と公開を行う 4 つの Web サイトを提供しており，その中の LinkData[2] ではテーブルデータを RDF データに変換するサービスを提供している．RDF 変換用のテーブルデータの雛型を Web サイト上で作成．ダウンロードし，RDF データの主語・目的語にあたる部分を埋めてアップロードすることで RDF データに変換が可能である．このサービスにより手持ちのテーブル型データを気軽に RDF 化することができる．これらのツールを利用することにより，今後さらにテーブル型 RDF データが増加していくと推察できる．

[1] http://linkdata.org/
[2] http://linkdata.org/home

2.2 グラフ構造データの分析

近年，世界中に広く普及したSNS(ソーシャルネットワークサービス) や生体科学における遺伝子構造など，グラフ構造をもつデータの分析に関する研究がされている [3][4].

分析対象となるデータが巨大な場合，すべてのデータを分析することはコストや時間面から難しいという問題が存在する．この問題に対し仲前ら [1][2] は，巨大グラフデータから部分的にグラフを抽出する手法として，ランダムウォークサンプリングを改良したサンプリング方法を提案している．入次数の大きいノードを訪れやすいというランダムウォークの性質を考慮したIRW(In-Degree Weighted Random Walk)[1] は，入次数に偏らないサンプリングが可能となる．IRW は入次数がわかることを前提としているが，巨大なグラフデータに対して事前に入次数を調べることは現実的ではないことから，Reservoir を用いたIRW の改善版であるIRRW を提案し，入次数を前提条件としないランダムサンプリングを可能にしている [2].

2.3 RDF データ分析

RDF データは，通常 SPARQL[3] と呼ばれるクエリ言語によって検索が行われるが，適切なクエリを作成する為にはデータ構造の理解が不可欠となる．そこで，後藤ら [5] は探索的検索アプローチによって LOD を理解・利用する DashSearchLD というシステムを提案している．探索的検索とは，探索目的を少しずつ明確化しながら新しい知識を獲得していく学習や調査のような情報検索である．探索的閲覧によって検索空間を遷移しつつ，絞込み検索によって検索絞り込むという行為を繰り返すことにより，検索空間の理解と情報要求の具体化を行い，データ集合の理解に繋がるとしている．DashSearchLD には，SPARQL Endpoint 機能を持つエンドポイントウィジェットと，RDF データのプロパティとその値を表示するメタデータウィジェットがあり，ユーザはこれらのウィジェットをマウスによって操作することで，SPARQL クエリを用いずにデータの探索的検索やプロパティ情報の獲得が可能となる．また，田代ら [6] は，RDF の特徴を考慮したデータ分析支援ツールとして，(1) 共通の述語を持つ主語の抽出・テーブル作成を行うツール．(2) 複数エンドポイント間の共通リソースの抽出を行うツール．(3) 時間情報に基づくデータ分析支援ツールを提案している．(1) は，最大公約数的に共通の述語を持つ主語の抽出を行い，行を主語，列を述語としたテーブルの出力を行う．(2) は，2 つの SPARQL エンドポイント間で共通するリソースを抽出することで，異なる LOD の連結可能性を検討する作業を支援する．(3) は，統計データやログデータのような RDF データから時系列データを抽出し，ヒストグラムとして可視化を行う．

RDF データを活用する上で必要となるのが重要リソースの把握である．SPARQL 検索によって，RDF データの一部を簡単に抽出することができるが，SPARQL は抽出したリソースを重要度の高い順にランク付けする機能を持っていない．検索結果が大量にあった場合，さらに情報を絞り込むためユーザの要望に応じたリソースのランキングを提供することは有用であるとして，一瀬ら [7][8] は DBpedia を対象に SPARQL 検索によって得られたリソースを，グラフ構造から重要度評価を行う PageRank アルゴリズム用い，ランク付けする方法を提案している．

3 LOD 構造判定方法

3.1 テーブル型と非テーブル型の判定

前述の通り，現在公開されている RDF データには，テーブル型データを RDF データに変換したデータとそうでないデータが存在する．前者は市町村が公開しているデータに多く見られる．一方，DBpedia の様な，多種多様なリソースを含む RDF データの場合には，テーブル型をとらないと仮定する．この仮定に基づき本稿では，RDF データがテーブル型か否かを判別することを目的とする．

グラフ構造を分析する方法としては，ランダムウォークサンプリングなどの方法が知られている [1][2]．このような探索を行う方法は，複雑ネットワークを構成しているデータには有効だが，テーブル型のデータ (図 1) の場合には，ほとんどの目的語がリテラル（数値あるいは文字列）であることが多いことから，ランダムウォーク等の探索方法を用いてもすぐに行き止まるため，有効に機能しないと考える．

テーブル型データの特徴として，本稿では以下の 4 点に着目する．

1. 同じプロパティが複数存在
2. 目的語として，リテラルまたは出次数が 0 のリソースを持つため，探索をしてもすぐに行き止まる
3. 各リソースの出次数が揃いやすい
4. 各プロパティはリソース毎に 1 回ずつ出現する

これらの特徴に基づきテーブル型か否かの判別を行い，テーブル型ではない場合のみ探索を行うことで，各 RDF データに対し効率的に起点ノードを発見することが可能と考える．

[3]http://www.w3.org/TR/rdf-sparql-query

名前	性別	年齢	職業	勤め先HP	住所	電話番号	E-mail
http://…A	M	30	○○	http://…	東京	111-1111	…@..jp
http://…B	F	30	△△	http://…	神奈川	222-2222	…@..jp
http://…C	M	30	□×	http://…	千葉	333-3333	…@..jp
http://…D	F	30	▽◇	http://…	埼玉	444-4444	…@..jp

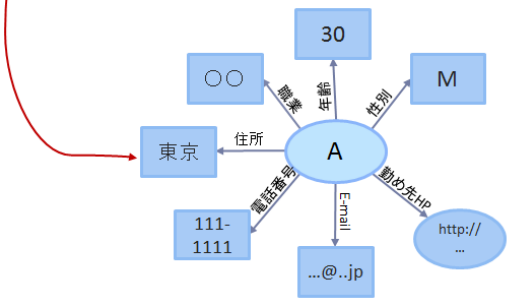

図 1: テーブル型データの RDF グラフ

3.2 データの抽出方法

本稿ではデータ型判断のためのデータ抽出法として，以下の手順をとる．

1. 対象となる RDF データからランダムに主語リソースを抽出し，探索の起点とする．
2. 起点から最良優先探索を行い，取得したノードについて以下の情報を記録する．
   - 出次数
   - 探索の STEP 数
   - プロパティ
3. 起点が持つ各プロパティの出現回数を求める．

ステップ 2 において，最良優先探索に用いるヒューリスティック関数として各リソース出次数を用い，出次数の大きいノードを優先的に探索する．

図 2 に示す例で A を起点とすると，子ノード B, C, D 中で出次数最大（2）の B を選択し，これを主語とするトリプルを SPARQL により求める．

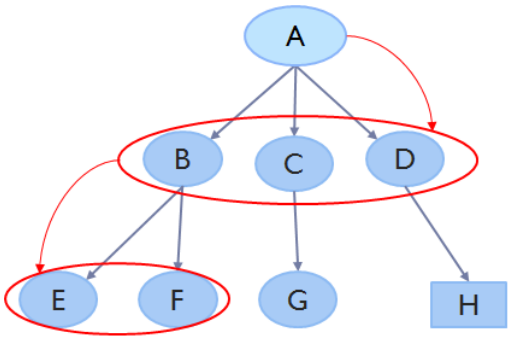

図 2: (例) 最良優先探索

4 型判定に関する予備実験

4.1 実験概要

本稿では，表 1 のデータセットを対象に，以下の 2 点に関する調査を目的として実験を行う．

- 調査 1：起点のプロパティに関する調査
- 調査 2：ステップ数の調査

調査 1 では，プログラム 1 回の試行でデータセットからランダムに起点リソースを 10 個抽出する．各起点に対し 3.2 節で述べた手順でプロパティを取得し，プロパティの出現回数を計算する．各データセットにおける調査回数は表 2 の通りである．非テーブル型はテーブル型に比べ，構造の特徴がわかりづらいため試行を 2 倍行った．

調査 2 では，最良優先探索により各起点から何ステップ進めるかを調査する．DBpedia Japanese 以外の 3 つのデータに関しては，探索可能なノードがなくなるまで探索を続け，DBpedia Japanese に関しては探索の上限を 30 ステップとした．

表 1 に示すデータセットにおいて，テーブル型と想定されるデータとして横手市 AED 設置場所 [4] 及び神奈川名所 LOD データセット [5]，非テーブル型と想定されるものとして横手市 AED 設置場所加工及び DBpedia Japanese を対象データセットしてそれぞれ選んでいる．

横手市 AED 設置場所加工データは，横手市 AED 設置場所データを元に，675 トリプルを削除し各主語リソースの出次数をまばらにした後，人工データ 67 トリプルを追加した．DBpedia Japanese[6] は，2013 年 9 月 4 日の以前に公開されたデータを使用した．

表 1: 使用データセット

データ	総トリプル数	主語リソース数
横手市 AED 設置場所	1,252	113
神奈川名所 LOD	451	45
横手市 AED 設置場所加工	644	140
DBpedia Japanese	32,633,660	3,626,642

表 2: 調査 1

データ	抽出起点数×試行回数
横手市 AED 設置場所	10 × 5
神奈川名所 LOD データセット	10 × 5
横手市 AED 設置場所加工	10 × 10
DBpedia Japanese	10 × 10

[4] 横手市情報政策課:http://linkdata.org/work/rdf1s843i
[5] kamogawa, SayokoShimoyama:http://linkdata.org/work/rdf1s2537i
[6] http://ja.dbpedia.org/

4.2 実験結果

図 3, 4, 5 に実験 1 の結果，表 3 に実験 2 の結果を示す．図は，縦軸がプロパティの種類数，横軸がプロパティの出現回数である．また，表 3 に示す平均及び標準偏差は，標本についてのものであり，母集団の不偏推定量ではない．

テーブル型である横手市 AED 設置場所データ (図 3) と神奈川名所 LOD データセット（図 4）の結果では，両者とも出現回数が 9 回もしくは 10 回のプロパティ数が多くなっている．

横手市 AED 設置場所データに関して，取得したリソースを観察すると，3 回目の試行以外の起点リソースは全て AED に関するリソースであり，共通プロパティが数多く見られた．そのため，10 個の起点に対し 10 回出現したプロパティの種類が多くなっている．これは 3.1 節に示したテーブル型データの特徴 1 に該当する．また，これらのプロパティは各起点リソースに 1 回ずつ出現していたため，特徴 4 にも該当する．3 回目の試行に関しては，10 個の起点の中で 1 つだけ AED に関するものではなく E-mail に関するリソースだったため，AED リソースとは異なるプロパティを持っていたた．そのため，他の試行とは異なり出現回数 9 回のプロパティ数が多くなっている．9 または 10 回出現したプロパティは，AED の名前，設置場所の住所，設置場所の郵便番号などで，すべての AED リソースで出現していた．出現頻度の少なかったプロパティは設置場所施設の開く時間・閉まる時間，外部リソースへのリンクなど，全ての AED リソースが持っているわけではない要素であった．また，今回抽出した全 50 個のリソースそれぞれの出次数は AED リソースで 10〜13，E-mail リソースは 2 であり，AED リソースは特徴 3 を満たしている．

表 3 より，調査 2 に関してはステップ数が全て 1 で終了したことがわかる．このことから 3.1 節で述べた「探索してもすぐ行き止まる」という特徴 2 が満たされていることがわかる．以上より，横手市 AED 設置場所データは，3.1 節で挙げたテーブル型データの特徴を満たすことがわかる．

神奈川名所 LOD データセットに関しては，探索した全 50 個の主語リソース中 49 個は名所に関するリソース，残り 1 個は動画情報を定義するリソースであった．各名所リソースにて，共通のプロパティが多く存在したため，横手市 AED 設置場所データと同様に 10 個の起点に対し，9 回もしくは 10 回出現したプロパティ数が多くなった．また，表 3 より，ステップ数も全て 1 であり，テーブル型データの特徴を満たしていると言える．

図 5 より，横手市 AED 設置場所加工データは，横手市 AED 設置場所データや神奈川名所 LOD データセットとは異なり，1 度しか出現しないプロパティが数多いことがわかる．これは複数のリソースが共通して持つプロパティが少ないということを意味している．複数回出現したプロパティも存在しているが，これは図 3 に示したとおり，加工前のデータが共通プロパティを多く含んでいたためである．また，表 3 より，ステップ数が必ずしも 1 ではないことがわかる．さらに，出次数の標準偏差がテーブル型データと判断した 2 つのデータよりも大きいこともわかる．以上よりこのデータはテーブル型ではないと判断できる．

DBpedia Japanese(図 6) では，共通プロパティが少ない傾向が横手市 AED 加工データよりも顕著に現れている．抽出された起点リソースは，人名や地名，学校名や神社などの施設が主である．同じ人名でも国籍や職業の違いから様々なプロパティが出現したため，出現回数の少ないプロパティが多く現れた．

複数回出現したプロパティは 2 種に大別される．1 つは大抵のリソースが保有するプロパティで,「http://www.w3.org/1999/02/22-rdf-syntax-ns#type(リソースのタイプ)」や「http://dbpedia.org/ontology/wikiPageID (wikipedia のページ ID)」などが該当する．もう 1 つは，ある 1 つのリソースが同じプロパティをいくつも持っている場合である．DBpedia Japanese には後者のパターンが多く見られた．また，表 3 より，出次数の標準偏差が他のデータよりもかなり大きいこともわかる．前述の通り DBpedia は大規模であるため探索の上限を 30 としたが，100 個の起点リソースのうち，ステップ数が 30 以内で終了したものは 7 個であった．このことより，DBpedia Japanese のデータセット内で多くのリンク関係が存在すると言える．以上のことから，本実験結果では DBpedia Japanese はテーブル型ではないと判断できる．

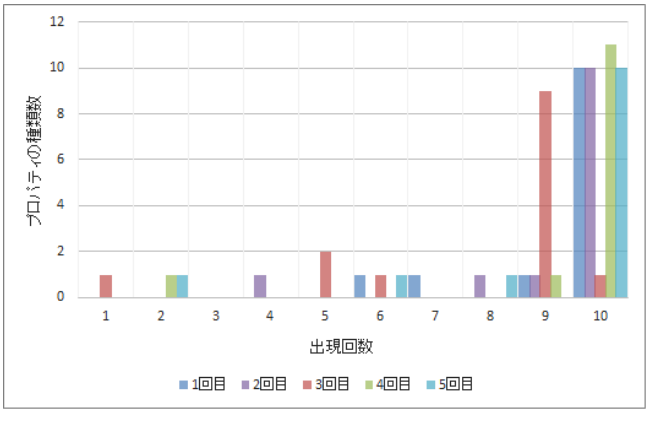

図 3: 調査 1 の結果：横手市 AED 設置場所

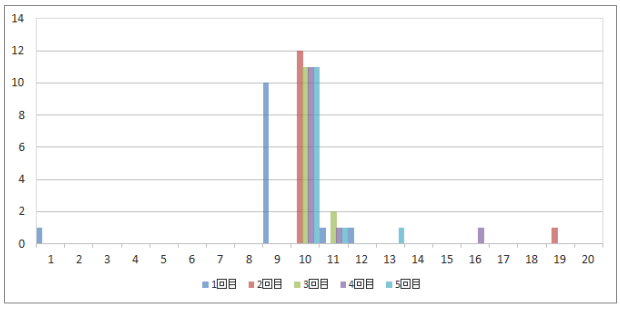

図 4: 調査 1 の結果：神奈川名所 LOD データセット

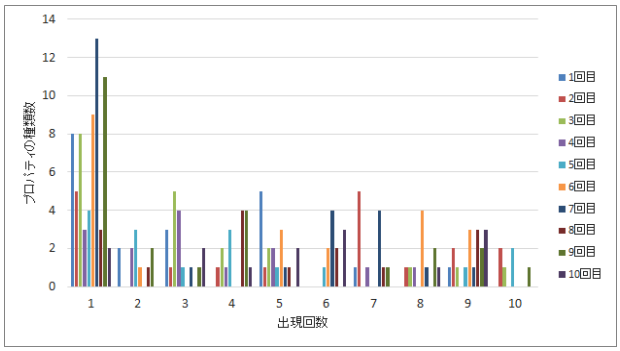

図 5: 調査 1 の結果：横手市 AED 設置場所加工

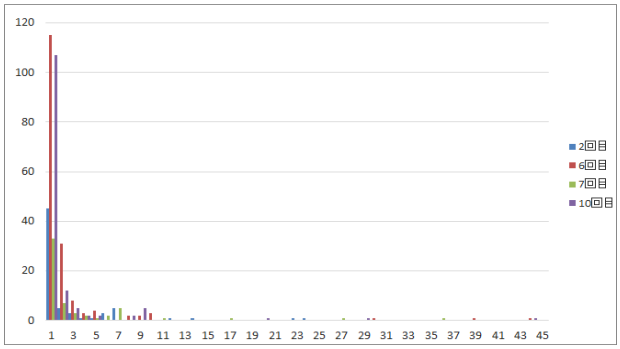

図 6: 調査 1 の結果：DBpedia Japanese

表 3: 調査 2 の結果

	出次数		STEP 数	
データセット	平均	標準偏差	平均	標準偏差
横手市 AED	11.76	1.59	1	0
神奈川名所 LOD	13.32	1.81	1	0
横手市 AED 加工	7.42	3.47	1.29	0.791
DBpedia Japanese	28.5	24.95	-	-

4.3 DBpedia Japanese に関する考察

DBpedia Japanese は日本で公開されている LOD の中で巨大なデータセットの一つであり，各データセットを繋ぐハブのような役割を果たしている[7]．しかし大規模な分，どの様なデータが含まれているかを知ることは困難であるため，その構造を把握することは有用であると考える．本節では，予備実験を通じて観察された DBpedia Japanese の特徴的な構造について考察する．

人名や地名などの様々なリソース“http://ja.dbpedia.org/resource/○○”が持つ，共通プロパティ“http://xmlns.com/foaf/0.1/isPrimaryTopicOf”はプロパティ“http:// xmlns.com/foaf/0.1/primaryTopic”と図 7 のように相互関係を持っていることがわかった．“primaryTopic”の主語リソースは，“http://ja.wikipedia.org/wiki/○○”であり，wikipedia のページである．また，目的語は“http://ja.dbpedia.org/resource/○○”である．両プロパティは同じ関係を逆向きに表現したものであるため，この様な循環的構造をとっている．このように構造が決まっているプロパティを探索過程で発見できれば，DBpedia Japanese の構造理解に役立てることができると考える．

また，本実験過程で得られた，“http://ja.dbpedia.org/resource/Category:○○”が主語として出現する場合，特有のプロパティを持つことがわかった．表 4 に示すプロパティ“core#related”は関連するカテゴリ“Category:○○”が目的語となる．プロパティ“core#broader”は主語リソースを包含する上位カテゴリ“Category:○○”が目的語となる．“pref#Label”は主語リソースのラベル (リテラル) が目的語となる．リソースによって出次数はばらつきがあるものの，プロパティの種類数にはあまり差異がなかったため“Category:○○”を主語としたときのトリプル構造を大雑把に表型と捉えることは可能と考えられる．すなわち，DBpedia Japanese には同種の情報が構成するテーブル型データが複数含まれ，それらの間につながりがあることが想定される．この点については今後調査を行う必要があると考える．

30 ステップ以内に探索が終了しなかったリソースに関して，展開されるリソースにある一定のパターンが存在する事が観察された．よく現れるパターンとしては，学問に関するリソースが連続して展開されるパターン，都道府県に関するリソースが連続するパターン，日本の歴代総理大臣が連続するパターン，欧米の地名から各国の大統領へ遷移するパターンなどが観察された．それらは元の起点リソースが一見全然関係ないものでも現れた．例えば起点リソース“http://ja.dbpedia.org/resource/極道の妻たち_危険な賭け”の場合，俳優や映画といったリソースから世界観，宇宙論・宇宙物理学，物理学…と遷移した．この様なパターンが頻繁に観察された理由として，本稿では出次数の大きいノードを選んでいく探索を行ったため，一度出次数の大きいノードが展

[7] http://linkedopendata.jp/?p=411

開されると，後は毎回同じパスが展開されることが挙げられる．例えば，DBpedia Japanese では,「日本」というリソースを目的語に持つリソースが多いため，このリソースが探索の過程で現れる確率は高く，さらに出次数が 133 と大きいため，展開されやすい．このため，その後に展開されるパターンが類似したものになる場合が多く発生した．DBpedia Japanese の構造を探る上でこのような決まったパターンが多く出現してしまうと，探索の妨げになる可能性があるため，今後対処法を検討する必要がある．

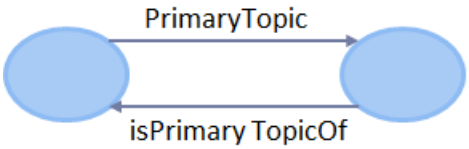

図 7: 相互リンクを持つリソース

表 4: “ Category:○○ ”特有のプロパティ

http://www.w3.org/2004/02/skos/core#related
http://www.w3.org/2004/02/skos/core#broader
http://www.w3.org/2004/02/skos/core#prefLabel

5　まとめ

本稿では，RDF で記述された LOD を探索的にブラウズする作業を支援するシステムの要素技術として，SPARQL クエリによるサンプリングに基づく LOD 構造の推定手法に関する基礎的な検定を行った．

本実験では，3.1 節で述べたテーブル型データの特徴を元にデータがテーブル型であるか否かを判断したが，データセットによっては，ある主語リソースが同じプロパティを複数持つ場合も観測されたため，プロパティの出現回数だけでは型の判定が難しくなる可能性があると考える．よって型判定の精度を高めるためにはプロパティの出現率も考慮する必要がある．今回使用したテーブル型のデータは 1 つのテーブルデータを 1 つの RDF データに変換したものと想定されるため，ステップ数の平均や標準偏差の情報だけでもテーブル型判と断することができた．しかし，複数のテーブルデータを 1 つの RDF データにまとめたようなデータが存在し表と表の間に繋がりある場合，ステップ数が必ずしも 1 ではなくなる．そういったデータセットにおいて表部分をどう発見するか今後の課題である．また，推定精度と実行時間の関係についても調査を行い，サンプリングの起点とするリソース数について検討することも必要と考える．さらに，探索的ブラウズの起点として提示すべきノードについての検討も今後の課題である．

参考文献

[1] 仲前晋太郎, 成凱: Blog における話題分析のためのランダムサンプリング手法の提案, *DEIM Forum*, D3-5, 2010

[2] 仲前晋太郎, 成凱: Reservoir を用いた巨大グラフのランダムサンプリング, *DEIM Forum*,D3-2,2011

[3] 鹿島久嗣: ネットワーク構造予測, 人工知能学会論文誌, Vol.22, No.3, pp.344-351, 2007

[4] 安田雪, 松尾豊, 武田英明: リンクマイニングによる研究者ネットワークの抽出:成長プロセスと国内外からの見え方, 第 *21* 回人工知能学会全国大会, 1B2-8, 2007

[5] 後藤孝行, 濱崎雅弘, 武田英明: DashSearch LD: 探索的検索の Linked Data への適用, 第 *26* 回人工知能学会全国大会, 3C1-OS-13a-3, 2012

[6] 田代航一, 高間康史: RDF データベースを対象としたデータ分析支援ツールの提案, 第 *5* 回情報アクセスと可視化マイニング研究会, SIG-AM-05-02, pp7-12, 2013

[7] 一瀬詩織, 小林一郎, 岩爪道昭, 田中康司: DBpedia における SPARQL 検索結果のランキング手法, 第 *27* 回人工知能学会全国大会, 2N5-OS-21b-4, 2013

[8] 一瀬詩織, 小林一郎, 岩爪道昭, 田中康司: DBpedia を対象にしたリソースのランキング手法における一考察, 情報処理学会第 *75* 回全国大会, 4N-9,2013