ARG WI2 No.8, 2016
WI2-2016-13

文書の多様性指標の提案とオンラインニュース記事の分析

須藤 明人† 鷲田 祐一†† 本田 秀仁† 和嶋 雄一郎††† 粟田 恵吾††† 植田 一博†

† 東京大学 †† 一橋大学 ††† 大阪大学 †††† 日本総合研究所

sudoa@iis.u-tokyo.ac.jp

概要 組織内で行われる議論や報道の多様性を知るためには，文書や発言の多様性を定量化できることが望ましい．しかし，従来の多様性の指標は，大規模な学習コーパスが必要であるため適用に限界があった．実際，報道やソーシャルメディアへの投稿については学習コーパスの日々の更新の負担が重く，議事録に関しては大規模なコーパスの入手が困難である．そこで本報告では学習コーパスを用いること無く文書の多様性が定量化できる手法を提案する．具体的には，生態系と文書の類似性を利用して，生物多様性の研究で用いられている Hill number を文書の多様性指標に応用した．インターネットで取得したニュース記事を用いた評価実験では，提案手法は従来手法を上回る性能を示した．

キーワード 文書の多様性, Hill number, ニュース記事

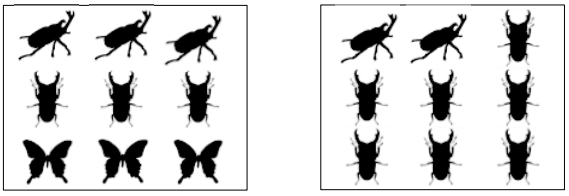
1 はじめに

本論文では，文書がどれくらい異なった内容を豊富に含んでいるかという多様性を定量化する指標を提案する．これまで，報道の多様性や会議での議論に含まれる意見の多様性の重要性が示唆されてきた．このような文章や会議の中の意見（トピックス）の多様性を，[Bache 13] に倣って「文章の多様性 (document diversity)」と呼ぶ．文書の多様性を定量的に測る指標があれば，報道や議論の多様性を直接測ることができ，メディアや会議の良し悪しを理解する一助になると考えられる．

これまで文書の多様性を与える指標の研究は我々の知る限り [1] のみである．Bache ら [1] は，あらかじめ大規模なコーパスを学習したトピックモデルを用いて文書を実数値ベクトルで表現し，そのベクトルから文書の多様性を算出する指標を提案した．具体的には，PubMed の大量の文書を Latent Dirichlet Allocation のトピックモデルで学習し，そのトピックモデルによって２つの論文や研究提案書を結合した擬似文書の多様性を定量化した．また，文書検索の分野では，内容の重複を避けてなるべくバラエティに富んだ検索結果をユーザーに与えることを目的に，明示的なアプローチと暗黙的なアプローチで研究が行われてきた [2]．多様な検索結果を与えるため，明示的アプローチでは検索クエリーのサブトピックと一致する文書を検索結果とし，暗黙的なアプローチでは検索済みの文書と重複しないような文書を検索結果とする．ただし，これらの手法は個々の文書の多様性を定量化するものではない．

[1] の手法で文書の多様性を評価するためには，過去に得られたコーパスを学習データとして用いてあらかじめトピックモデルを構築しておく必要がある．しかし，報道されるニュース記事，会議の議事録，ソーシャルメディアの投稿は，いずれもトピックモデルを構築するためのコーパスを容易に入手することができない．ニュース記事やソーシャルメディアの投稿のうち無償で入手できるのは最近のものだけで，現在から過去までの大規模なデータを得ようとすると有償になることが一般的である．会議の議事録については，過去に同じテーマの会議体が存在しないために過去の議事録が利用できないことが多い．さらに，報道やソーシャルメディアでは，日々新たに起こる出来事が新たなトピックになりうるので，過去の文書に加えて新しい文書も収集し続ける必要がある．例えば，2008 年当時のリーマン・ショックに関するニュース記事や分析記事においては，リーマン・ショックそのものがひとつの注目度の高いトピックとして扱われる必要があるが，過去のコーパスを学習しても “リーマン・ショック” というトピックをモデルが持つことはできない．ソーシャルメディアによってはウェブ API を通じて直近の投稿を無料で収集できる場合もあるが，投稿を収集するシステムを安定して稼働させるための技術力とサーバー代を含めた保守運用のコストが必要になる．

そこで本研究では，過去のコーパスを使うことなく文書の多様性を定量化することを目的に，生態学での多様性の指標のひとつである Hill number を文書の多様性の指標に用いることを提案する．文書はいくつかの種類の単語が集まってできている点が様々な種の個体が集まっている生態系と共通している．そのため，生態学の多様性の指標を文書の多様性に適用することは自然に行える．実際，以下に詳述するように，文書を bag-of-words 法で表現することで，Hill number を文書の多様性の指標に自然に拡張できる．



(a) 多様性が相対的に高い例　(b) 多様性が相対的に低い例

図 1　生態系おける多様性の比較.(a) は種の数が多い上に種の間の個体数が均等なので (b) よりも多様性が高いと考えられる.

Hill number が文書の多様性の指標として適切かどうかの評価には，先行研究 [1] に倣って二つの文書を結合した擬似文書を用いた. ただし，先行研究では学術論文から作成した擬似文書であるのに対し，本報告ではウェブ上のニュース記事を用いて擬似文書を作成した．これにより，ニュース記事から報道の多様性を測定できるかどうかの評価が行えると考えられる．

2　文書の多様性の指標

2.1　生態学分野における多様性の指標

ある生態系の多様性を定量化する方法は自明ではなく，これまで数多くの指標が提案されてきた [3]．広く用いられている指標としては，Shannon-Weiner エントロピーや Simpson インデックスがある．提案されてきた多くの指標には一長一短があり，生態学分野において，最良の指標に関するコンセンサスは存在しないと言われている [4].

数多くある生態系の多様性の指標に共通する性質が，多様性は種の豊富さ（Richness）と均等性（Evenness）の積に分解できることである [5, 6]．このとき，種の数が豊富で，それぞれの種の個体数が均等な生態系のほうが多様性が高いと評価される．図 1(a) の例では 3 種の個体が 3 匹ずつ均等に存在しているのに対し，図 1(b) では 2 種類しか存在しない上に個体数が 2 匹と 7 匹と偏っている．この例では図 1(a) のほうが生態系の多様性が高いと評価することが一般的である．

一長一短がある多様性指標を個々に用いていては統一的な視点で多様性を扱えないという課題を解決するため，パラメタを持つ族を多様性の指標とする方法がある．これは，多様性をスカラー量ではなく，様々な観点での多様性指標を要素とする複数次元のベクトル量として表現することに相当する．これまで複数の族が提案されてきたが [7]，Hill number [8] が多様性の指標として他の族に比べて優れていると [7] で報告されている．Hill number が多様性の指標が持つべき特徴を備えた，良い性質の指標であるとの報告もある [9].

Hill number は，種 i が個体の全数に占める割合が p_i であるとき，

$$ {}^qD = \left(\sum_{i=1}^{R} p_i^q \right)^{1/(1-q)} \tag{1} $$

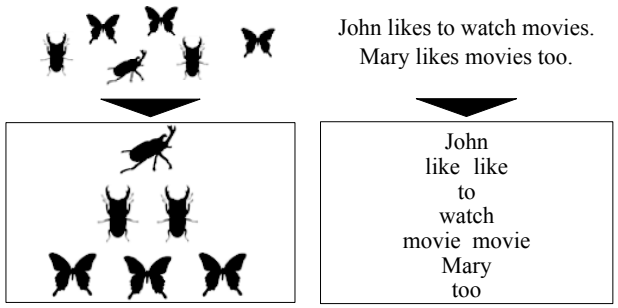
で与えられる. ここで，任意の実数を取る q が多角的な視点を与えるためのパラメタである．Hill number は $q = 1$ で Shannon-Weiner エントロピー，$q = 2$ で Simpson インデックスに一致する [8].

Hill number は生態系からランダムに個体を R 回選択する際に選ばれる種の数の期待値と解釈できる．ただし，種を選択する際の確率は p_i ではなく，p_i に q に応じた重み付けがされた確率である．$q = 0$ では Hill number は種の数に一致する．これは存在確率が種ごとに異なることを無視していることに相当する．$q = 1$ では，種の選ばれる確率が p_i のもとでこの操作を行った場合の種の数の期待値である. q が 2 より大きい時は，一般的に見られる種（p_i が大きな種）が選ばれる確率を p_i よりも大きな値に定めた上で，この操作を行って得られる種の数の期待値が Hill number の値に一致する．q が大きいほど，一般的に見られる種の選択確率が大きくなる．これは，q が大きいと，種のバランスが少し崩れるだけで多様性の値が大きく減少することを意味する．この Hill number の解釈については [10] に詳述されている．

2.2　文書の多様性の指標の提案

生態系と文書は，様々な種類の要素が複数集まって構成されている点が共通している．生態系の要素は生物の個体であり，それぞれの個体は生物種にグループ分けされている．例えば，図 2(a) のような生態系を考えることができる．そこでは，カブトムシ，クワガタ，チョウという生物種が存在し，それぞれ決まった個体数だけ存在している．この例では個体数は 1，2，3 である．一方，文書も同様に考えることができる．図 2(b) の “John likes to watch movies. Mary likes movies too.” という例では，文書は John，like，to，watch，movie，Mary，too という 7 種類の単語から構成されており，文書中に like と movie が 2 回，それ以外の単語は 1 回ずつ出現している.

このような異なる種類の複数の要素から構成されている集合を数理的に表現するには，それぞれの次元が要素の種類に対応し，その値は各次元に対応する種類が存在する個数であるようなベクトルを用いれば良い．生態系をそのようなベクトルで表現する場合，存在する種の数を次元数にするようなベクトルを考え，それぞれの次元の値を対応する種の個体数とすれば良いので，図 2(a) の例は，それぞれの次元がカブトムシ，クワガタ，チョウに対応するベクトルを用いて $(1, 2, 3)$ と表現できる．文書の場合は，単語の種類と次元を対応させてそれぞれの



(a) 生態系と種の多様性　　(b) 文書と単語の多様性

図 2 異種の要素が集まった集合の多様性

次元の値を単語の出現回数とすればよいので，図 2(b) の文書は $(1, 2, 1, 1, 2, 1, 1)$ と表現できる．このベクトル表現は，自然言語処理で広く用いられる bag-of-words 法での文書表現に他ならない．

Hill number による生態系の多様性の定量化は，この種の出現回数のベクトルを全個体数で割ることで得られるベクトルを特徴ベクトルとして生態系の多様性を表現していると解釈できる．種の出現回数のベクトルを全個体数で割ると，それぞれの次元の値は全個体に対するその種の個体が存在する割合になる．これは式 (1) の p_i に他ならない．従って，種の出現回数のベクトルとパラメタ q を与えれば，Hill number によって多様性が定まることになる．

種の出現回数のベクトルが多様性を与えるのは，種類の数とその偏りが多様性を与えることを考えれば自然である．先述の通り，生態系の多様性は種の数の豊富さ (Richness) と，種のあいだで個体数に偏りがないこと (Evenness) で決まる．このベクトルの次元数が Richness に対応し，個体数（または個体数の割合）の偏りのなさが Evenness に対応する．従って，種の出現回数のベクトルは，多様性を評価するための自然な特徴ベクトルであるといえる．

文書においても，出現する単語の種類が豊富で，それらの単語を偏りなく用いている文書ほど多様性が高いとして多様性を定義して良いと考えられる．直観的には，同じ単語が繰り返し用いられる文書と比べて，様々な単語が幅広く用いられる文書のほうが，豊富な内容を含んでいる可能性が高いと考えられるからである．このことはトピックモデルを考えるとより説得力が増すように思われる．トピックモデルは単語がトピックから生成されるという生成モデルであり，文書をトピックごとに分類するタスクをはじめとして非常に多くの活用事例がある．トピックモデルでは，学習データとして用いるコーパスの共起関係をもとに，それぞれの単語の各トピックへの帰属確率を推定しておく．文書が新しく与えられると，その文書が含む単語のトピックへの帰属確率がわかるので，その確率を用いて文書のトピックを決める．文書のあいだで，出現する単語のトピックへの帰属確率の偏りと，所属するトピックの順位が均等であるとすれば，トピックモデルの “トピック” の豊富さが文書の多様性を表すという仮定のもとで，単語の出現回数のベクトルの次元数と値の均等さで文書の多様性を定量化できることになる．

Algorithm 1 Calculate document diversity

```
U ← the document set
Q ← the set of the parameters q
for d in U do
    Do morphological analysis for d (optional)
    Remove stop words in d (optional)
    Remove words in d which are particular part-of-speeches (optional)
    Calculate bag-of-words vector v = (v_i)
    p ← v / Σ_i v_i
    for q in Q do
        ^qD = ( Σ_{i=1}^{R} p_i^q )^{1/(1−q)}
        D_d^q ← ^qD
    end for
end for
return {D_d^q}_{d∈D, q∈Q} as the diversity of d on q
```

以上より，文書の多様性は生態系の場合と同様に単語の種類の豊富さと均等さで与えて良いと考えられるので，本研究では文書の多様性を Hill number で与えることにする．文書が与えられたとき，その文書の多様性を Hill number で定量化するためには，その文書が含む単語の種類の数と，それぞれの単語が文書に占める割合を計算し，単語 i が文書にしめる割合を p_i として式 (1) に代入すればよい．これは，文書を bag-of-words 法で表現したベクトルを全単語数で割ったベクトルの要素を p_i として，Hill number に代入することと等価である．

文書の長さで Hill number を正規化した値が，多様性の指標として有効な場合もあると考えられる．偏りがある非常に長い文章と，まんべんなく単語が用いられている短い文書では，長い文章のほうが Hill number での多様性が高くなる傾向にある．たとえ偏りがあっても，文章が非常に長ければ単語の種類が増えるからである．Hill number を文書の全単語数で割った値を多様性として用いることで，1 単語あたりの多様性を評価でき，短くても満遍なく幅広い単語を用いている文書の多様性を高く評価できる．例えば，会議等で活発に発言するものの同じ内容ばかり繰り返す傾向にある人よりも，発言数は少なくても幅広い視野で発言のたびに異なる視点を与

表 1　ニュース記事からなる擬似文書の多様性の識別力の比較．評価指標には AUC を用いた．太字は文書長及び検索キーワードごとの最良の値を表す．

Document Type		Hill number (Proposed)						Bache			
Length	Keyword	q=0	q=1	q=2	q=3	q=4	q=5	k=10	k=30	k=100	k=300
[1000, 1500]	ggl alp || apl sdc	0.646	0.677	0.703	0.720	**0.721**	0.719	0.566	0.578	0.603	0.610
	ggl alp || apl ios 9	0.666	0.718	0.816	0.888	0.896	**0.898**	0.804	0.674	0.665	0.634
	ggl mcmc || apl sdc	0.576	0.648	0.704	0.742	**0.751**	0.750	0.526	0.432	0.547	0.468
	ggl mcmc || apl ios 9	0.743	0.795	0.849	**0.868**	0.866	0.862	0.739	0.664	0.635	0.661
	Average	0.658	0.709	0.768	0.804	**0.808**	0.807	0.659	0.587	0.613	0.593
[1500, 2000]	ggl alp || apl sdc	0.661	0.731	**0.771**	0.760	0.730	0.708	0.496	0.620	0.673	0.734
	ggl alp || apl ios 9	0.705	0.784	0.852	0.871	**0.873**	0.868	0.701	0.460	0.564	0.615
	ggl mcmc || apl sdc	0.634	0.682	0.726	0.755	**0.758**	0.754	0.548	0.382	0.494	0.519
	ggl mcmc || apl ios 9	0.825	0.909	**0.947**	0.938	0.907	0.894	0.701	0.770	0.678	0.768
	Average	0.706	0.777	0.824	**0.831**	0.817	0.806	0.611	0.558	0.602	0.659
[2000, 2500]	ggl alp || apl sdc	0.733	0.833	**0.901**	0.836	0.808	0.793	0.574	0.637	0.617	0.586
	ggl alp || apl ios 9	0.730	0.822	0.914	**0.920**	0.914	0.912	0.620	0.702	0.592	0.550
	ggl mcmc || apl sdc	0.655	0.720	0.859	**0.884**	0.882	0.876	0.492	0.573	0.602	0.542
	ggl mcmc || apl ios 9	0.829	0.874	0.908	**0.912**	0.902	0.890	0.831	0.764	0.663	0.692
	Average	0.737	0.812	**0.895**	0.888	0.876	0.868	0.629	0.669	0.618	0.592

える人の発言の多様性を高く評価したい場合は，全単語数で正規化した Hill number を用いると良い．

Algorithm 1 に Hill number を用いて文書の多様性を算出する手順を示した．bag-of-words を求める前処理として，対象とする文書の言語が膠着語である場合は形態素解析しておく必要がある．また，ストップワードや特定の品詞の除外を行っても良い．

3　評価実験

文書の多様性の指標としての妥当性を評価するため，本研究の指標と先行研究の指標 [1](以下，Bache の指標または Bache と呼ぶ）を実装して比較した．ただし，文書の多様性に関する正解データが存在しないため，現実に存在する文書をそのまま用いたのでは定量的な評価は行えない．そこで，二つの文書を結合した擬似文書を用いた評価手法を用いて，先行研究と本研究の指標を比較した．これは先行研究 [1] で用いられた評価方法である．

3.1　データセットと擬似文書

評価に用いた擬似文書は，インターネットから取得した英文のニュース記事を結合して作成した．豊富なトピックスが含まれるように作成した擬似文書には ‘多様性が大きい’ というラベルを，そうでない場合には ‘多様性が小さい’ というラベルを付与した．これにより，評価に必要な正解ラベル付きのデータを用意できる．本節の評価実験が [1] での実験と異なる点は，文書がニュース記事である点と，学習に使えるコーパスが多様性を評価したい対象の文書だけという点であり，他の条件は同じである．

以下に擬似文書の作り方を具体的に述べる．まず，ニュース検索エンジンにあらかじめ定めたいくつかのキーワードを入力し，ヒットするニュース記事の本文を複数取得する．これらのニュース記事群を ‘親文書’ と呼ぶ．次に，ふたつの親文書を任意に選び結合する．この文書を ‘子文書’ と呼ぶ．この際，異なるキーワードで検索した二つの親文書を結合した子文書のほうが多様性が高くなると考えられる．そこで，異なる検索キーワードで取得した親文書を結合した子文書には ‘多様性が大きい’ とラベル付けし，親文書が同じ検索キーワードである子文書には ‘多様性が小さい’ とラベル付けする．ここで，文書長と多様性には正の相関関係があると考えられるので，異なる文書長同士の文書を結合するとラベルが不正確になる恐れがある．この影響を考慮して，同水準の文書長の親文書を結合して子文書を作成する．このように作成した子文書をラベル付きの文書セットとして評価に用いる．

ニュース記事を取得する際に用いた検索キーワードは ‘google alphabet’, ‘google mcmc’, ‘apple self driving car’, ‘apple ios 9’ の 4 パターンである．2015 年 8 月 17 日の午後 3 時前後にこれらの検索キーワードを Google news 検索 (米国版) に入力して記事を取得した．図表中では ‘google’, ‘apple’, ‘alphabet’, ‘self driving car’ を ‘ggl’, ‘apl’, ‘alp’, ‘sdc’ とそれぞれ略記する．記事の日付が新しい順に検索結果が表示されるように設定した上で，検索結果の上位 100 件のニュース記事の本文を取得した．取得したニュース記事からある範囲の長さの記事のみを抽出し，キーワードごとにランダムに結合してキーワードと文書長の組み合わせごとに 100 の子文書を作成した．100 件のうち，‘多様性が高い’ と ‘多様性が

低い’というラベルを持つ子文書がそれぞれ 50 件ずつになるようにした．文書長の範囲は [1000, 1500], [1500, 2000], [2000, 2500] の 3 通りである．ここで [x,y] は，スペースを含めた文字数が x 以上 y 未満という意味である．ストップワードの削除は，python sklearn パッケージのストップワードリストを用いて行った．品詞を基準とした単語の削除は行わなかった．

3.2 手法のパラメーターの設定

式 (1) のパラメタ q は $q = 0, 1, 2, 3, 4, 5$ と設定をした．文書長での正規化は行なっていない．Bache の指標で多様性を与えるためには，データ行列と距離尺度の種類をあらかじめ定めておく必要があり，本実験では，文書トピック行列とコサイン類似度の逆数をそれぞれ選択した．これを選択したのは，最も精度が高い組み合わせと [1] で報告されているからである．トピックモデルの学習には，多様性の算出対象の 100 の文書を学習データに用いた．トピック数 k は 10, 30, 100, 300 とし，LDA の学習回数は 5000 にした．これらの設定は [1] で示された実験の設定と同じである．

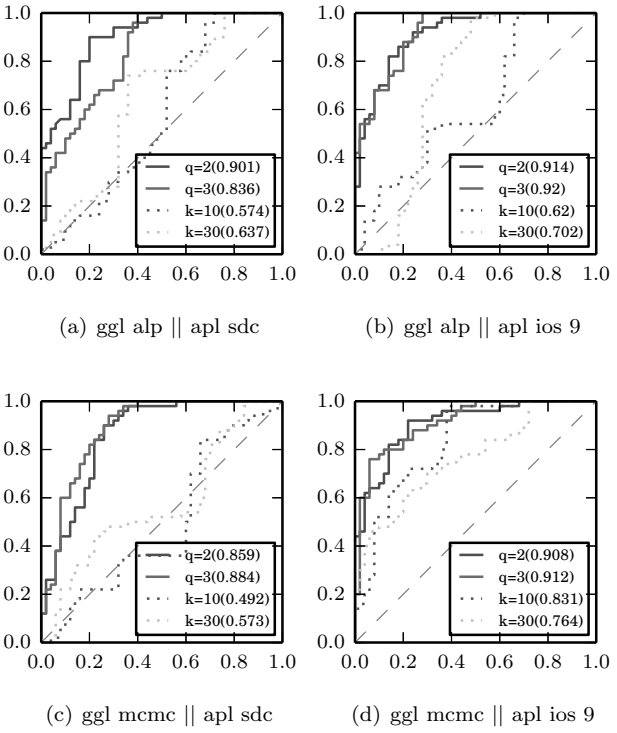
3.3 評価指標

評価指標も [1] に倣って Receiver operating characteristic curve (ROC 曲線) と，ROC 曲線の下側の面積である Area under the curve (AUC) を用いた．ROC 曲線は，識別機のラベルを判定する際の閾値を変化させた際の擬陽性率と真陽性率をプロットしたものであり，AUC はこの ROC 曲線の下側の面積である．あらかじめ閾値を定め，多様性の値がその閾値よりも高い時に，その文書を多様性が高いラベルの文書と識別することにすれば，多様性の指標を識別器とみなすことができるので，ROC 曲線と AUC で多様性の指標の性能を評価できる．このとき，AUC は，多様性が高いラベルの文書と多様性が低いラベルの文書をそれぞれランダムに選んで多様性を定量化した時に，多様性が高いラベルの文書の値が多様性が低いラベルの文書よりも大きくなる確率と解釈できる．

3.4 評価結果

表 1 に文書タイプ (表中では Document Type) と手法のパラメーターごとの AUC を示した．ここで文書タイプとは，文書長の範囲及びどの検索キーワードの親文書を組み合わせて作られた擬似文書かを表す．変化させたパラメーターは，Hill number の q と Bache のトピックの数 k である．表中の Average はキーワードの組み合わせについての平均である．また，文書長 [2000, 2500] について ROC 曲線を図 3 に示した．

$q = 0$ の文書長 [1000, 1500] の場合を除き，提案した指標の AUC の平均値 (表中の Average) はいずれの条件



(a) ggl alp || apl sdc　(b) ggl alp || apl ios 9

(c) ggl mcmc || apl sdc　(d) ggl mcmc || apl ios 9

図 3　キーワードの組み合わせごとの文書長 [2000,2500] の擬似文書の多様性の識別の ROC 曲線．縦軸は感度 (真の陽性率)，横軸は疑陽性率．凡例の括弧内は AUC.

でも 0.7 以上であり，0.709 から 0.895 の間であった．一方，Bache の AUC の平均値は 0.587 から 0.669 であった．[1] での実験と本実験の設定の違いは十分な学習コーパスの有無なので，比較手法のパフォーマンスが本実験で悪化している原因は学習コーパスの不足であると考えられる．この結果から，学習コーパスが限られる条件下では，提案した指標が従来手法よりも優れた指標であるといえる．

本研究の指標の結果を異なる q について比較すると，文書長 [1000, 1500] では $q = 4$ の，[1500, 2000] では $q = 3$ の，[2000, 2500] では $q = 2$ の提案した指標がそれぞれ最も AUC が高かった．多様性の指標として広く用いられる $q = 1$ (Shannon-Weiner エントロピー）や $q = 2$（Simpson インデックス）よりも $q = 3$ や $q = 4$ が文書長 [1000, 1500] と [1500, 2000] で高い AUC が得られており，q を変化させて多角的な分析ができる Hill number の利点を示唆する結果といえる．

4 おわりに

生態学分野の多様性の指標である Hill number を用いて，文書の多様性の指標を定量化する手法を提案した．大規模なコーパスが得られない条件下での，ラベル付きの擬似文書を用いた評価実験で，従来の文書の多様性の指標を上回る精度で多様性の高い文書を識別することが

できた．実験にはウェブから取得したニュース記事を用いた．

本研究を進めることで，将来的に次のような課題解決につながる可能性があると考える．まずニュース記事に関してだが，欧米各国に比べて日本の報道は多様性が欠けるという指摘がされることがある．実際に国内の報道は多様性が欠けているのか，どのような報道を増やすことで多様性を増やすことができるのかを，ニュース記事の多様性を提案手法で定量化することで明らかにできると考えられる．また，多様性と記事の閲覧数の関係を分析することで，より読者に好まれる記事を発信することにつながる可能性がある．例えば，速報記事は多様性が低いほうが好まれるが，分析記事は多様性が高いほうが好まれるという分析結果が得られれば，記事のタイプによって多様性を増減させることでより読者に好まれる記事が書けると考えられる．多様性と読者の好みの関係性は，ニュース記事にとどまらず雑誌，単行本，テレビといった文字や言葉のメディア全般を対象に分析できる．また，会議録のように人々の議論が記録された文書の多様性の分析を行うことで，組織の生産性や創造性を向上することにつながる可能性がある．未来についてのアイデアを生成する際に，多様な視点が独自性の高いアイデアにつながる可能性が [11] で指摘されているように，組織の構成員の多様性や個々人の発言の多様性が新事業開発や商品開発の現場での生産性と関係性があるかもしれない．[12] では，チームでアイデアを生成する際の「拡散的な議論」が既存の情報にとらわれない新しい視点の導入につながると述べられているが，提案手法を用いて会議やワークショップ等における議論の拡散や収束を測定することで，アイデア生成のプロセスをより生産的にできる可能性がある．他にも膨大なインターネット等の情報の取捨選択の際に多様性を確保することで，創造性豊かな組織作りに貢献できるかもしれない．

技術的な課題として，本研究で与えた指標は単語の意味や類似性を考慮していないことがあげられる．また，応用によっては適切な q の値をひとつに決めて分析することが求められる可能性があり，その場合は適切な q を選択する手法が必要である．表 1 では，文書長が短い場合には大きな q が良い識別率を与えており，文書長と適切な q の間に関係があることを示唆する．実際，生態系の多様性の評価には，サンプル数が少ないほど大きな q が良いという指摘がある [10]．従って，文書長は q の適切な選択のひとつのファクターになり得ると考えられる．今後は，以上の応用面及び技術面の両面で研究を行う予定である．

謝辞

本研究は，株式会社日立ソリューションズならびに株式会社ゼロストラクトとの共同研究として実施されました．ここに謝意を表します．

参考文献

[1] K. Bache, D. Newman, and P. Smyth, “Text-based measures of document diversity,” Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data miningACM, pp.23–31 2013.

[2] S. Kharazmi, M. Sanderson, F. Scholer, and D. Vallet, “Using score differences for search result diversification,” Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrievalACM, pp.1143–1146 2014.

[3] A.E. Magurran, Ecological diversity and its measurement, Springer Science & Business Media, 2013.

[4] A. Chiarucci, G. Bacaro, and S.M. Scheiner, “Old and new challenges in using species diversity for assessing biodiversity,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol.366, no.1576, pp.2426–2437, 2011.

[5] M.A. Buzas and L.-A.C. Hayek, “Biodiversity resolution: an integrated approach,” Biodiversity Letters, pp.40–43, 1996.

[6] H. Tuomisto, “An updated consumer’s guide to evenness and related indices,” Oikos, vol.121, no.8, p.12031218, 2012.

[7] B. Tóthmérész, “Comparison of different methods for diversity ordering,” Journal of vegetation Science, pp.283–290, 1995.

[8] M.O. Hill, “Diversity and evenness: a unifying notation and its consequences,” Ecology, vol.54, no.2, pp.427–432, 1973.

[9] R. Routledge, “Diversity indices: Which ones are admissible?,” Journal of theoretical Biology, vol.76, no.4, pp.503–515, 1979.

[10] N. Gotelli and A. Chao, “Measuring and estimating species richness, species diversity, and biotic similarity from sampling data,” Encyclopedia of biodiversity, vol.5, pp.195–211, 2013.

[11] 本田秀仁，鷲田祐一，須藤明人，粟田恵吾，植田一博，“未来に関するアイデア生成のエキスパートとノンエキスパートは何が違うのか?: 認知プロセスの分析,” 知識共創，pp.II2–1–II2–9，2015.

[12] 和嶋雄一郎，鷲田祐一，冨永直基，植田一博，“ユーザ視点の導入による事業アイデアの質の向上,” 人工知能学会論文誌，vol.28，no.5，pp.409–419，2013.