DEIM2022 G43-3

文書整理に用いる分類リスト順位付けの試み

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
†† 筑波大学　ビジネスサイエンス系　〒 112–0012 東京都文京区大塚 3-29-1
E-mail: †{miyagoshi.haruka.nw,umemura}@tut.jp, ††mitsuo@gssm.otsuka.tsukuba.ac.jp

あらまし　現在，様々な文書データが膨大に存在する．この中から目的とする文書を得るために検索を行った場合，検索結果は当てはまるものが膨大に示されるため，より目的に近い文書を得るには絞り込む必要がある．そこで，検索結果を分類することでより得たい文書を絞り込むことができるようにする．文書分類において分類後にどのような分類であるかを把握するためにも，ラベルが必要である．そこで本研究は，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定し，順位付けて提示する手法を検討した．文書分類に適したラベルを順位付けて示すことは，文書を利用する人の用途や意思によってラベルを選択する際のガイドにもなることから本研究は有用であると考える．そのため，機械的に適切なラベルを順位付けて提示する手法を，地域区分のラベルに絞り実験を行い，提案手法の妥当性を確かめた．

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

現在，パソコンを用いた文書作成が増えたこと，リモートワークやペーパーレス化に伴い文書の保管方法が紙からデータとなったことにより，様々な文書データが膨大に存在する．このような膨大な文書データを活用するためには，目的に応じた文書を抽出する必要がある．しかし，膨大にある文書から目的に応じた文書を探すことは大変である．そこで，目的とする文書データを得る方法として検索がある．

検索は，検索ワードを与え，この語に関連する文書を抽出できる．そのため，検索者の目的に応じた検索ワードを与えることで，検索者の目的とする内容に関連した文書を抽出することができる．しかし，検索結果には検索ワードと少しでも関連した内容の文書も抽出されるため，絞り込まれたとはいえまだ膨大な文書が示される．そのため，検索者がよりほしい文書まで絞り込む必要があり，人手で行うことは大変である．

そこで，検索結果を機械的に分類することを考える．そうすることにより，検索者がよりほしい文書を絞り込みやすくなる．例えば,「技術系の大学について知りたい」と考えている検索者が「工学大学」と検索したとする．この検索結果を,「関東」「近畿」「北陸」「東北」と言った八地方区分で分類することで，検索者の背景として住んでいる場所や興味のある場所から，より検索者の目的に関連した文書まで絞り込むことが可能となる．また,「この情報は知っている」「この情報は興味がない」というような消去法によっても，得たい文書まで絞り組むことが可能となる．分類の区分は「八地方区分」に関わらず，検索者が「技術系の大学にはどのような学部があるのだろう」という考えがある場合には,「学部や専攻」といった区分による分類を行うというように，目的によってどのような分類の区分で分類を行うかまで選べることで，得たい文書がより得やすくなる．

しかし，ここで考えられる問題として，検索結果の内容によっては分類の区分として適切でないような分類の区分を選択した場合に，得たい文書を得られる分類とならない場合も考えられる．例えば，分類結果が 1 つに偏ってしまう場合がある．この場合は，絞り込むことができないため分類の意味をなさない．そのため，ここでの分類は 1 つに偏って分類されないラベルを用いることが必要となる．

機械的な文書分類手法として，大きく分けて 2 種類の手法が存在する．1 つはクラスタリングと呼ばれる手法である．クラスタリングに関する研究として Zhao ら [1] や新納ら [2] などの様々なものが存在する．クラスタリングは，文書間の類似度を算出し，その類似度から類似した文書同士をまとめていき，文書グループ（クラスタ）を作ることで分類する手法である．この手法では，分類後にクラスタそれぞれがどのような文書集合であるかを把握することができない．そのため，クラスタへラベルを付ける手法（クラスタラベリング）が存在する [3] [4]．もう 1 つはカテゴライゼーションと呼ばれる手法である．カテゴライゼーションに関する研究として様々な研究が存在する [5] [6]．カテゴライゼーションは，事前に複数のラベルを手動で用意し，その中から文書の内容に最も近いラベルへ分類していく手法である．

検索結果に対してクラスタリングを行う手法が存在する [7] [8]．検索結果を分類し，得たい文書を得やすくするために，これらの手法はクラスタリング後にクラスタラベリングによるラベル付けや要約文書の作成，文書間の関係を示すなどといった工夫がなされている．ただし，それぞれのクラスタについてラベルや要約文があったとしても，それぞれの粒度がバラバラで得たい文書を選択しづらいという問題点は残る．

検索結果に対してカテゴライゼーションを行う手法が考えられる．しかし，カテゴライゼーションではラベルを手動で用意する必要があり，文書集合の内容に即したラベルを用意するこ

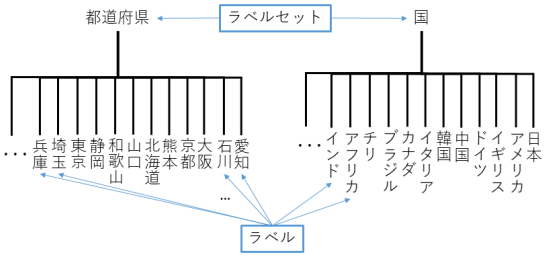


図 1　階層構造の例

とは大変である．どのような検索結果であっても特定のラベルを用いた分類を行うと，ラベルとそのラベルの付けられた分類結果の関連性が低く，得たい文書を得られない場合や分類が偏り得たい文書を絞り込めないという問題が起こる．そのため，分類したい文書に適したラベルを機械的に提示する必要がある．

そこで，本研究では文書集合の分類に適したラベルセットを順位付けて提示する手法を提案し，主観的な有用性を確認する．本研究におけるラベルは，2 階層であることを前提とする．イメージ図を図 1 に示す．ラベルの構造として，ラベルセットの中に粒度の揃ったラベルが存在する形をとる．例えば，ラベルセットが「都道府県」であれば，ラベルは「愛知」や「石川」などとなる．このような 2 階層のラベルを事前に様々な分野において大量に用意することを前提としている．本研究では，この大量に用意したラベルセットがそれぞれ，何番目に文書集合の分類に適しているかを示す手法を提案しており，提案手法の結果から得られた分類に適したラベルセットを 1 つ用い分類することを前提としている．本研究で行う実験は，主観的な有用性を確認しやすくする観点から地域区分のラベルに絞り行う．

2　関 連 研 究

得たい文書を得る手段として検索を行った場合に結果が膨大に示されるため，得たい文書により近い文書まで絞り組むことが大変である．そこで，検索結果に対して分類して示すことでより得たい文書を得やすくする手法として様々な手法が存在する [8] [9].

Web 検索結果に対してクラスタリング，クラスタラベリングを行った結果を既存の分類階層である Yahoo!カテゴリを利用し，それぞれのクラスタの上位概念として関連を持たせることで，検索結果を階層化する手法が存在する [8]．階層構造を持つ点で本研究と類似しており，得たい情報を得やすくする点で階層構造にして示すことは有用である．しかし，木村らの研究の結果では上位概念において類似したクラスタをまとめているが，上位概念の中でそれぞれのクラスタの粒度がしっかりと揃っているとは言えず，本研究とは異なる．また，上位概念に含まれるクラスタ数が少ない項目もあるため，分類後に得たい情報を絞り込むツールとしては多少疑問を持つ．

クラスタリングによって得られたクラスタへ人間がラベルを付けると様々な観点からラベルが付けられる．その点を考慮したクラスタラベリング手法として Eikvil らの手法 [9] がある．この研究は生物医科学論文に対してクラスタリングを行い，ラベリングを行っているため，ラベルを事前に用意しやすい．しかし，対象とする文書がどのようなものであるかが分からないと，ラベルを用意できないという問題が存在する．その点本研究では，文書の内容が分からない場合においても，様々な種類のラベルの中から分類したい文書集合に適したラベルを提案することができる点で異なっている．また，本研究は Eikvil らの研究における様々な側面を捉えるという点で，分類したい文書集合に対して分類に適した複数のラベルセットを示すことで，その文書集合の様々な側面を捉えることができるのではないかと考えている．

本研究と類似した目的に対して行われている研究が存在する．Teshima らの手法 [10] は文書中から抽出したキーワードを利用し，文書集合に適したラベルを示す手法である．平島らの手法 [11] は Wikipedia を利用した 2 階層のラベルを利用し，文書集合の分類に適したラベルを示す手法である．宮越ら [12] の手法は手動でラベルセットとラベルの構造を持つラベルを複数用意し，それぞれのラベルセットが文書集合の分類に適しているかを判定する手法である．

Teshima らの手法は，統計値に基づいた特徴によるキーワードの抽出を行い，複数得られたキーワードから文書の話題を捉えるようなトピックラベルを選定する．このトピックラベルを，文書集合内にある複数の話題ごとに分類した際に付けられるラベルとする手法である．本研究は手動でラベルを作成しているのに対して，この手法は文書中からキーワードを抽出し，さらにその中で話題を表すのに適したトピックラベルを抽出することで，ラベルを作成している．また，この手法は階層構造を持つようなラベルではないため，本研究とは異なる．ただ，検索結果に対して分類を行うためにラベルを提示する点においては同様である．

平島らの手法は，Wikipedia のカテゴリと見出し語に着目し，階層的なラベルを作成する．この階層的なラベルの中から，80:20 の法則（パレートの法則）を利用し文書の分類に適したラベルを提示している．80:20 の法則は，経済学者のヴィルフレド・パレートが考案した経済学上の法則であり，特定の要素の 20%が全体の 80%の成果を生み出しているという法則である．平島らの手法は，この法則を用い，見出し語の度数の大半が一部の見出し語が生み出していると考え手法を作成している．平島らが提示するラベルは Wikipedia の情報を利用しており，Wikipedia のカテゴリと見出し語は，本研究のラベルセットとラベルの構造と同様の構造を持つ．この点で，本研究と類似する．ただ，Wikipedia の情報にはラベルに適切でないと考えられるカテゴリや見出し語が存在するため，それらを除去する必要がある．そのため，本研究ではラベルセットとラベルを一般的な知識から手動で作成し，確実に粒度の揃ったラベルとなる

ようにしている．また，手法においては本研究と着目する範囲が異なる．平島らの手法は見出し語の頻度をカウントし，この頻度が降順となるようにラベルを並べ上位 20%の見出し語にのみ着目している．本研究においては，全てのラベルに着目しているため，平島らの手法とは異なっている．文書集合の分類に適したラベルを提示し，主観的な有用性を確認するという目的は本研究と同様であり，この目的に対して取り組むことは意義があると考える．

宮越らの手法では，ラベルの適正判定を行ったのちにラベルセットの適正判定を行っている．使用しているラベルセットは本手法と同様に地域区分に絞り手動で用意されている．宮越らの手法では文書分類に適したラベルセットか適していないラベルセットであるかの 2 値による判定である．本手法では適している順にラベルセットを示すことができるため，分類する際にどのラベルセットを選択すべきかが示される．本手法は，宮越らの手法をもとにしており．そのため，本手法の判定において閾値を設けることで，宮越らの手法と同様となる．

3 提案手法

本手法では，分類対象の何らかの特徴を持つ文書集合に対して分類に適しているラベルセットを順位を付けて示す．しかし，ラベルセット内のラベルの数はラベルセットによって異なる．そのため，偏りのない文書集合を用いて比較が可能となるようにする．そのため，偏りのない文書集合を作成する必要がある．

次に，3 つの手順によって，対象とする文書集合の分類に適したラベルセットを順位付けて示す．1 つ目は「ラベルの度数のカウント」，2 つ目は「ラベルの分類適性を判定」，3 つ目は「ラベルセットの分類適性の判定」という手順で行う．

3.1 手法に使用するデータについて

本手法では，偏りのない文書集合を利用し，分類に適しているラベルセットを順位付けて示す．そのため，偏りのない文書集合として，ランダムに文書データを抽出した文書集合を作成する．本研究では，この偏りのない文書集合を「種類を特定しない文書集合」と呼ぶ．また，分類対象となる文書集合を「種類を特定した文書集合」と呼ぶ.「種類を特定した文書集合」と「種類を特定しない文書集合」のイメージ図を図 2 に示す．

3.2 ラベルの度数をカウント

ある文書集合においてラベルセット内のラベルが文書中に存在する文書数をカウントする．このカウント結果をラベルの度数と呼ぶ．ラベルの度数を，全ての文書集合（「種類を特定しない文書集合」と「種類を特定した文書集合」）とラベルセットとの組み合わせにおいてカウントを行う．ラベルの度数のカウントのイメージ図を図 3 と図 4 に示す．図 3 の例において，「石川」のラベルは 1 つの文書内にのみあるため，この場合の「石川」のラベルの「ラベルの度数」は 1 となる．図 4 の例において,「石川」のラベルは 3 つの文書内にのみあるため，この場合の「石川」のラベルの「ラベルの度数」は 3 となる．

ラベルの度数を求めると同時に，ラベルセット内のラベルが

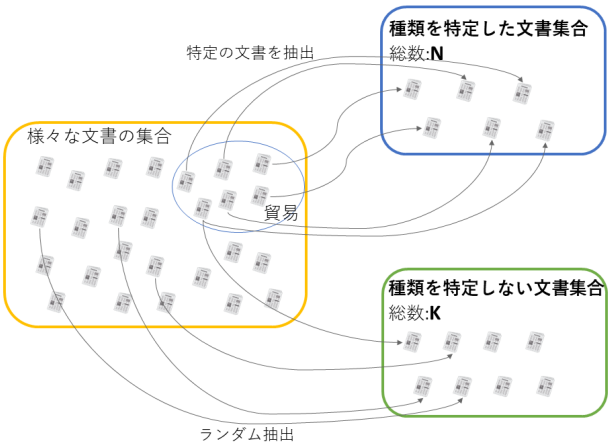
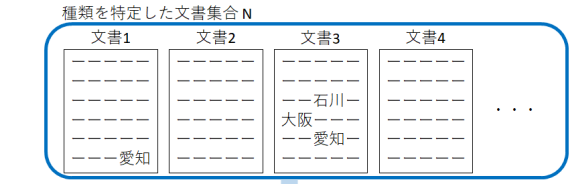


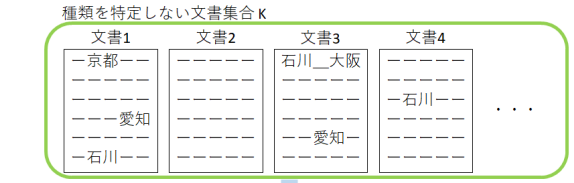
図 2　N と K についての図



ラベルセット
都道府県
愛知
石川
京都
大阪

文書集合	ラベル				1つも存在しない
	愛知 (j=1)	石川 (j=2)	京都 (j=3)	大阪 (j=4)	
文書1	○	×	×	×	
文書2	×	×	×	×	○
文書3	○	○	×	○	
文書4	×	×	×	×	○
n_j	2	1	0	1	
D_N					2

図 3　n_j についての図



ラベルセット
都道府県
愛知
石川
京都
大阪

文書集合	ラベル				1つも存在しない
	愛知 (j=1)	石川 (j=2)	京都 (j=3)	大阪 (j=4)	
文書1	○	○	○	×	
文書2	×	×	×	×	○
文書3	○	○	×	○	
文書4	×	○	×	×	
k_j	2	3	1	1	
D_K					1

図 4　k_j についての図

1 つも存在しない文書数をカウントする．このカウントは,「種類を特定しない文書集合」と「種類を特定した文書集合」において，すべてのラベルセットとの組み合わせに対して行う．図 3 の例では，2 つの文書においてラベルセット内のラベルが 1 つも存在しないため，このカウントは 2 となる．

3.3 ラベルの分類適正を判定

ランダムに集められた文書集合から推測した値と分類対象となる何らかの特徴を持つ文書集合で観測された値の差を，その差の分散で正規化し，シグモイド関数 σ を利用することで，0 から 1 の間でどの程度分類に適しているラベルであるかを判定できるようにした．この値をラベル判定スコアと呼ぶ．式で使用している N，K，n_j，k_j は宮越らの手法と同様の定義をしている．

$$\sigma(x) = \frac{1}{e^{-x}+1} \tag{1}$$

$$x_j = \frac{\frac{n_j}{N} - \frac{k_j}{K}}{\sqrt{\frac{(\frac{n_j}{N})(1-\frac{n_j}{N})}{N} + \frac{(\frac{k_j}{K})(1-\frac{k_j}{K})}{K}}} \tag{2}$$

種類を特定した文書集合の文書数を N，種類を特定した文書集合においてラベル名が出現した文書数を n_j とする．また，種類を特定しない文書集合の文書数を K，種類を特定した文書集合においてラベル名が出現した文書数を k_j とする．N と K のイメージを図 2 に示す．n_j のイメージを図 3 に示し，k_j のイメージを図 4 に示す．n_j と k_j は，対象とする文書集合が異なっているが算出方法は同様である．

式（1）はシグモイド関数である．シグモイド関数への入力値を式（2）の結果としている．式（2）はランダムに集められた文書集合から推測した値と分類対象となる何らかの特徴を持つ文書集合で観測された値の差を用いて作成した．式（2）の分母は差の分散を利用しており，ランダムに集められた文書集合から推測した値と分類対象となる何らかの特徴を持つ文書集合で観測された値は独立であると考え作成した．

3.4 ラベルセットの分類適性の判定

ラベルセット判定スコアを算出することによって，分類に適している順にラベルセットを示す．ラベルセット判定スコアを算出する前に条件を設けている．条件を満たさない場合は，ラベルセット判定スコアは 0 となる．

前条件は，文書集合にかかわりがあまりないラベルセットは分類に使用するには適切でないと考え設けている．文書集合において，ラベルセット内のラベルが 1 つ以上文書中に存在する文書数が，ラベルセット内のラベルの総数の二倍よりも少ない場合は，文書集合とラベルセットとの関わりはあまりないと判断する．この条件を,「種類を特定した文書集合」と「種類を特定しない文書集合」のどちらにおいても，ラベルセット内のラベルが 1 つ以上文書中に存在する文書数のほうが多い場合にのみラベルセット判定スコアを算出する.「種類を特定した文書集合」においての条件式を式（3）に示し,「種類を特定しない文書集合」においての条件式を式（4）に示す.

$$N - D_N > M \times 2 \tag{3}$$

$$K - D_K > M \times 2 \tag{4}$$

ラベルセット内のラベルが 1 つも文書内に存在しなかった文書数を D とする．種類を特定している文書集合の場合は D_N と表記し，種類を特定していない文書集合の場合は D_K とする．ラベルセット内のラベルの総数を M とする．

ラベルセット判定スコアはラベルセット内のラベルにおいて，どの程度適切であるかの割合を算出することによって求める．このようにすることで，ラベルセット内のラベルの総数が異なる場合においても比較できるようにしている．ラベルセット判定スコアを S とし，計算式を式（5）に示す.

$$S = \frac{\sum_{j=1}^{M} \sigma(x_j)}{M} \tag{5}$$

式（5) で算出されたラベルセット判定スコア S が 1 に最も近いラベルセットを最も適しているラベルセットとする.

3.5 手法の実行にあたり対処した内容

実際に計算する際には，2 つの例外処理を追加した．1 つ目は，種類を特定した文書集合においてラベル名が出現した文書数（n_j）が 0 もしくは，種類を特定しない文書集合においてラベル名が出現した文書数（k_j）が 0 の場合において処理を行った．どちらかが 0 の場合は，式（2）の分母である分散を上手く見積もることができないため，例外として式（1）と式（2）を計算せず，種類を特定した文書集合においてラベル名が出現した文書数が 0 であれば，そのラベル判定スコアを 0 とし，1 以上であればラベル判定スコアを 1 とした．2 つ目は，計算する際に式（1）の分母が発散してしまい計算が困難であることから，式（2）が 30 以上の場合はラベル判定スコアを 0，-30 以下の場合はラベル判定スコアを 1 とした．式（1）の分母が大きいほど式（1）の値は 0 に近づいていき，分母が小さいほど式（1）の値は 1 に近づくことから，x が 30 以上または -30 以下の場合はラベル判定スコアをそれぞれ 0 と 1 とした.

4 実験に使用したデータ

今回使用した文書は Ceek.jp News[1]が 2004 年 1 月から 2020 年 5 月までに収集した文書（ニュース記事）を使用している（32,301,089 件）．この文書データのうち，ニュース記事本文のみを抽出し使用する．分類対象となる集合として，特定の単語のいずれかを含む文書を 3000 件抽出した「種類を特定した文書集合」と Ceek.jp News によって抽出された全てのニュース記事本文の文書からランダムに 30000 件抽出した「種類を特定しない文書集合」を作成する.「種類を特定した文書集合」は，表 1 の 21 種類を作成し，それぞれの文書集合を作成するにあたり，表 1 の検索ワードのいずれかを含む文書によって作成した（りんご農家に関する文書集合においては，該当する文書が 3000 件存在しなかったため，1706 件の抽出となっている）.「種類を特定した文書集合」と「種類を特定しない文書集合」のイメージを図 2 に示す．図 2 の「様々な文書の集合」は Ceek.jp News が抽出したニュース記事本文を示す.

文書集合の分類に適切であるかを主観的に判定しやすいように，地域区分に限定してラベルセットを手動で 15 種類作成し

1 : http://news.ceek.jp/ より

表 1　種類を特定した文書集合を作成するために用いた単語の一覧

文書集合名	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
4 大大会	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

表 2　ラベルセットを構成するラベルの例

ラベルセット名	ラベル名
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 …

た．作成した，ラベルセットとそのラベルセットに対応するラベルの一部を表 2 に示す．

5　手　　順

まず，分類対象となる文書集合である「種類を特定した文書集合」と「種類を特定していない文書集合」，地域区分に限定したラベルセットとラベルを用意する．

手法の実行としてまず，ラベルの度数をカウントする．種類を特定しない文書集合においてラベルそれぞれがいくつの文書に存在するかを数える．また，種類を特定した文書集合においてもラベルそれぞれがいくつの文書に存在するかを数える．このカウントを 15 種類すべての記事集合に対して行う．このラベルの度数のカウントと同時に，ラベルセット内のラベルが 1 つも存在しない文書数をカウントする．このカウントは，種類を特定しない文書集合と種類を特定した文書集合 15 種類に対して，用意した全てのラベルセットとの組み合わせに対してカウントを行う．数えた値から，ラベル判定スコアを計算する（式（1）と式（2））．

これらの計算した値を利用して式（5）を計算する．全ての文書集合とラベルセットの組み合わせに対して行う．分類する文書集合ごとに計算結果が 1 に近い順に並べることで，その文書集合を分類するのに適切なラベルセットが順に示される．この結果と人間が直観的に適切だと考えるラベルセットと一致するかを確認する．

6　結果・考察

結果を表 3 に示す．文書集合とラベルセットに対応するベルセット適正判定値を示し，文書集合に最も適していると判定されたラベルセット判定スコアを赤色，2 番目に適していると判定されたラベルセット判定スコアを黄色，3 番目に適していると判定されたラベルセット判定スコアを緑色で示している．

サッカーワールドカップや 4 大大会，貿易といった海外とのかかわりが深い文書集合では，1 番から 3 番までに海外に関わる「州」「大陸」「国」のラベルセットが適切であると判定された．一方，その他の文書集合では，これらの海外に関わるラベルセットは下位に近い判定となった．

りんご農家に関する文書集合においては，りんごの生産量が最も多い青森をはじめとした主な生産地がある「東北地方」のラベルセットが最も適していると判定された．

災害に関する文書集合では，台風の勢力が弱まる前に上陸することによる影響を大きく受ける「九州地方」や「四国地方」のラベルセットが上位の判定となった．3 番目に適していると判定された「東北地方」では 2011 年に東日本大震災が発生し甚大な被害が出ていることから，災害に関する文書集合において分類対象として適していると判定されたと考えられる．

災害の中でも地震に関する文書集合として，震源地に関する文書集合の結果を見ると，2011 年の東日本大震災があった「東北地方」のラベルセットが最も適していると判定された．また，東日本大震災で帰宅困難者などが多く出た「関東地方」においては「東北地方」に続いて適切であると判定された．

みかんに関する文書集合においては，暖かい地方で栽培されるみかんの特性から,「九州地方」や「四国地方」といったラベルセットが適切であると判定されたと考えられる．

富士山に関する文書集合においては，富士山がある「関東地方」や「中部地方」が適切であると判定されと考えられる．ま

表 3　順位付けする手法によって得られたラベルセット判定スコアと適切であると判定された上位 3 位までを示す表

文書集合＼ラベルセット	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	東北地方	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	近畿地方	都道府県	関東地方
知事	0.986	0.973	0.982	0.986	0.091	0.004	0.025	0.814	0.812	0.992	0.792	0.849	0.983	0.985	0.954
国立大学	0.950	0.932	0.916	0.960	0.365	0.636	0.673	0.917	0.992	0.710	0.760	0.798	0.933	0.934	0.855
高校野球	0.999	0.999	0.999	1.000	0.001	0	0.000	0.914	0.881	1.000	0.799	0.737	0.910	0.993	0.985
りんご農家	0.459	0.755	0.300	0.456	0.030	0	0.118	0.380	0.726	0.999	0.698	0.492	0.302	0.689	0.785
サッカーワールドカップ	0.221	0.239	0.261	0.414	0.672	0.726	0.780	0.457	0.402	0.038	0.254	0.339	0.339	0.255	0.378
選挙	0.821	0.410	0.811	0.882	0.436	0.425	0.416	0.379	0.424	0.649	0.388	0.529	0.540	0.689	0.619
相撲	0.674	0.698	0.526	0.553	0.027	0.051	0.003	0.430	0.298	0.789	0.174	0.327	0.780	0.736	0.768
4大大会	0.047	0.113	0.001	0	0.403	0.380	0.392	0.002	0.005	0.000	0.002	0.003	0.169	0.032	0.002
ふるさと納税	0.850	0.925	0.992	0.910	0.007	0	0.041	0.655	0.693	0.955	0.502	0.585	0.950	0.965	0.907
Jリーグ	0.829	0.838	0.949	0.903	0.305	0.329	0.470	0.920	0.357	0.706	0.169	0.410	0.132	0.809	0.965
コシヒカリ	0.945	0.957	0.754	0.817	0.009	0.054	0.046	0.557	0.787	0.995	0.655	0.635	0.912	0.952	0.961
マンゴー	0.273	0.359	0.833	0.215	0.327	0.345	0.518	0.538	0.664	0.221	0.575	0.504	0.445	0.504	0.576
阿蘇山	0.912	0.854	1.000	0.975	0.018	0.255	0.093	0.605	0.997	0.804	0.960	0.809	0.488	0.884	0.690
みかん	0.908	0.893	0.996	0.988	0.062	0.183	0.272	0.871	0.846	0.797	0.725	0.656	0.907	0.949	0.786
牛肉	0.678	0.713	0.953	0.827	0.213	0.311	0.447	0.643	0.855	0.956	0.700	0.702	0.822	0.890	0.812
災害	0.973	0.951	0.994	0.982	0.155	0.311	0.386	0.877	0.896	0.981	0.908	0.768	0.967	0.981	0.828
カニ	0.996	0.835	0.393	0.463	0.025	0	0.073	0.599	0.827	0.763	0.708	0.545	0.744	0.727	0.201
震源地	0.684	0.978	0.990	0.975	0.268	0.326	0.212	0.941	0.996	1.000	0.955	0.783	0.981	0.987	1.000
フィギュアスケート	0.163	0.299	0.090	0.032	0.226	0.151	0.244	0.604	0.512	0.328	0.382	0.471	0.053	0.191	0.093
貿易	0.044	0.014	0.060	0.039	0.587	0.713	0.730	0.015	0.231	0.023	0.171	0.336	0.024	0.018	0.033
富士山	0.660	0.960	0.616	0.285	0.176	0.343	0.311	0.841	0.760	0.626	0.604	0.659	0.740	0.810	0.982

た,「政令指定都市」が 3 番目に適切であると判定された．理由としては，富士山が日本を代表する山であり，ニュース記事において政令指定都市にある山で起こった滑落事故などで事故にあった人物の登頂経験の例として示されることから，多くの文書に出現したことが要因なのではないかと考える．

相撲に関する文書集合では，巡業で訪れる地域を調べてみたところ近畿地方や中部地方，関東地方が多かったことから,「東北地方」や「近畿地方」「関東地方」のラベルセットが上位に判定されたことは適切であるといえる．

ズワイガニに関する文書集合では，ズワイガニの漁獲量上位の県が含まれる「中国地方」や「中部地方」「東北地方」といったラベルセットが上位になったと考えられる．3 番目に適切であると判定された「日本の八地方区分」に関しては，ズワイガニの漁獲量上位は日本海側に多く，複数の地方にまたがっていることから適切であると判定されたと考えられる．また，表 4 に示す全てのラベルセットの順位から，国外や太平洋側の地方に関するラベルセットは下位に判定されたことも適切な判定であったといえる．

Ｊリーグに関する文書集合は，3 番適していると判定された「政令指定都市」にチームが多いことから適切であると考えられるが，全国的にチームがあるため,「日本の八地方区分」のラベルセットなどが上位にならなかった点で改善が必要なのではないかと考えられる．また，Ｊリーグチームの分布を見ると太平洋側にチームが多いことから，ラベルセットとして，日本海側や太平洋側という区分についても検討の余地があると考える．

表 4　ズワイガニに関する文書集合におけるラベルセットの順位

順位	ラベルセット	数値
1	中国地方	0.996411559
2	中部地方	0.834563725
3	日本の八地方区分	0.827132373
4	東北地方	0.762516246
5	近畿地方	0.74421118
6	都道府県	0.727090582
7	気象庁による日本の区分	0.707745363
8	政令指定都市	0.598853399
9	衆議院比例代表制選挙区による日本の区分	0.545209503
10	四国地方	0.462594271
11	九州地方	0.392650969
12	関東地方	0.20112059
13	州	0.07314031
14	国	0.024649168
15	大陸	0

知事や国立大学，高校野球，コシヒカリ，ふるさと納税といった特に日本国内で広く関係がある文書集合では,「州」「大陸」「国」といった日本国外に関するラベルセット判定スコアが高くなった．一方順位に目を向けると，高校野球では，優勝校の少ない地方が上位で，優勝校の多い地域が下位となる結果であったことから，判定に疑問が残る結果となった．

7 ま と め

文書集合の分類に適したラベルセットを順位付けて提示する手法を提案した．結果として，宮越らの手法で適していると判定された中でどのラベルセットがより適しているかがを示すことができた．また，本手法で上位 3 つまでに示されたラベルセットはほとんど分類に適していると考えられる結果となったことから，本手法の有用性が示された．今後，地域区分に限定せずにラベルセットを作成し，実験することで，地域区分以外での分類においても有用性かを検討する必要がある．

文　献

[1] 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.

[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] Alexandrin Popescul and Lyle H Ungar. Automatic labeling of document clusters. *Unpublished manuscript, available at http://citeseer. nj. nec. com/popescul00automatic. html*, 2000.

[4] David Carmel, Haggai Roitman, and Naama Zwerdling. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 139–146, 2009.

[5] 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.

[6] 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.

[7] 村松亮介, 福田直樹, 石川博. 分類階層を利用した検索エンジンの検索結果の構造化とその提示方法の改良. 電子情報通信学会第 19 回データ工学ワークショップ, B, Vol. 6, , 2008.

[8] 木村塁, 戸田浩之, 田中克己. 検索結果スニペットのクラスタリングによる同姓同名人物の特定. *DEWS2006, 2C-i11*, 2006.

[9] Line Eikvil, Tor-Kristian Jenssen, and Marit Holden. Multi-focus cluster labeling. *Journal of biomedical informatics*, Vol. 55, pp. 116–123, 2015.

[10] Ryota Teshima, Masayuki Okabe, and Kyoji Umemura. Finding effective query strings from results of primary search. In *Proceedings of the 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pp. 305–308, 2014.

[11] 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.

[12] 宮越遥, 吉田光男, 梅村恭司. 文書整理に用いる分類リスト選別の試み. 第 13 回データ工学と情報マネジメントに関するフォーラム, 2021.