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

平島　峻成† 　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 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: †s163367@edu.tut.ac.jp, ††yoshida@cs.tut.ac.jp, †††umemura@tut.jp

あらまし　膨大な数の資料の中から目的の資料を探し出せないという問題がしばしば起こる．こうした場合，前もって膨大な数の資料を分類することで目的の資料を探すことは容易になるが，分類ラベルが必要となる．これがないと膨大な数の資料の内容を読み解く必要があり，資料の内容に関する知識がないと難しい．本研究では新聞記事検索結果に対して，Wikipedia カテゴリを用いて分類ラベルを付与することを目的とする．具体的には，まず新聞記事検索結果中のカテゴリ毎の見出し語の頻度を求める．全体の頻度の大部分が，全体の一部の見出し語が生み出しているようなカテゴリは分類ラベルとして用いない．実験の結果，文書集合「カンパニー」に対してカテゴリ「東証一部上場企業」，文書集合「キャンパス」に対してカテゴリ「学校記事」など適切な分類ラベルが付与されることを確認した．

キーワード　ラベルセット，80:20 の法則，Wikipedia カテゴリ

1　は じ め に

膨大な数の資料の中から目的の資料を探し出せないという問題がしばしば起こる．こうした場合，前もって膨大な数の資料を分類することで目的の資料を探すことは容易になるが，分類ラベルが必要となる．これがないと膨大な数の資料の内容を読み解く必要があり，資料の内容に関する知識がないと難しい．資料の内容を読み解けたとしても， どのような基準で分類すればよいか判断することは人間にはコストの高い作業である．ここで，文書集合には話題を共通とする部分集合のグループが複数あると仮定し，このグループそれぞれの話題を表現する文字列のことをラベルとする．さらに，ラベルは文書集合に複数存在するので， 使われ方が似ているラベルをまとめたものをラベルセットとする．文書集合からこうしたラベルセットが求まれば， 内容に関する知識を得る手助けとなるだけでなく，文書集合の分類にも役立つと考えられる．

例えば，複数の資料がある場合，分類するためにフォルダ分けをするタスクが考えられる．この場合，フォルダに適切な名前をつける事が難しい．このタスクを実現する方法として，まずクラスタリングが考えられる．クラスタリングとは，文書に出現する語の分布の類似性に基づいて文書を自動的に分類することである．しかし，クラスタリングには文書集合を分類できるが，ラベルが付かないという問題がある．橋下ら [1] はクラスタリングを行ったのち固有表現抽出の手法 [2] によってラベルを付けているが，複数のクラスタで同名のラベルがついている．フォルダ分けのタスクを考えた場合，フォルダ名が同じものが複数存在するということなので分類には適さない．もう一つ考えられる方法として，文書集合の部分集合に共通する話題 (ラベル) の抽出と分類が挙げられる [3]．しかし，抽出されたラベルの使われ方が異なるためどんな分類にしたのか人間にとって理解しづらいという問題がある．他にも，漢字一文字とカタカナ列を同等に頻度で評価しこれに基づいて文献を分類する手法 [4]，人手で作成したキー概念を用いて新聞記事を分類する手法 [5]，分野ごとにキーワードを自動抽出しそのキーワードの出現頻度の偏りを用いて文書を分類する手法 [6] 等が提案されている．しかしこれらはいずれも十分な精度での分類はできていなかった．

そこで本研究では Wikipedia カテゴリに着目した．Wikipedia には記事タイトル (見出し語) に対して分類ラベルであるカテゴリが割り当てられている．例えば，カテゴリ「学校記事」は「東京大学」「京都大学」のような見出し語に割り当てられている．本研究では，文書集合に対する分類ラベル生成における Wikipedia カテゴリの利用法について検討する．具体的には，80:20 の法則を用いる．全体の見出し語の頻度の大部分が，全体の一部の見出し語が生み出しているようなカテゴリは分類ラベルとして使用しない．これにより，一部の見出し語のみ頻度が高い偏ったカテゴリが除去され，全体を俯瞰したバランスの良い分類ラベルを使用することができる．実験の結果，文書集合「カンパニー」に対してカテゴリ「東証一部上場企業」，文書集合「キャンパス」に対してカテゴリ「学校記事」など，文書集合の生成に使ったキーワードと親和性のある分類ラベルが付与されることを確認した．

2　Wikipedia カテゴリ

Wikipedia には記事タイトル (見出し語) に対して分類ラベルであるカテゴリが割り当てられている．表 1 に Wikipedia カテゴリの例を示す．カテゴリ「存命人物」「学校記事」とその見出し語は意味的にまとまっており，文書集合に対する分類ラベルとしてこれらのカテゴリは有用であると言える．また，カテゴリ「企業関連のスタブ」とあるが，スタブとは主題について

十分な説明がないなど，成長していない項目のことである．しかし今回のタスクでは見出し語が意味的にまとまっていることが重要であり，その項目の情報量は重要ではない．従って，カテゴリ「〇〇のスタブ」も有用であるといえる．また，カテゴリ「整理が必要な項目」のように，見出し語間で意味的にまとまっていないカテゴリも存在する．このようなカテゴリは 4.2.1 項の方法で除去する．

表 1　Wikipedia カテゴリの例

カテゴリ	見出し語	総見出し語数
存命人物	アイシャ・ラドワーン	184561 個
	士郎正宗	
	高橋留美子	
	村上もとか	
学校記事	北海道大学	24785 個
	京都大学	
	九州大学	
	筑波大学	
企業関連のスタブ	ESRI	18190 個
	UCS	
	NEC グループ	
	総合車両製作所	
整理が必要な項目	哲学	2766 個
	松田聖子	
	イタチョコシステム	
	非線形物理学	

3　提 案 手 法

本研究では，ある話題に関する新聞記事集合を入力した時に，それに適した分類ラベルである Wikipedia カテゴリを出力する (図 1)．分類ラベル生成における Wikipedia カテゴリの最も単純な利用法として，見出し語の頻度の総数が最も高いカテゴリを分類ラベルとすることが考えられる．しかし，一部の見出し語のみ突出して頻度が高いカテゴリが分類ラベルになる場合がある．このような分類ラベルでは，例えばフォルダ分けの際に扱いづらい．

そこで見出し語の頻度だけではなく，80:20 の法則と呼ばれる理論を用いる．これは，全体の数値の大部分は全体を構成するうちの一部の要素が生み出しているというものである．この理論の例として,「ビジネスにおいて売上の 8 割は全顧客の 2 割が生み出している」,「商品の売上の 8 割は全商品銘柄のうちの 2 割で生み出している」などが挙げられる．

まず新聞記事検索結果中における見出し語の頻度を求める．全体の頻度の大部分が，全体の一部の見出し語が生み出しているような (80:20 の法則に当てはまる) カテゴリは分類ラベルとして用いない．これにより，一部の見出し語のみ頻度が高い偏ったカテゴリが除去され，全体を俯瞰したバランスの良い分類ラベルを使用することができると考えた．

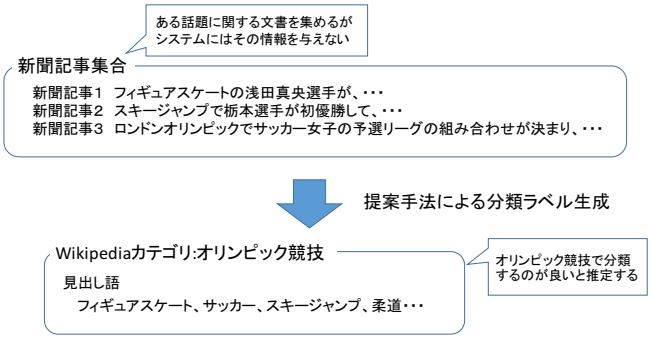

図 1　提案手法の入力と出力

4　実　　験

4.1　使用する文書集合

新聞記事 (2004 年 1 月 1 日～2018 年 5 月 31 日の共同通信) より「カンパニー」「キャンパス」などのキーワードで検索した結果のうち，時系列ランダムに 1000 件を選び，それを文書集合とする．以下の 18 個の文書集合を使用した．

カンパニー，キャンパス，オリンピック，芸能人，紅白歌合戦，アルゴリズム，ブラック企業，漫画，コンピュータ，料理，銀河，羽生結弦，パソコン，ロボット，ソフトウェア，テレビ，受賞

それぞれの文書集合の名前は検索キーワードを表している．文書集合はそのキーワードが文書中に現れる文書によって構成されている．これらの文書集合を考えるにあたり検索キーワードは文書集合の話題を表現していると考える．ただし，システムには検索キーワードは与えない．

4.2　実 験 方 法

4.2.1　カテゴリの除去

本研究で使用した Wikipedia カテゴリ 189,990 個のうち以下の 1)～3) の基準でカテゴリを除去する．除去後のカテゴリ数は 166,332 個となった．

1) 見出し語の種類が少ないカテゴリは，例えばフォルダ分けを考えた際に扱いづらい．そのため，見出し数が 3 以下のカテゴリを除去する．

2) カテゴリ「整理が必要な項目」「書きかけの節のある項目」など、分類ラベルとして扱いづらいカテゴリが存在する．そのため，カテゴリの記述パターンが「名詞」か「名詞+助詞+名詞」以外のカテゴリを除去する．これにより,「日本」「宇多田ヒカル」のような名詞単独のカテゴリや,「日本の野球選手」のような名詞+助詞+名詞のカテゴリのみ使用できる．

3) 4.2.2 項の実験を行なっているうちに，文書集合の話題に関わらずラベルセット候補になるカテゴリが存在することが判明した．そのため，複数の文書集合で候補となったカテゴリは事前に除去する．18 個の文書集合のラベルセット候補のうち，9 個以上の文書集合でラベルセット候補となるカテゴリを事前に除去した．除去したカテゴリを表 2 に示す．

表 2　複数の文書集合でラベルセット候補となったカテゴリの例

カテゴリ	見出し語
数字	7，8，9
仮名文字	ん，い，お
哲学の和製漢語	社会，哲学，原子
人間	個人，人類学，人間
時間	音楽，時間，時空

4.2.2　ラベルセット候補の生成

初めに，文書集合中の Wikipedia カテゴリ毎の見出し語の頻度を求める．1 文書中に見出し語が含まれている場合，その見出し語の頻度を 1 とし，含まれていない場合は 0 とする．ただし 1 文書中に同じ見出し語が複数含まれている場合でも頻度は 1 とする．これを全ての文書 (1,000 件) で行う．

次に，全カテゴリを見出し語の頻度の合計が上位 50 件のカテゴリを対象に，以下の式を求めて上位 10 件のカテゴリをラベルセット候補とする．

$$head = \frac{\text{上位 20\%の見出し語の頻度の総和}}{\text{全体の見出し語の頻度の総和}} \quad (1)$$

4.3　実 験 結 果

文書集合「カンパニー」のラベルセット候補を表 3 に示す．「東証一部上場企業」や「企業関連のスタブ」などがラベルセット候補となっている．カテゴリ「東証一部上場企業」の各見出し語の頻度を図 2 に示す．突出して頻度が高い見出し語がなく，バランスが良いのでフォルダ分けの際に扱いやすい．また，見出し語として「トヨタ自動車」「ソニー」などの企業名がまとまっており，文書集合の生成に使ったキーワード「カンパニー」と親和性があるので分類ラベルとして妥当であると考えた．ドキュメント「キャンパス」「オリンピック」「芸能人」「紅白歌合戦」のラベルセット候補を表 4〜表 7 に，各見出し語の頻度を図 3〜図 6 に示す．これらも文書集合「カンパニー」と同様に，頻度のバランスが良く，かつ文書集合の生成に使ったキーワードと親和性がある分類ラベルを得ることができた．また，4.1 節に示した文書集合のうち,「ブラック企業」「漫画」「コンピュータ」「銀河」「羽生結弦」「パソコン」「ロボット」「ソフトウェア」「テレビ」でも同様に文書集合の生成に使ったキーワードと親和性がある分類ラベルを得ることができた．

しかし，文書集合「料理」「アルゴリズム」では文書集合の生成に使ったキーワードと親和性がある分類ラベルを得ることはできなかった．文書集合「料理」のラベルセット候補を表 8 に示す.「料理」の分類ラベルとして親和性のあるカテゴリが存在しない．カテゴリ「時間の単位」の各見出し語の頻度を図 7 に示す.「年」「日」など，料理とは関連性の低い見出し語が大半を占めている．このように提案手法が望むように作用しない文書集合も存在する．

5　拡張固有表現

固有表現とは，固有名詞，日付表現，時間表現の総称である．具体的には組織名・人名・地名・日付表現・時間表現・金額表現・割合表現・固有物名の 8 種類である．Sekine ら [7] は，固

表 3　ラベルセット候補 (文書集合「カンパニー」)

カテゴリ	head
東証一部上場企業	0.524768
企業関連のスタブ	0.547945
情報	0.688805
技術	0.712215
製造	0.731707
工学関連のスタブ	0.747619
企業	0.752399
市場	0.777193
経済	0.803073
シンボル	0.865801

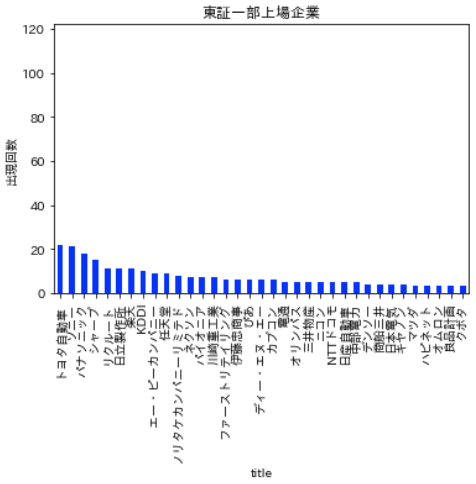

図 2　東証一部上場企業

表 4　ラベルセット候補 (文書集合「キャンパス」)

カテゴリ	head
学校記事	0.518359
認識論の概念	0.555256
知識	0.559896
社会	0.618926
大学関連のスタブ項目	0.644543
教育	0.680905
教育に関するスタブ	0.737500
学校教育	0.744511
行政区画の単位	0.778689
学部	0.789744

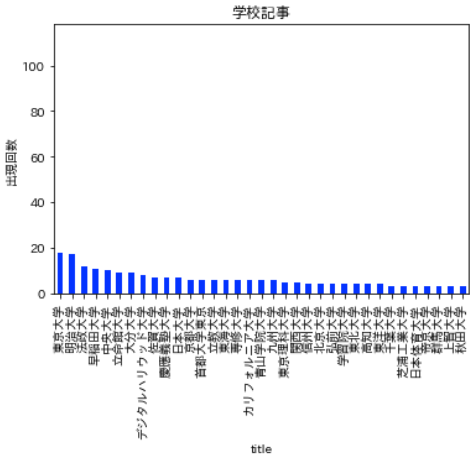

図 3　学 校 記 事

表 5　ラベルセット候補 (文書集合「オリンピック」)

カテゴリ	head
オリンピック競技	0.526221
教育	0.573333
頭字語	0.605405
G8 加盟国	0.630021
共和国	0.686327
思考	0.733746
スポーツ関連のスタブ項目	0.794366
スポーツ用語	0.800448
政治	0.810631
教育に関するスタブ	0.823423

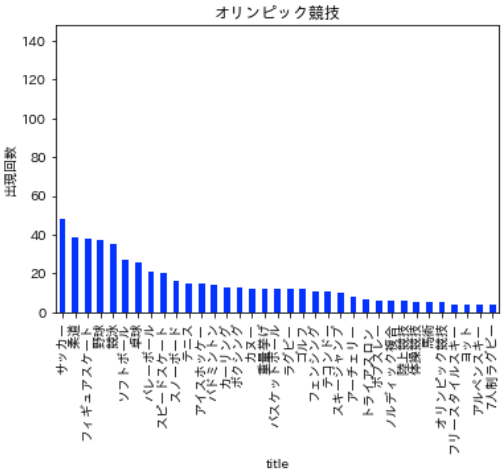

図 4　オリンピック競技

表 6　ラベルセット候補 (文書集合「芸能人」)

カテゴリ	head
日本の男優	0.455859
日本の女優	0.470064
日本のアイドル	0.478261
東京都区部出身の人物	0.486047
日本の司会者	0.508156
日本のタレント	0.518519
お笑い芸人	0.535385
吉本興業	0.542268
感情	0.543520
コミュニケーション	0.665127

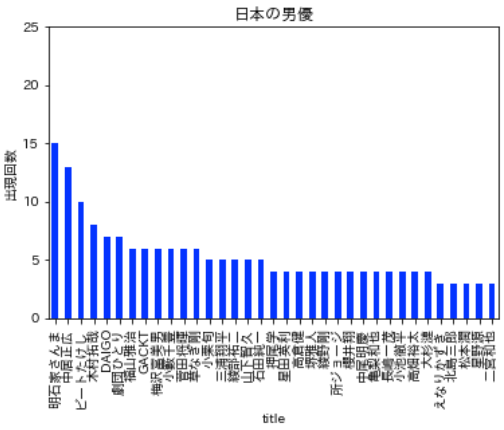

図 5　日本の男優

表 7　ラベルセット候補 (文書集合「紅白歌合戦」)

カテゴリ	head
東京都出身の人物	0.519737
日本の女優	0.617925
日本の男優	0.619593
東京都区部出身の人物	0.619870
日本のタレント	0.621086
日本の女性声優	0.629268
日本のアイドル	0.635623
日本の女性歌手	0.653061
NHK 紅白歌合戦出演者	0.654036
感情	0.656051

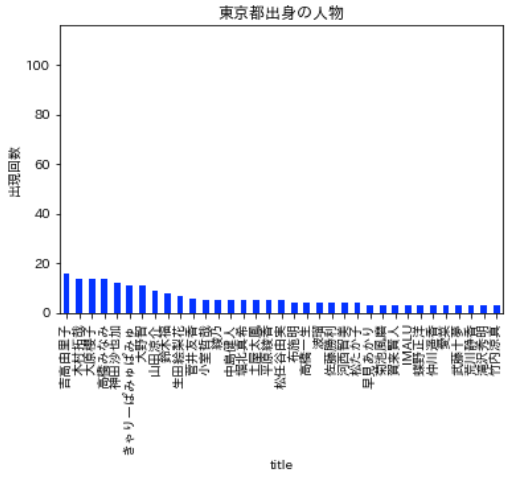

図 6　東京都出身の人物

表 8　ラベルセット候補 (文書集合「料理」)

カテゴリ	head
時間の単位	0.436416
人名の曖昧さ回避	0.519187
感情	0.559367
心理学関連のスタブ	0.666667
教育に関するスタブ	0.684322
フランス語の語句	0.727273
日本語の男性名	0.755319
賞	0.776952
日本の歴史関連のスタブ項目	0.795796
映画	0.817043

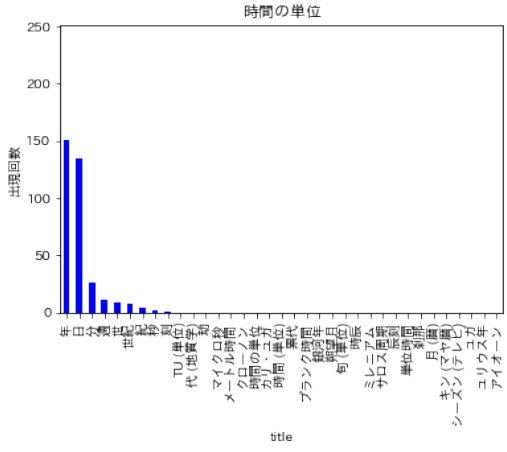

図 7　時間の単位

有表現を 200 種類に定義してこれを拡張固有表現とした．また，Sekine らは Wikipedia 記事に対して拡張固有表現のラベルを機械学習により自動的に付与した [8]．

拡張固有表現は機械学習と人手による作業により定義されており，分類ラベルとしての有用性が高い．そのため我々は研究当初，新聞記事検索結果に対する分類ラベルとして拡張固有表現を利用するタスクを行っていた．しかし，拡張固有表現「人名」に「世界」「人生」のように関連性が低い語が例外的に属しているケースが数多く存在し，望ましい結果を得ることができなかった．

一方で，Wikipedia カテゴリは見出し語に対する分類と関連キーワードを示すように設計されているので，このような例外は少ないと予想される．そのため，本研究では分類ラベルとして Wikipedia カテゴリを用いた方がよいと分かった．

6 ま と め

本研究では，文書集合に対する分類ラベル生成における Wikipedia カテゴリの利用法について検討した．具体的には，80:20 の法則を用いた．全体の頻度の大部分が，全体の一部の見出し語が生み出しているようなカテゴリは分類ラベルとして使用しない．これにより，一部の見出し語のみ頻度が高い偏ったカテゴリが除去され，全体を俯瞰したバランスの良い分類ラベルを使用することができる．実験の結果，文書集合「カンパニー」に対してカテゴリ「東証一部上場企業」，文書集合「キャンパス」に対してカテゴリ「学校記事」など文書集合の生成に使ったキーワードと親和性のある分類ラベルが付与されることを確認した．

今後の課題は，文書集合の生成に使ったキーワードと分類ラベルが，どの程度親和性があるかといった評価指標を検討することが挙げられる．また，文書集合の生成に使ったキーワードと親和性のある分類ラベルを生成できなかった場合の改善も今後の課題である．

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] 山田寛康，工藤拓，松本裕治. SupportVectorMachine を用いた日本語固有表現抽出. 情報処理学会論文誌，Vol. 43，No. 1，pp. 44–53，2002.

[3] 手島亮太. 統計値だけに基づくことを特徴とするトピックラベル抽出. 豊橋技術科学大学修士論文，2015.

[4] 梅田茂樹，細野公男ほか. 漢字カタカナ列の頻度情報に基づいた日本語文献の自動分類. 第 32 回情報処理学会全国大会論文集，4T-10，pp. 1687-1688，1986.

[5] 亀田弘之，藤崎博也. テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム. 情報処理学会論文誌，Vol.28，No.11，pp. 1103-1111，1987.

[6] 田村淳，渡辺道枝，原良憲，笠原祐. 統計的手法による文書自動分類. 第 36 回情報処理学会全国大会論文集，6U-5，pp. 1305-1306，1988.

[7] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In Proceedings of LREC, pp.1818–1824, 2002.

[8] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Neural Joint Learning for Classifying Wikipedia Articles into Fine-Grained Named Entity Type. Proc. 30th Pacific Asia Conference on Language, Information and Computation, pp.535–544, 2016.