多様性の解析を用いたニュース記事に対するコメント集合の分析

宇野 毅明・武富 有香（国立情報学研究所・情報学プリンシプル研究系）
小林 亮太（東京大学 新領域創成科学研究科）
橋本 隆子（千葉商科大学 情報学部）
久保山 哲二・申 吉浩（学習院大学 計算機センター）

概要：本稿では，ニュース記事に対するコメント，および，ニュース記事のあるカテゴリのニュース記事に対するコメントの多様性を測り，その結果を紹介する．記事やカテゴリに対して人々がどのような反応をしているかを読み解くための材料として，新たな方向性を提案する．異なる記事に対するコメントは，単一の記事に対するコメントよりも多様性が増すであろう，という仮説から，あるカテゴリの複数の記事に対するコメントの多様性を比較することで，カテゴリの多様性を評価する．
キーワード：ソーシャルメディア，コメント，べき乗則，多様性

Analysis on Comments to News Articles by Diversity Analysis

Takeaki Uno / Yuka Takedomi (Principles of Informatics Research Division, National Institute of Institute)
Ryota Kobayashi (Graduate School of Frontier Science, Tokyo University)
Takako Hashimoto (Faculty of Commerce and Economics, Chiba University of Commerce)
Tetsuji Kuboyama / Yoshihiro Shin (Computer Centre, Gakushuin University)

Abstract: We analyze the diversity of the comments posted to news articles, and news articles in some categories. We think this helps interpretation of the behavior of the persons in a society from a new view point. From a hypothesis that the diversity of comments to two different news articles is larger than that to one news article, we compare the diversity of several number of articles in the same category, to observe the diversity of the comments to the articles in the category.
Keywords: social media, comment, power law, diversity

１．まえがき

近年，社会や群衆の意見や気持ちを抽出する方法としてソーシャルメディアが注目されている．今までは取得がほぼ不可能であった「その場そのときの人々の気持ち」が，完全ではないものの，大量に取得できるようになっている．アンケートで取得した意見は，心と気持ちが整った状態で書かれたものであり，ソーシャルメディアのような「何かの情報や体験に触れたそのときの意見や気持ち」は書かれない．気持ちが動いたとき，そのときの意見をそのままに書いたデータは，人類が初めて手に入れるものと言っても過言ではないだろう．

ここでは，ソーシャルメディアの中でも，ニュース記事に対するコメント，特に株式会社ヤフーが運営するヤフーニュースのコメント欄に着目したい．一般にヤフコメ，と略されるものである．ツイッターなどのメディアでも意見を述べている人は多いが，その意見が一体何に対して向けられているのか，その判別の自動化は難しい．通常はキーワードで絞り込むのであるが，例えば「選挙」を含むツイートを集めたとしても，あるツイートの意見が選挙制度なのか，選挙の結果なのか，選挙に興味のない人が多いことなのか，わからない．一方で，ニュース記事に対するコメントは，確実に，その記事，あるいはその記事に関連する事柄に対する意見である．このようにカテゴリを簡単に絞り込むことができるという意味で，ニュース記事のコメントは大変有用である．また，ヤフコメは非常に数が多く，一日に 10 万以上のコメントが発信されている．また，それぞれのコメントに，いいね，悪いね，がつけられるようになっており，多くの賛同を得ているか，賛同と批判はどちらが多いか，を知ることもできる．この質的なことの評価があることと，量的なアドバンテージがあることが，ヤフコメのメリットである．

ヤフコメには，様々なことが書かれる．記事に対する意見がしっかりと書かれていることもあるし，記事の書き方や取材，マスコミに対する意見であることもある．政治家や芸能人を褒める，

あるいは悪口を言うものもあるし，記事に関連する自分の体験を語っているものもある．この多様さ，猥雑さが，人々の自然な気持ちの表れであると考えてもよいだろう．また，世論調査とまったく違う意見がマジョリティや，いいねを多く集める意見になることもある．多くの人々がうっすらと思っている意見と，少数だが強い思いを持っている人の意見は大きく異なることもあり，その差が，この違いに現れているとも考えられる．全体的な調査では拾えないような，目立たないが大事なこと，あるいは助けを求めるメッセージなどが現れている可能性がある一方で，何も考えずに反射的に書いたであろうコメントで溢れている場合もあり，こういったものをすべて含んだものとして，貴重な分析対象であり，魅力あるデータであると考えられる．

このように，豊かな情報が隠れていると考えられる一方で，コメントに対する情報技術を用いた分析は，必ずしも深い意味をくみ取れるとは限らない．一般にコメントは丁寧に書かれているとは限らず，不完全であることも多く，単純な情報処理技術での扱いは困難である．頻度を基軸にしたキーワード抽出や，個別のコメントに対するセンチメントの分析などで，おおまかな様相を獲得する程度のことしかできていない．

本稿では，コメントに対する単語の多様性の解析からヤフコメの様相を分析したい．ここでは単語の多様性とは意味的な多様性ではなく，単に文字として異なる単語がいくつ，どのような割合で入っているか，という意味である．コメントの単語の多様性は，コメント，意見，感想などの多様性を表していると考えられる．同じ意見でも，物事を説明している場合は，説明の仕方，根拠の付け方は多様であるため，多様性が高くなると考えられ，一方で感情的，紋切り型の意見は，感情を表す言葉は説明に比べると多様でないため，多様性が相対的に低くなると考えられる．２つのコメント集合があるとき，２つを１つに混ぜたら一般には多様性が上がると考えられるが，１つにしても多様性があがらなかったら，両者は同じようなことに対して同じよう意見を述べている可能性が高いと考えられるだろう．このように，多様性，という普段はあまり見ない側面からコメント集合を見ることにより，キーワードなどからは得られない情報が獲得できると考えられる．

本稿では，一般性の高さを考慮し，記事カテゴリの多様性を考えたい．あるカテゴリの記事に対するコメントの多様性を評価するには，そのカテゴリの記事のコメントの多様性の平均を見る方法がある．一方で，コメントの内容自体は多様なのではあるが，どの記事に対しても同じような反応しか見られない，という場合もある．このような非多様性は，１つの記事のコメントからは分析できない．そこで，１つの記事のコメントの多様性と，複数の記事のコメントを合わせたものの多様性を比較することで，ある記事群，たとえばあるカテゴリの記事に対するコメントの多様性を測ることとする．

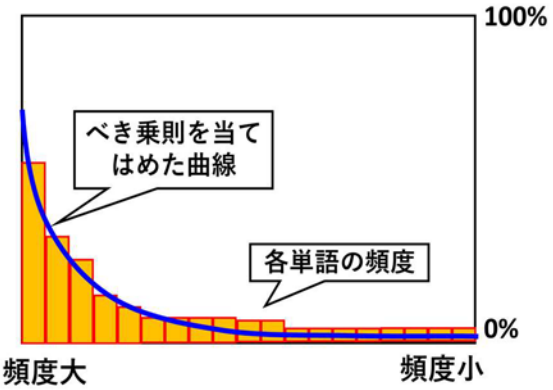

図 1：頻度分布にべき乗則を当てはめた様子
Figure 1: fitting power law to word frequency

２．コメントの多様性

ニュース記事に対するコメントの多様性は，例えばであるが，健全性の一つのバロメーターと考えられるだろう．炎上したり，誹謗中傷が起こったりすると，皆が同じようなことを言う．エコーチェンバーのような状態，あるいは 2 極化などがおこっても，同じようなコメントが増え，多様性は下がるだろう．コメントの集合が多様性を持つかどうか，が，いろいろな意味を見いだす手伝いをしてくれる．多様性はいろいろな意味合いを持つが，本稿では，このような，コメント集合，あるいはニュース記事からなんらかの意味を取り出すために役立つように，多様性というものを考え，モデル化したい．

コメントの多様性は，意見が多くなれば上がる．いろいろものが出てくればあがる．しかし，炎上しているような状態では，みなが同じようなことを言い，多様性が低いと感じるだろう．つまり純粋に意見の種類が増えれば，多様性が自動的に増える，というものでもない．例えば，ある記事のコメントが炎上していて，コメントが 1000 件ついていたとする．翌日に，同じように炎上していて 3000 件に増えたとする．炎上の内容・意味が変わらず，炎上の仕方が変わらなければ，1000 件のコメントと，3000 件のコメントの多様性は同程度であってほしい．単に異なる意見の数とすると，きっと 3000 件のコメントの多様性は，1000 件のものの 3 倍となってしまう．このような単純な測り方では，記事や炎上の意味を調べたい，とする研究の助けにはならず，より目的にかなったモデルが必要となる．

多様性を図るためには，もう一つ大きな問題がある．我々は，同じような内容のコメントが溢れていたら，多様ではないと考える．しかし，コメントの意味や内容の理解，内容の類似性の測定は，

現在の情報技術では非常に難しい難題である．特に本稿が扱うヤフコメのような，あまり整っていない，正確に精緻に書かれたわけではない文章であればなおさらである．たとえ，人間が分析したとしても，類似性を客観性を持って判断できるかどうかは疑問である．

これらの問題に対処するために，我々は過去の研究において，以下の方法を考案した．本稿は，その紹介と，この方法でヤフコメの多様性を測った場合にどのようなことがおこるのかを観察し，それを分析して考察することである．まず，内容の類似性については，意味を取る，というアプローチを諦め，内容の多様性に相関が高そうな，他の尺度を用いることとする．具体的には，コメント集合が含む単語の多様性を用いる．いくつかのコメントの内容が類似していれば，それらに共通して用いられる単語も自然と増えるであろう，その結果，多様性が減少するであろう，というものである．実際になんらかの証明や実証がなされているわけではないが，異なる話題，異なる意見が述べられているのであれば，より多くの単語が用いられるため，単語の共通性は平均的には低くなると考えられるだろう．

このような単語の共通性，つまりコメント集合にどのように単語が現れているか，という情報から，多様性を図る方法としては，単語の頻度分布を用いるものがある．単語の頻度分布とは，多様性を測定したいコメント集合に対して定義されるものであり，以下のように定義される．コメントの集合 X に対して，単語 w の出現回数を，X に含まれるコメント，に w が何回現れるか，で定義する．たとえば，X が以下の 4 つのコメントからなる集合とする．

「これは良いと思った」
「これはありがたいと思った」
「良いというのだから良いのだろう」
「あまり信じられない」

このとき，「良い」という単語は最初のコメントに 1 回，3 番目のコメントに 2 回現れているので，X には 3 回現れている，となる．単語 w の X における出現階数を，X の総単語数で割ったものが，w の X における頻度である．X の総単語数とは，X に含まれるコメントの単語数を全て足したものであり，上記の例では，1 つめのコメントが 5 つ，2 つめが 5 つ，3 つめが 8 つ，4 つめが 4 つであるので，総単語数は 22 となる．単語をどのように切り分けるか，については議論があると考えるが，情報技術としては，形態素解析プログラムを使って得られるものを単語とすることが多く，本稿でもその流儀に従うものとする．頻度分布は，全ての単語の頻度を大きい順に並べた数列のことである．例えば，a という単語の頻度が 0.1，b が 0.1，c が 0.5 で d が 0.3 であれば，頻度分布は 0.5, 0.3, 0.1, 0.1 となる．この頻度分布の多様性を計算する尺度がいくつか提案されており，ここではそれらを用いることにしたい．それら多様性の尺度については次節で解説する．

3．頻度分布の多様性測定手法

ここでは，頻度分布から多様性を測る尺度として，BPK，エントロピー，ジニ係数，その他に，頻度分布にべき乗則の関数を当てはめたときのべき数の値で評価するもの[3]がある．多様性の測定について詳しくは [2]を参照されたい．

べき乗則当てはめ

単語の頻度分布はべき乗則に従う，という結果が知られている[4]．つまり，この頻度の列は，多くの場合，aY^{-k} という関数で近似できることがわかっている．つまり，ある程度自然に集められた文章の集合であれば，ある適切な a, k を選ぶことにより，X 番目の単語の頻度が aX^{-k} に近い値になる，ということである．このように近似できることをべき乗則に従う，と言い，多くの現象，例えば年収の分布，論文の引用数の分布，などがこの法則に従うことが知られている．また，自然言語の文章集合の単語頻度では，一般に k の値は 1.0 程度になることが知られている[4]．文章の多様性が低く，少数の単語が高頻度で現れている場合は，k の値は大きくなり，文章が多様であり多くの単語が低頻度で現れている場合，k の値は小さくなる．頻度分布に対して，べき乗則をあてはめ，関数 aX^{-k} が頻度分布にもっとも近くなるように a, k と調整し，そのときの k の値で多様性を測るのがこの方法である．実際の方法に関しては，例えば[1]を参照されたい．k が大きければ，少しの単語の頻度が大きい，ということであり，多様性は小さく，k が小さければ，多くの単語が平均的に頻度が高いということであり，多様性が大きいと解釈する．

BPK

今回紹介する尺度の中で最もシンプルなものが BPK である．BPK は頻度分布の最初の数字，つまり，もっとも多く現れている単語（最頻出単語という）の頻度で，そのコメント集合の多様性を測るものである．これは，直感的に言えば，「多様性が低ければ，同じ単語がたくさん現れているはずである．よって一番多く現れる単語がどれくらい現れているかを比較すれば，多様性がわかるだろう」と説明できる．一般に文章の単語頻度はべき乗則に従うので，コメント集合 X の多様性が高く，いろいろな単語が現れていれば，X において高頻度の単語の頻度は，あまり高くならない．逆に X の多様性が低ければ，ある種の単語が集中的に現れている，ということであり，つまり X において高頻度な単語は相対的に頻度が高くなる(図 1 参照)．また，頻度分布が Zipf 則に従うので，BPK の値，つまり一番頻度の高い単語の頻度

が高いのに，残りの単語の頻度が平均的に高く多様である，というようなことはあまりおきず，BPK が低いのに単語全体は多様性が低い，ということもあまりないだろう，ということが導かれる．

ただし，BPK には，BPK の値が実際の多様性と大きく乖離してしまう危険性がある点が，弱点として指摘できるだろう．たとえば，たまたま，コメント集合の最頻出単語に表記揺れなどがあり，二通りの書き方があったとしよう．例えば「にほん」と「日本」，「良い」と「よい」などである．このようなとき，事実上，これらの単語は 2 つの異なる単語として扱われてしまい，BPK の値は，2 番目に頻度の高い単語の頻度となってしまう．1 番目と 2 番目の単語の頻度は，30%程度ことなることも多く，値が大きく変わってしまうケースが存在しうる．この他の要因で，最頻出単語の頻度が，なんらかの要因で偶発的に変わることがあり，その意味でリスクのあるものだと言えるだろう．今回は，このような理由から，BPK については紹介にとどめ，分析の道具としては利用しないことにする．

ジニ係数

ジニ係数は，貧富の格差を測るためによく使われているが，頻度分布に用いた場合，多様性の大きさを測っていると見なすこともできる．単語の頻度がみな同じような値であれば，多様性が高く，ジニ係数は小さくなり，多様性が低くなれば，ジニ係数は大きくなる．

頻度分布に対するジニ係数は，以下のように定義できる．まず，単語を頻度が低い順に並べ，頻度が小さい順に足していく．その合計が増加するカーブ(図 2 の「頻度の累計」カーブ)と，単語の頻度が全て同じである場合の合計増加カーブ(図 2 の「全頻度が同じ場合の頻度の累計」の直線)を考える．こうしてできる，図 2 の A の部分の面積を，A と B の面積の合計で割ったものとなる．具体的な計算方法の詳細については，ここでは割愛する．

今回扱っているようなべき乗則に従うような分布に対して，このように係数を計算すると，単語数が多くなるにしたがって，頻度の低いものの割合が増え，ジニ係数は大きくなる（つまり多様性が減ると解釈される）．つまり，同じようなコメント集合であっても，コメントの量が増えれば，自動的に多様性が下がると解釈されてしまう．つまり，コメント数に対する安定性がない．通常ジニ係数を計算するときは，データ，ここでは単語，を定数個，例えば 10 個のグループに頻度順で分割し，それぞれのグループの頻度（グループに含まれる単語の頻度の合計）を用いてジニ係数を計算する．頻度順のグループとは，最初のグループ

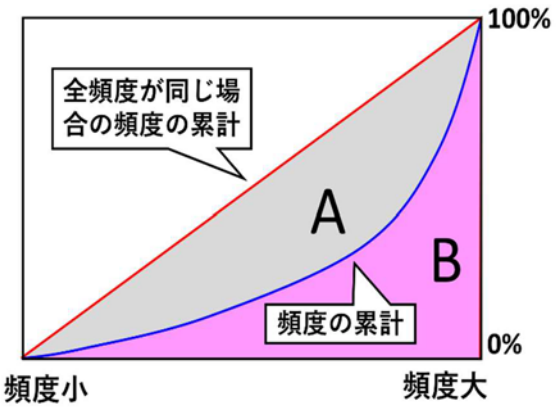

図 2 ：ジニ係数の計算方法
Figure 2: Computation of Gini coefficient

が頻度上位 10%，次のグループが上位 10%から 20%，というようにわけるものである．これにより，データの数が増えてもグループ数の数は変わらないので，上記のデータ数に対する不安定性を排除できる．しかし，単語の多様性の場合，人が内容の多様性を感じるであろう，頻度の高い単語が最初のグループに全て入り，混ぜ合わされてしまうため，意味的に重要な部分が全て潰されてしまう可能性がある．そのため本稿では，比較を行う際にはデータの数をそろえることによって安定した比較を試みている．

エントロピー

エントロピー（平均情報量）は，どれだけ情報が含まれているかを表す指標である．確率分布に対して定義されるものであるが，単語の頻度が確率であると見なせば，頻度分布に用いて問題はない．統計物理でもエントロピーは使われており，系の乱雑さを表す尺度として用いられる．

エントロピーは，全ての単語について，その単語の（単語頻度）×（−log 単語頻度）を足したもので定義される．頻度のばらつきが小さいほど，つまり多様であるほど，エントロピーは大きくなる．

エントロピーは全ての単語頻度が等しいときに最大となり，その値は，log(単語数) となる．基本的にはエントロピーは単語数が大きくなるほど大きくなりやすい．そのため，エントロピーを log 単語数で割って正規化した（最大値が 1 となるようにした）値をここでは用いる．

4．ヤフコメの分析

今回は，2020 年 12 月のヤフーニュースの記事からコメントが多く付いた記事をいくつかピックアップし，その記事についたコメントを多様性の面から解析した．

取得した記事からいくつかのカテゴリを定義し，それぞれのカテゴリに所属する記事を集めた

表 1 ：各カテゴリの記事に対するコメントの多様性尺度の平均
Table1: Mean diversity values for comments in each category.

	コロナ	感染者数	企業	皇室関係	事故火事	犯罪訴訟	外交国際	国内政治	スポーツ	天候災害
エントロピー	0.6535	**0.6570**	0.6434	0.6338	0.6212	**0.6698**	0.6503	0.6524	**0.6613**	0.6557
ジニ係数	0.7106	0.7051	0.7267	0.7370	0.7402	**0.6868**	0.7144	0.7120	**0.6958**	**0.7034**
べき乗則	**1.4242**	1.4325	1.4514	1.4738	1.5201	1.4708	**1.4213**	**1.4275**	1.4797	1.4374

記事集合を作る．そして，カテゴリ内の記事をランダムに 10 個取り出し，そのコメントをランダムに 160 個取り出して，各多様性尺度での評価値の平均値を計算した（表 1）．多様性の尺度は，エントロピー，ジニ係数，べき乗則当てはめを用い，それぞれの尺度での多様性上位 3 つを太字に，下位 3 つを赤字にしてある．.

また，各カテゴリ内の記事 x 編についてのコメントを混ぜ，その中から 640 コメントをランダムに取りだし，その多様性を解析する，ということを，x=2 から 10 の場合について，それぞれ 10 回ずつ行い，その平均を取った．図 3 にその結果を示す．横軸は記事数であり，分かりやすいよう，縦軸の多様性の値は，記事数 x の多様性の値を記事数 10 の場合の多様性の値で割った，正規化した値を表示している．つまり，記事数 10 のときに比べてどれくらい大きいか小さいか，が縦軸になっている．カテゴリの多様性が大きいほど，複数の記事を用いた場合に値が大きく，あるいは小さくなるため，グラフ上で曲線が急勾配であるものほど，カテゴリの多様性が高いこととなる．

カテゴリについては，ヤフーニュースのカテゴリである「国内」「国際」「経済」「エンタメ」「スポーツ」「IT」「科学」「ライフ」「地域」の分類を参考とした．ただ，1 つの記事が複数のカテゴリに所属することも多く，カテゴリが広すぎると考えられるため，より細分化し，より明確なカテゴリを定義することとした．具体的には以下のように定義した．

・タイトルに「コロナ」を含む記事
・コロナの感染者数に関する記事
・企業に関する記事
・皇室に関係する記事
・事故や火災に関する記事
・事件，犯罪，訴訟に関する記事
・外交や国際的な事柄に関する記事
・国内の政治に関する記事
・スポーツに関する記事
・天気，天候，災害に関する記事

他にも多くのカテゴリが考えられるが，特徴的なもの，ある程度数がそろい，コメントが多く付いているものを選んでいる．

５．考察

まず表 1，各カテゴリのコメントの平均多様性を見てみる．どの指標でも，カテゴリの順位はだいたい同じではあるが，正確に同じではない．興味深いところは最も多様性が高いものは評価が揺らいでいるが，多様性が低いものは企業，皇室関係，事故火事，とどの尺度でもほぼ同じものが下位となっていることである．下位のカテゴリについて，それらの記事にたいするコメントがなぜ多様でなくなるのかはわかりにくいが，他のカテゴリに比べ，多角的な視点から意見や感想を述べることが相対的に難しいのではないか，と推測される．

次に，図 3，同一カテゴリの複数の記事のコメントの多様性を見る．まず，どのカテゴリ・尺度においても，多様性の数値はほぼ単調に減少・増加していることから，この方法はある程度安定的であると考えられる．ただし，スポーツは単調ではなく乱れており，なんらかの要因があると考えられる．この 3 つの尺度では，べき乗則当てはめが最も安定的であると見られるが，一方で天候災害では値が不安定であり，当てはめを行うプログラムになんらかの弱点があると考えられる．

複数の記事での多様性が高い，つまり記事ごとのコメントの違いが大きいものは，企業，天候災害，国内政治，スポーツである．これらは，記事事に内容が大きく異なり，読者の多様な反応を引き出していると考えられる．一方で，多様性が低いものは，皇室関連と感染者数である．このカテゴリでは，どのような記事に対しても，同じような意見，感想を述べているのであろうことがうかがえる．企業の記事は，単独の記事ではコメントが多様ではないが，カテゴリとしては多様となっており，興味深い．

「多くのコメントが同じようなことを言っている」ということは，人間にとっては判別しやすいのであるが，このように数理的，あるいは機械的に評価することは簡単ではない．多様性という，意味とは一見関連が低いものに着目することで，このような観察が得られることが，この方法の興味深いところであると考えている．

６．むすび

本稿では，単語多様性の面から，ニュース記事のコメントを分析する手法を使い，ヤフーコメン

トを分析したときにどのようなものが見えるのかを紹介した．ニュース記事に対するコメントは，社会や個人，マイノリティの意見や感情を抽出する意味で新しい可能性を持っており，今後の広い場面での利用が期待される．本稿がそのような研究の動機付けになることを期待する．

謝辞

本研究は，科学研究費補助金基盤研究 A，課題番号 19H01133，学術変革研究，課題番号 20H05962 の補助を受けた．

参考文献

[1] Clauset, A., Shalizi, C. R., & Newman, M. E.. Power-law distributions in empirical data. SIAM review, 51(4), 661-703, 2009.
[2] Morales, P. R., Lamarche-Perrin, R., Fournier-S'Niehotta, R., Poulain, R., Tabourier, L., & Tarissan, F.. Measuring diversity in heterogeneous information networks. Theoretical Computer Science, 859, 80-115, 2021
[3] Pavlov, Anton, and Boris V. Dobrov. "Detecting Content Spam on the Web through Text Diversity Analysis." SYRCoDIS. 2011.
[4] Zhang, Z. K., & Zhou, T.. Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes. Scientific reports, 3(1), 1-7, 2013

図３：べき乗則当てはめ，エントロピー，ジニ係数による各記事カテゴリのコメントの多様性の評価
Figure 3: Evaluation of the diversity on comments to the news in each category, by power-law fitting, entropy, and Gini coefficient.