DEIM Forum 2021 H11-3

文書整理に用いる分類リスト選別の試み

宮越　遥† 吉田　光男† 梅村　恭司†

† 豊橋技術科学大学　情報・知能工学系　〒 441–8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
E-mail: miyagoshi.haruka.nw@tut.jp, yoshida@cs.tut.ac.jp, umemura@tut.jp

あらまし　多くの文書の中から目的とする文書を探すことは大変である．ここで，文書を分類しておくことで目的とする文書を探すことが容易となる．一方，文書を分類するためには分類するための箱が必要となる．ある特徴を持つ文書集合を分類することを考えた時にその文書に適したラベルで区分けされた箱を用意する必要がある．この箱を用意するには，分類する文書の内容を把握し，それに関する知識を必要とする．本研究では，どのような区分で文書を分類することが適切であるかを機械的に判断させることを行った．その際に，地名別の分類に焦点を絞り，提案方法が妥当であることを，分類意図を持ったデータを用意して確かめた．

キーワード　分類ラベル, 文書整理, 文書分析, 文書要約

1 はじめに

多くの文書の中から目的とする文書を探すのは大変であるが，文書を分類しておくことで目的とする文書を探すことが容易となる．文書を分類するときに分類するための箱を考える．ある特徴を持つ文書集合を分類することを考えた時に，文書集合に対して的を射ていない分類の箱を用意し分類したとしても，探したいものが見つかるような分類とはならない．それゆえ，この文書に適したラベルで区分けされた箱を用意する必要がある．

ラベルで区分けされた箱はカテゴリと見出し語からなる.「国」をカテゴリとした場合は，見出し語は「日本」や「タイ」などの国名となり,「都道府県」をカテゴリとした場合は，見出し語は「愛知」や「石川」となる．こうすることで階層的となり，分類後に目的とする文書を探しやすくなると我々は考えている．このカテゴリと見出し語は事前に用意し，分類する文書の集合に適切であるカテゴリをコンピュータに推定させることを目的とする．今回,「国」や「都道府県」などの地域区分別に分類することを考える．また，結果の確認を行いやすいように，分類する文書の集合にニュース記事を用い，設定した単語が含まれるニュース記事を集めたニュース記事集合を作成し使用する．

適切なラベルは分類の仕方によって大きく異なるため，分類の箱には粒度があると考えられる．また，部分集合または階層をなすと考えられる．良い分類のためには適切な階層を選ぶのが良いと考えられる．

本研究の目的は，どのような区分で文書を分類することが適切であるかを機械的に判断させることにある．

2 関連研究

文書分類には大きく 2 つの手法が存在し，1 つはクラスタリング，もう 1 つはカテゴライゼーションという手法である．クラスタリング手法においての文書分類手法には，橋本らの研究 [1] や新納らの研究 [2] などが存在する．また，クラスタリングの評価として Zhao らの研究 [3] などが存在する．クラスタリング手法では分類されたものにラベルはないが，分類された集合の意味が取れ，分類したものに対しては要約を得ることができる．しかし，人間が分類された集合を見て，一目でどのような集合であるかを判断するのは難しい．さらに，今回の目的は分類ラベルを用意することであり，分類された集合にラベルを付けることが目的でないクラスタリング手法とは目的が異なる．もう 1 つの手法であるカテゴライゼーションにおいて，上嶋らの研究 [4] や石田らの研究 [5] などが提案されている．カテゴライゼーションは，文書に対して最も適切であると思われるカテゴリを付与するまたはあるカテゴリにまとめるべきと判断された文書を紐づけるという手法である．この手法ではあらかじめカテゴリを用意する必要があり，多くのカテゴライゼーション手法の提案では，すでに用意されたカテゴリと文書を用いて実験を行っているため，どのようなカテゴリが必要であるかという観点がそもそもない．このことから，今回の提案手法はどちらの分類手法ともに異なるものである．

カテゴリや見出し語などの分類ラベルとして固有表現を利用することが多くある．関根らが固有表現を 200 種類定義した拡張固有表現というものがある [6]．この拡張固有表現を利用したラベルを自動的に付与する研究 [7] が存在することから，拡張固有表現の中から文書集合に適切な分類ラベルを選択することは本提案の利用法である．また，森羅プロジェクト [8] において用いられている属性も分類ラベルとすることができることから，多くのラベルから文書集合に適切な分類ラベルを選択することはもう一つの利用法である．

本研究と同様の目的で手法の提案を行っている研究として，平島らの研究 [9] がある．平島らの研究では分類に適切と思われるカテゴリを選択する手法として，80:20 の法則を用いた head の値が用いられていたが，本提案では異なる選択手法を用いている．また，平島らの研究では，Wikipedia カテゴリを利用しているが，Wikipedia カテゴリにおいてカテゴリと見出し語の対応が直観的でないものが多く存在し，前処理などを考える必要がある．そのため，今回は「国」のカテゴリに「日本」や「タイ」などの国名が見出し語として対応するような一般的な知識

から分類ラベルを作成している．

3　提案方法を検討するためのデータ分析

分析に使用したデータについて順に述べる．場所に関係するカテゴリと見出し語に対応関係が存在するようなカテゴリと見出し語を作成した．これらのカテゴリは，地域別で分類を行おうとしたときに使用する候補として考えられるものを用意した．表 1 に今回実際に用意したカテゴリと，そのカテゴリに含まれる見出し語の例を示す．

今回の研究に使用する文書集合として Ceek.jp News[1] が 2004 年 1 月～2020 年 5 月までに収集したニュース記事から記事を抽出してニュース記事集合を作成した．ニュース記事集合として，種類を指定しないニュース記事集合と種類を指定したニュース記事集合を作成した．種類を指定していないニュース記事集合は重複しないニュース記事をランダムに 30000 件抽出したものである．種類を指定したニュース記事集合は指定した単語を本文内に 1 つでも含むニュース記事を抽出し，重複を取り除いた中から 3000 件ランダムに抽出したものである．例えば，知事に関するニュース記事集合では「知事」が 1 つでも含まれるニュース記事を抽出し，リンゴ農家に関する記事集合は「りんご農家」「リンゴ農家」「林檎農家」のいずれかを 1 つでも含むニュース記事を抽出する．指定した単語を表 2 に示す．リンゴ農家の記事集合のみ 3000 件見つからず，1706 件のみの抽出となった．

分析は，次のような手順で行った．ニュース記事集合，カテゴリと見出し語をそれぞれ用意し，あるニュース記事集合に，あるカテゴリの見出し語がいくつの記事に存在するかをカウントする．この結果を見出し語の度数とする．カテゴリ内の見出し語の度数が多い順に見出し語を並べ，縦軸を見出し語の度数，横軸を見出し語とするグラフを作成する．このグラフはニュース記事集合とカテゴリの組み合わせ全てについて作成する．作成したグラフを，ニュース記事集合ごとに集め比較し，またテゴリごとに集め比較する．ニュース記事集合ごとに集めた時の例を図 1 に，カテゴリごとに集めた例を図 2 に示した．

見出し語の頻度を利用することで，ある話題のニュース記事を集めた集合からどのような観点で分類すべきであるかを検討することができると考えた．また，この見出し語の頻度をグラフ化し，ニュース記事の話題ごとに集めての比較やカテゴリごとに集めての比較をすることで，見出し語の度数の順位による変化がそれぞれの集めた中でどのように異なるかや，話題を無視して集めた記事集合での結果を見比べることでそれぞれの話題やカテゴリによって特徴が見られるのかを分析することとした．

表 1　カテゴリを構成する見出し語の例

カテゴリ名	見出し語
国	日本, アメリカ, タイ …
大陸	ユーラシア, アフリカ, 北アメリカ …
州	アジア, ヨーロッパ, オセアニア …
都道府県	北海道, 愛知, 京都 …
政令指定都市	京都, 大阪, 名古屋 …
衆議院比例代表制 選挙区による区分	九州, 東京, 東海 …
日本の八地方区分	関東, 東北, 中部 …
気象庁による日本の区分	関東甲信, 北陸, 東海 …
東北	福島, 岩手, 秋田 …
関東	千葉, 神奈川, 埼玉 …
中部地方	石川, 長野, 愛知 …
近畿	三重, 和歌山, 兵庫 …
中国地方	山口, 広島, 鳥取 …
四国地方	香川, 愛媛, 徳島 …
九州地方（沖縄を含める）	熊本, 鹿児島, 大分 …

表 2　種類を指定したニュース記事集合を作成するために用いた単語の一覧

ニュース記事集合名	検索ワード
知事	知事
国立大学	国立大学
高校野球	高校野球
リンゴ農家	リンゴ農家, りんご農家, 林檎農家
サッカーワールドカップ	サッカーワールドカップ
選挙	選挙
相撲	相撲
テニスとゴルフ	全米オープン, 全英オープン, 全仏オープン, 全豪オープン
ふるさと納税	ふるさと納税
J リーグ	（半角） J リーグ,J1,J2,J3, （全角） J リーグ,J １,J ２,J ３
コシヒカリ	コシヒカリ
マンゴー	マンゴー
阿蘇山	阿蘇山
みかん	みかん, ミカン, 蜜柑
牛肉	牛肉, 和牛
災害	地震, 津波, 洪水, 豪雪, 噴火, 台風, 豪雨, 落雷
ズワイガニ	ズワイガニ, 越前ガニ, 松葉ガニ, 加能ガニ, 香箱ガニ, 越前蟹, 松葉蟹, 加能蟹, 香箱蟹
震源地	震源地
フィギュアスケート	フィギュアスケート
貿易	貿易, 輸入, 輸出
富士山	富士山

4　データ分析の結果と問題点

図 1 に示すようにニュース記事集合ごとにグラフを集め，その中のグラフを比較すると，カテゴリに含まれる見出し語の総数が異なることから比較しにくいことがわかる．例えば，国立大学に関するニュース記事でのグラフを集めた時に，国のカテゴリ（図 3）と東北のカテゴリ（図 4）を見ると，見出し語の総数が異なり，比較することが難しい．また，このニュース記事集合をこのカテゴリで分類することが適切であると思われる

1：http://news.ceek.jp/ より

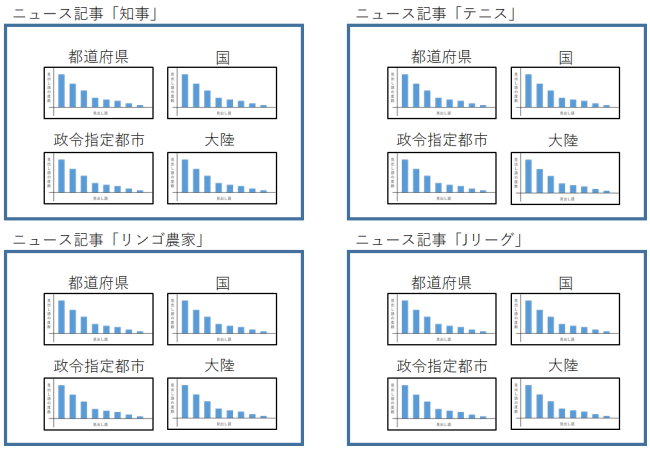

図 1　記事ごとにまとめる例

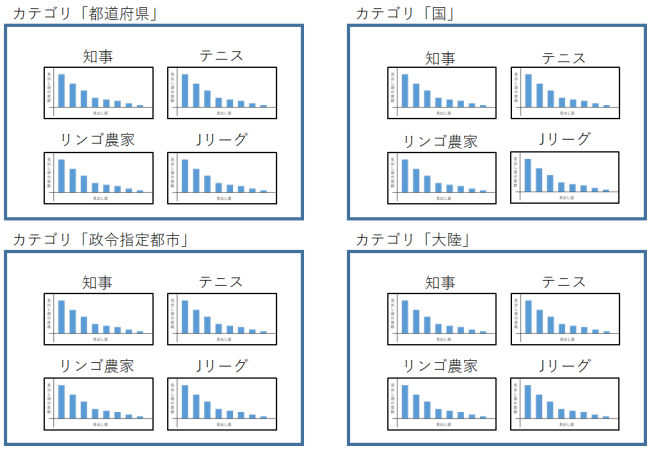

図 2　カテゴリごとにまとめる例

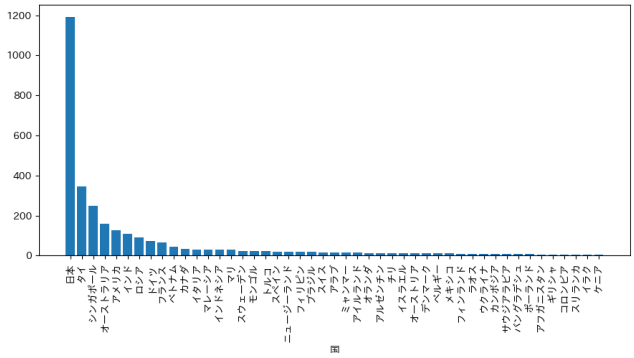

図 3　国立大学に関する記事集合における国のカテゴリのグラフ

グラフを見ると，見出し語の度数がある程度偏りが少ないものも見られる中（例えば図 5），1 つがとびぬけて多いもの（例えば図 6）などがあった．また，このカテゴリで分類することが適切でないと思われるカテゴリでも見出し語の度数に偏りがあるもの（例えば図 7）とないもの（例えば図 8）とあり，平島ら [9] の 80:20 の法則を用いた手法では適切なカテゴリであるかの比較を行うことが難しいと分かる．

また，見出し語の総数に差が存在する時の判断も難しい．国のカテゴリ（図 3）と都道府県のカテゴリ（図 9）を見比べたときに，都道府県のほうがよさそうに見えるが，東北地方のグラフ（図 4）と見比べるとどちらが良いか判断がつかない．グラフからはよさそうに見えたとしても比較が難しいため，実際にこのカテゴリが分類に適しているかは判断がつかない．

以上より，ある記事集合だけを取り出してカテゴリを決める

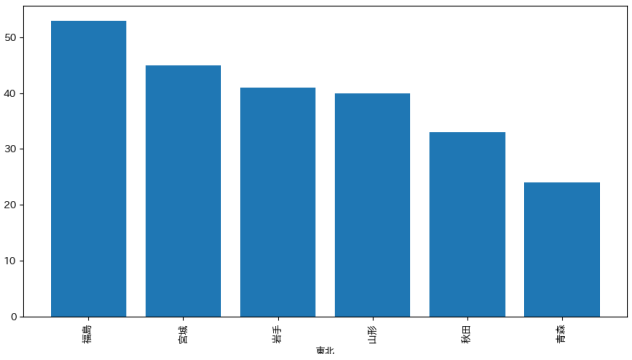

図 4　国立大学に関する記事集合における東北のカテゴリのグラフ

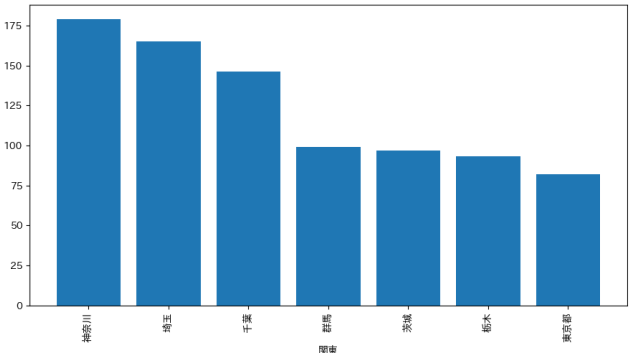

図 5　高校に関する記事集合における関東のカテゴリのグラフ

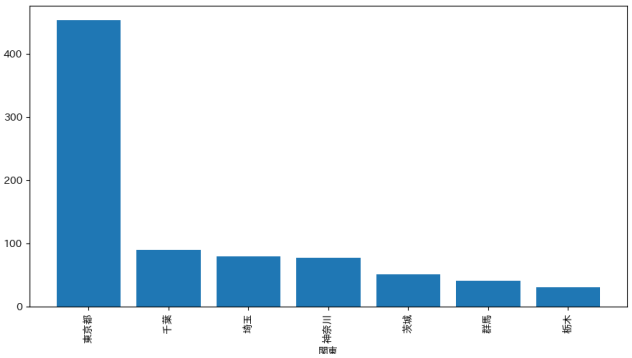

図 6　知事に関する記事集合における関東のカテゴリのグラフ

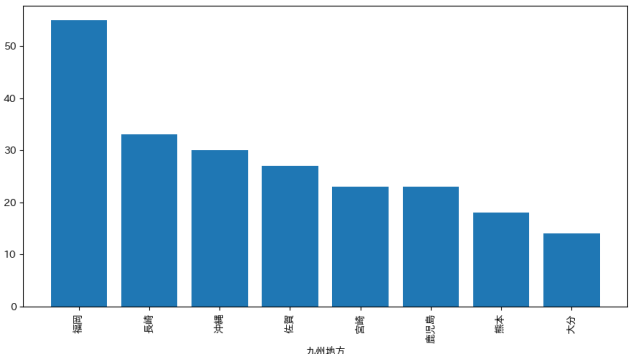

図 7　カニに関する記事集合における九州のカテゴリのグラフ

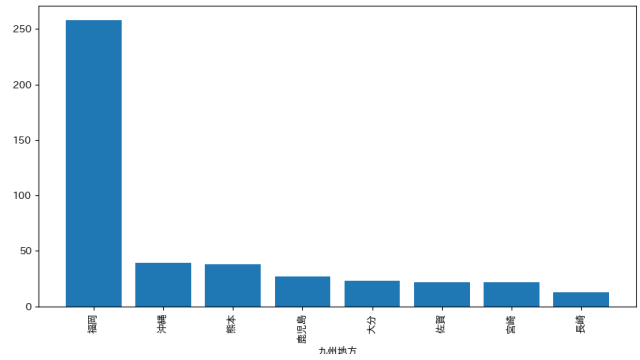

図 8　相撲に関する記事集合における九州のカテゴリのグラフ

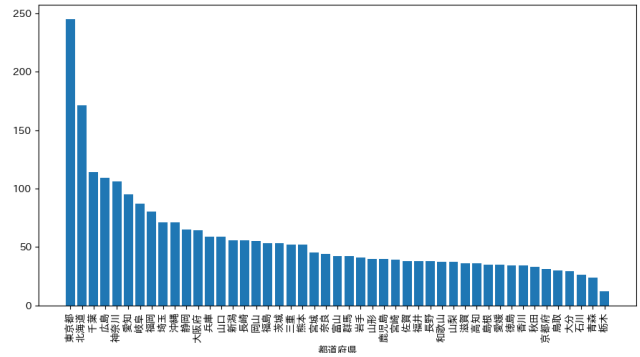

図 9　国立大学に関する記事集合における都道府県のカテゴリのグラフ

ことは難しいと判断した．そこで多くの記事集合に対して，カテゴリごとにグラフを集め（図 2），その中のグラフを比較した．このことで，同じカテゴリを比較しているため，見出し語の個数は同じとなる．ここで，ニュース記事が異なることでどのように見出し語のグラフが異なっているかを見る．また，ここでニュース記事によってこのカテゴリが適切であるかを確認するために，ランダムに取得したニュース記事集合における見出し語の度数をグラフ化したものを用意し，この結果とも見比べた．ニュース記事によって見出し語の度数が 1 つの見出し語に集中したものや，複数の見出し語の度数がある程度同じくらい存在するものもあり，ニュース記事集合によって何らかの変化があることは見て取れるが，どのニュース記事集合においてこのカテゴリによる分類が適切かを判断するのは難しい．また，グラフを比較するとニュース記事集合によってグラフが大きく異なることが分かる．複数のグラフを見比べることで、文書集合を分類するのに適切なカテゴリであるかを判断できる可能性が出てきたが、カテゴリを選ぶにはまだ工夫が必要である．

5　提 案 手 法

分析結果をもとに，カテゴリ内の見出し語の順位かカテゴリの選別を行う方法を提案する．提案手法はカテゴリが文書を分類するのに適しているかを判定する手法である．まず，カテゴリ内の見出し語が文書集合に適したラベルであるかを判定し，これを基に，そのカテゴリで分類するのに適しているかを判定する．見出し語の判定は，見出し語が種類を指定していない記事集合内の記事にどれだけ出現しやすいかに注目する．種類を指定した記事集合がこの出現しやすさよりも出現しやすい場合にこの記事集合に適した見出し語であると判定する．カテゴリの判定は，次の 3 つの条件をすべて満たしたときに適していると判定する．1 つ目は，適していると判定された見出し語がカテゴリ内に含まれる見出し語の半数より多い場合．2 つ目は，カテゴリ内の見出し語が，種類を指定したニュース記事集合に十分出現する傾向にある場合．3 つ目は，カテゴリ内の見出し語が，種類を指定していないニュース記事集合に十分出現する傾向にある場合．これらの方法を，全体の準備を述べた上で，順に説明する．

全体の準備として，記事数のいろいろなカウントを行う．偏りが無いように種類を指定しないニュース記事集合と，種類を指定したニュース記事集合を用意する．用意したニュース記事集合において，見出し語の度数を求める．同時に，カテゴリ内にある見出し語が 1 つも見つからなかった記事数をカウントする．そして，種類を指定したニュース記事集合において全ての見出し語の度数を求める．同時に，カテゴリ内にある見出し語が 1 つも見つからなかった記事数をカウントする．準備は以上である．

見出し語が文書集合に適したラベルであるかの判定は次のように行う．

$$n_j - N \cdot p_j > 0 \tag{1}$$

ただし，対象とするカテゴリ内の見出し語の総数を M とし，この見出し語を辞書順に並べた時の位置を j とした．j の最大値は M となる．種類を特定したニュース記事集合に関する記号は，記事数を N,「見出し語の度数」の観測値を n_j と表している．種類を特定していないニュース記事集合に関する記号は，記事数を K,「見出し語の度数」の観測値を k_j と表し，ある記事に j 番目の見出し語が存在する確率を $p_j = k_j/K$ と表している．式（1）は種類を指定していないニュース記事集合において求めた見出し語の度数から，種類を指定したニュース記事集合の見出し語の度数を推定し，種類を指定したニュース記事集合での実際の見出し語の度数と比較する．推定した値よりも実際の見出し語の度数が大きいものは分類するのに適していると考える．あるカテゴリ内の見出し語が式（1）で適していると判断された数が，見出し語の数よりも多い場合，1 つ目の条件を満たす．

2 つ目の条件を次に示す．

$$N - D_N > M \times 2 \tag{2}$$

ただし，M，N は式（1）と同様である．種類を指定しているニュース記事集合内のニュース記事本文に，あるカテゴリに含まれる見出し語が 1 つも存在しなかったニュース記事数を D_N と表す．式（2）はあるカテゴリに存在する見出し語の数の 2 倍よりも，このカテゴリの見出し語がどれか 1 つでも含まれるニュース記事数が多い場合，2 つ目の条件を満たす．

3 つ目の条件を次に示す．

$$K - D_K > M \times 2 \tag{3}$$

ただし，M，N は式（1）と同様である．種類を指定していないニュース記事集合内のニュース記事本文に，あるカテゴリに含まれる見出し語が 1 つも存在しなかったニュース記事数を D_K と表す．式（3）はあるカテゴリに存在する見出し語の数の 2 倍よりも，このカテゴリの見出し語がどれか 1 つでも含まれるニュース記事数が多い場合，3 つ目の条件を満たす．

6　実験の条件と手順

使用したニュース記事は第 3 節で示したものと同様の Ceek.jp News が 2004 年 1 月〜2020 年 5 月までに収集したニュース記事を使用した．種類を指定しないニュース記事集合はランダムに 30000 件のニュース記事を取得したものとし，種類を指定したニュース記事集合は，第 3 節の表 2 で示す，検索ワードのいずれかを含むニュース記事を 3000 件取得したものとした（リンゴ農家に関するニュース記事集合のみ 1706 件のみの抽出）．使用しているカテゴリと見出し語は第 3 節で示したデータ分析に使用したものと同様のものを使用した．ニュース記事集合においての処理は，全てのカテゴリごとに行った．

7　結果・考察

実験の結果を表 3 に示す．この手法において，分類するのに適切であると判断したものに色を付けた．縦にニュース記事集合，横にカテゴリとしてある．プラスとマイナスは見出し語が分類するのに適していると判定した数と適切でないと判定した数を示している．プラスは，カテゴリ内の見出し語が分類するのに適していると判定したもの（式 1 を満たしたもの）であり，マイナスは，カテゴリ内の見出し語が分類するのに適切でないと判定したもの（式 1 を満たさなかったもの）を示している．この表は，ニュース記事集合とカテゴリが重なる位置に色があれば，そのニュース記事集合において，このカテゴリが適切であることを示している．「サッカーワールドカップ」に関するニュース記事集合においては，「国」「大陸」「州」のカテゴリが適切であるとしている．

ニュース記事集合を分類するのに適切であると思われるカテゴリではプラスの数が多くあった．しかし一方で，プラスの多いものの中に分類するのに適切ではないと思われるカテゴリが存在した．また，分類するのに適切であるカテゴリであると考えていたカテゴリであってもマイナスが多いもの存在した．

リンゴ農家の記事集合では，東北と都道府県が選ばれていた．日本国内のリンゴの生産量の多い県は青森，長野，岩手，山形，福島，秋田，北海道，群馬，宮城，岐阜と続いている [2]．この順序からも分かるように東北地方の県が多いことから，東北が選ばれた点は良い結果であったと考える．また，東北地方以外にもリンゴを生産している県は多くあることから，都道府県が選ばれた点も良い結果であったと考える．

見出し語の度数のグラフ（図 10）を見ても日本での最大生産地である青森が最も多く存在している．また，種類を指定していないニュース記事集合における東北のカテゴリでは図 11 に示すように 3 万件の記事の内，青森を除く県においては見出し語の度数が 200 強（3 万件中 200 件見つかったとして約 0.67%）であるが，リンゴ農家に関する記事集合において，図 10 に示す東北のカテゴリの結果では 1706 件の記事の内，見出し語の度数が 50 強（1706 件中 50 件見つかったとして約 2.9%）であった．このことからも，リンゴ農家に関する記事では東北のカテゴリが良いことが分かる．

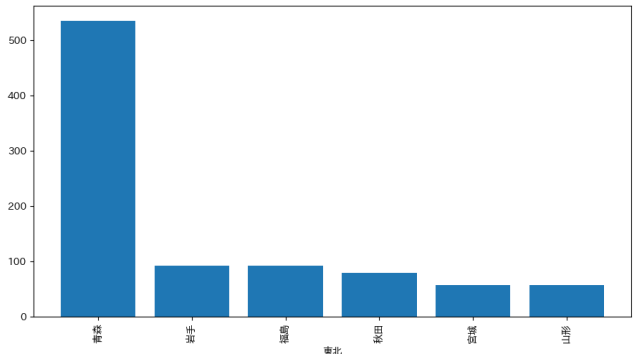

図 10　リンゴ農家に関する記事集合における東北のカテゴリのグラフ

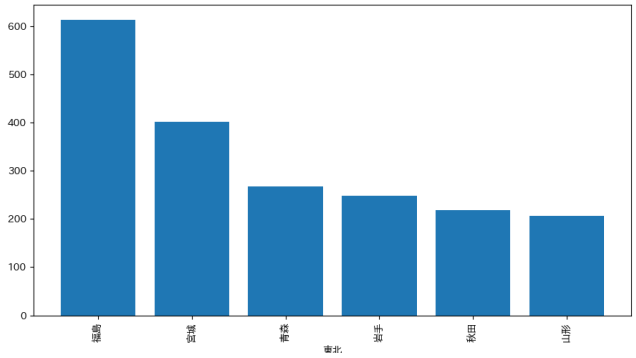

図 11　種類を指定していない記事集合における東北のカテゴリのグラフ

サッカーワールドカップの記事集合では国や州，大陸のカテゴリが良いと判断された．また，そのほかのカテゴリは日本国内の分類であるため，サッカーワールドカップは日本国内の分類ではなく国に関わる分類が適当であると判断されているといえる．

サッカーワールドカップの記事集合と同様に貿易に関するニュース記事集合でも国や州，大陸が良いと判断された．日本の記事であることから，日本と貿易する相手国の国名や大陸名が出現することが予想できる．このことから，より良い分類であると選ばれたのではないかと考える．

J リーグに関する記事集合では政令指定都市が選ばれた．J リーグのチームは政令指定都市に多くあり，チーム名に政令指定都市名が含まれている場合もあったことから，選ばれたと考えられる．J リーグは日本のサッカーのリーグであるが，州のカテゴリが選ばれていた．分類に適切でないと思われたが，サッカーについて調べると FIFA クラブワールドカップという

2：りんご大学 https://www.ringodaigaku.com/study/statistics/statistics.html より (viewed 2020-12-22)

表 3　提案手法によって得られた，ニュース記事を分類するのにそれぞれのカテゴリが適切か不適切かを示した表

category	中国地方		中部地方		九州地方		四国地方		国		大陸		州		政令指定都市		日本の八地方区分		東北		気象庁による日本の区分		衆議院比例代表制選挙区による日本の区分		近畿		都道府県		関東	
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Chizi_3000	5	0	9	0	8	0	4	0	28	162	0	6	0	7	16	4	6	1	6	0	8	2	10	0	7	0	47	0	7	0
KokurituDaigaku_3(	5	0	8	1	8	0	4	0	67	123	5	1	6	1	18	2	7	0	4	2	7	3	8	2	7	0	43	4	6	1
KoukouYakyu_3000	5	0	9	0	8	0	4	0	2	188	0	0	0	7	18	2	7	0	6	0	9	1	8	2	6	1	45	2	6	1
RingoNouka_3000	4	1	6	3	1	7	2	2	14	176	0	0	1	6	6	14	7	0	6	0	9	1	7	3	3	4	30	17	7	0
Sakka-Wa-rudokap	1	4	1	8	2	6	1	3	123	67	5	1	6	1	6	14	3	4	0	6	1	9	3	7	2	5	8	39	1	6
Senkyo_3000	4	1	3	6	8	0	4	0	81	109	3	3	3	4	7	13	4	3	4	2	5	5	7	3	3	4	32	15	5	2
Sumou_3000	3	2	6	3	4	4	2	2	13	177	1	5	0	7	12	8	2	5	4	2	2	8	3	7	6	1	29	18	4	3
4-major-tennis-con	0	5	1	8	0	8	0	0	71	119	2	4	2	5	0	20	0	7	0	6	0	10	0	10	1	6	3	44	1	6
Hurusatonouzei_30	4	1	9	0	8	0	4	0	5	185	0	0	0	7	12	8	6	1	5	1	7	3	8	2	7	0	45	2	7	0
J-League_3000	4	1	8	1	7	1	3	1	63	127	3	3	4	3	18	2	4	3	2	4	3	7	7	3	0	7	31	16	6	1
Koshihikari_3000	4	1	9	0	6	2	3	1	9	181	0	6	0	7	9	11	6	1	6	0	7	3	8	2	7	0	43	4	7	0
Mango_3000	0	5	3	6	7	1	0	4	72	118	2	4	3	4	8	12	5	2	1	5	6	4	5	5	3	4	18	29	3	4
Asozan_3000	5	0	7	2	8	0	4	0	10	180	2	4	1	6	9	11	7	0	4	2	10	0	7	3	4	3	37	10	4	3
Mikan_3000	5	0	8	1	8	0	4	0	19	171	1	5	3	4	17	3	7	0	5	1	9	1	7	3	7	0	45	2	7	0
Beef_3000	4	1	7	2	8	0	4	0	45	145	2	4	3	4	12	8	7	0	6	0	9	1	8	2	7	0	42	5	5	2
Saigai_3000	5	0	9	0	8	0	4	0	40	150	2	4	3	4	16	4	7	0	6	0	10	0	9	1	7	0	46	1	6	1
Crab_3000	5	0	7	2	2	6	2	2	9	181	0	0	0	7	11	9	5	2	5	1	5	5	5	5	4	3	28	19	2	5
Singenti_3000	4	1	9	0	8	0	4	0	49	141	2	4	2	5	15	5	7	0	6	0	10	0	8	2	7	0	45	2	6	1
Figure-skating_300	1	4	3	6	1	7	0	4	41	149	1	5	2	5	9	11	4	3	2	4	4	6	3	7	0	7	8	39	1	6
Trade_3000	0	5	0	9	0	8	0	4	117	73	5	1	6	1	0	20	1	6	0	6	0	10	3	7	0	7	0	47	0	7
Fuji_3000	2	3	9	0	4	4	0	4	40	150	2	4	2	5	16	4	7	0	5	1	7	3	9	1	5	2	33	14	7	0

大会があり，出場資格は大陸によって定められた大会で優勝したチームとなっている．アジアでは AFC チャンピオンズリーグという大会での結果次第となっている．J リーグのチームはこの大会に出場しており，多少「州」による区分に関係しているため，選ばれた可能性がある．

カニに関するニュース記事集合では，カニの漁獲として聞かないような四国や九州，関東が選ばれていなかった．また，カニの中でもズワイガニについてのニュース記事集合であるため，日本海側や北海道などで漁獲される点からも，これらのカテゴリが選ばれなかったのは良かったのではないかといえる．一方，カニの出荷量を考えた時に政令指定都市ではカニの出荷はさほどなかったことから，政令指定都市に関しては適切でないと考えていたが，適切であると判断されていた．

テニスとゴルフに関する記事集合ではすべてのカテゴリにおいてマイナスが多かった．テニスとゴルフに関する記事集合は国のカテゴリによって分類することが妥当だと考えていた．また，種類を指定していない記事集合における国のカテゴリ（図 12）とテニスとゴルフに関する記事集合における国のカテゴリのグラフ（図 13）を比べても分かるように，見出し語の度数の偏りが少なく，テニスとゴルフに関するニュース記事では国名が記事中で多く見つかることからも分類することが妥当であると考えていた．しかし，国のカテゴリに関してもマイナスが多かったことから，日本選手と対戦する選手の出身国や入賞する

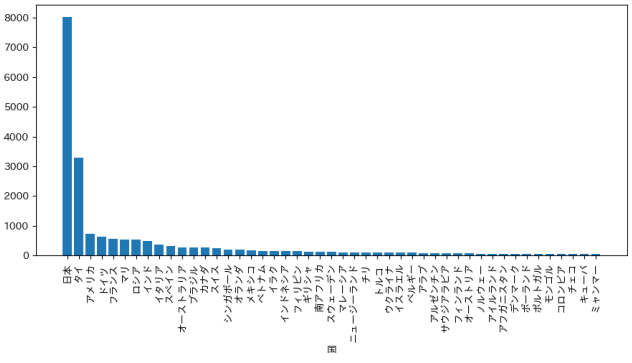

図 12　種類を指定していない記事集合における国のカテゴリのグラフ

選手の出身国が限られているのではないかと考えられる．

フィギュアスケートのニュース記事集合では日本の八地方区分が分類するのに適しているカテゴリであると判定された．フィギュアスケートの大会名に日本の八地方区分の区分名が含まれるものがあったことから，このカテゴリが選ばれたことは適切であったと考える．ただし，政令指定都市に関しては，フィギュアスケート選手の出身地やイベントの開催地に政令指定都市が多くあった．この点から適切であると選ばれる可能性を考えていたが，選ばれなかった．

他のニュース記事に関しては日本国内の記事が多いためか，日本国内に関するカテゴリがすべて選ばれるニュース記事集合が多くあった．

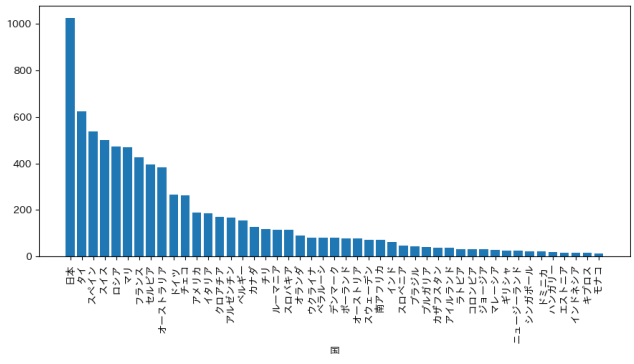

図 13　テニスとゴルフに関する記事集合における国のカテゴリのグラフ

8　今後の展開

同じ語で異なる意味を持つような単語を区別する試みとして，語義の曖昧性解消（Word Sense Disambiguation:WSD）がある．語義の曖昧性解消は古くから存在し [10]，近くの共起を用いる手法 [11] なども存在する．今後，本提案で得られたドキュメントが属する分類ラベルを利用することによって，文の前後からではなく文書集合から語義の曖昧性解消にもつながると考えられる．

9　おわりに

今回，どのような地域区分で分類することが適切であるかを判定することを考え，地域区分に関するカテゴリを用意し，カテゴリが分類するのに適しているか，適していないかを判定する手法を考えた．提案手法は，実際の新聞記事での見出し語の数を分析して考案した．結果，分類に適切であると考えていたカテゴリのほとんどが適切であると判断された．そして，日本国内か海外の国名などで分類するべきかははっきりとわかるような結果であった．

また，結果からこの手法において改善の余地がある点も存在した．日本国内に関するカテゴリのほとんどが良いと判断されたニュース記事集合においてはカテゴリの中に分類に適切でないと考えられるものも存在した．また，適切であると考えていたカテゴリにおいて適切でないと判断されたものも存在した．

また，以下のことを今後検討しなければならない．日本国内に関する分類の場合は国内に関する分類のカテゴリがすべて選ばれたことから，今後この点の改善方法を考えていく必要がある．また，分類するのに適切であると判断されるカテゴリが複数存在するものもあるため，1 つのみを提示する手法の検討も今後必要となると考えられる．

この手法の内容に関しても，プラスと判断された見出し語の数が，カテゴリ内の見出し語の数の半数より多い場合でよいのかも今後検討する必要があると考える．その他には，国のカテゴリにある「タイ」という見出し語が人名の一部や，「5 位タイ」などの別の用語に含まれるものがカウントされていることが分かっているため，今後この点をどのように解消するかも考える必要があるといえる．

文　献

[1] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道. 文書クラスタリングによるトピック抽出および課題発見. 社会技術研究論文集, Vol. 5, pp. 216–226, 2008.

[2] 新納浩幸, 佐々木稔. Nmf による重み付きハイパーグラフを用いたアンサンブル文書クラスタリング. 自然言語処理, Vol. 14, No. 5, pp. 107–122, 2007.

[3] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 515–524, 2002.

[4] 上嶋宏, 三浦孝夫, 塩谷勇. 同義語, 多義語の考慮によるテキストカテゴライゼーションの精度向上. 第 14 回データ工学ワークショップ DEWS2003.

[5] 石田栄美ほか. 人の価値観を表すカテゴリを対象にした複数カテゴリへの自動分類の試み. 文化情報学: 駿河台大学文化情報学部紀要, Vol. 16, No. 2, pp. 53–68.

[6] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *LREC*, 2002.

[7] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Neural joint learning for classifying wikipedia articles into fine-grained named entity types. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters*, pp. 535–544, 2016.

[8] 小林暁雄, 中山功太, 安藤まや, 関根聡. Wikipedia 構造化プロジェクト「森羅 2019-jp」. 言語処理学会 第 26 回年次大会 発表論文集, pp. 1029–1032, 2020.

[9] 平島峻成, 吉田光男, 梅村恭司. 新聞記事検索結果に対する分類ラベル生成における wikipedia カテゴリ情報の利用法. 第 11 回データ工学と情報マネジメントに関するフォーラム, 2019.

[10] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. Word-sense disambiguation using statistical methods. In *29th Annual meeting of the Association for Computational Linguistics*, pp. 264–270, 1991.

[11] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992.