地域連想語辞書の構築に関する研究

晃舜祥恵　　森田和宏　　泓田正雄　　青江順一
徳島大学大学院　先端技術科学教育部

1. はじめに

近年,インターネット上のコンテンツや広告などの情報を地域別に配信し分ける,エリアターゲティングという機能が登場した[1]. エリアターゲティングは, IPアドレスからアクセス元の地域を特定し,その地域に適した情報を配信するものが主流である. しかし, ユーザが常に現在地の情報に関心を持っているとは限らないという問題点がある. 解決策として, 文書分類技術によりユーザが閲覧するページやTwitterのつぶやきの地域を特定し, 適した情報を配信することが有効と考えられる.

文書分類の手法は,機械学習のように統計的手法を用いる手法が主流である[2]. しかし, 統計的手法を用いるためには,学習データとして大量の文書を用意する必要があり, 内容も的確でなければならない. 統計学的手法を用いない文書分類手法の1つに,辻らの分野連想語を用いる手法がある[3]. 分野連想語とは, 特定の分野を連想できる単語のことをいう. 文書中の分野連想語を抽出することにより, 文書の話題を特定できる. 文書を地域別に分類するには,地域を連想できる単語を収集し辞書に登録する必要がある. 金木ら[4]は, 文書を地域別に分類する手がかり語として, 地名とランドマークを用いている. しかし, 同様の単語は特産品や伝統行事など他にも多数存在する.

そこで本研究では,地名や特産品などの特定の地域を連想できる単語を自動的に収集し,地域連想語辞書を構築することを目的とする. 地域は, 都道府県に設定する. 本稿では, 既存の地名施設名データから地名・施設名辞書を構築する手法と, その辞書を用いて Wikipedia から地域連想語候補を収集する手法,地域連想語候補の地域決定手法について述べる. また, 収集した地域連想語の精度実験及び地名・施設名辞書と地域連想語辞書を用いた文書分類の比較実験をおこない, その結果について述べる.

2. 関連研究

文書を地域別に分類する研究として, 金木らの地名辞書を用いた文書の地名分類が挙げられる[4]. 金木らは, 住所データから取得した地名とランドマークを用いて地名辞書を構築し,特定方式を考案して文書を分類している. ランドマークとは, “東京タワー”のように土地上の目印となる建物などをいう. しかし, 地域と関連する単語は, 地名やランドマーク以外にも特産品や伝統行事など多数存在する. 本研究では, Wikipedia のデータから特産品や伝統行事などの単語も収集する.

3. 分野連想語と地域連想語

3.1 分野連想語

分野連想語とは,特定の分野を連想できる単語のことをいう[3]. 例えば,野球の分野連想語として,“ホームラン”がある. 分野連想語は, 連想できる分野に対する得点が付与されて分野連想語辞書に登録される. 得点は, 分野を連想できる度合いによって設定され,文書分類の際に用いる. 手法として, まず, 文書中の分野連想語を全て抽出する. そして, 分野毎に得点を集計する. 最後に, 得点が最多である分野を文書の分野と特定する.

3.2 地域連想語

地名や, 特産品, 施設名のように特定の地域を連想できる単語のことを地域連想語と定義する. 例えば, 徳島県の地域分野連想語として, “徳島市”や“鳴門金時”, “阿波おどり会館”がある. ただし, 一般名詞や人名は特定の地域を連想できないため, 地域連想語としない. 以降, 地域を＜＞内に記述する. 地域連想語も, 分野連想語と同様に地域連想語辞書へ登録され,文書分類の手法も分野連想語を用いる場合と同様とする.

4. 提案手法

4.1 提案手法の概要

本研究では, まず, GSK 地名施設名辞書[5]を用いて地名・施設名辞書を構築する. 次に, Wikipedia から特産品や学校名,伝統行事など地域連想語候補となる単語を収集する. そして,Web の情報を用いて地域連想語候補の地域決定をおこなう. 地域連想語辞書の構築手法を図1に示す.

4.2 地名・施設名辞書の構築

4.2.1 地名辞書

GSK 地名辞書には, 全国の住所 117,075 件, 住所の表記揺れ, 読み仮名, 緯度経度などが登録されている. 本手法では住所のみを用いる. 住所中の都道府県名や市区郡町村名, 字名を短単位地名として抽出し, 元住所の都道府県の地名としてする. また, 短単位の地名を連結して, 複合

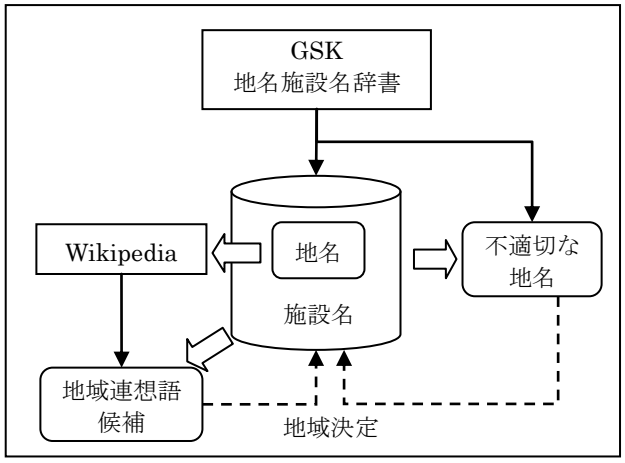

図 1. 地域連想語辞書の構築

地名を作成する．理由は，短単位地名より複合地名の方が地域を連想する度合いが高いからである．例えば，複合地名“東京都新宿区”は短単位地名“東京都”よりも＜東京都＞を強く連想できる．得点は，複合地名に 30 点，連結されていない短単位地名に 10 点を付与する．ただし，一般名詞と 4 ヵ所以上の地域に存在する地名，数字や漢数字を多く含む地名，1 文字の地名は特定の地域を連想できないため，辞書に登録しない．一般名詞の判別には，辞書中の名詞[6][7]を用いる．

4.2.2 地名の地域決定

“松村”と“小山市”，“みゆき町”などから地名接尾語（“県”，“市”，“町”など）を除いた“小山”や“みゆき”などは，文書中の人名と一致しやすい．また，“北上市”から地名接尾語を除いた“北上”は，地名以外の意味で用いられる．このような地名は，文書分類の際に誤分類を招く要因になると考えられる．そこで，2 文字の地名や地名接尾語がない地名を対象として，地域を決定し直す．地域決定には，Web の情報を用いる[8]．Web の情報を用いることで，文書中で地名以外の意味で用いられやすい地名は＜地域なし＞に，特定の地域を連想できる地名はその地域に判定されると考えたからである．以下に手順を示す．

Step1. BingAPI より文書を取得

入力語を検索クエリとして，完全一致検索をおこなう．検索結果のタイトルとサマリを連結したものを文書とする．取得件数は最大 100 件とし，取得件数が 10 件以上の場合は Step2 へ進む．取得件数が 10 件未満の場合は，＜地域なし＞と判定し，終了する．

Step2. 地域ごとに文書分類

地域決定の対象を除いた地名・施設名辞書を用いて，3 章で述べた文書分類をおこなう．地域候補毎に判定件数を集計する．この時＜地域なし＞の件数も集計する．

> 阿波踊り（あわおどり）は徳島県（旧・阿波国）を発祥とする盆踊りである。日本三大盆踊りのひとつ。
> ---以下省略---
>
> （網かけ部：地名）

図 2. 阿波踊りのページ本文

Step3. 入力語の地域を判定

判定件数が最多の地域候補を入力語の地域と判定する．最多の地域候補が複数となる場合は，複数の地域候補を入力語の地域とする．ただし，3 地域までを上限とする．

4.2.3 施設名辞書

GSK 施設名辞書には，全国の施設名 1,000 件，表記揺れ，緯度経度などが登録されている．今回は，施設名と表記揺れを辞書登録する．得点は，全て 10 点とする．

4.3 Wikipedia を用いた地域連想語候補の収集

4.3.1 地域連想語候補の収集手法

Wikipedia から特産品や伝統行事などの単語を収集する手法について述べる．Wikipedia では，特定の地域に関連する項目の場合，本文の序盤に所在地が記述されることが多い．例を図 2 に示す．この特徴を用いて地域連想語候補を収集する．ただし，人名は連想語候補として取得しない．収集手順として，以下の処理を Wikipedia の全ページについておこなう．

Step1. ページのタイトルを取得

Wikipedia ページのタイトルを取得する．

Step2. 本文を取得

本文を 1 文だけ取得する．

Step3. タイトルが人名であるか判別

取得文に生年月日を表す語[9]や職業名が含まれる場合，ページが人物について書かれていると判断し終了する．タイトルが人名ではないと判断された場合は，Step4 へ進む．

Step4. 文書分類

地名辞書を用いて取得文を文書分類する．どこかの地域に判定される場合は，タイトルをその地域の連想語候補として決定し，終了する．＜地域なし＞と判定される場合は，Step2 に戻り次の 1 文を取得する．本文 3 文目が＜地域なし＞と判定される場合は，タイトルは連想語候補ではないと判定し，終了する．

4.3.2 地域連想語候補の地域決定

4.3.1 項で収集した地域連想語候補の中には，特定の地域を連想できない単語も存在する．例として“日本経済新聞”の Wikipedia 本文を図 3 に示す．適切な地域連想語を収集するために，Wikipedia から収集した地域連想語候補に対して，4.2.2 項で述べた地域決定をおこなう．Step2

日本経済新聞（にほんけいざいしんぶん、題字：日本經濟新聞、英語：The Nikkei）は、東京都千代田区に本社を置く日本経済新聞社の発行する新聞（経済紙）であり、広義の全国紙の一つ。　---以下省略---

（網かけ部：地名）

図 3. 日本経済新聞のページ本文

の文書分類に用いる辞書は，地名・施設名辞書とする．また，Step3 では，判定件数が最多の地域と Wikipedia 本文から判定された地域が一致する場合は地域を決定し，一致しない場合は＜地域なし＞とする．地域が決定した地域連想語は，地名・施設名辞書のデータと併せて地域連想語辞書に登録する．得点は全て 10 点とする．収集した地域連想語の例を表 1 に示す．

5. 実験

5.1 地域連想語の精度実験

Wikipedia から収集した地域連想語の精度を確認する実験をおこなった．500 語の地域連想語を対象とし，正解不正解の判定を人手でおこなった．正解率は 99.4%と良好な結果が得られた．不正解の地域連想語は，一覧ページのタイトルであった．Wikipedia の本文の形式から一覧ページを判断する必要がある．

5.2 文書分類の比較実験

提案手法の有効性を確認するために，地名・施設名辞書のみと地域連想語辞書とで文書分類の精度を比較した．地名・施設名辞書には，地名 1,119,530 語と施設名 2,068 語が登録されている．地域連想語辞書には，地名・施設名辞書の内容に加えて Wikipedia データ（2012/1/4）から収集した地域連想語 72,756 語が登録されている．実験対象は，Yahoo!ニュース記事の徳島県に関する記事 464 件(2011/7/1～2011/9/30)とする．分類結果が＜徳島県＞の場合は正解，それ以外は不正解とした．また，正解率を以下の式より求める．

$$\text{正解率}[\%] = \frac{\text{正解数}}{\text{実験対象の総数}} \times 100$$

実験結果を表 2 に示す．両辞書の正解率を比較すると，地域連想語辞書の方が高い結果となった．よって，提案手法の有効性を確認できたといえる．正解率の差が小さい理由は，対象とした文書がニュース記事で，地名が頻繁に記述されていたためと考えられる． しかし，文書から抽出された連想語数は，地名・施設名辞書では 472 語であったが，地域連想語辞書では 572 語となった．よって，収集した地域連想語が文書分類に有効であることがわかる．問題点として，文書中に Wikipedia から収集した地域連想語の異表記が存在した．Wikipedia のリダイレクトを用いることで，別名を取得することができると考える．

表 1. 提案手法より収集した地域連語の例

地域	地域連想語
＜北海道＞	白い恋人，登別温泉，ルタオ
＜大阪府＞	通天閣，だんじり祭り，うめだ花月
＜徳島県＞	阿波踊り，鳴門金時，四国放送
＜沖縄県＞	シーサー，琉球ガラス，エイサー

表 2. 実験結果

	正解数	不正解数	正解率
地名・施設名辞書	395 件	69 件	85.12%
地域連想語辞書	403 件	61 件	86.85%

6. まとめと今後の課題

本稿では，地域連想語辞書を構築するために，既存の住所データから地名・施設名辞書を構築する手法と，その辞書を用いて Wikipedia から地域連想語を収集する手法について述べた．また，地域連想語の精度実験と文書分類の比較実験より，提案手法の有効性を確認できた．

今後の課題として，一覧ページを除去し，収集した地域連想語の別名を収集する必要がある．また，ブログ記事や Twitter のつぶやきを対象に地域別分類をおこないたい．

参考文献

[1] 株式会社ジェイ・キャスト：http://www.j-cast.co.jp/

[2] 橋本力，黒橋禎夫：“基本語ドメイン辞書の構築と未知語ドメイン推定を用いたブログ自動分類法への応用”，自然言語処理，Vol.15 ，No.5，pp.73-97，2008

[3] 辻孝子，泓田正雄，森田和宏，青江順一：“複合語の分野連想語の効率的決定法”，自然言語処理，Vol.7，No.2，pp.3-26，2000

[4] 金木雄太，山田剛一，金川博之，中川裕志：“地名辞書を利用した地名の曖昧性解消と文書の地域分類”，人工知能学会第 19 回全国大会，2E1-03，2005

[5] GSK 地名施設名辞書：http://www.gsk.or.jp/catalog/GSK2008-A/catalog.html

[6] MeCab 用の IPA 辞書：http://mecab.sourceforge.net/

[7] 形態素解析辞書 UniDic（MeCab 用）：http://www.tokuteicorpus.jp/dist/

[8] 安原寛之，森田和宏，泓田正雄，青江順一：“分野連想語を利用した未知語に対する分野の自動推定”，人工知能学会第 23 回全国大会講演論文集，3C4-3，2009

[9] 柴木優美，永田昌明，山本和英：“Wikipedia からの大規模な人オントロジー構築”，情報処理学会研究報告， NL-198 ， pp.1-8，2010