F-016

オンライン協調コミュニティの特徴付けによるコミュニティ分類手法の提案
A Coordinated Online Community Classification Method using Characterization

松崎 穂高[1)]　軽部 勲[1)]　平山 淳一[1)]
Hodaka Matsuzaki　Isao Karube　Junichi Hirayama

1　はじめに

オンライン議論は情報共有の場として有益である一方で、フェイクニュースや極端な意見の拡散により誤った世論を形成したり、国内外の分断に繋がる危険な側面をもつ。近年、ソーシャルメディア上では，多数のユーザーによる協調したオンライン行動により当該活動を広範囲かつ迅速に拡散するといった組織的な試みがオンライン討論に影響を及ぼしている。ネットワーク科学的手法や機械学習手法により協調行動するコミュニティを検出する研究はここ数年で加速している。一方で、協調行動の意図や悪意の程度、裏に潜む組織への解釈は依然として困難な課題として残っている。

協調行動するコミュニティの非本物性、有害性、組織性、時間分散性といった観点で特徴づけることで、協調行動についての解釈を支援する場合がある。本研究は、検出した協調コミュニティの特徴付けにより、情報操作を意図する協調コミュニティを抽出するフレームワークを提案する。具体的には、ネットワーク科学的手法により検出した協調リポスト関係にあるコミュニティを対象に、リポスト投稿に含まれるプロパガンダおよび共有リンクの数・種類といった特性を算出する。算出した特性値に基づき、階層型クラスタリングを用いてコミュニティを分類・評価する。コミュニティの特徴付けにより、これまで見落としていた協調行動に関する深い洞察を提供することができる。さらに、クラスタリング手法を用いたコミュニティ分類法は、人手によるコミュニティの解釈を半自動化する効果や特徴付けに用いた特徴量の妥当性を検証することに寄与する。

2　データセット

本研究は、2023 年に福島第一原発の ALPS 処理水放出をめぐって発生した国際的な情報発信の動向を背景に、情報拡散の構造を明らかにすることを目的とする。一部の海外メディアや SNS アカウントが、処理水に関する懸念を地球規模の環境問題として強調し、日本側の対応をめぐる批判や反応を促す投稿が見受けられた。これにより、SNS 上では国内外から多様な視点による活発な議論が展開された。こうした対立する情報環境を踏まえ、本研究では 2021 年 4 月から 2023 年 9 月までに X（旧 Twitter）上に投稿された処理水関連の 354,882 件の投稿をキーワードベースで収集した。特にリポスト行動に注目し、協調的な情報拡散コミュニティの構造と傾向を分析した。表 1 は、分析に使用した X のデータ名および取り得る値の例を示す。

3　提案手法

本研究は，まず、X 上で協調行動するコミュニティを Nizzol ら [1] が提案するネットワーク科学的手法を用いて検出する。その後、複数の指標に基づきコミュニティを特性評価する方式を提案する。また，特性評価結果の検証として，クラスタリング手法による分類方式を提案する。これは，検知対象の不審なコミュニティの傾向や特徴を捉えられているかを検証する役割を果たす。本研究の提案手法の概要を図 1 に示す。

表 1　分析に使用した X データの例

Feature name	Example (UserA)
User ID	10211
username	userA
Posted date	2022-12-22 13:00
Urls	https://t.com/hoge
Text	Fukushima water is treated water.
reposted status	retweeted
reposted status ids	172895
reposted status user ids	13551

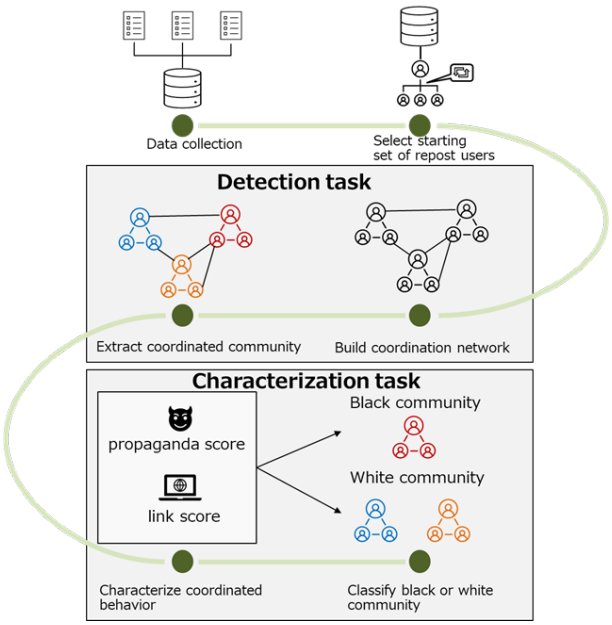

図 1　提案手法の概要

3.1　コミュニティの特徴付け

本研究では，複数の特性評価に基づき検出した協調コミュニティを特徴付けた。具体的には，協調したリポスト投稿に含まれたプロパガンダおよび投稿内で共有されたリンクの数・種類といった側面に基づいて，各協調コミュニティを特徴付けた。

3.1.1　プロパガンダスコアの算出

情報操作を意図した協調コミュニティは、しばしばプロパガンダを含む投稿を増幅させる。我々の研究グループは、過去に自然言語モデルを用いて SNS などのテキスト情報に含まれるプロパガンダ技法を検出する技術を開発している [2]。本研究は、このプロパガンダ検出法を用いてコミュニティ全体の有害性を評価する。具体的

1)　株式会社 日立製作所 Hitachi, Ltd.

には、分析対象のコミュニティ内でリポストした投稿について 9 つのプロパガンダ技法 (恐怖・偏見へ訴える、理由の単純化、誹謗中傷、そっちこそどうなんだ主義、全体主義、議論の 2 極化、疑い、誇張・過少化、含み言葉) がそれぞれ含まれている確率を算出する。

3.1.2 リンクスコアの算出

情報操作を意図した協調コミュニティは、国営メディアや政府支援のインフルエンサーによる投稿を拡散し、画像・動画・記事などを伴う情報でオーディエンスへの影響を強める傾向がある。本研究では、こうしたコミュニティの特徴を捉えるため、リポスト内に含まれる URL リンク数 R とその多様性に着目した指標となるリンクスコア（LV）を提案する。リンクの数 N と種類 T に加重 (w_1,w_2) をつけて計算することで (式 (1))、拡散行動の積極性や多様性を測定することが可能となる。

$$LV = \frac{w_1(\frac{N_i}{R_i}) + w_2(\frac{T_i}{N_i})}{w_1 + w_2} \tag{1}$$

3.2 協調コミュニティのクラスタ分類

本研究では、協調行動するコミュニティの特性を捉えるために算出した 10 個の指標を用い、それらの類似性に基づいてクラスタリングを行う手法を提案する。具体的には、各コミュニティの特性値に基づき、ユークリッド距離と Ward 法を用いた階層クラスタリングを実施した。Ward 法はクラスタ間の分散を最小化しつつ均一なクラスタを形成しやすいため、データの特徴を適切に保持しつつ分類を行うのに適している。

4 評価実験

4.1 評価項目

提案手法の性能を検証するため、以下の 2 つの観点で評価を実施した。

評価①：階層型クラスタリングにより検知対象とするブラックコミュニティ (BC) と非検知対象とするホワイトコミュニティ (WC) を分類できているか

評価②：提案したコミュニティの特性値の取り得る値の傾向が BC と WC で異なるか

4.2 実験条件

本研究は、検出する協調コミュニティのサイズが 20-90 を分析対象とする。分析対象のコミュニティに対して、BC または WC かラベリングする。以下の観点を BC の特徴と判断し、ラベリングを実施した。

- リポスト投稿の内容が何らかの工作意図や悪意を含む
- 投稿に含まれる共有リンク先が悪質な画像・動画・記事の類
- リポスト元アカウントが国営メディア/インフルエンサー/反日系アカウント
- コミュニティ内に不審アカウント（凍結済み，フォロー・フォロワーが極めて少ない）を多数含める

4.3 結果と考察

検出した協調コミュニティは全 3118 であり、そのうち 68 の協調コミュニティを分析対象とした。また、ラベリングしたコミュニティの内訳は BC：11、WC：57 である。図 2、図 3 はそれぞれ BC および WC の特性値をレーダチャートで示したものである。以下の主要な結果を得た。

評価①：階層型クラスタリングによる協調コミュニティの分類精度は再現率 81.8%、適合率 40.9%を達成した。一方で、どのクラスタがブラッククラスタかの解釈は人手による後付け的なアプローチであるため、実運用を想定した自動検出アルゴリズムを設計する必要性を確認した。

評価②：BC と WC で特性値の取り得る値の傾向が異なることを確認した。BC は「恐怖・偏見へ訴える」のプロパガンダスコアおよびリンクスコアが比較的高く，WC は「誹謗中傷」および「疑い」のプロパガンダスコアが比較的高い傾向にある。

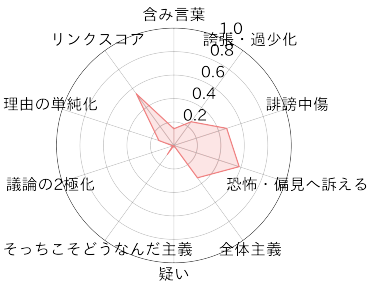

図 2 BC のレーダーチャート

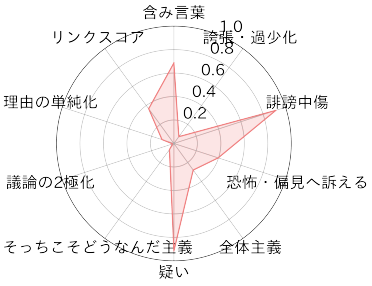

図 3 WC のレーダーチャート

5 おわりに

近年、SNS 上での情報操作を目的とした組織的な協調行動を検出する重要性が増している。本研究は、リポストの協調性やプロパガンダ、共有リンクの特性に着目し、不審なコミュニティの特徴を明らかにすることを目的とした。2023 年 8 月の ALPS 処理水放出に関する X の投稿を対象に階層クラスタリングを適用した結果、68 の協調的コミュニティを再現率 81.8%、適合率 40.9%で分類することに成功した。今後は、異常度に基づく自動検出手法の開発や大規模データに対応する手法の導入が課題である。本研究は、悪質な協調行動の検出とその応用の発展に寄与することが期待される。

参考文献

[1] Nizzoli L, Tardelli S, Avvenuti M, Cresci S, Tesconi M.: Coordinated behavior on social media in 2019 UK general election, In: Proceedings of the international AAAI conference on web and social media, vol 15, pp 443–454, 2021.

[2] Morio, G., Morishita, T., Ozaki, H., Miyoshi, T.: Hitachi at SemEval-2020 Task 11: An empirical study of pre-trained transformer family for propaganda detection, In: Proc. Fourteenth Workshop on Semantic Evaluation, pp. 1739–1748, 2020.