ユーザの興味を反映したニュースサイトの多観点スタンス分析

立浪 紀彦[1] 吉岡真治[1] 神門典子[2] James Allan[3] 伊藤正彦[4]

[1] 北海道大学 [2] 国立情報学研究所 [3]University of Massachusetts Amherst
[4] 北海道情報大学

s01240684b@eis.hokudai.ac.jp, yoshioka@ist.hokudai.ac.jp,
kando@nii.ac.jp
allan@cs.umass.edu, imash@do-johodai.ac.jp

1 はじめに

近年、ニュースは、新聞やテレビ放送だけでなく、多様なニュースサイトからも発信される様になってきている。また、これらのニュースを読む手段としては、Google News[1]のようなニュースアグリゲーションサイトを利用するユーザが増えている。これはユーザにとって便利である反面、従来のように記事の配信元が普段どのような報道スタンス（右寄り、左寄り）から論じているかを認識して、記事を解釈することを難しくしている。また、多くのユーザーやニュースサイトは、すべてのトピックについて、一貫して同じ様なスタンスを持つわけではないため、ユーザの興味を反映させた上で、ニュースサイトの分析を行う必要がある。

本研究では、個別の興味を持ったユーザに対して、多様なニュースサイトの報道スタンスについての複数の観点に基づく情報を提供することで、ニュースの読解を助けることを目的としている。本論文ではまず、我々がこれまで行ってきた、話題に対する特定の賛否に注目したニュースサイトの分析 [1] について紹介すると共に、多次元尺度構成法（MDS）[2] による多次元データの可視化結果をユーザの興味によりインタラクティブに操作可能とする SIRIUS[3] を紹介する。さらに、この手法を用いて 2016 年のアメリカ大統領選の記事データを対象に実験を行い、その結果について考察を行う。

2 ニュースサイトの比較分析

我々はこれまで、特定のトピックに対するニュースサイトの記事の賛否の割合に注目し、アメリカの 2016 年大統領選のトランプ氏とクリントン氏のような異なるトピックに対する賛否の割合を比較することで、ニュースサイトのスタンスを分析する方法を提案した [1]。しかし、賛否の割合による表現はニュースサイトの大まかな報道スタンスを表現するためには十分であるが、ある候補者に対するニュースサイトのスタンスが似たようなサイト同士でも政策によっては賛否が異なる場合があるため、それぞれのニュースサイトの特徴を理解するには不十分であった。

3 SIRIUS による多次元データの可視化

多次元データの可視化の代表的な手法として、多次元尺度構成法 (MDS:Multi Dimensional Scaling)[2] が知られている。この手法では、データ間の距離を用いた距離行列に関して、変換前後で各データ間の距離の差が最も小さくなるように変換する。しかし、データを分析するにあたって異なる興味を持ったユーザが、同じ定義の距離を用いた場合にそれぞれの目的に応じた結果が得られないことがあるため、興味に応じた適切な距離を定義し、新しい可視化結果を得る方法が提案されている。SIRIUS[3] は、この距離の定義をユーザが類似している（距離が近い）、または、類似していない（距離が遠い）と考えるペアの情報を与えることにより、ユーザーの興味を反映した重みを各次元に付与した距離を計算し、その距離関数を用いた重み付き MDS による可視化を行う。また、データの転置行列に対して、同様に MDS による次元圧縮を行うことで、属性間の類似性を示すプロットが可能となり、ユーザの曖昧な類似性に関する基準を理解するためのフィードバックを行うことができる。ユーザはこの結果を見

[1] https://news.google.com

ながら、さらなる操作を続けていくことで、より興味を反映した可視化結果を得ることができる。

4 提案手法

4.1 ユーザの興味を反映したニュースサイトの可視化

我々の従来のシステム [1] では、主に、比較対象となる話題（大統領選挙の 2 人の候補）に対する賛否の割合の違いに基づいた可視化を行っていたが、ニュースサイトを詳細に分析すると、全ての話題において賛否のスタンスが一貫しているようなサイトのみではなく、話題（外交・経済など）ごとに、賛否の割合が違うようなサイトも多く存在した。これらの話題ごとの違いを分析するためには、話題を考慮したスタンスの違いを議論することが必要となる。また、分析するユーザの興味（例えば、外交にのみ興味がある）を考慮したニュースサイトの比較のための可視化を行うことで、より詳細な分析が可能となる。

このユーザの興味を扱うために、ニュースサイトに対して、全ての話題に対する賛否ではなく、トピックを考慮した賛否の情報を付与するとともに、SIRIUS を用いることで、ユーザの興味を反映した距離を用いた分析を可能とする枠組を提案する。

4.2 News Genre Polarity-based Stance Matrix

前節で述べた可視化を行うためには、ニュースサイトの記事全体の賛否の情報ではなく、トピックごとに分類した賛否の情報を利用する。本研究では、汎用的なトピックとして、多くのニュースサイトにおいて用いられているジャンルの情報をトピックとして利用する。このジャンルごとの賛否の情報を以下のような News Genre Polarity-based Stance Matrix（NPSM）により表現する。

$$NPSM = \begin{pmatrix} pos_1 & pos_2 & \cdots & pos_m \\ neu_1 & neu_2 & \cdots & neu_m \\ neg_1 & neg_2 & \cdots & neg_m \end{pmatrix}$$

m はニュースのジャンルの数、pos_i、neu_i、neg_i はそれぞれジャンル i の肯定的、中立的、否定的な記事の割合を表す。

4.3 記事とジャンルのマッピング

記事をジャンル別に集計するにあたり、それぞれの記事に対して、ジャンルの情報を付与する必要がある。多くのサイトではサイトごとに独自のカテゴリ構造を持っており、それに沿って記事を分類・整理している。本稿ではいくつかのニュースサイトを参考に、7 つのジャンル (政治 (pol)・国内 (dom)・国際 (int)・経済 (eco)・科学 (sci)・ライフスタイル (lif)・エンターテインメント／スポーツ (e/s)) を設定し、一つの記事に対し、これらのジャンルのうちどれかひとつを割り当てることとした。

これらのニュースサイトでは、記事の URL は単なる id ではなく、カテゴリに相当する名前をディレクトリや URL の一部に含んだり、記事のタイトルに相当するような文字列を URL に含むものがほとんどであった。本文を収集するコストが非常に高いため、今回は、この URL のみを用いたジャンルの分類を行う。

具体的には、URL 中のディレクトリ名にカテゴリの情報を含むものは手動で設定したキーワードとのパターンマッチングによりジャンル分類を行い、さらにジャンルと URL 文字列に関する教師データを作成し、それ以外のものについては、URL からサイト名を取り除いたものを用いた BERT[4] によるジャンル分類システムを構築する。

4.4 SIRIUS によるプロット

上記のジャンル分類を行うことで、既存のシステムで利用していた賛否の割合のデータを NPSM に変換することができる。この行列を一行のベクトルに並べなおすことにより、サイトごとの SIRIUS の入力データとする。その結果、ニュースサイトの類似性を表現する可視化と、類似性を検討する際の賛否の情報の関係性を同時に可視化して、ニュースのインタラクティブな分析を行う。

5 実験

5.1 GDELT からのメタデータの収集

アメリカのニュースサイトを中心に、報道の偏りと質に関する情報を提供している Media Bias Chart 5.1[2] に掲載されていたサイトに関して、[1] と同様に、

[2] https://www.adfontesmedia.com

GDELT[3] から記事データを収集した。この記事データのうち内容にドナルド・トランプ氏あるいはヒラリー・クリントン氏は関係があるか否か、記事の配信された日付、URL、GDELT が本文に使われている単語から計算した Polarity の値を収集した。データは 77 サイトについて、2016 年 9 月 1 日～2016 年 11 月 31 日の期間内ものである。

5.2 BERT によるジャンル予測

記事の URL 中のディレクトリ名にカテゴリの情報がないものについて、ジャンル分類を行った。ここでの学習時の記事はサンプル数を確保するためにトランプ氏・クリントン氏に関係するとされていない一般的な記事も教師データの一部として用いた。

事前学習のモデルには 2018 年に Google によって公開された 12 層、768 次元の隠れ層、ヘッド数 12 の Uncased モデル[4]をファインチューニングして利用した。ファインチューニングの際には教師データを 6:2:2 の割合で訓練セット、開発セット、テストセットとし、URL に含まれる内容を表す文から各ジャンルに予測される確率を算出し、その確率が最も高かったものをその記事のジャンルとして割り当てた。テストデータに対する各種評価指標の値を表 1 に示す。

表 1: 記事のジャンル分類におけるテストデータに対する予測の性能評価

	precision	recall	f1-score	support
pol	0.96	0.86	0.91	8304
dom	0.74	0.79	0.76	2970
int	0.90	0.87	0.88	3286
eco	0.76	0.86	0.81	2833
sci	0.86	0.87	0.87	3147
lif	0.84	0.93	0.88	2882
e/s	0.88	0.89	0.88	3169
accuracy			0.87	26591

5.3 SIRIUS によるプロット

SIRIUS[3] により、ユーザによる操作を反映した対象（ニュースサイト）と属性（賛否とジャンル）のインタラクティブな可視化を行う。ここではドナルド・トランプ氏に関する、BERT によるジャンル予測が行われた記事も含めた記事のデータを用いた。分類方法と記事数の内訳は表 2 の通りである。

表 2: 分類方法別記事数

手動で分類した記事数	24564
BERT によって分類した記事数	120440
全記事数	145004

5.3.1 距離の操作

距離の操作に関して、図 1 は右寄りのサイトと左寄りのサイトの特徴を分析することを想定し、Media Bias Chart 5.1 にて、右寄りと評価されたサイトのうち *theamericanconservative,newsmax, theblaze,washingtontimes,infowars,politico,thehill* の 7 つを平面の角に、左寄りと評価されたサイトのうち *cnn,nytimes,washingtonpost,alternet, counterpunch* の 5 つを反対の角に移動させたのちに再度プロットしたものである。結果からは、操作によって近い位置に移動されたサイト同士の距離は近い位置関係を維持しており、他のニュースサイトも類似度の高いサイトの近くにプロットされた。属性に関しては否定的（neg）なジャンルがそれぞれ離れた位置にプロットされており、ニュースサイトの右寄り左寄りの違いが各ジャンルの否定的な記事の割合によって特徴付けられるものと考えられる。

5.3.2 重みの操作

重みの操作の関して、ユーザが賛否に関わらず特定のジャンルへの関心がある状況を想定し、国際（int）ジャンルの記事の重みを大きく変更し再プロットした。結果を図 2 に示す。図から対象に関しては、*foreignpolicy.com* のノードが大きく、他の対象と離れた位置にプロットされた。国際ジャンルに強く興味を持っているものとして現れたと考えられる。属性に関しては否定的かつ国際（neg-int）が他の属性と離れた位置にプロットされた。

6 おわりに

本論文では多様なニュースサイトの報道スタンスの特徴や類似性をユーザの興味を反映しながら分析する

[3]https://www.gdeltproject.org
[4]https://github.com/google-research/bert

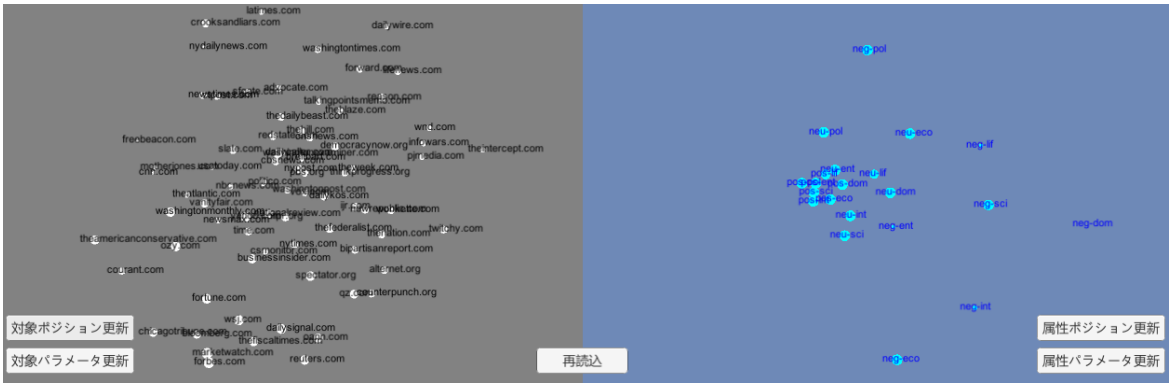

図 1: 右寄りのサイトと左寄りのサイトをそれぞれ近い位置に移動した後に再プロットした結果

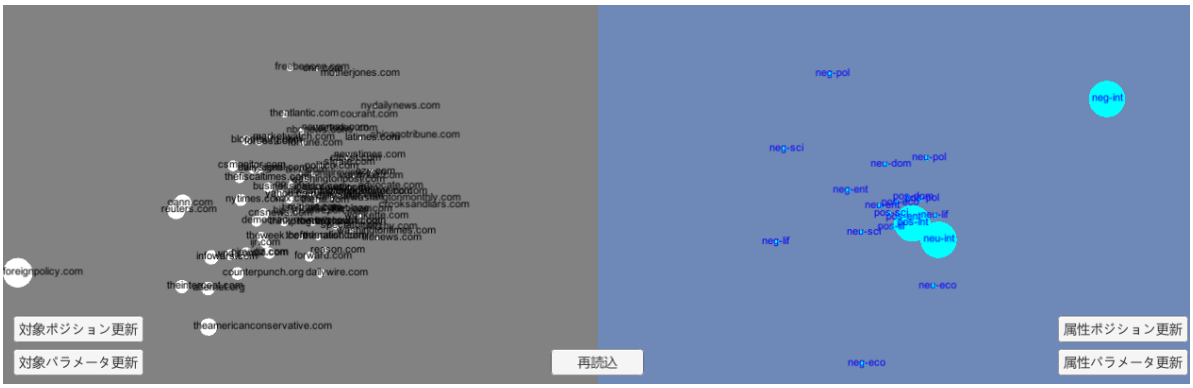

図 2: 国際（int）トピックの重みを大きくして再プロットした結果

手法を提案した。今後の研究としては、ニュースサイトの特徴をより具体的に分析するために、トピックに関わるイシュー別のスタンスなどを考えている。

謝辞

本研究の一部は、JSPS 科研費 18H03338 の助成と北海道大学国際連携研究教育局ビッグデータ・サイバーセキュリティグローバルステーションの支援を受けた。ここに記して謝意をあらわす。

参考文献

[1] Masaharu Yoshioka, Myungha Jang, James Allan, and Noriko Kando. Visualizing polarity-based stances of news websites. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018.*, pp. 6–8, 2018.

[2] A. Lundervold, H. Hauser, A. Lundervold, and C. Turkay. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization Computer Graphics*, Vol. 18, No. 12, pp. 2621–2630, dec 2012.

[3] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North. Sirius: Dual, symmetric, interactive dimension reductions. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 25, No. 1, pp. 172–182, Jan 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. cite arxiv:1810.04805Comment: 13 pages.