情報処理学会論文誌　データベース　Vol.5　No.3　141–148 (Sep. 2012)

Wikipediaカテゴリを用いたWikipediaとGeoNames間のリンク発見とメンテナンス

吉岡 真治[1,a)]　劉 亦奇[1]　神門 典子[2]

受付日 2012年3月20日, 採録日 2012年5月9日

概要：近年，地理情報を扱う情報システムの増加にともない，地理情報に関するデータベースへのニーズが高まっている．GeoNamesは，Open Dataとしては，最大規模の地理情報データベースである．本データベースをLinked Open DataとしてWikipediaの情報を媒介として関連づけることにより，Web Ontologyの開発などに役立てられている．ただし，GeoNamesとWikipediaの間のリンクについては，自動的なリンク発見の試みがいくつか行われているものの，十分な数のリンクが付与されている状態ではない．本論文では，Wikipediaのカテゴリ情報を使うことで，精度良くWikipediaのページに対応するGeoNamesのエントリを発見する方法を提案する．また，本手法は，既存のリンク中の不適切なリンクを発見する際にも利用可能であることを示す．本手法の成果については，すでに，GeoNamesの管理者に報告しており，その成果の一部は，GeoNames中のリンク情報として公開・修正が行われている．

キーワード：地理情報データベース，Linked Open Data，Wikipedia，リンク発見，データのメンテナンス

Discovery and Maintenance of Links between Wikipedia and GeoNames by Using Wikipedia Category

Masaharu Yoshioka[1,a)]　Yiqi Liu[1]　Noriko Kando[2]

Received: March 20, 2012, Accepted: May 9, 2012

Abstract: Recently, due to the higher demand for geographic information system, it is necessary to have a good geographical database for such systems. GeoNames is one of the largest geographical database as Open Data. This database is also used for constructing web ontology by adding links to the Wikipedia page as a part of Linked Open Data. Even though, here are several attempts to find links automatically, the number of links between GeoNames and Wikipedia is not sufficient. In this paper, we propose an automatic link discovery method to use Wikipedia categories to identify the correspondence between Wikipedia page and GeoNames entry. We also propose to use this method for inappropriate link detection. Link data obtained in this paper is already sent to the manager of GeoNames and a part of the result is used for updating the site.

Keywords: geographic information database, Linked Open Data, Wikipedia, link discovery maintenance of data

1. はじめに

近年，Wikipedia[*1]など有用なユーザ生成コンテンツが，数多く，Web上に公開されるようになってきている．また，これらのコンテンツの有用性をさらに高めるために，複数のコンテンツの間の関係（Linked Dataと呼ぶ）を記述することにより，よりその有用性を高めようとするLinked Open Data [1]と呼ばれる活動がさかんになっている．特に，Wikipediaからデータを抽出して，様々な関係を取り出して利用可能としたDBPedia[*2]のデータを中心に様々

[1] 北海道大学
Hokkaido University, Sapporo, Hokkaido 060–0814, Japan
[2] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] yoshioka@ist.hokudai.ac.jp

[*1] http://wikipedia.org
[*2] http://dbpedia.org

なデータが作成され Linked Data として利用可能な形で公開されている．

これらの Linked Data には，半自動に作成されるものから，ユーザ参加型で作成されるものまで様々なものがあり [2]，様々な形での情報利用の可能性を提示している．

一方，モバイルデバイスの普及などにともない，地理情報を扱う情報システムやデータベースへのニーズが高まっている．これらのニーズにこたえる形で，Linked Open Data の活動に参加している地理情報のデータベースである GeoNames [*3] に対する注目が高まっている．GeoNames は，世界各国の 800 万以上の地名に関するデータベースであり，Open Data としては，最大規模のものであり，Web Ontology である，GeoWordNet [3] や YAGO2 [4] などにおける地理情報の基本データとして活用されている．しかし，GeoNames と Wikipedia の間の対応関係が付けられたデータは，2012/2/1 現在で，330,017 件しか存在せず，十分に対応関係が付けられているという状態ではない．

この問題に対し，本論文では，英語版の Wikipedia と GeoNames のデータの対応関係を自動的に発見する方法を提案するとともに，提案手法が，人手などによって作成された Linked Data の不適切なリンクを発見することができ，リンク情報のメンテナンスに役立つことを示す．

このような，異種データ間の対応関係を発見する方法については，文献データベースなどを中心に，様々な研究が行われている [5] が，基本的には，対応関係を判定するにあたって，適切な手がかりを見つけることが重要である．

これまでの研究では，Wikipedia に記述されている緯度・経度の座標情報を利用したリンク発見手法 [4] が行われてきている．しかし，座標情報に頼った場合には，座標情報が設定されていない Wikipedia のページとの対応関係がとれないといった問題や，座標が近い同名のエントリが必ずしも適切なエントリではないといった問題があるため，一定レベルの不適切なリンクを含む可能性があった．

これに対し，本論文では，Wikipedia のカテゴリ情報を用いることにより，Wikipedia のページがどの国のどの地方についての記述であるかを判断するとともに，どのようなクラス（山，川，街など）の情報であるかを推定することにより，座標情報に頼らない，リンクの発見手法を提案する．また，クラス間の対応関係の知識を用いることにより，既存の座標情報に注目して作成されたリンク中に存在する不適切なリンクが発見できることについても述べる．

本論文の構成は，以下のとおりである．2 章では，GeoNames と Wikipedia の紹介を行うとともに，座標情報を中心としたリンク発見の手法として，YAGO2 の手法を紹介する．3 章では，リンク発見のための，基本的なアプローチについて述べるとともに，具体的なリンク発見の手法について述べる．さらに，4 章では，実際に行ったリンク発見と不適切なリンクの発見に関する結果について述べるとともに，その考察を行う．5 章では，本論文のまとめを行う．

2. Wikipedia と GeoNames

2.1 Wikipedia

Wikipedia [*4] は，Wiki をベースとして作成された百科事典であり，幅広い分野に関する項目を網羅している．この Wikipedia のデータの特性に注目して Wikipedia マイニング [6] による情報・知識の発見の研究や，Wikipedia オントロジ [4], [7] の構築などの研究が多く行われている．

ここでは，本研究に関連する Wikipedia に関する特徴を説明する．

(1) カテゴリによる分類
Wikipedia は百科事典であり，その項目は，カテゴリにより分類されている．英語版の Wikipedia のカテゴリには，「Geography of the United States」といった対応する地理情報の国名や場所の推定に役立つカテゴリや，「Mountains of the United States」といったクラスの推定に役立つカテゴリが存在する．Wikipedia オントロジの研究では，主に，後者のカテゴリ情報を用いて，ページのクラス推定などを行っている．また，このカテゴリには，階層関係があり，明示的に指示されていないカテゴリでも親カテゴリを参照することで，より上位レベルのカテゴリ情報との関係を判断することができる．

(2) InfoBox による定型的な情報の記述
Wikipedia のページには，ページに記述する内容に対応する形で属性情報を整理して表示する InfoBox が存在する．地理情報に関する多くのページでは InfoBox からその位置が緯度経度の座標で獲得することができる．

(3) リダイレクトリンクの利用
Wikipedia のページには，表記にぶれがある場合には，リダイレクトリンクにより，代表表記を見つけることが可能である．

2.2 GeoNames

GeoNames は，Creative Commons の Attribution ライセンスで開発されている地名情報に関するデータベースである．2012 年 2 月 1 日時点で，8,105,590 件の世界中の地名の情報が存在している．各地名の情報には，複数の属性が定義されているが，今回の提案手法示で用いる属性についてのみ，**表 1** に示す[*5]．

ここで，地名の Feature Code（FC）には，9 つの大分類と，さらに，詳細な 656 の小分類に分かれている．**表 2** に

[*3] http://www.geonames.org/

[*4] http://www.wikipedia.org/

[*5] 詳細については，http://www.geonames.org/を参照のこと

表 1　GeoNames のデータ構造（概要）
Table 1　Data fields of GeoNames.

属性名	内容
ID	地名の ID
名前	地名の ASCII 表記
別名	地名の各国語表記
国名	所属する国名
行政単位	所属する行政単位で 4 レベル レベル 1（日本の都道府県）， レベル 2（市区町村），···
Feature Code（FC）	地名のタイプ
座標	地名の緯度・経度の座標

表 2　GeoNames の Feature Code
Table 2　Feature Codes of GeoNames.

コード	説明	例
A	国，州， 県など	ADM1：行政単位レベル 1 の中心地 （日本の場合は，県庁など）···
H	川や湖など	LK：湖，STM：川，SEA：海，···
L	公園など	PRK：公園，MILB：軍事基地 AMUS：アミューズメント施設，···
P	町や村など	PPL：街や村など，PPLC：国の首都， PPLW：破壊された街 ···
R	道路や 鉄道など	RD：道，RR：鉄道，TNL：トンネル， OILP：オイルパイプライン
S	ビルや 農場など	AIRP：空港，BLDG：ビル， FRM：農場，···
T	山や丘など	MT：山，ISL：島，HLL：丘， VLC：火山，···
U	深海	MTU：海の山，TRNU：海溝， SHFU：海洋棚，···
V	森など	TREE：木，FRST：森，···

大分類とその大分類に属する小分類の例を示す．

2.3　座標情報を用いたリンク発見

Wikipedia と GeoNames の間のリンク発見の手法としては，YAGO2 [4] の研究で用いられている Wikipedia の InfoBox から獲得した緯度・経度の座標情報と名前を利用する方法がある．この手法では，座標情報を持つ Wikipedia のページに対して，以下の条件を満たす GeoNames のエントリとの間にリンクを設定する．

- Wikipedia の名前と GeoNames の名前が完全に一致するエントリ．
- Wikipedia と GeoNames の座標から求まる距離が 5km 以内のエントリ．ただし，複数存在する場合には，最も距離が近いもの．

この手法により，84,349 件のリンクを発見したと述べている [4]．

本手法では，名前の完全一致を行っているために，表記のぶれ（たとえば，現地語表記と英語表記の違いなど）に対応できない．また，Wikipedia のページに座標情報が設定されている必要があるという制限のため，十分な数のページに対して対応関係を発見することができていない．

また，単純に距離で対応するエントリを選択しているために，たとえば，「Achensee」というオーストリアの湖のページの座標情報が，「Achensee」の鉄道の駅と近いために，湖ではなく，鉄道の駅と対応づけるといった間違いが起こる．

3.　Wikipedia カテゴリを用いた Wikipedia と GeoNames 間のリンク発見

3.1　基本方針

2.3 節で紹介したリンク発見の手法には，十分な数のページを発見できないだけではなく，不適切なリンクを発見する可能性がある．

そこで，本手法では，まず，対応するリンクの数を増やすために以下の基準でリンク発見を行う．

(1) 名前のマッチング
名前のマッチングを行う際には，GeoNames の別名や Wikipedia のリンク情報を利用する．また，Wikipedia のページでは，地名などの曖昧性解消の際に，「Ueno, Mie」，「Ueno, Tokyo」のように，「,」を使って，追加情報を付与する場合がある．よって，これらの「,」以降の情報を除去したうえで，マッチングを行う．

(2) 座標情報を用いない場所の特定
多くの場合，国名や行政単位レベル 1（日本の都道府県やアメリカの州）が定まると，地名とその対象が一意になる可能性が高くなる．よって，これらの所属情報を Wikipedia のカテゴリ情報から推定することで，座標情報を用いないで，Wikipedia のページと GeoNames のエントリの対応関係の妥当性を調べることができる．ただし，このような形で対応関係をつける候補を増やすと，本来，地名でないものが対応したりする可能性がある．

次に，不適切な対応関係を取り除くための手段について述べる．2.3 節で述べたように，不適切な対応関係の多くは，Wikipedia のページの内容から判断できるクラス（山，川，街など）と GeoNames の FC の間の問題である．よって，Wikipedia のページの内容からクラスを判断し，その判断と FC の間の対応関係を利用し，その妥当性を検証する．

Wikipedia のページの内容からクラスを判断する方法としては，Wikipedia オントロジの研究などで用いられているカテゴリ情報を利用することとした．また，このカテゴリ情報を利用することにより，先に述べた地名以外のデータとのマッチングを抑制することも期待できる．

3.2　Wikipedia ページからの所属情報の生成

前節で述べたように，本手法では，座標情報ではなく，Wikipedia ページから，そのページがどの国のどの地方に

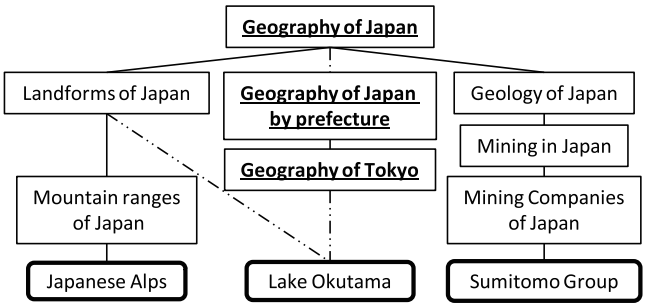

図 1　Geography of Japan のサブカテゴリと所属する Wikipedia ページの例

Fig. 1　Subcategories of “Geography of Japan” and their Wikipedia pages.

所属している地理情報を述べているかを推定して利用する．

具体的には，英語版の Wikipedia に存在する「Geography of ～」というカテゴリを利用する．英語版の Wikipedia では，地理情報についてのカテゴリの多くは，「Geography of ～」をその上位カテゴリに持つ．**図 1** に「Geography of Japan」の例を示す．本図中で，四角いノードは，カテゴリを表し，一番下の丸四角のノードは Wikipedia のページを表す．また，図中の点線は，サブカテゴリが直接関係付けられているものを示し，鎖線は間にいくつかのサブカテゴリが存在することを示す．

また，図の「Geography of Japan by prefecture」の例のように，多くの国のカテゴリにおいては行政単位レベルに対応したカテゴリが存在する．本手法では，これらのカテゴリを用いて所属情報を決定する．具体的には，「Geography of Japan」の下位カテゴリにある Wikipedia のページを日本に所属していると考える方法である．これを再帰的に繰り返し，「Geography of Japan」の下位カテゴリである「Geography of Tokyo」のさらに下位カテゴリのものは，日本の東京都に所属すると考える．たとえば，図 1 の例からは，「Japanese Alps」は日本の地名と判断し，「Lake Okutama」は東京都の地名と判断する．

一方，多段階に，下位カテゴリを検索すると「Mining Companies of Japan」といった地名と関係ないカテゴリに到達して，結果として，「Sumitomo Group」のように地名と関係ないページを候補に含む場合がある．よって，これらの問題による影響を減少させるために，各「Geography of ～」からは最大 4 段下までのカテゴリのみに限定することとした．

Wikipedia のページの所属情報を推定するのに有用なもう 1 つの情報は，曖昧性解消のための付加情報である．前節でも述べたように，Wikipedia のページにおいて，地名に曖昧性がある場合に，「,」以降にその曖昧性解消に有用な情報を追加する．本手法では，この情報と先のカテゴリから得られる情報を組み合わせて利用する．たとえば，「Geography of Japan」の下位カテゴリに存在する「Ueno, Tokyo」は，日本の東京の Ueno と判断する．

3.3　Wikipedia ページからのクラス推定と FC との比較

本手法において，Wikipedia から抽出するクラスの詳細度は，FC との対応がとれるレベルであることが望ましい．Wikipedia オントロジの研究では，すべてのカテゴリをクラスのように扱う場合もあるが，本手法では，もう少し，抽象的なクラスを利用する．

実際に，どのようなクラスが必要なのかを分析するために，GeoNames が提供している Wikipedia と GeoNames のリンク情報を利用した．本分析には，2011 年 9 月 1 日現在のデータである 147,226 ペアを利用した．

具体的には，各々の GeoNames の FC（たとえば，PPL：街，MT：山など）に対して，対応する Wikipedia のページのカテゴリ情報を収集し，有用なカテゴリの記述を選定した．この記述の選定の際には，単純に頻度を用いるのではなく，GeoNames の FC の説明も参考にして，手作業で対応関係を作成した．

この対応関係を作成する際に，GeoNames と Wikipedia のページの間にある不整合が発見された．たとえば，「Ubinas」（ペルーの山）は，Wikipedia のページを見ると火山であると判断できるが，GeoNames では，「VLC：火山」ではなく「MT：山」として，分類されている．これは，多くのユーザにとって，GeoNames の詳細な FC の違いを意識して入力することが困難であり，結果として，類似したあるいは，上位カテゴリと考えられる FC が設定されることがあることを示していると考えた．

本手法では，上記のような問題もふまえ，各 FC に対して，個別に対応するクラスを設定するのではなく，類似していると考えられる FC をひとまとめとして，その FC のグループに対して，対応するクラス情報を推定することとした．

クラス情報の推定には，Wikipedia オントロジにおいてよく用いられるカテゴリ中のキーワードを用いることにした．また，実際に十分な数のリンクの存在しない FC があったため，656 の FC 中 363 の FC について，54 グループに分けたうえで対応関係を設定した．**表 3** に代表的な FC のグループとそのキーワードを示す．

3.4　リンク発見システムの作業手順

Wikipedia と GeoNames のリンク発見は以下の手順で行う．

(1) 地名に関する Wikipedia ページ候補の発見

3.2 節で述べた方法を用いて，「Geography of ～」の「～」の部分を国名に限定することにより，少なくとも国名が判断できた Wikipedia のページに限定をすることが可能となる．

表 3　GeoNames の Feature Code と対応する Wikipedia カテゴリ中のキーワード
Table 3　Keywords of Wikipedia categories for estimating GeoNames Feature Codes.

ADM1, ADM2, ADM3, ADM4, ADMD, PPL, PPLA, PPLA2, PPLA3, PPLA4, PPLC, PPLF, PPLG, PPLL, PPLQ, PPLR, PPLS, PPLW, PPLX, PPLX（街，首都，県庁所在地など）	Census-designated place, Community, Commune, Cities, Suburbs, Departments, Governments, Parishes, Administrative, Neighbourhoods, Neighborhoods, Provinces, Constituencies, Populated Place, States, Goverment, Governorates, Woredas, Cantons, Prefectures, Subprefectures, Comarcas, Wards, Districts, Governorate, Counties, Capital, Municipalities, Divisions, Communities, Villages, Town, Municipality, Frazioni, Boroughs
AIRQ, AIRB, AIRF, AIRP（飛行場）	Air Force, Air Base, Airport, Airfields, Air Bases, Airfield, Airports
HLL, HLLS, MT, MTS, VLC, PK, CONE（山，丘）	Ranges, Ridge, Peak, Peaks, Hill, Nunataks, Moutains, Cone, Cones, Nunatak, Mountains, Moutain, Mountain, Range, Hills, Volcano, Volcanoes, Ridges
ISL, ISLS, ISLET, ISLF, ISLM, ISLT, ISLX（島）	Isle, Island, Isles, Islands
MN, MNAU, MNC, MNCR, MNCU, MNFE, MNN, MNQ（鉱山）	Mines, Mining, Mine

(2) ページ候補の所属情報の推定
カテゴリ情報を用いて，所属情報を推定する．ただし，1 つの Wikipedia のページは複数のカテゴリを持つため，複数の国にまたがる海外線や自国外の基地や植民地などの場合に，複数の国名がページに設定される可能性がある．本手法では，Wikipedia のページ名に曖昧性回避のための情報が書いてあり，その情報を用いて，所属情報が推定できる場合には，そちらを優先する．それでも，所属情報が確定できない場合には，複数の候補をそのまま利用するが，行政区分レベル 1 の所属情報が推定されている場合には，それに対応する国名のみに限定する（例，US：アメリカ or CA：カナダの場合は，両方残すが，US，WA：アメリカのワシントン州 or CA：カナダの場合は，US，WA：アメリカのワシントン州のみとする）．

(3) GeoNames との対応候補の生成
GeoNames の別名を含む名前と Wikipedia のリダイレクトを含む名前を用いて，名前の対応関係を判定するとともに，Wikipedia のページの所属情報と GeoNames の国名，行政単位の情報を比較することにより，リンクの候補を生成する．

(4) 明らかに地名でないページの除去
カテゴリから，明らかに人名や映画の名前といった地名でない情報であることが判断できる場合には，そのページを候補から削除する．具体的には，films, albums などのキーワードを含むカテゴリを持つページを候補から削除する．

(5) Wikipedia カテゴリと FC の対応表による妥当性チェック
3.3 節で述べた Wikipedia カテゴリと FC の対応表を用いて，妥当なリンクのみに限定する．

4. 本手法によるリンク発見と不適切なリンクの発見

4.1 リンク発見の実験

3 章で述べた手法を用いて，実際にリンク発見の実験を行った．リンク発見に用いたデータは，GeoNames については，2012 年 2 月 1 日現在の 8,105,590 件（330,017 件に Wikipedia とのリンクが存在[*6]）であり，Wikipedia については，英語版の 2012 年 1 月 4 日分のダンプである．

まず，最初に，「Geography of 国名」のカテゴリから，Wikipedia のカテゴリを調べることによりから 1,131,546 件のページを地名の候補として抽出した．このページに対して，GeoNames の名前と所属情報の対応関係を調べるとともに，明らかに地名でないと判断したページを削除した結果，421,585 件の Wikipedia のページに対して，1,472,438 件の GeoNames とのリンクが候補として作成された．この候補に対して，カテゴリと FC の対応表を用いて妥当なリンクを選択すると，308,284 件のリンクが得られた．

本システムが生成するリンクの妥当性を検証するために，GeoNames が配布している GeoNames と Wikipedia のリンクを正解とした 330,017 件を用いた評価を行った．本正解データは，Wikipedia と GeoNames のエントリが 1 対 1 で対応づけられている．これに対し，本手法では，1 つの Wikipedia のページに対して，複数の GeoNames のエントリを対応づける場合があるため，その結果は，完全一致，候補に正解を含む，異なる，新規の Wikipedia ページに対するリンクの 4 種類に分けることとなる．その結果を**表 4** に示す．表中の () 付きの値は，1 対多のリンクを含むことにより得られたリンクの総数を示す．

新規については，正解判定ができないため，GeoNames 提供のデータのみを正解と判断した場合の精度と再現率を

[*6] GeoNames のサイト運営者の Marc Wick 氏に問い合わせたところ，当時，座標情報を中心としたリンクの自動発見を行っていた．現在，我々の初期段階の研究成果 [8] などもふまえ，リンク発見の手法を改良中である．

表 4　GeoNames のリンクデータと本手法で作成したリンクとの比較
Table 4　Comparison between GeoNames links and automatic generated links.

	完全一致	候補に含む	異なる	新規
すべての候補	140,509	70,343 (155,234)	3,694 (5,308)	93,738 (125,534)
精度重視	127,114	0	1,245	62,660

表 5　ランダムサンプリングによる新規発見リンクの評価
Table 5　Evaluation of newly found links by random sampling.

正解	不正解			
	まったく別の場所	関連する地名	曖昧性解消	GeoNames の問題
679	14	2	4	1

計算すると，各々 70.0%，63.9%と満足のいく結果とはならなかった．

そこで，正解と判断されなかったリンクの情報を分析したところ，次のような場合に誤りが多く発生することが確認された．

(1) 1 つの GeoNames のエントリと複数の Wikipedia のページとのリンク
(2) 複数の GeoNames のエントリと 1 つの Wikipedia のページとのリンク
(3) Wikipedia のページの所属情報として判定されたエリア内に同じ名前の地名が複数存在する場合

これらの条件に合致するデータは，再現率を重視する場合には重要であるが，精度の低下をもたらす．よって，これらの条件を満たすエントリを候補から削除することにより，精度重視のリンクを生成した．このリンクでは，GeoNames と Wikipedia の関係は，1 対 1 に限定されるため，「候補に含む」は 0 となり，Wikipedia のページ数とリンク数は一致するため () 付きのリンクの総数は記載していない．精度重視の設定の場合，精度は 99.0%となり，非常に高い精度を得ることができた．再現率は 38.5%と低いが，既存のデータ総数の約 2 割弱に相当する 62,660 件のデータを発見できていることから，本手法の有用性は十分にあると考えている．

ただし，今回用いた精度はあくまでも，これまでに分かっていたデータに対する精度であり，新規に発見したデータについて，同じような性能を持つ保証は存在しない．そのため，この 62,660 件のデータから 700 件のランダムサンプリングを行い，その内容について分析を行ったところ**表 5** に示すように，3%（21 件）のデータが不適切と判断された．しかし，これらのリンクには，Wikipedia のページと直接の対応関係は読み取るのが困難であるが，対象のページに関連すると考えられる地名が 2 件，判定時に，曖昧性解消のページが間に想定された件数が 4 件，GeoNames の座標情報が不適切なため，対応関係が判定できないものが 1 件含まれるため，実質は 2%（14 件）が本手法の原因による誤りである．関連する地名に関する不正解については，後述する不適切なリンクの発見の議論との関係が深いため，そこでまとめて議論することとする．

表 6　リンク不一致の分析
Table 6　Analysis of disagreement between two types of links.

	提案手法（適切）	提案手法（不適切）
オリジナル（適切）	741	183
オリジナル（不適切）	303	18

4.2　不適切なリンクの発見

先に述べたリンク発見の評価では，GeoNames のリンク情報に不適切なリンクはないという前提で行ったが，実際に結果が異なるデータについて，その内容を比較した場合に，GeoNames に記述されているリンクの情報自体が不適当であるという場合が数多く存在した．

よって，この手法を用いて，結果が異なる状況になったデータを分析することで，効率の良い不適切なリンクの発見が行えると考え，我々のシステムが不正解と判定した 1,245 件の不一致について検討を行うこととした．その結果を**表 6** に示す．

この分析における判断基準は，下記のとおりである．

(1) GeoNames と本手法の結果得られたリンクの両方を適切と考える場合（741 件）
GeoNames のエントリにおいては，1 つのエントリは 1 つの FC を持つ形式で登録されている．そのため，1 つの場所に複数の役割があった場合に，同じ名前の類似した FC を持つエントリが複数存在する場合がある．たとえば，「Sapporo-shi」というエントリには，行政単位レベル 2（市区町村）というエントリと，行政単位レベル 2 の中心地という 2 つのエントリが存在する．このような場合に，どちらかが正しくて，どちらかが間違いとはいえない．街に関する FC 以外では，空軍基地と空港や，灯台と気象観測所のように，いくつかのグループが存在する．
基本的には，Wikipedia のエントリの定義に相当する，第 1 パラグラフの内容から判断できるものを優先すべきであると考えるが，それでも曖昧性が残る場合がある．このような場合に，現時点では，両方正解と判断して分析を行った．
(2) GeoNames に登録されている不適切なリンク（303 + 18 = 321 件）
緯度・経度の座標による近さを優先しているために，本来の Wikipedia のページの記述とは，直接関係ない，あるいは，その一部であると思われる FC に属する GeoNames と対応づけている．たとえば，前者の例

としては，列車の駅，交番，図書館などでは，街の名前と同じ名前が用いられる場合が多かった．後者の例としては，空港と空港ビルや，公園と公園管理事務所といった組合せが存在した．これらの誤りは本手法で提案しているクラス情報の推定により判断できたものであり，本手法によるエラー発見の有用性を示していると考えている．

(3) 本手法における不適切なリンク（183 + 18 = 201 件）
本手法の誤りの原因は，大きく 3 つに分けられる．

- 地名の曖昧性解消の失敗（157 件）
  本手法では，精度重視の設定を行う際に，曖昧性が低いものを選ぶことにしたが，その中でも，同地域にある別の地名と対応づけてしまい，誤りとなった．
- Wikipedia ページのクラスの判定ミス（11 件）
  クラスの判定を行うために，単純な単語とのマッチを行ったために，「〜 in the United States」といった固有名詞の記述から「States」を抽出して，街，県庁所在地などのクラスを付加してしまい，誤りとなった．
- Wikipedia のリダイレクトを用いたことに起因するミス（28 件）
  Wikipedia のリダイレクトには，単なる別名だけではなく，古い名前，関連して記述されている内容（たとえば，湖の記述に対してそこにあるダム）などが記述されている．古い名前などで，別名と認識できるようなものは正解と判定したが，Wikipedia のページだけでは，その別名と判断できないようなものは不正解と判断した．
- GeoNames の座標情報が不適切（5 件）
  GeoNames の座標情報が不適切と思われるため，判断できない．

全体には誤りの少ない GeoNames のデータから，これだけ高い確率で，不適切なリンクを発見できたことから，本手法はリンクのメンテナンスに使うための有用なツールであると考えられる．

今後の性能改善のためには，Wikipedia のリダイレクトを使う際の基準と，クラス推定の際のカテゴリの利用法についての再検討が必要であるとと考えられる．

4.3 実験のまとめと考察

本論文で提案したリンク発見の手法は，地名に関する一般的な性質（同一の地名は，特定の地域に複数存在しない）や，簡単なクラス推定の手法の組合せであるが，十分に，新しいリンクを発見できる可能性を示すことができた．現在，この成果については，GeoNames のサイト運営者である Marc Wick 氏に報告しており，その成果を反映する形で，データの更新が行われている．

また，本手法が十分な分類精度を持つことから，リンク元とリンク先のデータの齟齬に注目した不適切なリンクの発見が可能であることを示すことができた．人手で作成するリンクデータと，このようなエラー発見のシステムと組み合わせることで，その品質のメンテナンスが可能になると考えられる．

また，本手法では，精度を優先したために，精度重視の条件を満たさなかった約 30,000 件の新規発見リンクについては，分析の対象から外した．しかし，今回行ったエラーの分析結果をもとに，さらなる精度向上のための仕組みを構築することによって，これらのデータからのリンク発見についても検討を続ける必要がある．

また，Linked Open Data として，GeoNames と Wikipedia を対応づける際の問題点として，情報の粒度の問題があることが確認された．具体的には，Wikipedia の 1 ページに複数の地名の内容を記述することが可能であるという問題と，1 つの地名に複数の役割が存在するという問題の 2 種類が存在する．Linked Open Data におけるアイデンティティ [9] をどう扱うかという問題であり，この問題については，GeoNames の管理者と継続して議論していく予定である．

5. おわりに

本論文では，Wikipedia のカテゴリ情報を活用することにより，精度良く Wikipedia のページと GeoNames のエントリの対応関係をリンクとして発見する方法を提案した．その結果，62,660 件の新規データをランダムサンプリングで 98%の精度で発見できることを確認した．また，本手法により，既存のリンクデータのエラーを発見できることが確認された．本手法の成果は，GeoNames データベースに取り込まれる予定であり，このデータベースの性能向上や，Wikipedia への座標情報の付加といった様々な応用が考えられる．

謝辞　本研究の一部は，科研費基盤研究（B）21300029 ならびに NII 共同研究により行われた．また，GeoNames のサイト運営者である Marc Wick 氏には，GeoNames でのリンク発見のシステムや本発表で作成したデータについてのコメントをいただいた．ここに記して謝意を表す．

参考文献

[1] Bizer, C., Heath, T. and Berners-Lee, T.: Linked Data – The Story So Far, *International Journal on Semantic Web and Information Systems*, Vol.5, No.3, pp.1–22 (2009).
[2] 江渡浩一郎，濱崎雅弘：集合知による Linked Data の構築，人工知能学会誌，Vol.27, No.2, pp.181–188 (2012).
[3] Giunchiglia, F., Maltese, V., Farazi, F. and Dutta, B.: GeoWordNet: A Resource for Geo-spatial Applications, *The Semantic Web: Research and Applications*, Aroyo, L., Antoniou, G., Hyvonen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L. and Tudorache, T. (Eds.), Lecture Notes in Computer Science, Vol.6088, pp.121–136, Springer Berlin/Heidelberg (2010).

[4] Hoffart, J., Suchanek, F., Berberich, K. and Weikum, G.: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Technical Report Research Report MPI-I-2010-5-007, Max-Planck-Institut fur Informatik (2010).
[5] 相澤彰子，高須淳宏，大山敬三，安達　淳：異種データベース間でのレコード照合に関する研究動向，NII ジャーナル，Vol.8, pp.43–51 (2004).
[6] 中山浩太郎，原　隆浩，西尾章治郎：人工知能研究の新しいフロンティア：Wikipedia，人工知能学会誌，Vol.22, No.5, pp.693–701 (2007).
[7] 玉川　奨，桜井慎弥，手島拓也，森田武史，和泉憲明，山口高平：日本語 Wikipedia からの大規模オントロジー学習，人工知能学会論文誌，Vol.25, No.5, pp.623–636 (2010).
[8] Liu, Y. and Yoshioka, M.: Construction of large geographical database by merging Wikipedia's Geo-entities and GeoNames, 人工知能学会第 25 回セマンティックウェブとオントロジー研究会 (2011). SIG-SWO-A1102-03.
[9] 武田英明：Linked Data とアイデンティティ，人工知能学会誌，Vol.27, No.2, pp.171–180 (2012).

吉岡 真治　（正会員）

1996 年東京大学大学院工学系研究科精密機械工学専攻博士課程修了．同年より学術情報センター助手．2000 年より国立情報学研究所助手．2001 年より北海道大学助教授．現在，北海道大学大学院情報科学研究科准教授．情報検索への知識処理技術の応用，設計環境の知能化等の研究に従事．博士（工学）．ACM，人工知能学会，電子情報通信学会，言語処理学会各会員．

劉 亦奇

2007 年中国電子科技大学コンピュータサイエンス専攻卒業．同年四川省建設庁入庁．2009 年北海道大学情報科学研究科研究生入学．2012 年北海道大学情報科学研究科コンピュータサイエンス専攻修士課程修了．研究分野は地理情報処理．

神門 典子　（正会員）

国立情報学研究所情報社会相関研究系教授．総合研究大学院大学複合科学研究科情報学専攻教授を併任．1994 年慶應義塾大学大学院文学研究科図書館情報学専攻博士課程修了．博士（図書館・情報学）．学術振興会特別研究員を経て，同年 7 月学術情報センター研究開発部助手．シラキウス大学客員研究員，デンマーク王立図書館情報大学客員研究員等を経て，1998 年学術情報センター助教授，2000 年国立情報学研究所助教授，2004 年より現職．探索や学習のための情報アクセス技術，情報アクセス技術の評価，探索的検索の認知的研究，テキストからの主観情報抽出，知識背景に応じた文章作成支援の研究等に従事．

（担当編集委員　相良 毅）