観光領域の Linked Data を対象とした横断的知識ベースの構築法

槇 俊孝† 髙橋 和生‡ 若原 俊彦‡

†福岡工業大学大学院 〒811-0295 福岡県福岡市東区和白東 3-30-1

E-mail: † {bd15002, mgm16105}@bene.fit.ac.jp, wakahara@fit.ac.jp

あらまし Linked Data は，Uniform Resource Identifier (URI)によりウェブ上のリソースを識別し，そのリソースのメタデータを記述したデータであり，オープンデータとして公開したものを Linked Open Data (LOD)という．LOD の公開件数は年々増加しており，日本国内においても地域や施設などに関する LOD が多数公開されている．しかし，現在公開されている LOD は，述語やリンクなどのデータ構造に課題があり，汎用的に用いることが難しいと考えられる．このため，本稿では，Linked Data における述語の統一と潜在的リンクの推定により，横断的知識ベースを構築する Resource Propagation Algorithm (RPA) を提案する．RPA は，URI のリンクが全く存在しない Linked Data でも，キーワード特性を考慮してキーワードやカテゴリ，市区町村のリンクを推定可能である．実験の結果，孤立状態にあったリソースが減少し，Linked Data の汎用的利用が可能になる見通しが得られた．

キーワード Linked Data，オープンデータ，知識ベース，リンク推定，語彙，観光

1. はじめに

Linked Data は, Resource Description Framework (RDF) [1]に基づいて主語，述語，目的語の 3 つ組(triple)でウェブ上に存在するリソースのメタデータを体系的に記述したデータである． Linked Data は，ウェブ上のリソースを Uniform Resource Identifier (URI) により識別し，また，各リソースの関係性をそれぞれの URI により参照することが望ましい [2]．2009 年に米国のオバマ政権がオープンガバメント[3]を表明して以降，行政を中心として公共データのオープンデータ化が進み，Linked Data をオープンデータとして公開した Linked Open Data (LOD)が脚光を浴びている．日本における事例としては，福井県鯖江市が「データシティ鯖江」[4]をスローガンとして,観光地や避難場所などの LOD を積極的に公開している．また，電子情報通信学会は，学会誌や論文誌，研究技術報告，企業誌などの文献メタデータを Linked Data として蓄積し，I-Scover SPARQL Endpoint [5]を提供している．さらに Wikipedia のデータベースを LOD に変換した DBpedia [6]は, 様々な領域における LOD を横断的にリンクするクロスドメインとして重要な存在となっている．

LOD の公開件数は，世界的に増加しているが，様々な課題が浮上している．例えば，LOD STATS の調査によると，LOD STATS が認識している 9,960 件のデータセット (約 1,500 億 triples) のうち 6,971 件のデータセットはデータ構造やアクセス環境に問題があることを示している[7]．また，日本国内で公開されている LOD においても様々な課題が存在する．図 1 は，LinkData.org [8]で公開されており，ダウンロード数が多い地域関係の 100 件の LOD を可視化したグラフの一部である．

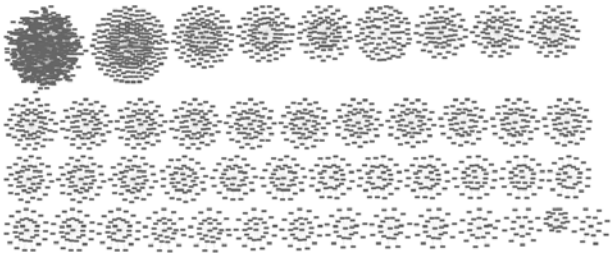

図 1 LinkData.org 上の LOD におけるグラフ構造の一部

図 1 は，「京都市観光スポットリスト_2013」や「さばえトイレ情報」などの triple が含まれており，ここでは URI 形式の主語と目的語を source と target として可視化した．同図のグラフは，6,872 nodes，5,564 edges から構成されており，1,615 件のコンポーネントが存在する．つまり，各リソースの意味概念が共有されず孤立状態にあることを示している．このため，複数のデータセットを用いたサービスの開発が困難となり，データセットに合わせて個別にサービスを開発する他ない現状にあると考えられる．本来，LOD は，相互にリンクすることでリソースの意味概念を共有し，意味概念の再開発を避けて効率的に知識ベースを構築できるものであり，セマンティックウェブ（データのウェブ）の形成に寄与するものである．

このため，本研究では，孤立状態にある LOD の潜在的なリンクを推定し，知識ベースを構築することを目的とし，観光領域の LOD を対象として有効性を検証する．本稿の構成は次の通りである．第 2 章で Linked Data における URI の性質を議論し，第 3 章で関連研究を述べる．第 4 章で提案アルゴリズムについて述べ，第 5 章で実験と考察を述べた後に，第 6 章で本稿の内容を纏める．

2. Linked Data における URI の性質

Linked Data は，XML スキーマ定義言語 (XSD) [9]を基盤としてデータ型が定義されており，Web Ontology Language (OWL) [10]によって構造体のデータ型を新たに定義可能である．Linked Data における triple は，主語と述語が URI 型 (xsd:anyURI)で記述され，目的語が URI 型や文字列型 (xsd:string)，整数型 (xsd:integer)，小数型 (xsd:decimal)などで記述される．

Linked Data は，RDF クエリ言語である SPARQL [11]により取り扱うことができ，triple を指定することで検索や分析が可能である．SPARQL は，count 関数や sum 関数の他，if 関数や replace 関数，regex (regular expression) 関数などの様々な関数が使用でき，RDF データの高度な利活用が期待される．例えば，DBpedia Japanese を用いて「福井県にある名湯百選の名称と所在地」を取得するクエリを図 2 のように記述できる．“?”から始まる文字列は変数を表しており，triple の条件に該当する値が代入される．本クエリでは，「福井県」と「名湯百選」をキーワードとしリンク構造に基づいて検索し，rdfs:label (名称)と property:所在地に該当する各値を変数に代入して select により出力している．本クエリを実行すると name:"芦原温泉"@ja，address:"福井県あわら市"@ja が得られる．regex 関数を用いることで全文検索も可能であり，図 3 に示すクエリでも同結果を得られる．但し，URI による完全一致検索と，文字列による部分一致検索では，後者は処理コストが高いため応答時間を要する．図 2 のクエリ応答時間は平均 0.09 秒であるのに対し，図 3 のクエリ応答時間は平均 0.56 秒であり，応答時間に 6 倍以上の差異が生じることが分かる．ビッグデータを取り扱う場合，この差異はさらに拡大することが想定されるため，可能な限り完全一致で検索できることが望ましく，URI によりリソースを参照することが重要であると考えられる．また，URI は，一般的な関係データベースにおける ID に相当し，データアクセスの効率化だけでなく冗長性の削減と一意性の確保に寄与する．さらに，URI はウェブ上のリソースを参照可能なため，LOD 間に横断的なリンクを設定することができ，ウェブ上に大規模な知識ベースを構築することが可能となる．

以上のことから，Linked Data の triple における目的語は可能な限り URI 型で記述することが望ましく，正規形に変形できないラベルやコメント，住所，緯度，経度などを文字列型や小数型などで記述することが最良であると考えられる．このため，本研究における潜在的なリンクの推定では，URI 型リソースの関係性を導出することする．Linked Data のグラフ構造に基づいてリンクを推定するために，述語のマッピング機能と述語に対応したエッジ重み（伝搬定数）を定義した語彙基盤を実装する．

```
prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>
prefix property:<http://ja.dbpedia.org/property/>
prefix ontology:<http://dbpedia.org/ontology/>
prefix resource:<http://ja.dbpedia.org/resource/>

select
  ?name
  ?address
where {
  ?subject
    rdfs:label ?name;
    property:所在地 ?address;
    ontology:wikiPageWikiLink resource:福井県, resource:名湯百選.
}
```

図 2　URI による検索クエリの例

```
prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>
prefix property:<http://ja.dbpedia.org/property/>

select
  ?name
  ?address
where {
  ?subject
    rdfs:label ?name;
    rdfs:comment ?comment;
    property:所在地 ?address.

    filter(regex(?comment, "福井県") = true)
    filter(regex(?comment, "名湯百選") = true)
}
```

図 3　文字列による検索クエリの例

3. 関連研究

本章では，RDF データの構築において重要な語彙基盤と，リンク推定に関する関連研究について述べる．

3.1. 語彙基盤

World Wide Web Consortium (W3C)は，RDF データの述語統一と構造化のために，RDFS や OWL，WGS84 Geo などの様々な語彙を提供している．日本では，情報処理推進機構が共通語彙基盤[12]の整備を進めている．述語は，RDF データの triple において主語と目的語の関係を意味付けする機能を担う．関係データベースにおけるテーブルのカラムと同様であり，データ型の定義や使用回数などの制約を設定できる．現在公開されている多くの LOD は，個々に述語が定義されており，述語の統一が進んでいない状況がある．図 1 に示したデータセットの場合，664 種類の述語が用いられており，ラベルや住所などの類似した述語が再定義されている．本研究では，観光領域における述語の統一化と潜在的リンクの推定を目的として，観光語彙基盤の整備を進めている．従来の語彙基盤には存在しない述語のマッピングやエッジ重みを定義する伝搬定数の機能を有している．また，非ネスト構造で簡単に RDF データを記述できる特徴がある．

3.2. Silk

Silk は，Julius Volz 氏や Christian Bizer 氏らが開発したセマンティックウェブのためのリンク発見フレームワークであり，2 つの異なるデータセット間にリンクを生成できる[13]. 文字列や数値，日付などの各類似度

に基づいて双方向のリンクを生成でき，そのリンクの述語を個別に指定できる．例えば，データセット X において要素 A，B が述語 C でリンクされているが，データセット Y において要素 B, A がリンクされていない場合，要素 B, A を owl:sameAs のような述語 C’でリンクするものである．

本研究で提案する Resource Propagation Algorithm (RPA)は，任意のデータセットと DBpedia をリンクするものであり，リンクの述語として tour:キーワード，tour:カテゴリ，tour:市，tour:区，tour:町，tour:村，及び tour:都道府県の 7 種を全自動で推定可能である．なお，tour の prefix は，本研究で提案する観光語彙基盤を示している．Silk における文字列の類似度は，Dice 係数を応用した jaroSimilarity により評価されるため，日本語を取り扱うためには事前に分かち書きが必要である．また，リソースの意味概念を表すキーワードやカテゴリを推定することが難しいと考えられる．これに対して本研究で提案する RPA は，意味概念の推定に特化したものであり，後述のキーワード特性に基づいた TF-IDF により精度良くキーワードを推定できる．

3.3. 潜在的リンクの推定

ノード間のリンクを推定する手法として，Jaccard 係数[14]や，ラベル伝搬を応用したリンク伝搬[15]などがある．従来のリンク推定を Linked Data に適用することを想定したとき，図 1 に示したように十分なリンクが存在するデータセットが少ないため，精度良く推定することが難しい．このため，本研究では，述語のマッピング，概念推定，地域推定を順に施した後に，表層的な文字列では推定できない潜在的なリンクを推定する各機能を実装した RPA を提案する．RPA は，一般的なラベル伝搬アルゴリズムと同様に隣接ノードは同じクラスに属するという仮定に基づいており，ノードにラベルを付与する点は同じであるが，教師データを必要としないため，既存の様々な Linked Data に対して適用しやすいと考えらえれる．

4. Resource Propagation Algorithm

RPA は，Linked Data の潜在的リンクを推定して，知識ベースを生成する新しいアルゴリズムであり，図 4 に示すように 4 つの機能から構成され，それぞれの機能は観光語彙基盤，DBpedia，及び IPAdic を用いている．

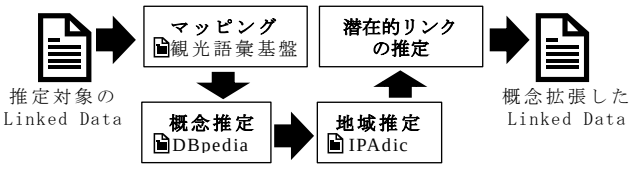

図 4　RPA の概略図

4.1. 観光語彙基盤

観光語彙基盤は，著者らが整備を進めている，観光領域の RDF データを記述するための述語セットである．観光語彙基盤は，主に以下の特徴を有している．

- 観光領域の述語を提供

観光に関するウェブ上の全てのリソースに対してメタデータを記述できるようにすることを目的とし，記事型をマスタとして画像や動画などのリソースのメタデータを記述できる述語を提供する．

- 非ネスト構造

述語をドメインにより管理しているため，非ネスト構造の RDF データを作成できる．LinkData.org や自治体独自のデータカタログサイトで公開されている LOD は，非ネスト構造で記述されているものが多いため，容易に述語を対応付けることが可能である．

- URI の識別子

知識ベースとして用いることが可能な RDF データを作成できるように，URI 型を基本とした述語構成とする．URI は，参照可能であることが条件であるため，表記揺れの発生を抑制する効果も期待できる．

- 日本語の述語

日本語表記の述語を提供する．述語を統一化して Linked Data を記述することで，RDFS や OWL，WGS84 Geo などの他の語彙に変換することも可能となる．

- マッピング

正規表現に対応した述語のマッピング機能を提供しており，様々な LOD を統合して利用可能となる．

- 伝搬定数

主語と目的語の関係性の強さを示すエッジ重み（伝搬定数）により，柔軟に RDF データをグラフデータに変換することができる．

観光語彙基盤は，以下の名前空間で公開しており，“tour”を接頭辞として用いることを想定している．

http://www.tourism.property/#

図 5 は，観光語彙基盤を用いて「沖田中央公園」のリソースを記述した例である．“dbpedia:公園”のように URI で目的語を記述することで，他の LOD におけるリソースの概念を継承できる．

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix tour: <http://www.tourism.property/#>.
@prefix dbpedia: <http://ja.dbpedia.org/resource/>

<http://www.tanoshingu.org/沖田中央公園>
  tour:名称 "沖田中央公園"@ja, "Okita central park"@en;
  tour:概要 "沖田中央公園は，福岡県糟屋郡新宮町にある
          セントラルパークである．"@ja;
  tour:カテゴリ dbpedia:公園;
  tour:キーワード dbpedia:公園, dbpedia:自然, dbpedia:噴水;
  tour:記事型.
```

図 5　観光語彙基盤を用いた Linked Data の記述例

4.2. マッピング

RPA は，始めに Linked Data の述語をマッピングする．例えば，以下のように類似した述語を“tour:名称”に統一する．変換対象の文字列は，観光語彙基盤の各述語で定義されており，任意値を設定可能である．

- http://www.w3.org/2000/01/rdf-schema#label
- http://purl.org/dc/terms/title
- http://imi.go.jp/ns/core/rdf#表記
- http://schema.org/name
- http://linkdata.org/property/rdf1s2442i#名称

4.3. キーワード推定

キーワードは，リソースの概念を単語，あるいは単語の組み合わせで表現したものであり，複数のキーワードを用いてリソースを判別できることが望ましいと考えられる．本研究では，電子情報通信学会の I-Scover SPARQL API を用いて図 6 に示すようにキーワード特性を評価し，その特性値を導入した式(1)の TF-IDF によりキーワードを推定する．εは，キーワードの特性値であり，表 1 に示す文字列のパターンに応じた各式により算出する．n_{t,r_s}は，リソースr_sに対応する用語tの出現回数であり，$f_{r_s}(t)$は，用語tに対応するリソースr_sの件数である．N は，全リソースの件数である．τ_{t,r_s}は，リソースr_sにおける用語tの評価値であり，任意の閾値以上の用語をキーワードとして同定できる．

$$\tau_{t,r_s} = \varepsilon \cdot \frac{n_{t,r_s}}{\sum n_{t,r_s}} \left(\log \frac{N}{f_{r_s}(t)} + 1 \right) \qquad (1)$$

表 1　キーワード特性

文字列のパターン	ピーク時文字数	ピーク以下	ピーク以上
{英数字（小文字）}	16	ε = 0.070x − 0.120	ε = − 0.052x + 1.832
{英数字（大文字）}	3	−	ε = − 0.121x + 1.363
{カタカナ}	7	ε = 0.200x − 0.400	ε = − 0.093x + 1.651
{漢字}	4	ε = 0.330x − 0.320	ε = − 0.257x + 2.028
{ひらがな, 漢字}	5	ε = 0.238x − 0.190	ε = − 0.207x + 2.035
{カタカナ, 漢字}	8	ε = 0.223x − 0784	ε = − 0.188x + 2.504
{英数字, カタカナ}	8	ε = 0.215x − 0.720	ε = − 0.104x + 1.832
{英数字, 漢字}	9	ε = 0.176x − 0.584	ε = − 0.072x + 1.648
{ひらがな, カタカナ, 漢字}	9	ε = 0.143x − 0.287	ε = − 0.148x + 2.332
{英数字, カタカナ, 漢字}	10	ε = 0.166x − 0.666	ε = − 0.107x + 2.070

表 2　各パターンにおけるキーワードの文字数

文字列のパターン	総計	文字数	網羅率
{英数字（小文字）}	94,234	3 - 33	93.3%
{英数字（大文字）}	13,706	3 - 11	93.7%
{カタカナ}	16,912	3 - 12	95.9%
{漢字}	55,081	1 - 6	93.5%
{平仮名, 漢字}	7,759	2 - 8	90.3%
{カタカナ, 漢字}	41,652	4 - 12	92.2%
{英数字, カタカナ}	4,164	4 - 16	91.8%
{英数字, 漢字}	2,299	3 - 19	91.0%
{平仮名, カタカナ, 漢字}	2,981	5 - 14	90.5%
{英数字, カタカナ, 漢字}	3,480	6 - 17	90.4%

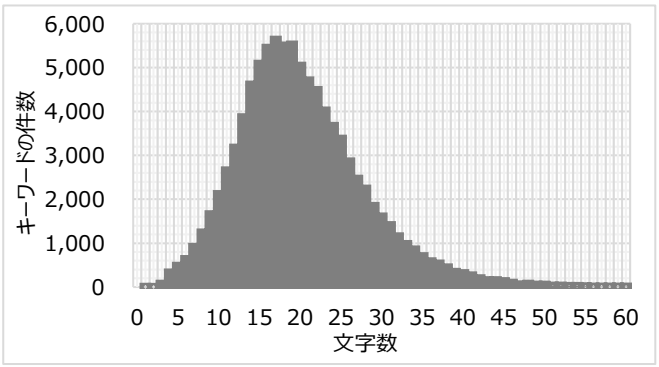

図 6　英数字（小文字）で構成されたキーワードの文字数と件数の分布

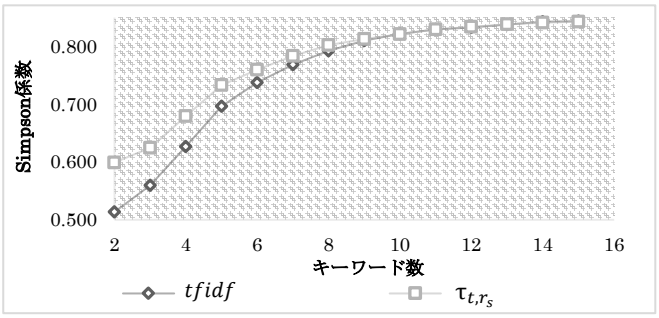

図 7　Simpson 係数によるキーワード推定の評価

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix tour: <http://www.tourism.property/#>.
@prefix dbpedia: <http://ja.dbpedia.org/resource/>

dbpedia:観光
   tour:名称 "観光"@ja;
   tour:キーワード dbpedia:レジャー, dbpedia:観光圏;
   tour:説明 "観光（かんこう）は、一般には、楽しみを目的とする
         旅行のことを指す。"@ja;
   tour:参考 <http://ja.wikipedia.org/wiki/観光>;
   tour:ライセンス <https://creativecommons.org/licenses/by-sa/3.0/>;
   tour:提供者 <http://wiki.dbpedia.org/about/dbpedia-community>;
   tour:作成者 <http://www.tourism.property/aboutUS>;
   a tour:用語型.
```

図 8　DBpedia から生成した用語辞書の例

τ_{t,r_s}は，表 2 に示す各パターンにおける文字数によって候補となるキーワードを制限しているため，0 以上 1 以下の実数となる．2016 年電子情報通信学会総合大会で発表された論文のうち，論文キーワードが概要文に含まれている 2,846 件の論文を対象としてキーワード推定の精度を評価したところ図 7 の結果が得られた[16]．なお，論文キーワードを正解データとし，I-Scover に登録されている約 33 万件のキーワードを辞書として使用している．同図の結果より，キーワード特性を考慮した TF-IDF は，一般的な TF-IDF よりもキーワード推定の精度が高いことが分かる．

RPA では，キーワード推定のために DBpedia のラベルデータから生成した 506,543 語の辞書を用いており，この辞書は図 8 に示すように観光語彙基盤に基づいている．

4.4. 地域推定

形態素解析エンジンで用いられている IPAdic を用いて，推定キーワードに含まれる市区町村を判定し，tour:都道府県，tour:市，tour:区，tour:町，tour:村の各述語に対応する目的語を推定する．推定キーワードが上述のいずれかの述語に対応する場合は，その推定キーワードをキーワード群から除外する．

4.5. カテゴリ推定

観光語彙基盤では，1 つの主語に対して最大 1 件のカテゴリを記述する制約を設けている．推定されたキーワード群において使用回数が多い順にキーワードをカテゴリの候補とし，各主語における推定キーワードと一致したらその推定キーワードをカテゴリとする．

4.6. 潜在的リンクの推定

RPA は，先述の各推定を実施した後に，グラフ構造に基づいて潜在的なキーワードのリンクを推定する．一般的なラベル伝搬アルゴリズムと同様に，隣接ノードは同じクラスに属すると仮定して，キーワードのラベルを伝搬する．まず，triples における主語と目的語をノードとし，観光語彙基盤に基づいて述語に対応するエッジ重み（伝搬定数）を設定し，無向グラフとして取り扱う．次に，多様な LOD に対応できるように教師データを式(2)に基づいて自動的に設定する．deg_iは，ノードiに接続されたエッジの総数（次数）であり，$max\,deg$は，ノード群における最大の次数である．任意の閾値以上のノードを教師データとして設定する．式(3)は，ラベル予測値の更新式であり，$edge_{i,j}$はノードi，j間のエッジ重みである．

$$C_i = \frac{log(deg_i)}{log(max\,deg)} \tag{2}$$

$$p_{j,k} + \frac{edge_{i,j} * p_{i,k}}{\sqrt{deg_i}} \rightarrow p_{j,k} \tag{3}$$

$p_{i,k}$は，伝搬元のノードiにおけるラベルkのラベル予測値である．$p_{j,k}$は，伝搬先のノードjにおけるラベルkのラベル予測値である．教師データとして設定された$p_{i,k}$は 1.0 の値がセットされており，ノードiからノードjまでの最短経路を経て再帰的にラベル予測値が伝搬される．本推定は，マルチラベルに対応しており，1 つのノード（主語）に対して複数のラベル（キーワード）が推定される．下限値を設けており，任意の値以下となった場合は，該当の伝搬を停止することで高速化を図っている．

先行研究において，福岡県糟屋郡新宮町の LOD を対象として潜在的リンクの推定精度を評価した結果，図 9 の結果が得られている[17]．正解データとして新宮町 LOD を用いており，ランダムにキーワードリンクを削除して推定対象の Linked Data を生成した．同図の横軸は，この Linked Data と正解データを Dice 係数により評価した値である．縦軸は，この Linked Data の潜在的リンクを推定して得られた Linked Data と正解データを Dice 係数により評価した値である．同図より，Linked Data のリンク構造が改善していることが分かる．なお，この結果は tour:カテゴリの伝搬定数を 0.5，tour:キーワードの伝搬定数を 0.95 に設定したときの結果である．

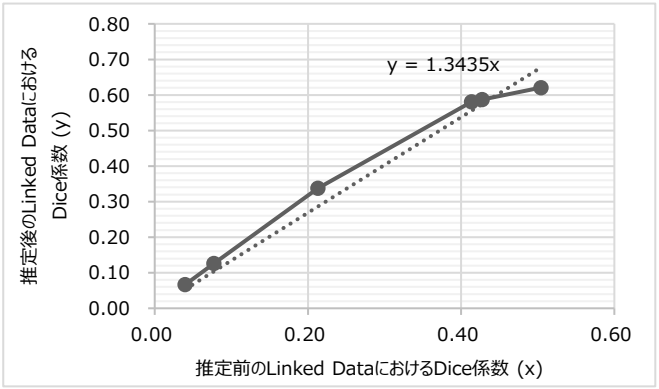

図 9　潜在的リンクの推定における性能評価

5. 実験

本実験では，LinkData.org で公開されており，ダウンロード数が多い 100 件のデータセット (108,942 triples)を対象とし，RPA によるリンク推定の効果を検証する．概念推定における各主語の最大キーワード数を 5 件とし，$0.3 \leq \tau_{t,r_s}$とする．また，潜在的リンクの推定においては，各主語に対する最大キーワード数を 3 件とし，また，$0.3 \leq C_i$，下限値 0.1 としてラベル予測値が 0.3 以上のラベルをキーワードとする．これにより，各主語に対して最大 8 件のキーワードが付与されることとなる．図 10 は，RPA により推定された Linked Data のグラフ構造の一部であり，推定前は図 1 に示したグラフ構造の通りである．

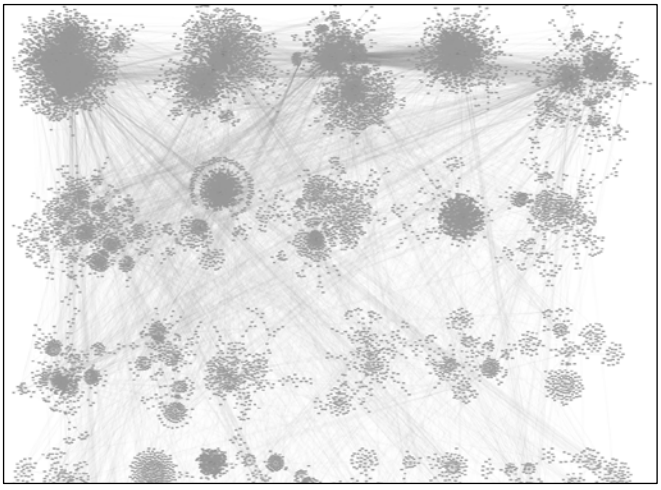

図 10　RPA により推定された Linked Data のグラフ構造

```
<http://www3.city.sabae.fukui.jp/ls/001>
    <http://linkdata.org/property/rdf1s283i#city> "鯖江市"@ja ;
    <http://linkdata.org/property/rdf1s283i#title> "きらめきロード中河"@ja ;
    <http://linkdata.org/property/rdf1s283i#llocation> "上河端町、浅水川堤防沿い"@ja ;
    <http://www.w3.org/2003/01/geo/wgs84_pos#lat> "35.95254"^^xsd:float ;
    <http://www.w3.org/2003/01/geo/wgs84_pos#long> "136.207561"^^xsd:float ;
    <http://linkdata.org/property/rdf1s283i#feature> "水辺"@ja ;
    <http://linkdata.org/property/rdf1s283i#season> "春"@ja ;
    <http://linkdata.org/property/rdf1s283i#description> "鯖江市の東部を流れる浅水川の堤防
    沿いの通り。桜並木と地域の人々が植えた水仙が美しい花を咲かせます。4 月上旬
    ～中旬の頃が特に美しい景観となります。"@ja ;
    <http://linkdata.org/property/rdf1s283i#url> <http://www3.city.sabae.fukui.jp/ls#1> ;
    <http://linkdata.org/property/rdf1s283i#imageurl>
        <http://www3.city.sabae.fukui.jp/ls/image/No1.jpg> ;
    <http://linkdata.org/property/rdf1s283i#imagelargeurl>
        <http://www3.city.sabae.fukui.jp/ls/imagelarge/No1.jpg> .
```

図 11　RPA による推定前の triples の例

```
<http://www3.city.sabae.fukui.jp/ls/001>
    <http://linkdata.org/property/rdf1s283i#season> "春"@ja;
    tour:ウェブページ <http://www3.city.sabae.fukui.jp/ls#1>;
    tour:カテゴリ dbpedia:景観;
    tour:キーワード dbpedia:中旬, dbpedia:堤防, dbpedia:景観, dbpedia:桜並木;
    tour:名称 "きらめきロード中河"@ja;
    tour:市 dbpedia:鯖江市;
    tour:市区町村 "上河端町、浅水川堤防沿い"@ja, "鯖江市"@ja;
    tour:特記事項 "水辺"@ja;
    tour:画像 <http://www3.city.sabae.fukui.jp/ls/image/No1.jpg>,
              <http://www3.city.sabae.fukui.jp/ls/imagelarge/No1.jpg>;
    tour:経度 "136.207561"^^xsd:float;
    tour:緯度 "35.95254"^^xsd:float;
    tour:説明 "鯖江市の東部を流れる浅水川の堤防沿いの通り。桜並木と地域の人々が
    植えた水仙が美しい花を咲かせます。4 月上旬～中旬の頃が特に美しい景観となり
    ます。"@ja.
```

図 12　RPA による推定後の triples の例

RPA により 6,872 nodes から 18,324 nodes に増加し，また，5,564 edges から 39,174 edges に増加した．これにより，推定前は 1,615 件のコンポーネントが存在したが，推定後は 235 件に減少し，孤立状態のリソースが改善されたことが分かる．Triples の件数は，108,942 triples から 141,554 triples に増加しており，32,612 triples のリンクが増加したことになる．図 11 と図 12 は，それぞれ推定前と推定後の triples の一部である．推定前は，ウェブページと画像の URI のリンクのみであったが，推定後はカテゴリやキーワード，市のリンクが追加されたことが分かる．また，推定前は 664 種類の述語が存在したが，マッピングにより 430 種類に減少したことから，メタデータの取り扱いが容易になったと考えられる．

6. むすび

近年，Linked Open Data (LOD) の公開件数が世界的に増加しており，日本国内でも増加傾向にある．Linked Data における URI は，リソース間を横断的にリンクするために必要なものであり，データのウェブを構築するために不可欠である．しかし，多くの LOD は，十分な URI のリンクが存在せず，機械判読が課題となり二次利用が難しい状況にあると考えられる．このため本研究では，リソースの意味概念を推定し，キーワードやカテゴリの潜在的リンクを推定する Resource Propagation Algorithm (RPA)を提案した．RPA は，述語のマッピング，概念推定，地域推定，潜在的リンクの推定の 4 つの機能から構成されている．LinkData.org に登録されている LOD を対象として RPA の有効性を検証したところ，リソースのキーワードやカテゴリのリンク数が増加し，DBpedia を介して各リソースが横断的にリンクされたことから，RPA は LOD の知識ベース化に有効であると考えられる．

謝　辞

本研究は JSPS 特別研究員奨励費 17J09765 の助成を受けたものである．

参　考　文　献

[1] Stefan Decker, Prasenjit Mitra, Sergey Melnik, “Framework for Semantic Web: An RDF Tutorial”, IEEE Internet Computing, Volume 4, Issue 6, DOI: 10.1109/4236.895018, pp. 68-73, 2000.

[2] Christian Bizer, Tom Heath, Tim Berners-Lee, “Linked Data -The Story So Far”, International Journal on Semantic Web and Information Systems (IJSWIS), Volume 5, Issue 3, pp. 1-22, 2009.

[3] Wendy R. Ginsberg, “The Obama Administration’s Open Government Initiative: Issued for Congress”, CRS Report for Congress, pp. 1-32, 2011.

[4] Data City Sabae , “データシティ鯖江とは”, 鯖江市, http://data.city.sabae.lg.jp/data-city-sabae/ (Accessed on January 10).

[5] 五味弘，“I-Scover で始める文献探しの旅”，電子情報通信学会, 情報・システムソサイエティ誌, Volume 22, Issue 3, pp. 25-28, 2017.

[6] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann, “DBpedia – A Crystallization Point for the Web of Data”, Web Semantics: Science, Services and Agents on the World Wide Web, Volume 7, Issue 3, pp. 154-165, 2009.

[7] “LOD STATS”, http://stats.lod2.eu/ (Accessed on January 10).

[8] 一般社団法人リンクデータ，“LinkData.org”, http://linkdata.org/ (Accessed on January 10).

[9] World Wide Web Consortium, “W3C XML Schema Definition Language”, https://www.w3.org/TR/xmlschema11-1/, Accessed on February 13, 2018.

[10] World Wide Web Consortium, “OWL Web Ontology Language”, https://www.w3.org/TR/owl-features/, Accessed on February 13, 2018.

[11] World Wide Web Consortium, “SPARQL Query Language for RDF”, https://www.w3.org/TR/rdf-sparql-query/, Accessed on February 13, 2018.

[12] 情報処理推進機構，“共通語彙基盤概要”, https://imi.ipa.go.jp/doc/IMI_Overview_v2.pdf, Accessed on January 10，pp. 1-9，2015.

[13] Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov, “Silk – A Link Discovery Framework for the Web of Data”, 18th International World Wide Web Conference, 2009.

[14] David Liben-Nowell, Jon Kleinberg, “The Link Prediction Problem for Social Networks”, Journal of the American Society for Information Science and Technology, Volume 58, Issue 7, pp. 1019-1031, 2007.

[15] 鹿島久嗣，加藤毅，山西芳裕，杉山将，津田宏治，“リンク伝搬法：リンク予測のための半教師付き学習法”，人工知能基本問題研究会，Volume 73，pp. 19-24，2009.

[16] 槇俊孝，古賀大騎，髙橋和生，若原俊彦，小舘亮之，曽根原登，“Linked Data の知識ベースを拡張する Resource Propagation Algorithm の特性”, 信学技報, Volume 117, no. 389, LOIS2017-69, pp. 111-116, 2018.

[17] Toshitaka Maki, Kazuki Takahashi, Toshihiko Wakahara, Akihisa Kodate, Noboru Sonehara, “Resource Propagation Algorithm to reinforce Knowledge Base in Linked Data”, The 20th International Conference on Network-Based Information Systems NBiS-S9 (NBiS2017) pp.476-483, 2017.