


```

<?xml version="1.0" encoding="UTF-8"?>
<terms>
  <general>
    <term>名勝</term>
  </general>
  <synonyms>
    <term>名勝</term>
    <term>名勝地</term>
    <term>景勝</term>
    <term>景勝地</term>
    <term>特別名勝</term>
  </synonyms>
</terms>

```

図 3 Broadic による「名勝」の異表記取得

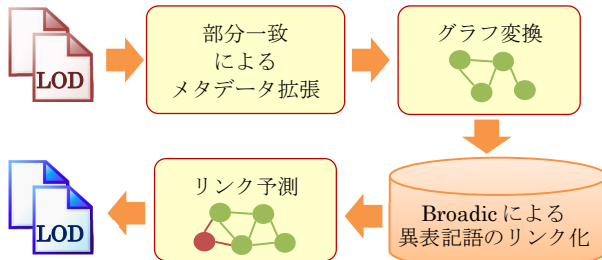


図 4 MLPA への Broadic 導入手法

4. 表記揺れ補償辞書 API「Broadic」の概要

表記揺れとは、ある語彙に対して同義語または類義語が存在することである。例えば、「名勝」という語彙に対して「名勝地」や「景勝」等の異表記が存在する。基本的に計算機は「名勝」、「名勝地」、「景勝」をそれぞれ別の語彙として扱うため、図 1 に示したクエリの「名勝」を「景勝」と記述した場合は「称名滝」の結果が得られない。このため、計算機に表記揺れを考慮させる機能が必要である。Broadic は、インターネット百科事典 Wikipedia [13]と電子情報通信学会の文献検索システム I-Scover [14]のデータを用いて構築した表記揺れ補償辞書 API である。これを用いると例えば、「名勝」の表記揺れの語彙を以下の URI により図 3 に示すように XML 形式で取得できる。

<http://ictlab.ce.fit.ac.jp/broadic?名勝>

5. MLPA への Broadic 導入手法

図 4 に MLPA への Broadic 導入手法の概略図を示す。始めに、予め設定したカテゴリ属性の語彙に対して部分的に一致するデータにメタデータを追加する。例えば、「福岡工業大学」というデータに対して「福岡」や「大学」を追加する。次に {ノード 1, ノード 2, エッジ重み} を 1 組としてグラフデータに変換する。このとき、語彙が同一でも URI が異なるものが多く存在するため、語彙のみをノードに割り当てる。そして、このグラフデータに含まれる語彙の表記揺れを Broadic により補償し、異表記の語彙との新しいエッジ (リンク) を作成する。最後にリンク予測を行い、リンクを拡張した LOD を出力する。

6. 評価実験

本評価では、LinkData.org に 2016 年以降に登録された LOD からランダムに 20,000 件のデータを取得し、第 5 章で述べた手法によって LOD のリンク予測を行う。図 5(a)は、LOD に対して部分一致によりメタデータを拡張してグラフ化したものである。また、図 5(b)は、図 5(a)の LOD に対して Broadic を用いずにリンク予測を実施した結果である。

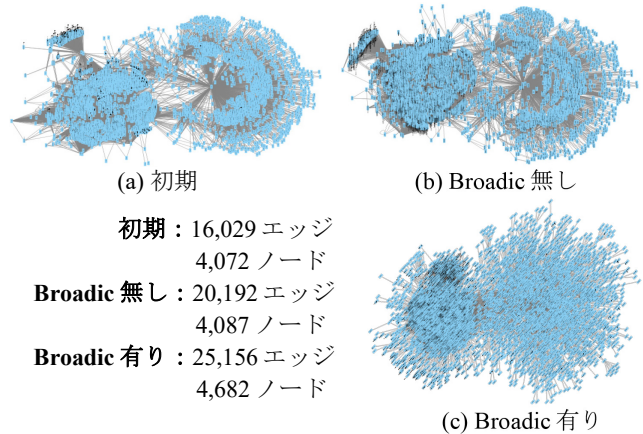


図 5 Broadic の有無による MLPA の性能比較

本評価実験では、15 種類のカテゴリ属性の語彙を 50 ノードに教師データとして与えた。このため、初期状態よりも 15 ノードだけ増加し、リンク数は 4,163 エッジだけ増加し、グラフ構造が密になった。更に Broadic を用いることで同義異表記語が追加されてノード数が増加し、図 5(c)のように密なグラフ構造となることを確認した。例えば、「中学校」に対して「学校」や「施設」等のリンクが拡張された。

7. おわりに

本稿では、表記揺れ補償辞書 API である Broadic を MLPA に導入し、性能改善を図る手法について述べた。評価実験の結果、Broadic によってリンク数が 1.25 倍だけ増加したため、Broadic がリンク予測に有効であることを確認した。今後の課題は、述語の種類によりエッジの重みを変化させ、MLPA の精度向上を図ることである。

謝辞

本稿を執筆するにあたり、LinkData.org に登録されている LOD を利用致しました。本 LOD の作成者、並びに一般社団法人リンクデータに深く感謝申し上げます。

参考文献

- [1] Michael Hausenblas, "5-star Open Data", <http://5stardata.info/>, Accessed on June 27, 2016.
- [2] 総務省, "統計におけるオープンデータモデル事業—地域振興とビジネスの活性化に向けて—", 2015/04/24, http://www.soumu.go.jp/main_content/000389540.pdf, Accessed on June 27, 2016.
- [3] Toshitaka Maki, Kazuki Takahashi, Toshihiko Wakahara, Akihiro Yamaguchi, Yu Ichifuji, Noboru Sonehara, "A New Multiple Label Propagation Algorithm for Linked Open Data", IMIS 2016, Big Data and Data Management, July 2016.
- [4] 横 俊孝, 若原俊彦, "I-Scover 文献メタデータを用いた時系列・技術要因分析システム", 電子情報通信学会論文誌, vol. J99-D, no. 10, October, 2016.
- [5] Tom Heath, Christian Bizer, "Linked Data: Evolving the Web into a Global Data Space", Morgan & Claypool Publishers, ASIN: B009KC1YM2, Feb 2011.
- [6] Jorge Pérez, Marcelo Arenas, Claudio Gutierrez, "Semantics and complexity of SPARQL", ACM Transactions on Database Systems, vol. 34, no. 16, 2009.
- [7] DBpedia Community, "DBpedia Japanese", <http://ja.dbpedia.org/>, Accessed on June 27, 2016.
- [8] DBpedia Community, "Virtuoso SPARQL Query Editor", <http://ja.dbpedia.org/sparql>, Accessed on June 27, 2016.
- [9] W3C, "RDF Schema 1.1", <https://www.w3.org/TR/rdf-schema/>, Accessed on June 27, 2016.
- [10] 一般社団法人リンクデータ, "LinkData.org", <http://linkdata.org/>, Accessed on June 27, 2016.
- [11] 横 俊孝, 高橋和生, 若原俊彦, 山口明宏, 曾根原 登, "地域の Linked Open Data の現状課題とその解決法に関する考察", 信学技報, LOIS 研究会, 2016/07/15.
- [12] 横 俊孝, 高橋和生, 若原俊彦, 山口明宏, 一藤 裕, 曾根原 登, "CityData の LOD を対象としたマルチプルラベル伝搬アルゴリズムの適用手法", 信学技報, vol. 116, no. 23, pp. 101-106, May, 2016.