ニュース記事を対象としたトピック解析を用いたバイアス発見方式

柳瀬 愛里† 木村 侑斗† 萩本 新平† 中田 亮佑† 中村 洋太† 本田 くれあ†
仲程 凜太朗† 中西 崇文†

†武蔵野大学データサイエンス学科 〒135-8181 東京都江東区有明 3-3-3
E-mail: †{s1922073, s1922060, s1922027, s1922078, s1922023, s1922031, s1922069}@stu.musashino-u.ac.jp, takafumi.nakanishi@ds.musashino-u.ac.jp

あらまし 本稿では，ニュース記事を対象としたトピック解析を用いたバイアス発見方式について示す．本方式は，ある事柄について示すニュース記事群を対象として，トピック抽出を行い，各トピックを軸として各ニュース記事がどのように偏っているかを示すことを可能とする．本方式は，抽出したトピックについて類似度計量を行い，類似度が遠いトピック同士を極端な事象を表す観点として定義する．本方式は，各ニュース記事と導出された観点との関係を提示することにより，ニュース記事に含まれる偏り(バイアス)を発見することが可能となる．本方式により，ニュース記事で表される事象の様々な観点を俯瞰的に把握することが可能となる．

キーワード トピック抽出，バイアス発見，ニュース記事，Word Mover’s Distance

1. はじめに

ロイター研究所が 2020 年 5 月に世界 40 か国を対象に行なったニュースに関する調査[1]によると，日本のマスコミのうち，発行部数の多い新聞と公共放送 NHK を含む全国 5 つの放送局を中心とした情報発信について，過去 7 年間で急激に読者や視聴者の数が減少していることが分かっている．それに対し，Yahoo!ニュースのような Web サイト上での情報発信は急激に読者や視聴者の数を増やしている．

一方，総務省が出した令和元年度の情報通信白書[2]によると，メディアが発達した現代では，テレビとインターネットを並行利用する行動様式が生まれただけでなく，対面メディア(人と対面でコミュニケーションを取る中，情報を得られる状態のことを示す)・マスメディア・ソーシャルメディアが重層的・複合的に併存している状況にあることが示された．遠藤ら[3]の研究では，多くの者が日常的に，これらのメディアを重層的に利用していることから，マスメディアの情報がネットを通して伝わったり，ネットの情報がマスメディアへと伝わったりするため，これらを分けて論じることができない状況にあると指摘し，こうしたソーシャルメディアと残存メディアが重層的に相互利用しながら世間を形成する現代のメディア環境を「間メディア社会」と定義している．「間メディア社会」では，これまで大きな社会的関心ごとにならなかった出来事が，メディア間の相互作用が緊密化することにより，大きく取り上げられるようになる．その影響として挙げられる例が，ネット炎上や世論の二極化である．メディアとしてのインターネットは，誰もが情報の発信者となり，利用者となる「双方向化」を実現し，あらゆる人が様々な情報や知識を世界に共有することを可能にしてきた．つまり、ある事象をいろいろな人たちが，各々の視点で捉え直し，新たな情報や知識，あるいは考えや意見などを共有している．一方で、インターネットの特徴である情報源の最適化により引き起こされる，「エコチェンバー」や「フィルターバブル」などの現象，そして米国の法学者サンスティーン[4]が指摘した，同じ思考や主義を持つ者同士をつなげやすいというインターネットの特徴から生み出される現象「サイバーカスケード」によって，ネット炎上や世論の二極化が起きていると考えることができる．

これらの背景から，ユーザが今読んでいるニュース記事が，他のニュース記事と比較してどのように位置づけされるのかを俯瞰できるようなメディアが実現されれば，ユーザはニュース記事を全体の世論の流れや，そのニュースに包含する情報の偏り(本論文では以下バイアスと定義する)を理解しながらニュースを理解することが可能になると考えられる．このようなメディアが実現することにより，「エコチェンバー」や「フィルターバブル」などの現象，如いては「フェイクニュース」などの発見につながると考える．

本稿では，ニュース記事を対象としたトピック解析を用いたバイアス発見方式について示す．本方式は，上記のようにバイアスを認識できていないことにより起こる現象を解消する手段として，ある話題に関するニュース全体像を俯瞰的に可視化し，ユーザが着目しているニュースはどのような立ち位置であるかを提示することを可能とする．本方式は，ある事柄について示すニュース記事群を対象として，トピック抽出を行い，各トピックを観点として各ニュース記事がどのよ

うに偏っているかを示すことを可能とする.本方式は，抽出したトピックについて類似度計量を行い，類似度が遠いトピック同士を極端な事象を表す観点として定義する．本方式は，各ニュース記事と導出された観点との関係を提示することにより，ニュース記事に含まれるバイアスを発見することが可能となる．本方式により，ニュース記事で表される事象の様々な観点を俯瞰的に把握することが可能となる.

本方式を実現することにより，ユーザが着目しているニュースを中心としてその様々な観点を俯瞰する全体像を可視化することが可能となり，ユーザが自身で情報の偏りを認識できるようになり,「間メディア社会」特有のネット炎上・世論の二極化などの問題を解消することに繋がっていくと考えられる.

本論文の目的は，ユーザが支持しているニュースの立ち位置を他のニュースとの関係性を元に可視化し，ニュース自体にバイアスがあるか認識できるようにすることである．そこでまず 2 章では，関連研究を列挙すると共に，本研究の立ち位置を明らかにする．3 章では，本方式で用いる Word Mover's Distance(WMD)と潜在的ディリクレ配分法(LDA)に関して述べ, 4 章では本方式のシステム構成に関して述べる．5 章では，ある事象を対象としたニュースを利用した実験から示された本方式の評価を提示し，6 章で本稿をまとめる.

2. 関連研究

本章では，本方式に関連する研究としてニュースのトピック差異推定やニュース全体像の可視化に関する研究について紹介する．ニュースへのコメント可視化に関する研究について示す.

2.1. ニュースへのコメント可視化に関する研究

山口ら[5]は，他者の意見や考えの視点を網羅的かつ俯瞰的に知るための手法として，オンラインニュースサイト上のニュースに寄せられたコメントから類似意見を抽出することで，議論の全体像を議論ツリーという構造で表示する手法を提案している.この研究では，それぞれのコメント文に対し単語の出現頻度ベクトルを作成し，任意の文章同士の類似度を比較した際に，一定の閾値を超える類似度が出力された場合にリンクの付与することで，議論ツリーと呼ぶ木構造を構築していく．また，類似度推定に用いる閾値と利用する単語の出現ベクトルのセットを複数導入することにより，コメント群全体から複数のトピックを設定し，Web 上に投稿されたユーザのコメントをいくつかのトピックに分類することによって，どのような観点の意見が存在するのかを俯瞰するために必要なデータを収集できることが確認できている．この研究では，あるニュースに対するユーザのコメントを対象とし，類似意見を集約しながら可視化をしていくことで，他者のコメントがどのような意見や考えの元に述べられたものか可視化し提示することができる．本研究では，ニュース記事そのものと類似する内容を要素として持つニュース記事群を対象とした分析を行い，ニュース記事自体がどのようなバイアスを持っているのかを可視化している.

2.2. Twitter でのユーザ反応を元にしたニュースを示す特徴語抽出と関連ニュースの提示に関する研究

池田ら[6]は，Twitter で投稿されたニュースに対する反応としてリプライおよび引用リツイートを利用し，ニュース自体の特徴語とニュースに対する反応の際に用いられた特徴語を抽出して，ニュースに対する反応の特徴や他のニュースとの関連性をわかりやすく可視化するインターフェイスを実現した．これを用いることで，ユーザは閲覧しているニュースに対する反応語を知り，ニュースにおいて何が注目されているのかを知ることができ，さらに閲覧しているニュースと関連するニュースを知ることによって，それぞれのニュースの位置づけを理解することができると考えられている．この研究では関連するニュースを特徴語から推定し，ユーザに提示するという点で本研究と似ている．一方，この研究では関係性の可視化をニュース単位で行っているのに対し，本研究では関係性の可視化をニュースの記事群単位で行う.

2.3. ニュース記事が持つセンチメント解析による差異の提示に関する研究

濱砂ら[7]は，ニュース記事を対象に書き手のセンチメントを抽出し，同一話題に関して抽出したセンチメントをサイトごとに分析することで，各ニュースサイトの観点の差異を提示できるセンチメントマップのプロトタイプを作成し，実験で各ニュースサイトのセンチメントの相違をグラフにより提示することを確認した．本研究では，感情解析は行わずニュース記事に含まれる語を基準にニュース記事群同士の差異・関係性を可視化している.

2.4. ニュースへのコメント可視化に関する研究

西川ら[8]は，ニュースを理解する上で必要となる予備知識や専門知識を持たない人を対象に，ある事象に関するニュース記事をトピックごとに階層構造で表示し，トピックの関係性をキーワードの相違によって明確にすることで，テーマの理解を促すことを目的とした可視化インターフェースの構築を行った．本研究では，ニュース記事群内でどのような語が使われているかをユーザが確認でき，また，記事群同士の類似度計算を行うことにより，ニュースの観点の関係性を俯瞰

することが可能となる。

3. 本研究で用いる既存研究

3.1. Word Mover's Distance (WMD)

WMD とは，文章間の距離を計算するための手法である．Word2Vec で得られる単語の分散表現(単語ベクトル)から文書の分散表現を作成し，ある文章 A がある文章 B に変換する際の対応付けの変換コストが最も低い場合の変換コストの和を文書間距離と考える．
WMD は以下の様に表せる．

$$\min_{\mathbf{T}\geq 0} \sum_{i,j=0}^{n} \mathbf{T}_{i,j}\ c(i,j)$$

$$\textbf{subject to}: \min_{\mathbf{T}\geq 0} \sum_{i,j=0}^{n} \mathbf{T}_{i,j} = d_i \quad \forall \in \{1,\dots,n\}$$

$$\sum_{i,j=0}^{n} \mathbf{T}_{i,j} = d_i \quad \forall \in \{1,\dots,n\}$$

語i,jの分散表現ベクトルをそれぞれ$\mathbf{x}_i, \mathbf{x}_j$とし，それらの距離は$c(i,j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2$である．変換行列$\mathbf{T}$は成分$T_{ij}$により構成され，語の出現頻度を示す．全体のコストはT の各成分に対応した語同士の距離をかけたものの総和となり，コストが最小となる T を定め，そのときのコストが距離となる
本方式では，データベース内でカテゴリ分けされたニュース記事群の観点とユーザが着目するニュース間の距離（類似度計量）を導くために WMD を用いている。

3.2. 潜在的ディリクレ配分法 (LDA)

LDA とは，文書群から話題を取り出す手法，トピックモデルの一種である．LDA において，文書は潜在的なトピックの組み合わせとして表現され，各トピックは単語の分布によって特徴づけられる．LDA は，文書内のトピック分布にしたがってトピックを選択し，トピック内の単語分布に従ってキーワードを選択するプロセスでドキュメントが生成されることを前提としている．LDA に用いられる変数は，Nは文書内の単語数，αは分布前のトピック K 番目の次元を示すハイパーパラメータ，βは単語分布のパラメータ，θはトピック分布，zはトピック集合，wは単語集合である．またα，βを指定すると，トピックの混合分布θ，トピック集合z、単語集合wは，以下のように表される

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

ここで，θとβは潜在的なパラメータであり，文書が観測された際にこのパラメータの推定を行うことによって文書のトピック推定が可能である．

本方式では，ニュース記事データベース内でカテゴリ分けされたニュース記事群を対象にトピック解析を行う処理に LDA を用いて，ある事柄に関するニュース記事群がどのような観点から記述されているかを導出する際に利用する．

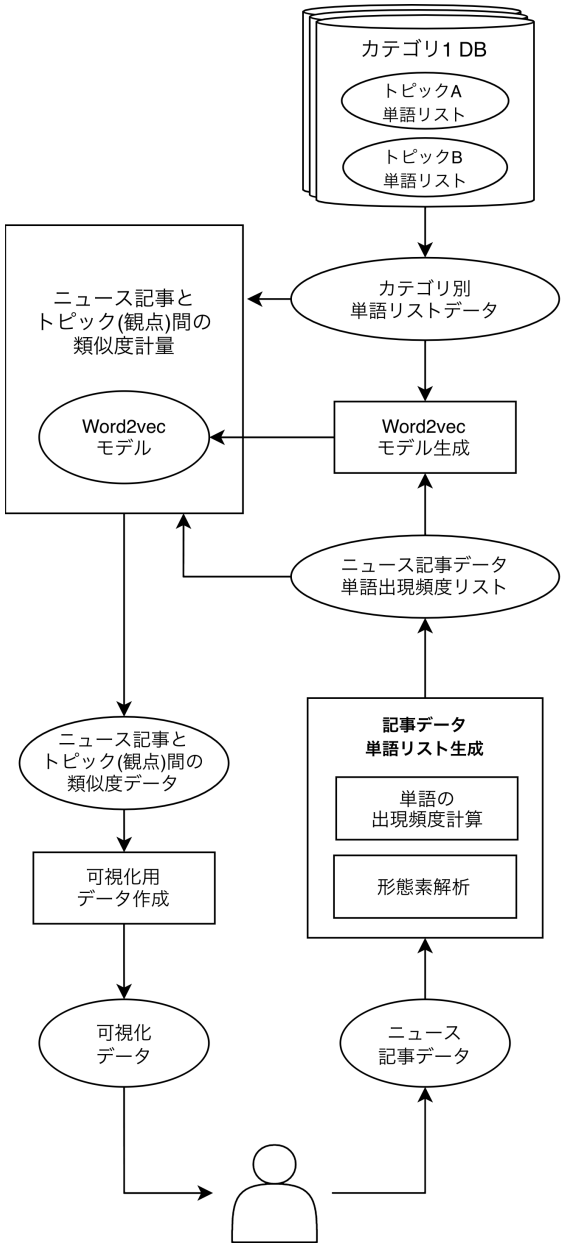

図 1 提案システムの全体像

4. トピック解析を用いたニュース記事を対象としたバイアス発見方式

本節では，ニュース記事を対象としたトピック解析を用いたバイアス発見方式について述べる．

4.1. 提案手法の概要

本手法が提案するシステムの全体像を図 1 に示す．本方式では大きく分け，ニュース記事データベース構

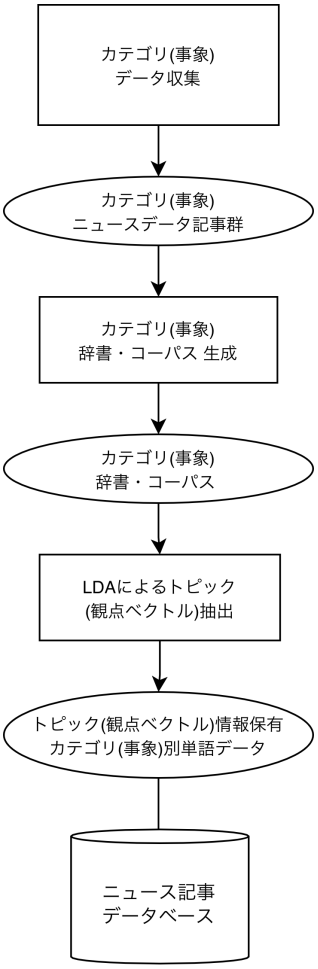

図 2 ニュース記事データベース構築

築, 記事データ単語リスト生成, Word2vec モデル生成, ニュース記事とトピック(観点)間の類似度計算, 可視化用データ作成の 5 つの機能によって構成される.

4.2. ニュース記事データベース構築

ニュース記事データベースの構築方法を表したものを図 2 に示す.

ニュース記事データベースは, カテゴリデータ収集部, カテゴリ辞書・コーパス生成部, LDA によるトピック(観点ベクトル)抽出部の 2 つから構成されている.

4.2.1. カテゴリ(事象)データ収集部

ここでは, NEWSAPI[9]で利用されている日本のニュースサイトの中から, 特に有用であると判断されたサイトを対象としてニュース記事のスクレイピングを行い, カテゴリ別にニュースデータを収集した. カテゴリに対していくつかのキーワードを設定し, キーワードがニュース記事のタイトルに含まれる場合, 該当するカテゴリに属すると判定し, 収集を行った.

4.2.2. カテゴリ(事象)辞書・コーパス生成部

収集したカテゴリニュースデータ記事群の本文を対象に形態素解析とストップワード除去を行い, 本文中で使われている単語を抽出した後, これらの単語を用いて辞書とコーパスの生成を行う. この 2 つのデータを用いてトピック抽出を行う.

4.2.3. LDA によるトピック(観点ベクトル)抽出部

ここでは, カテゴリ別に収集したニュース記事が, それぞれどのような観点から述べられているかを明らかにするため, カテゴリに含まれるニュース記事群に対して LDA を用いたトピック抽出を行う. カテゴリから抽出されたトピックは, それ自体がカテゴリで記述されている事象を説明・記述する基準を示す重要な観点であり, ユーザに対し, この観点と共に着目したニュース記事の位置づけを示すことで, バイアスの検出を行うことが可能となる.

4.3. 記事データ単語リスト生成

ユーザが着目するニュース記事データに対して, 形態素解析と単語の出現頻度計算を行い, 記事に含まれる単語出現頻度をリストにまとめる. ここで得られたニュース記事データ単語出現頻度リストを用いて, トピックと記事間の類似度計量を行う.

4.4. Word2vec モデル生成

形態素解析とストップワード除去を行ったカテゴリ別単語リストデータとユーザが入力したニュース記事データから生成されるニュース記事データ単語リストの 2 つを利用して Word2vec のモデルを生成する. ここで得られたモデルは, 記事やトピック同士の類似度計量に用いることが可能である.

4.5. ニュース記事とトピック(観点)間の類似度計算

ニュース記事とトピック(観点)間の類似度計算では, ユーザがシステムに入力したユーザが着目しているニュースとカテゴリ別単語リストデータを Word2vec モデルで分散表現ベクトルへと変換し, ユーザが入力した着目しているニュース記事が属するカテゴリを WMD による類似度計量から選定する. その後, カテゴリ内の観点との類似度計量も行い, ユーザの着目しているニュース記事がどの観点と近しいかを定量的に判断することを可能とする. これらの結果を 4.6 節で示す可視化用データ作成機能に受け渡すことにより, ユーザが入力したニュースの位置づけをそのカテゴリの観点を用いて可視化することが可能となる.

4.6. 可視化用データ作成

ユーザが入力したニュース記事と導出された観点との関係を提示するために, 4.4 節で導出した結果をネットワークで可視化する. この可視化によって, ユーザが着目しているニュースとカテゴリ内に存在する同様の事象について述べられている記事を対象とし, トピック抽出により求められた観点を基軸として, そ

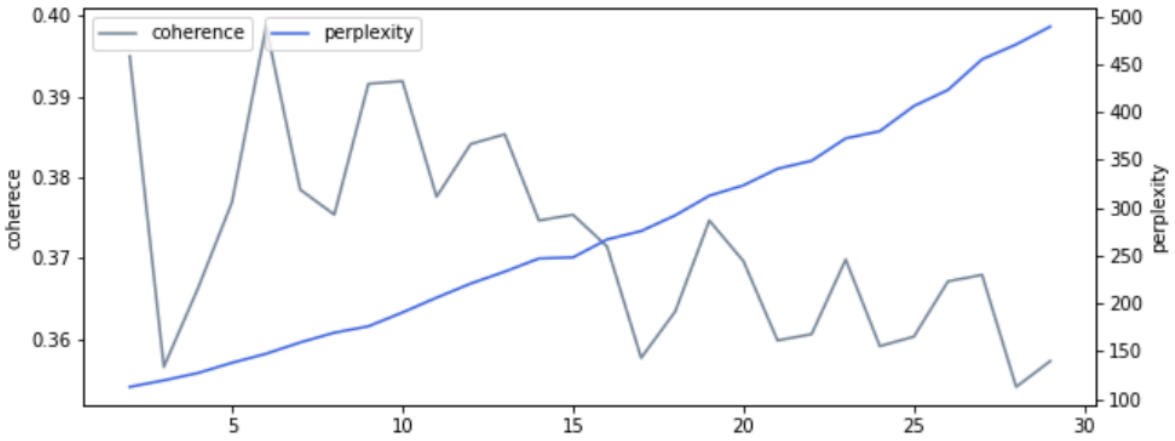

図 3 カテゴリ数の変化に伴う Perplexity 値と Coherence 値の推移

の記事の位置づけを俯瞰的に確認することを可能にする．

5. 評価実験

本節では，提案方式を用いて，ユーザが着目しているニュースが属しているカテゴリを判別した後，そのカテゴリ内のトピック(観点)を基軸として可視化を行うことで，カテゴリ内においてそのニュースがどのような観点から述べられているかを確認するために．こ実験システムを構築し実験を行った．

5.1. 実験方法

本実験では，カテゴリ内のトピック数推定が上手くいくか，「日本学術会議」に関するニュース記事 2 つ[10][11]に対し，正確なカテゴリ選定とそのカテゴリ内の観点を基軸とした可視化を行うことができるかの 2 つを検証した．カテゴリ(事象)データ収集部を用いて 72 件の「日本学術会議」に関するニュース，40 件の「地震」に関するニュース 139 件の「新型コロナウイルス」に関するニュース， 293 件の「大統領選挙」に関するニュース，合計 544 件のニュース記事を取得した．各カテゴリ内の記事本文に対して LDA を用いてトピック抽出を行うことによりニュースカテゴリの観点を抽出した．この時，トピック数は事前実験により求められた Perplexity と Coherence の値を参考に推定した．ユーザが着目しているニュースとして，ニュース記事[10][11]を入力として与え，上記で設定した 4 つのニュースカテゴリのうち，「日本学術会議」のカテゴリとして，この 2 つのニュース記事が正しく判別されるかを確認した．さらにニュース記事[10][11]は，日本学術会議という事象が持っている観点の内，どれとの距離が近いかを類似度計量により導出し，観点を基軸として，その記事の位置づけを可視化した画像の出力を行った．

概要 (概要説明)	会議，学術，任命，会員，問題，推薦，6 人，日本学術会議，政府，首相
拒否 (任命拒否)	任命，会議，学術，会員，推薦，日本学術会議，政府，首相，6 人，拒否
科学 (日本学術会議と科学)	任命，会員，会議，学術，政府，推薦，日本学術会議，ない，科学，6 人
介入 (人事介入による問題)	任命，会議，会員，学術，問題，6 人，人事，説明，政府，推薦
政府 (政府による説明)	任命，学術，会議，会員，日本学術会議，推薦，政府，首相，問題，説明
首相 (首相による説明)	任命，会議，推薦，会員，首相，学術，説明，日本学術会議，6 人，問題

表 1 LDA を用いて日本学術会議の記事群から抽出された各観点を示す上位 10 単語

5.2. 実験結果

トピック数と LDA の精度検証に用いる Perplexity と Coherence という指標の関係を示した結果の一例を図 3 に示す．ここにおいて，Perplexity はモデルの予測精度を測る指標で、トピックモデルを設定した環境下において，どれほどの精度で候補となる単語群から，その場に適した単語を選択できるかを示している．Coherence は単語間の類似度を基準に「人間にとってトピックモデルがどれほどわかりやすいか」という曖昧で定義が難しい事柄を示している．今回は収集した日本学術会議に関するニュース記事群に対して、トピック数を 3~15 つに変化させた際にどのように 2 つの値が推移したかを図で示した．この図から，Perplexity はトピック数が増えるに従って値が大きくなり，Coherence はトピック数が 3 つ，もしくは 6 つの時に高い値を取ることが読み取れる．Perplexity の値が小

さく，Coherence の値が大きい程良いモデルとされているため，この図 3 の結果から，日本学術会議のトピック数は Coherence の値が最も高い 6 つに設定した．その他のカテゴリにも同様の方式を用いることで．トピック数の選定を行った結果，地震と新型コロナウイルスの説明に必要なトピック数は 6 つ，大統領選挙の説明に必要なトピック数は 3 つに設定した．続いて，ニュース記事[10][11]と日本学術会議に含まれる各トピック(観点)との類似度計量の結果を表 2,3 に示す．Word Mover's Distance では，出力された値が小さいほど類似度が高いとされているため，ユーザが着目したニュース記事[10]と類似度が最も高い観点は任命拒否に関するものであるのに対し，類似度が低い観点は日本学術会議と科学や人事介入に関するものであることが読み取れるのに対し，ユーザが着目したニュース記事[11]と日本学術会議に関する観点 6 つ間はほとんど同じ類似度であることが読み取れる．この結果をネットワーク図で示したものを図 4,5 で示す．図中に濃い灰色で示された図形は「日本学術会議」の観点の位置づけを示すものである．

5.3. 考察

本実験結果から，Perplexity と Coherence という指標は，ニュースが持つ観点を抽出する際に行うトピック数の推定に有用であるため，これを活用することによって，現在手動で設定しているトピック数を自動で設定することが可能になると考えられる．また，同じ「日本学術会議」に関するニュース記事であっても，任命拒否・概要説明以外のトピック 4 つに注目して日本学術会議に関する説明しているニュース記事[10]や，どの観点にも依ることなく記述されているニュース記事[11]があることがわかった．このことから，本方式を用いることにより，ニュースで重視されている観点と着目したニュースの関係性を可視化し，着目したニュースがバイアスを持っているか否かを定量的に判別することが可能になると示された．

6 おわりに

本稿では，ニュース記事を対象としたトピック解析を用いたバイアス発見方式について述べた．本方式は，ある事柄について記述されたニュース記事群を対象として，トピック抽出を行い，入力されたニュース記事が各トピックを観点として，どのような偏りを持っているか示すことを可能とする．これにより，各ニュース記事と導出された観点との関係を提示することにより，ニュース記事で表される事象の様々な観点を俯瞰的に把握することが可能となり，ニュース記事に含まれる偏り(バイアス)をユーザに認識させることも可能となる．

Topic	科学	政府	拒否	概要	首相	介入
WMD	0.0127	0.0105	0.0170	0.0160	0.0120	0.0145

表 2　ユーザが着目している記事[10]とトピックの WMD 距離

Topic	科学	政府	拒否	概要	首相	介入
WMD	0.0191	0.0199	0.0215	0.0186	0.0182	0.0221

表 3　ユーザが着目している記事[11]とトピックの WMD 距離

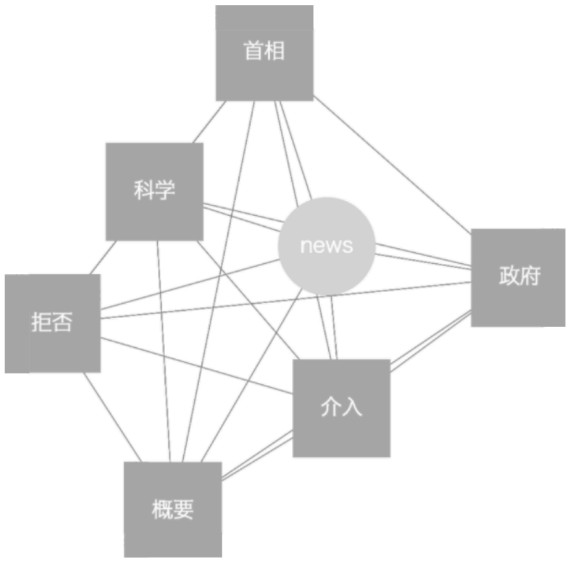

図 4 ニュース記事[10]のネットワーク図

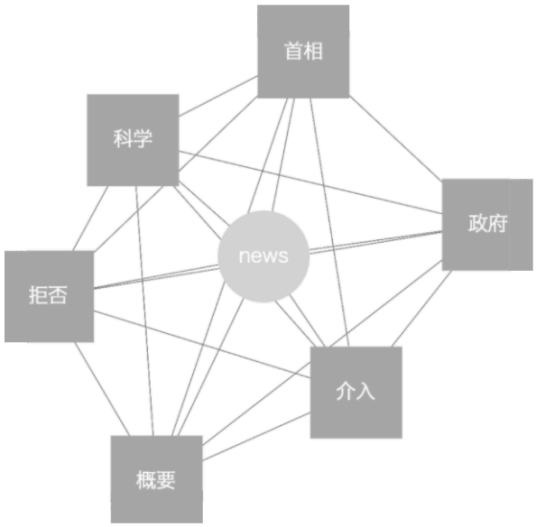

図 5 ニュース記事[11]のネットワーク図

また，本稿では，本方式を実現する実験システムを構築し，有効性を示した．その結果から「日本学術会議」に関する問題の記事について，関連記事との位置づけを俯瞰的に示すことが可能であると示された．

今後の展開としては，Perplexity と Coherence を用いたトピック数の自動設定機能の実装，様々な分野のニュース記事における本方式の適用と有効性の検証，本方式を用いた新たなニュースアプリとしての開発などの展開が挙げられる．

参　考　文　献

[1] ロイター研究所，”Country and Market Data”, https://www.digitalnewsreport.org/survey/2020/country-and-market-data-2020/.

[2] 総務省，平成 30 年度版情報通信白書 ,https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h30/html/nd252550.html.

[3] 遠藤薫，”間メディア民主主義と<世論>”，社会情報学第 5 巻 1 号，2016.

[4] キャス・サンスティーン，”インターネットは民主主義の敵か”，毎日新聞出版，2003.

[5] 山口雄也，伏見卓恭，”オンラインニュースサイトにおける類似意見の抽出”，研究報告数理モデル化と問題解決(MPS), 29, pp.1-2, 2019.

[6] 池田将，牛尼剛聡，” Twitter の反応を用いたニュース全体像の理解支援のための可視化手法”，研究報告データベースシステム(DBS)，5，pp.1-5, 2019.

[7] 濱砂佳貴，河合由起子，熊本忠彦，田中克己，”センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案”，第 19 回データ工学ワークショップ(DEWS2008)論文集，B6-4，2008.

[8] 西川奈都月，盛山将広，内藤峻，松下光範，”初学者を対象としたニュース記事中のトピックの関係性に基づく可視化インタフェースの提案”, SIGAM，15(10), pp.62-67. 2017.

[9] NEWSAPI, https://newsapi.org/ .

[10] DIGITAL『首相の任命拒否「想定にない」　学術会議めぐる政府文章』朝日新聞，2020 年 10 月 27 日(最終閲覧日:2021 年 2 月 10 日)，https://digital.asahi.com/articles/ASNBW6THJNBWUTFK01Q.html.

[11] TOKYO WEB『日本型アカデミーとしての「学術会議」に誇りを　宇山智彦・北海道大教授』東京新聞，2020 年 12 月 22 日(最終閲覧日:2021 年 2 月 10 日)，https://www.tokyo-np.co.jp/article/75942.