検索結果整理のためのラベルセット選出計算高速化と Wikipedia カテゴリからのラベルセット選出

細野 湧城　高本 綺架　廣中 詩織　梅村 恭司
豊橋技術科学大学
{hosono.yuki.ql, takamoto.ayaka.nx, hironaka.shiori.ru, umemura}@tut.jp

概要

我々は情報収集のために検索することが多いが，適切なキーワードを使わないと，膨大な数の検索結果が表示されてしまう．そこで，適切なキーワードを思いつけないときでも検索結果を絞り込めるように，検索結果をうまく分類できるラベルセットを提示することを考える．本研究では，ラベルの含まれる文書数（文書頻度）をもとに算出した適性度を用いて，Wikipedia カテゴリをもとに生成した複数のラベルセットを順位付ける．ラベルセットの数が多いため適性度の計算に時間がかかるという問題があったが，前処理を工夫し，文書頻度の計算に Suffix Array を用いたアルゴリズムを利用することで，短時間で順位付けできた．

1　はじめに

我々は情報収集をするとき，膨大な数の文書に対して検索することで情報を探すことが多いが，適切な検索キーワードがわからないと，膨大な数の検索結果が表示されてしまう．そのため，我々はキーワードを追加して検索結果の絞り込みを行う．しかし，目的の情報に関しての知識が少ないと追加のキーワードを考えるのは難しいため，サジェストを用いることがある．サジェストは検索ワードや多数のユーザーの検索ログなどからキーワードを提案する機能である [1,2]．サジェストにより提案されるキーワードは，検索結果を絞り込むためによく使われているキーワードであり，様々な粒度のキーワードが提案される．本研究では，カテゴリに対応するキーワード集合（ラベルセット）を大量に用意し，その中から検索結果とマッチングのとれたものを複数提案することを考える．

検索結果には複数の話題が含まれているため，検索結果の文書集合を共通の話題を含む複数のグループに分類できれば，検索結果の整理や絞り込みに役立つと考えられる．グループ内の文書に共通する話題がキーワード（ラベル）で表現されるとき，ラベルには情報の抽象性の粒度があり，様々な粒度のラベルがある．粒度の小さいラベルは，それよりも粒度が大きいラベルによってさらにグループ化できること [3] や，小さい粒度のラベルによる小グループを，粒度が大きいラベルを用いて大グループにまとめることで，分類された文書をより整理しやすくなること [4] がわかっている．我々は，粒度のそろったラベルで検索結果が整理されている方が，ユーザにとって理解しやすくなると考えている．

宮越ら [5] は，粒度のそろったラベルセットを事前に用意し，その中から検索結果の整理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ている．宮越らは，検索結果には複数の話題が含まれているため，共通の話題を含む複数のグループへ検索結果を偏らず分類できるラベルセットが良いラベルセットであると考えている．実験に用いられたラベルセットは地域区分に関するものに限定されていたが，検索結果の整理に適したラベルセットを提示できていた．

本研究では，あらかじめ用意するラベルセットをつくるために Wikipedia[1)] を用いる．Wikipedia は多種多様な分野の記事を網羅しており，シソーラス辞書の作成にも用いられている [6] ため，この記事名をラベルとすることで多様な話題に対応できることを期待している．また，Wikipedia の記事はカテゴリによって分類されているため，カテゴリによってグループ化される記事の粒度はある程度そろったものとなっている．Wikipedia カテゴリを用いて，粒度をある程度そろえたラベルセットを大量に生成する．

本研究では，Wikipedia のカテゴリを用いて生成

1） https://ja.wikipedia.org/

した複数のラベルセットの中から，検索結果の分類に適するラベルセットを順位付けて提示する．さらに，検索結果とラベルセットのマッチング方法を工夫し，その結果の実行時間も報告する．

2 関連研究

検索結果を分類し，ラベル付けすることで文書の内容を提示する研究は多数存在する．淀川ら [7] はクラスタラベリングによって重要語を特定することで，内容の整理を試みた．村松ら [8] らはクラスタラベリングによって付与されたラベルに対して，既存の分類階層である Yahoo!カテゴリを利用した上位語を求めることで，分類した検索結果を階層化させて整理している．これらの研究により付与されたラベルは粒度が不均一である．我々は粒度が均一なラベルのほうが整理に適していると考えている．

Wikipedia を利用して検索結果を整理する研究も多数存在する．Ugo ら [9] は Wikipedia のリンク構造を用いたクラスタラベリングを行い，検索結果の内容にそぐうラベルの提示を試みた．平島ら [10] は Wikipedia カテゴリに対してパレートの法則を用いたり，分類に不適切なカテゴリの除去したりすることで，検索結果の内容にそぐうカテゴリを提示できている．

3 使用データ

3.1 文書集合

Ceek.jp News[2] が 2004 年 1 月から 2020 年 5 月に収集したニュース記事から抽出した記事集合を実験に用いる．検索結果に適するラベルセットを提示する実験をするため，検索結果を模した，21 種類の選別理由のある文書集合 RR と，選別理由のない文書集合 D を用意する．用意した文書集合は宮越ら [5] が実験で用いたものと同一である．選別理由のある文書集合 RR は，本研究における分類対象であり，記事集合から特定の単語リストのいずれかを含む文書をランダムに最大 3,000 件抽出したものである．選別理由のない文書集合 D は，検索対象となる全文書集合を代表するドキュメントの集合であり，記事集合からランダムに 30,000 件抽出したものである．

3.2 ラベルセット

多種多様な話題に対応できるよう，我々は Wikipedia の記事名とその所属カテゴリからラベルセットを作成する．まず，各ラベルセットがそのカテゴリに所属する記事名をラベルの集合として持つように，すべての Wikipedia のカテゴリをラベルセットに変換する．次に，分類に不向きであるため，ラベルセットに属するラベルが 200 種類を超えるラベルセットを削除する．人名などが羅列されたラベルセットが取り除かれることも期待している．さらに，各ラベルセットに属するラベルから，4 文字未満のラベルを削除する．ラベルと検索結果のマッチングをする際，「京都」が「東京都」に含まれてしまうような，意図しないマッチングを防ぐことを期待している．削除するラベルを 4 文字未満としたのは，これらのラベルを削除してもラベルセットに与える影響は小さいと見込んだためである．

本研究では，2022 年 8 月 20 日に作成された Wikipedia 日本語版のデータベース・ダンプ[3] から取得した 314,422 個のカテゴリと，1,974,022 種類の記事データをもとにラベルセットの集合 LSS を生成した．生成した LSS には 305,462 個のラベルセットと，1,441,413 種類のラベルが属していた．

4 実験方法

Wikipedia カテゴリを用いて生成したラベルセットを使って，各 RR に適したラベルセットの適性度を計算し，順位付ける．順位付けには宮越ら [5] が提案した手法を用いる．宮越らの場合よりラベルセット数が増えているため，適性度の計算の前に分類に適する見込みのないラベルセットを削除することにより処理対象を減らし，実行時間の短縮を図る．さらに，適性度の計算の際にはラベルが 1 回以上出現する文書数（文書頻度）を計算する必要がある．文書頻度を求めるために，我々は df_k を高速に数えるアルゴリズム [11] を用いる．このアルゴリズムは，分析対象となる文書集合に対して，前処理の段階で Suffix Array を用いた頻度表を作成することにより，複数の単語の文書頻度を，文書集合を逐一参照することなく求められる．このアルゴリズムを利用して計算できるように，宮越らの使用した一部の条件式を変更する．本研究ではこのアルゴリズムの C 言語実装を Python から呼び出して利用した．

2） http://www.ceek.jp/

3） https://dumps.wikimedia.org/jawiki/

その他の処理は Python で実装した．

処理全体の流れは以下の通りである．初めに，D で出現する傾向にあるラベルとラベルセットを特定し，分類に適する見込みがないラベルセットを削除する．次に，RR で出現する傾向にあるラベルセットを特定する．最後に，特定したラベルセットに対して，宮越ら [5] で提案された手法を用いて適正度を算出し，順位付けする．

4.1 ラベルセットとラベルの絞り込み

初めに，D で出現する傾向にあるラベルセットを特定する．この処理は検索対象を定めた時点で行う処理である．まず，LSS に含まれる全てのラベルの集合 L を求める．その後，L の全ラベルに対してラベルの文書頻度を求める．ここでは D におけるラベル l の文書頻度を $df(l;\mathrm{D})$ と表す．次に式 (1) を満たすラベルセット LS を特定し，満たさない LS は LSS から削除する．ここで，N_D は D の文書数であり，本研究では 30,000 である．また，M_{LS} はラベルセット LS に属するラベル数である．なお，式 (1) は宮越らの論文 [5] 内の式 (4) を，文書頻度による推定で置き換えた式である．

$$1-\prod_{l\in LS}\left(1-\frac{df(l;\mathrm{D})}{N_\mathrm{D}}\right)>\frac{2M_{LS}}{N_\mathrm{D}} \tag{1}$$

その後，文書頻度が 2 以上のラベルの集合 L_2 を求め，L_2 の要素を 1 つも含まないラベルセットを LSS から削除する．最後に，削除後の LSS をもとに L を再度求める．ここまでの処理により，ラベルセットは 5,253 個，ラベルは 97,708 種類に絞り込まれた．

次に RR で出現する傾向にあるラベルセットを求める．まず，L の全ラベルに対して RR における文書頻度 $df(l;\mathrm{RR})$ を求める．次に，L_2 から $df(l;\mathrm{RR})$ が 1 以上のラベルを取り出した部分集合 RL_1 を特定する．最後に RL_1 の要素を含む LSS の部分集合 LSC を求める．RR によって LSC に含まれるラベルセット数は異なるが，21 種類の RR で平均 4072 個に絞り込まれた．

4.2 ラベルセットの順位付け

宮越ら [5] の提案手法を用いて適正度を算出し，順位付けを行う．初めに，式 (2) を満たさない，分類に不適切なラベルセットを削除する．ここで，N_RR は RR の文書数であり，「りんご農家」の RR では 1,706，それ以外の RR では 3,000 である．なお，式 (2) は宮越らの論文 [5] 内の式 (3) を，文書頻度による推定で置き換えた式である．

$$1-\prod_{l\in LS}\left(1-\frac{df(l;\mathrm{RR})}{N_\mathrm{RR}}\right)>\frac{2M_{LS}}{N_\mathrm{RR}} \tag{2}$$

次に，ラベルの適正度を求める．まず，ラベルセット LS 内の各ラベルに対し式 (3) を求め，x_l を式 (4) のシグモイド関数に代入した値 $\sigma(x_l)$ をラベルの適正度とする．

$$x_l=\frac{\frac{df(l;\mathrm{RR})}{N_\mathrm{RR}}-\frac{df(l;\mathrm{D})}{N_\mathrm{D}}}{\sqrt{\frac{\left(\frac{df(l;\mathrm{RR})}{N_\mathrm{RR}}\right)\left(1-\frac{df(l;\mathrm{RR})}{N_\mathrm{RR}}\right)}{N_\mathrm{RR}}+\frac{\left(\frac{df(l;\mathrm{D})}{N_\mathrm{D}}\right)\left(1-\frac{df(l;\mathrm{D})}{N_\mathrm{D}}\right)}{N_\mathrm{D}}}} \tag{3}$$

$$\sigma(x_l)=\frac{1}{1+e^{-x_l}} \tag{4}$$

最後にラベルセットの適正度を求める．ラベルセット LS の適正度は LS 内の全ラベルの適正度の平均であり，式 (5) の s_{LS} で表される．この適正度に従いラベルセットの順位付けを行う．

$$s_{LS}=\frac{\sum_{l\in LS}\sigma(x_l)}{M_{LS}} \tag{5}$$

4.3 実行時間の計測方法

ラベルセットの順位付けは検索時に行うことを想定しているため，処理速度が求められる．そこで，本研究では RR を読み込んでからラベルセットの順位付けを行うまでの実行時間を調べることで，検索時に行うに足る処理速度か判断する．本稿では，各 RR に対して 10 回実行した時間の平均を報告する．

実験に用いた PC の OS は Windows 10 Home であり，CPU は Intel Core i7-9700，メモリは 32GB である．Python 3.8.10 で実行し，C 言語で実装されたアルゴリズムのコンパイルは GCC 9.4.0 で行った．また，時間計測には，C 言語で実装されたアルゴリズムの前処理は time モジュールの `pref_counter` 関数を，その他の処理は `process_time` 関数を用いた．

5 実験結果・考察

我々は，21 種類の RR に対して，4 節で説明した方法で適正度を求め，ラベルセットを順位付けした．本節では RR「J リーグ」に対するラベルセットの順位付け結果のみ示す．その他 20 種類の RR に対する順位付け結果は付録 A に示した．また，本節では各 RR における実行時間も示す．この実行時間をもとに，実際の検索で用いるに足るかを判断する．

RR「J リーグ」に対するラベルセットの順位付け結果を表 1 に 10 位まで示す．ラベル数が 10 以上の

表 1　J リーグに対するラベルセットの順位付け結果

順位	ラベルセット	適正度	ラベル数
1	日本のラグビー競技施設	0.98	1
2	セリエ A_(サッカー) の日本人選手	0.97	10
3	1886 年に成立した国家・領域	0.96	1
4	日本のサッカー選手名を冠したカテゴリ	0.94	15
5	プレミアリーグの日本人選手	0.93	10
6	誤った GND 識別子が指定されている記事	0.92	1
7	J リーグクラブ	0.89	60
8	カナダの貨物自動車メーカー	0.85	1
9	FIFA ワールドカップ日本代表選手	0.84	78
10	サッカー・ブンデスリーガ_(ドイツ) の日本人選手	0.82	26

ラベルセットは，日本人のサッカー選手や J リーグに関するラベルセットであるため，J リーグの内容と関連のあるラベルセットだと考える．一方，ラベル数が 1 のラベルセットはラベルセット名だけではラベルの内容が不明であった．ラベルセット内のラベルを用いて検索結果を分類することを考えると，ラベル数 1 のラベルセットは分類に不適切であり，ラベルの内容も名前だけで予測しづらいため，ラベル数が 1 のラベルセットは取り除かなければならないと考える．なお，平島らの研究 [10] では，ラベル数が 3 以下のラベルセットがあらかじめ取り除かれている．

次に，各 RR における実行時間を表 2 に示す．RR「震源地」以外では，平均約 3.6 秒で順位付けできたため，実際の検索時に行う処理として許容できる速さだと考える．RR「震源地」で遅くなった原因は，RR「震源地」には文書の 7 割以上を英語が占めているものが存在するためである．文書頻度を求める際に用いたアルゴリズムの実装は日本語で記述された文書に対して使うことを想定している．このアルゴリズムは反復度という特徴量を用いているが，英語は反復度が日本語よりも大きいため [12]，英語が多く含まれる文書集合は適していなかった．アルゴリズムの実装を反復度が大きい単語を考慮するように変更することにより，英語が多く含まれる文書集合に対しても，他の文書集合と同様に数秒程度で処理を終えられると考える．

表 2　各 RR における実行時間

RR の名前	実行時間［秒］
知事	3.0
国立大学	6.1
高校野球	2.7
りんご農家	2.6
サッカーワールドカップ	3.7
選挙	3.3
相撲	2.6
四大大会	2.9
ふるさと納税	3.7
J リーグ	2.7
コシヒカリ	3.4
マンゴー	5.1
阿蘇山	3.8
ミカン	4.4
牛肉	3.9
災害	2.9
ズワイガニ	3.4
震源地	44.9
フィギュアスケート	3.2
貿易	3.8
富士山	4.1

6　おわりに

本研究では，Wikipedia カテゴリを用いて生成した複数のラベルセットの中から，検索結果の分類に適するラベルセットをほぼ数秒で順位付けて提示することができた．上位に選ばれたラベルセットのうち，ラベル数が 10 以上のラベルセットをみると，文書の内容に関連するラベルセットを提示できたと考える．今後の課題として，ラベル数が少ないラベルセットの削除は行うことが必要であるとわかっている．

参考文献

[1] Zhiyong Zhang and Olfa Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th International Conference on World Wide Web*, pp. 1039–1040, 2006.

[2] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 795–804, 2011.

[3] 平川秀樹, 木村和広. 概念体系を用いた概念抽象化手法と語義判定におけるその有効性の評価. 情報処理学会論文誌, Vol. 44, No. 2, pp. 421–432, 2003.

[4] Hiroyuki Toda and Ryoji Kataoka. A search result clustering method using informatively named entities. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, pp. 81–86, 2005.

[5] 宮越遥, 吉田光男, 梅村恭司. 文書整理に用いる分類リスト順位付けの試み. 第 14 回データ工学と情報マネジメントに関するフォーラム. G43-3, 2022.

[6] 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia マイニングによるシソーラス辞書の構築手法. 情報処理学会論文誌, Vol. 47, No. 10, pp. 2917–2928, 2006.

[7] 淀川翼, 加登一成, 伊東栄典. 単語の分散表現を用いた文書クラスタのラベル推定. 人工知能学会第二種研究会資料, Vol. 2019, No. SWO-049, p. 03, 2019.

[8] 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. 第 19 回データ工学ワークショップ. B6-3, 2008.

[9] Ugo Scaiella, Paolo Ferragina, Andrea Marino, and Massimiliano Ciaramita. Topical clustering of search results. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp. 223–232, 2012.

[10] 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. G2-4, 2019.

[11] Kyoji Umemura and Kenneth Church. Substring statistics. In *Computational Linguistics and Intelligent Text Processing*, pp. 53–71, 2009.

[12] Yoshiyuki Takeda, Kyoji Umemura, and Eiko Yamamoto. Deciding indexing strings with statistical analysis. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, pp. 79–85, 2002.

表 3　すべての RR に対するラベルセットの順位付け結果

ラベルセット / RRの名前	鳥取県副知事	高等専門学校	1世紀の各年	1886年に成立した国家・領域	副大統領候補	鐵道の操り人形だと疑われるユーザー	スコットランドのテニス選手	環境保護	日本のラグビー競技施設	1784年に記載された甲殻類	サトウキビ属	住宅関連のナビゲーションテンプレート	マクドナルドユーザー	韓国の海域	1994年設置の日本の市町村	シスメックス	忍野村の企業
ラベル数	1	2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
知事	0.997	0	0.572	0	0.990	0.473	-	0.980	-	0	0	0.950	-	0.757	0	-	0
国立大学	0	0.998	0.993	0.520	0	0.526	-	0.753	-	0	0	0.532	0	0.809	0	-	0
高校野球	0	0	0.983	-	-	0.739	-	-	-	-	-	0	0	-	0.960	-	-
りんご農家	0	-	0.980	-	-	0.452	-	0.746	-	-	0	0.909	0	0.834	0	-	-
サッカーワールドカップ	0	-	0.422	1.000	0.637	0.906	-	0	0	-	0	0.933	0.945	-	0	-	0.486
選挙	0	-	0.180	0.416	0.998	0.473	-	0.473	-	-	-	0	0	0	-	-	0
相撲	0	-	0.180	0	-	0.988	-	-	-	-	-	0.418	0	-	0	0	-
ミカン	-	-	0.811	0.301	-	0.987	-	0.893	-	0.844	0.958	0.921	0.805	-	0	-	-
四大大会	-	-	0.064	0.990	-	0.257	1.000	-	-	-	-	0	0.616	-	-	-	-
ふるさと納税	0.807	0	0.899	-	-	0.093	-	0.995	-	0.973	0	0.990	0.697	-	0	-	0
富士山	-	0	0.979	0	0	0.970	-	0.999	-	-	0	0.981	0.380	0	0.700	-	0.486
Jリーグ	-	-	0	0.964	-	0.623	-	0	0.982	-	-	0	-	-	0.700	-	0
コシヒカリ	0	-	0.691	0	-	0.893	-	0.960	-	0.979	0.844	0.942	0.380	0	0	-	0
ズワイガニ	1.000	0	0.738	-	-	0.844	-	0.753	-	1.000	-	0.997	0.616	0.583	0	-	-
マンゴー	0	-	0.840	0.680	0	0.967	-	0.803	-	0.876	0.999	0.699	0.985	-	-	-	-
阿蘇山	-	0	0.962	0	-	0.526	-	0.595	-	-	0	0.999	-	0.583	0	-	0.486
牛肉	0.851	-	0.500	0	0.637	0.973	-	0.473	-	0.951	0.900	0.888	0.993	0	-	-	-
災害	0	-	0.778	0	0.406	0.666	-	0.686	-	0.798	0.798	0.950	0.380	0.971	0.935	-	-
震源地	-	-	0.257	0.785	0.637	0.000	-	-	-	-	-	0.532	0.380	0	0.998	-	-
フィギュアスケート	0	-	0.029	0.416	0	0.991	-	-	0	-	-	0	0	-	-	1.000	-
貿易	-	-	0.500	0.680	0.977	0.011	-	0.473	-	0	0.798	0.532	0.616	0.757	0	-	0.996

A　すべての RR に対する順位付け結果

表 3 にすべての RR に対するラベルセットの順位付け結果を示した．ここに示したラベルセットは，各 RR に対して最も適正度が高いラベルセットのみ取り出したものである．行は RR を，列はラベルセットを示している．各セルは RR に対するラベルセットの適正度を示している．赤く塗りつぶされたセルは，該当する RR（行）において最も適正度が高いことを示している．セルの値がハイフン（-）となっているラベルセットは，該当する RR において LSC に含まれず，適正度の計算対象外であったことを示している．また，セルの値が 0 となっているラベルセットは，該当する RR において式 (2) を満たさず，適正度の計算対象外であったことを示している．

表 3 をみると，フィギュアスケートの RR に対して，シスメックス（坂本花織選手などが所属する企業）というラベルセットの適正度が最も高いことが分かる．そのため，シスメックスのラベルセットにより，フィギュアスケートの内容の一部を捉えられていると考える．しかしながら，マンゴーやみかんの RR に対してサトウキビ属の適正度が高いなど，記事の内容とはおよそ的外れなラベルセットが分類に適すると判定されたものも存在することが分かった．さらに，ハイチ系日本人や，仏教系政党といった，ラベルセット名だけでは内容が予測しづらいラベルセットが存在することも分かった．検索結果の内容にそぐわなかったり，ラベルセット名から内容が予測しづらいラベルセットを削除するために，平島らの研究 [10] のようにラベル数が 3 以下のラベルセットを取り除くべきだと考える．