D-005

人間関係と誹謗中傷検出によるオンラインハラスメント対策
Online harassment countermeasure by detecting human relations and slanders

尚　鉄淞 †
Tiesong Shang

周 娟 ‡
Juan Zhou

高田 秀志 ‡
Hideyuki Takada

1　はじめに

オンラインコミュニティは，ユーザがウェブ上でやり取りできるプラットフォームである．互いの共通の関心事や話題についてやりとりできるので，多くの人々が利用している．しかし，その普及に伴い，オンラインハラスメントという新たな問題が徐々に認識されるようになった．オンラインハラスメントは，インターネット上における中傷および嫌がらせである．知っている人から，パソコンや携帯電話などの端末でネットを経由して一方的で継続的に物理的・精神的苦痛が加えられ，被害者にとっては深刻な苦痛を与えられることがある．また，リアルな交友範囲から離れ，現実世界で互いに知らない人々が，異なる意見を持っているために，他の人を傷つけることを狙って，個人またはグループで，敵対的な行動を意図的に何度も行うこともある [1].

Pew Research Center が 2017 年に発表した調査によると，インターネットを利用するアメリカの成人のうち，66 %がオンラインハラスメントを目撃したことがあり，41 %は自身で経験したことがある．具体的には，中傷，セクシャルハラスメント，身体的な脅威などである [2]．また，62 %の人がこれを大きな問題と認識している．オンラインコミュニティの良い環境を保つには，オンラインハラスメントを検出し，発信者を特定することが重要であると考える．オンラインハラスメントの検出には，オンラインでの嫌がらせメッセージと関係者，特に加害者と被害者を検出することが必要である．

本研究では，オンラインハラスメントを減少させるため，人間関係と誹謗中傷の検出によるオンラインハラスメントの対策を提案する．ソーシャルネットワークにおける不適切なコンテンツを発見し，さらに，被害と加害の関係を見つける．これによってオンラインハラスメントの重症度を測定する．これにより，管理者が早期介入できることが期待される．

2　既存研究

オンラインコミュニティにおける誹謗中傷の検出に関する既存の研究には，主にテキスト抽出など，いわゆる自然言語処理などにより不適切な単語を検出する方法などがある [3][4]. しかし，既存の方法では人間関係を考慮していないため，オンラインハラスメントの検出が難しく，被害者を検出することができない．そのため，定量的にオンラインハラスメントの重症度を測定することが難しい．

3　手法

3.1　手法概要

オンラインハラスメントでは，複数の加害者が特定の対象者に対してメッセージを送信すること，および，一人の加害者が特定の対象者に対して複数のメッセージを送信することが考えられる．そのため，一つのコメントだけでオンライハラスメントが発生しているかどうかを判断することが難しい．本手法では，まず，メッセージの内容によって誹謗中傷しているかどうかを判断し，加害者と被害者を特定する．次に，加害者と被害者を中心に，以前の投稿記事や友達へのリンクなどのすべての関連情報を取得し，ソーシャルグラフを構築する．このソーシャルグラフを利用し，誹謗中傷と嫌がらせに関するメッセージを送信あるいは受信した回数に基づいて，誹謗中傷と嫌がらせが含まれるメッセージが全体のメッセージに占めた割合を計算する [5]．これにより，オンラインハラスメントが発生しているかを判断し，その重症度も測定する．

3.2　データソース

本研究では，Baidu が提供するコミュニケーションプラットフォームである Baidu Tieba を利用する．この SNS では特定の話題に対してスレッドを立ち上げて掲示板形式で意見をかわすことができる．幅広いトピックを扱えるので，各ユーザがトピックを自由に設定することができ，歌手や，オンラインゲーム，自分の名前などのトピックを設定することができる．このような SNS の特徴は，トピックの数が多いこと，一つのトピックの中では同じ関心を持つ人々で構成されていること，メンバが固定していること，使用頻度が高いことである．

本手法ではソーシャルグラフの構築のために，以下の 3 種類の機能から情報を取り出す．

- 記事：あるトピックに属している各話題
- コメント：記事に対する返事
- アートマーク：特定の人に送るメッセージ

これらの情報に基づき，メッセージの発信者と受信者の関係性を見出す．さらに，各関係へ重みを与えることで，ソーシャルグラフを構築する．

3.3　ソーシャルグラフ分析

ソーシャルグラフは，メッセージの数，メッセージの発信者，メッセージの受信者を表現している有向グラフである．ソーシャルグラフは，ノードの集合と有向辺の集合で構成される．ノードはユーザを表し，有向辺はメッセージの発信元と宛先を表す．

図 1 はソーシャルグラフの一例である．ユーザ A はユーザ B，D，E から合計 11 件のメッセージを受信している．また，ユーザ A はユーザ C とユーザ E に合計 7 件のメッセージを発信している．

† 立命館大学大学院情報理工学研究科
‡ 立命館大学情報理工学部

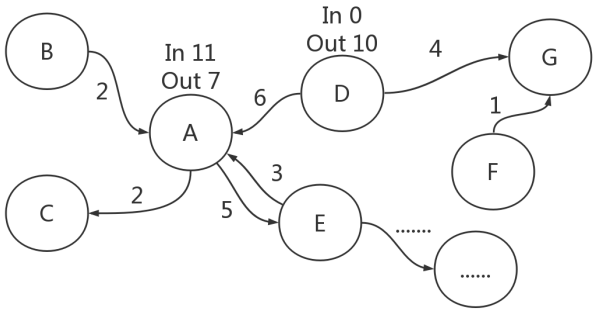

図 1　ソーシャルグラフ

3.4　誹謗中傷の検出

オンラインコミュニティでは誹謗中傷を含む言葉を特定することが難しい．なぜならば，オンラインコミュニティでは非構造的な言葉がよく使われているからである．また，スペルミスがあると検出できない場合もある．

したがって，本研究では，ルールベースのアプローチを用いる．ルールベースのアプローチでは，誹謗中傷を表す単語またはパターンを事前に定義する．また，bag-of-opinion モデル [6] を用いた共起情報を利用することで，単語ごとに悪口度を与え，文が誹謗中傷を表しているかを判定する．これは，オンラインコミュニティに対する適切な検出アプローチである．

3.5　計算方法

Wulczyn らの研究では，不適切なメッセージを発信する人をカテゴリー化している．その結果，40 %の発信者において，侮辱的な投稿平均数は 5 未満であることが示されている [7]．すなわち，このような人は，常に不適切なメッセージを発するとは考えにくい．また，重症度の観点から言えば，コミュニティに対して悪い影響がそれほど大きくないと考える．また，1 %のユーザは 10 %の不適切なメッセージを投稿するという研究もある [7]．したがって，重症度の低い一般ユーザと攻撃性の強いユーザの計算方法が異なる．まず，一般ユーザからのオンラインハラスメント指標は，不適切なメッセージがすべてのメッセージに占める割合と，不適切なメッセージ数の積とする．一般的には，割合のみを指標として採用することが多いが，割合が同じでも発したメッセージ数が異なるとコミュニティへの影響が異なるため，影響度を考慮するために，不適切なメッセージ数を重みとして掛ける．以上より，ユーザ j からユーザ i へのすべてのメッセージ数を a_ij，ユーザ j からユーザ i への不適切なメッセージ数を b_ij とすると，すべての一般ユーザから受けたオンラインハラスメント指標は，$\sum_{n=1}^{N} b_{ij}b_{ij}/a_{ij}$ で表される．

次に，攻撃性の強いユーザ k に対して，オンラインハラスメント指標は不適切なメッセージ数 c_{ik} とする．

ユーザ i が受けたオンラインハラスメント指標 p_i はそれらの和である．

$$P_i = \sum_{n=1}^{N} b_{ij}^2/a_{ij} + \sum_{m=1}^{M} c_{ik} \quad (1)$$

これにより，オンラインハラスメントが発生したかどうかを確認する．同時に，p_i の大きさによって，重症度も測定する．

4　実装

システムアーキテクチャを図 2 に示す．まず，Scrapy spider を利用して，サーバから特定のオンラインコミュニティのコンテンツを収集する．次に，ユーザの関係性を保存し，ソーシャルグラフを構築する．また，Jieba という中国語のテキストセグメンテーションを利用して，データクリーニングを行ってから，誹謗中傷を含むメッセージを検出する．得られた誹謗中傷メッセージの数をソーシャルグラフに反映する．

最後に，数式 (1) を用いて，計算した p の大きさでオンラインハラスメントがあるかどうかを判断する．

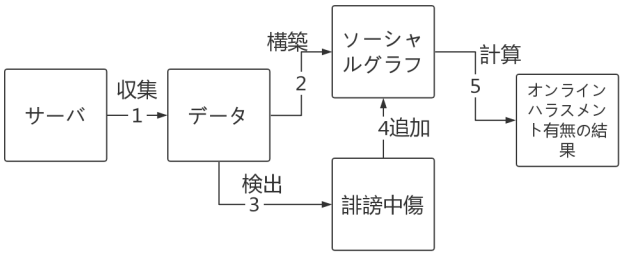

図 2　システムアーキテクチャ

5　おわりに

本稿では，人間関係と誹謗中傷の検出によるオンラインハラスメントの対策を提案した．本手法は，ソーシャルグラフを利用して，オンラインハラスメント発生の診断と重症度の測定を行う．

今後は評価実験を行い，検出精度と有効性を評価する予定である．

参考文献

[1] 文部科学省.「ネット上のいじめ」 に関する対応マニュアル・事例集 (学校教員向け)[J]. 2008.

[2] Duggan M, Rainie L, Smith A, et al. Online harassment. Pew research center[J]. 2017.

[3] 株式会社ドワンゴ／株式会社ニワンゴ「ネットでのいじめなどに関する実態調査」 http://dwango.co.jp/pi/ns/2013/1202/index3.html

[4] 三島浩路, 本庄勝. 技術的観点からのネットいじめ対策 [J]. 電子情報通信学会 通信ソサイエティマガジン, 2015, 9(2): 102-109.

[5] Dinakar K, Reichart R, Lieberman H. Modeling the detection of Textual Cyberbullying[J]. The Social Mobile Web, 2011, 11(02).

[6] Qu L, Ifrim G, Weikum G. The bag-of-opinions method for review rating prediction from sparse text patterns//Proceedings of the 23rd International Conference on Computational Linguistics.

[7] E. Wulczyn, N. Thain, and L. Dixon,“Ex machina: Personal attacks seen at scale,” arXiv preprint:1610.08914, 2016.