203-9in

カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出

Detection of Harmful Entries on Informal School Website Based on Maximization of Category Relevance

新田 大征[*1] 桝井 文人[*1] プタシンスキ ミハウ[*1] 木村 泰知[*2] ジェプカ ラファウ[*3]
Taisei Nitta Fumito Masui Michal Ptaszynski Yasutomo Kimura Rafal Rzepka

荒木 健治[*3]
Kenji Araki

[*1]北見工業大学 情報システム工学科
Department of Computer Science, Kitami Institute of Technology

[*2]小樽商科大学 社会情報学科
Department of Information and Management Science, Otaru University of Commerce

[*3]北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

In this paper we discuss the practicality of the method to detect harmful information on the Internet. In particular, we focus on "cyber-bullying", defined as slandering other people through BBS or e-mails. To deal with the problem members of Parent-Teacher Association (PTA) perform Web site monitoring activities called "net-patrol". Unfortunately, searching for harmful information on the Web manually comes with high cost in time and fatigue of PTA members. We propose a method for maximization of relevance of categories for searching harmful information automatically. In the proposed method categorized seed words are used to calculate semantic orientation score from PMI-IR for each category. We conducted an experiment to detect harmful information in cases where the cyber-bullying entries cover 50% (fair condition) and 12% (real world condition) of the whole data. The results show that the proposed method performs much better than baseline settings for both Precision and Recall.

1. はじめに

「ネット上のいじめ」が新たないじめの形として社会問題化している.「ネット上のいじめ」とは，携帯電話やパソコンを通じてインターネット掲示版などに特定の子どもへの誹謗中傷を書込んだり，嫌がらせメールを送ったりするなどしていじめを行なうものである [文部科学省 2008].

このようないじめに対処するために学校関係者や一部の保護者などが主体となってネットパトロール活動を行なっている. ネットパトロールとは掲示板を巡回，監視することである. その際，誹謗中傷のような有害な書込みを発見した場合，該当掲示板の管理人あるいはプロバイダに書込みの削除を依頼する. 実際にネットパトロールによって発見され，削除を依頼された有害な書込み例を表 1 に示す.

このようなネットパトロール活動は主に人手で行なわれており，多数の掲示板に記述された膨大な書込みの中から有害情報を含む書込み (有害書込み) を探し出すには計り知れない労力と時間を要する. また，このような作業をするための人的コストや作業従事者の身体的精神的影響も懸念されるところである.

上記の問題に対処するために，我々は有害書込み候補を見極める有害極性判定手法を提案した [松葉 2011]. この手法は Turney[Turney 2002] の関連度判定手法 (PMI-IR) を拡張して有害書込みとの関連度である有害極性値を算出し，少数の種単語を用意することで大量の有害書込みの候補を効率よく発見できる手法である. [松葉 2011] では 50%の有害書込み混合率のテストデータにおいて，有害書込みを 83%の精度で判定できることを確認しているが，現実の有害書込み混合率における有効性を確認するには至っていなかった.

表 1: ネットパトロールによって削除を依頼された有害な書込み例

- 調子乗りすぎいっぺん殺らなあかんで
- 新田キモイつかキショイほんま死んで
- ンな奴どつき回したれ
- 性格わるーい　ぶちゃいくー笑
- ＞＞ 17 あの女、昔、モバだったかグリに登録してたヤリマンじゃん。
- すぐにヤレる。01234567890。めっちゃカワイイで

本研究では，松葉らの手法に種単語のカテゴリ化と関連度の最大値を取得する考えを導入したカテゴリ別関連度最大化手法を提案する. 本手法は，少数の種単語を複数のカテゴリに分類し，各カテゴリとの関連度の最大値を有害極性値とすることで，有害書込みをより効率的に発見できる. さらに，現実的な有害書込み混合率を持つテストデータを用いて提案手法の評価実験を行ない，本手法の実用性を検証する.

以下，2 章で有害書込みの抽出に関する研究について述べ，3 章で提案手法について説明する. また，4 章で現実の有害書込み混合率についての調査結果，それに基づくテストデータの作成，提案手法の評価実験とその結果について述べ，5 章で評価実験についての考察を行なう.

連絡先: 新田 大征，北見工業大学 情報システム工学科，北海道北見市公園町 165

2. 関連研究

有害書込みの抽出に関する研究例はこれまでにも複数存在する．石坂ら [石坂 2010] は，巨大電子掲示板「2 ちゃんねる」を対象とし，悪口表現辞書を構築している．彼らは，悪口表現を「バカ」や「マスゴミのクズ」などの特定の他者に対して直接侮辱や誹謗中傷している単語，句と定義し，これらの悪口表現の使われ方，すなわち悪口表現に連接する単語の繋がりやすさを考慮し，周辺単語列から悪口表現を抽出することを試みている．

池田ら [池田 2010] は，人手で有害と無害に分けられた学習用文書を用いて，単語の出現頻度の偏りによる有害判定キーワードリストを構築している．文脈によって有害と無害に分かれる単語は係り受け関係を利用して対処し，分類性能を向上させた．しかし，Web 文書では「爆破」と「爆ー破」のように少しだけ文字を変えた表現も多く，この手法では人手で学習用文書を作成する膨大な手間が問題となる．

藤井ら [藤井 2010] は，文章中の 2 単語間の距離を利用し，過度な性的描写を含む文章を判別するシステムを提案している．彼らは，文章中に含まれている有害な意味にも無害な意味にも成り得る単語であるグレーワードと，単語単体で有害な意味になる単語であるブラックワードの距離が近いほど有害性が高いと判定している．

橋本ら [橋本 2010] は，隠語の有害語意を検出する手法を提案している．彼らは，単語の周りに出現する語である周辺語によって隠語の語意が決定されると推測し，周辺語との共起頻度を算出することで有害語意を検出している．

本研究では，周辺単語列を考慮せず有害書込み候補単語列を有害判定にする．また，Web 検索ヒット数を判定基準に利用するため，人手で学習用文書を構築するコストがかからない．さらに，係り受け関係を有害判定の対象とするため，ある単語の前後全ての単語に注目する必要がなく，処理コストも小さい．

3. 提案手法

本章では，我々が提案するカテゴリ別関連度最大化手法の概要を説明する．提案手法は，Turney[Turney 2002] が提案した評価極性判定手法を拡張し，掲示板の書込みと種単語との関連度を算出する．また，種単語を複数のカテゴリに分類，それぞれとの関連度を算出し，その中の最大値を書込みの有害極性値としている．本手法は (1) フレーズ抽出，(2) 有害語検出とカテゴリ化，(3) 関連度最大化による有害極性判定という三つの処理から構成される．以下，各処理について説明する．

3.1 フレーズ抽出

掲示板に書込まれる有害書込みには，単語単体で有害化するものと，単語単体では有害ではないが組み合わせによって有害化するものがある．例えば,「性格が悪い」という有害書込みの場合,「性格」や「悪い」という単語単体では有害とは言えないが，これらが係り受け関係を持つことで有害化している．そのため，単語のみで有害書込みを検出する場合，係り受け関係によって有害化した有害書込みは検出できない．

この問題に対処するために，係り受け関係を持つ形態素*1組み合わせに対する極性判定を行なう．具体的には，書込みを形態素解析し，名詞を中心とした係り受け関係にある形態素ペアを「フレーズ」と定義し，これらを判定対象として書込みから抽出している．対象とするフレーズとフレーズの例を表 2 に示す．

表 2: 本手法が対象とするフレーズとその例

フレーズ	フレーズの例
名詞-名詞	サル顔 → 身体や性質を揶揄する
名詞-動詞	新田を殺す → 相手を脅す
名詞-形容詞	性格が悪い → 身体や性質を非難する

3.2 有害語検出とカテゴリ化

本処理では，単体で有害性を有する単語を「有害語」として検出する．しかし，有害語は一般的な単語として認知されていない文字列や表記を伴う場合が多く，特別な配慮をしなければ形態素解析誤りなどを招き，適切に処理できない可能性が高い．我々は，文部科学省による有害語の定義に基づいて学校非公式サイトの書込みを調査し，得られた名詞，動詞及び形容詞の合計 255 語を有害語として形態素解析器の辞書に登録した．

さらに，有害語を卑猥語，暴力誘発語，誹謗中傷語のカテゴリに分類した．そして，各カテゴリの頻出上位 3 語の合計 9 語を有害極性単語とし，種単語として登録した．登録した有害極性単語は卑猥語の「セックス」,「ヤリマン」,「フェラ」，暴力誘発語の「死ね」,「殺す」,「殴る」，誹謗中傷語の「うざい」,「きもい」,「不細工」である．

3.3 関連度最大化による有害極性判定

この処理では，フレーズが持つ有害極性および有害性を有害極性単語の各カテゴリとの関連度を算出することにより定量化する．有害極性単語の各カテゴリとの関連度を測る尺度としては自己相互情報量（PMI）を用いる．ここでの PMI は，フレーズと有害極性単語の各カテゴリに登録されている単語 3 語の共起頻度を示す．そして，共起頻度の算出には Web 検索（IR）を用いる．Web 上には多様なページが存在し，そこには様々な単語が書込まれている．そのため，Web 検索を用いることによって高い網羅性を得ることができる．

フレーズと有害極性単語の各カテゴリとの関連度は式（1）で求める．p_i は書込みから抽出されたフレーズ，w_j は有害極性単語の 1 カテゴリに登録されている 3 単語であり，$hits(p_i)$ や $hits(w_j)$ は p_i や w_j はそれぞれを検索単語としたときの Web 検索ヒット件数，$hits(p_i\&w_j)$ は，p_i と w_j が同じ Web ページに出現するサイトの検索ヒット件数を示す．そして，$PMI-IR(p_i, w_j)$ は p_i と w_j の関連度である．

$$PMI - IR(p_i, w_j) = log_2 \left\{ \frac{hits(p_i \& w_j)}{hits(p_i) hits(w_j)} \right\} \quad (1)$$

フレーズと有害極性単語の関連度のうち，最大値をフレーズの有害書込みとの関連度とする．そして，書込みから抽出された全てのフレーズに有害書込みとの関連度を算出し，その中の最大値を書込みの有害極性値である $score$ とする．$score$ は式（2）で求める．

$$score = max(max(PMI - IR(p_i, w_j))) \quad (2)$$

*1 本研究では，形態素と単語を同じ意味として扱う．

ベースライン [松葉 2011] では，フレーズと有害極性単語 1 単語との関連度を算出し，その和を *phrase* の有害書込みとの関連度としていたが，本手法では有害極性単語を 1 単語ではなく 1 カテゴリに登録している 3 単語としている．これにより，フレーズと有害極性単語の 1 カテゴリに登録されている単語全てが同じ Web ページに出現するヒット件数を取得することになり，有害性が強いフレーズの有害書込みとの関連度のみを高めている．また，フレーズと有害極性単語 1 カテゴリとの関連度を算出し，その最大値を *score* とすることで，全ての有害極性単語と同じ Web ページに出現するが，それぞれとの関連度は小さいフレーズによって *score* が高くなることを防いでいる．

例として,「可愛いけど性格が悪い女」という書込みの *score* 算出方法について述べる．まず，この書込みから「可愛い-女」,「性格-悪い」,「悪い-女」というフレーズが抽出される．次に,「可愛い-女」と「セックス，ヤリマン，フェラ」,「死ね，殺す，殴る」,「うざい，きもい，不細工」との関連度を算出し，最大値をフレーズの有害書込みとの関連度とする．同様に「性格-悪い」と「悪い-女」に対して有害書込みとの関連度を算出する．最後に，抽出されたフレーズの有害書込みとの関連度のうち，最大値を書込みの *score* とする．

このようにして算出した *score* をもとに，全ての書込みを有害極性値が高い順に並び替える．そして閾値 n を設定し，上位 n 件の書込みを有害，それ以外を無害と判定する．

4. 評価実験

提案手法の評価実験を行ない，ベースラインと結果を比較した．以下，4.1 節で評価実験を行なうための予備調査について説明し，4.2 節で実験環境について述べ，4.3 節で評価実験の結果を報告する．

4.1 予備調査

提案手法の実用性を評価するためには，現実の学校非公式サイトと同じ有害書込み混合率のテストデータを作成する必要がある．そこで，現実の学校非公式サイトを調査し有害書込み混合率を算出した．調査した学校非公式サイトは表 3 に示す 3 つの学校非公式掲示板であり，調査期間は 2012 年 1 月 27 日〜2012 年 1 月 30 日の 4 日間で取得した書込み 2,222 件である．

表 3: 調査した学校非公式掲示板

学校非公式掲示板	総書込み数	有害書込み数	割合 (%)
掲示板 (1)	600	75	12.5
掲示板 (2)	736	90	12.2
掲示板 (3)	886	100	11.3

調査の結果，掲示板 (1) では総書込み数 600 件に対し有害書込みは 75 件，出現率は 12.5%であった．同様に，掲示板 (2) では 736 件中 90 件の 12.2%，掲示板 (3) では 886 件中 100 件の 11.3%であった．以上を総合し，我々は現実の学校非公式サイトにおける有害書込み混合率を 12%程度であると結論付けた．

4.2 実験環境

まず，有害書込み混合率 50%のテストデータに対する提案手法とベースラインの判定性能を比較した．次に，有害書込み混合率 12%のテストデータを作成し，それに対して提案手法とベースラインを適用して判定性能を比較した．

有害書込み混合率 50%のテストデータは，実際にネットパトロールによって収集された掲示板への書込みデータと，松葉らが独自に収集した書込みデータ (三重県域に限定されたサイトから収集したもの)2,998 件 (有害書込み 1,508 件，無害書込み 1,490 件) である [松葉 2011]．このテストデータに対して提案手法及びベースラインを適用し，フレーズを抽出できた書込みの有害極性値を算出した．そして，有害極性値に基づいて全書込みをランキングし，50 件毎に閾値 n を設定して上位 n 件における判定性能を評価した．

また，有害書込み混合率 12%のテストデータは，予備調査の結果をもとに有害書込み混合率 50%のテストデータから有害書込み 60 件，無害書込み 440 件を無作為に取り出した合計 500 件のテストデータ 5 セットを用意した．作成した各テストデータに対して提案手法及びベースラインを適用し，フレーズを抽出できた書込みの有害極性値を算出した．そして，有害極性値に基づいて全書込みをランキングし，10 件毎に閾値 n を設定して上位 n 件における判定性能を評価した．

評価基準として精度 (式 3) と再現率 (式 4) を用いた．精度とは，上位 n 件のうち正しく有害と判定できた書込み数の割合であり，再現率とは，本来の有害書込み数のうち正しく有害と判定できた書込み数の割合である．実験では，各テストデータにおける精度及び再現率の平均を実用性の評価基準とする．

$$\text{精度} = \frac{\text{システムが正しく有害と判定した書込み数}}{\text{システムが有害と判定した書込み数}} \tag{3}$$

$$\text{再現率} = \frac{\text{システムが正しく有害と判定した書込み数}}{\text{全ての有害書込み数}} \tag{4}$$

4.3 実験結果

有害書込み混合率 12%，50%のテストデータにおける提案手法とベースラインの精度及び再現率を図 1 に示す．横軸，縦軸は各閾値における精度と再現率である．

有害書込み混合率 50%のテストデータに対し，ベースラインでは精度は 49%〜72%，再現率は 3%〜100%であり，提案手法では精度は 49%〜90%，再現率は 5%〜100%であった．

また，有害書込み混合率 12%のテストデータに対し，ベースラインでは精度は 11%〜30%，再現率は 8%〜100%であり，提案手法では精度は 10%〜48%，再現率は 13%〜100%であった．

5. 考察

実験の結果，ベースラインと比較して提案手法の方が全体的に性能が高いことがわかった．精度と再現率の相関曲線の形を見ると，ベースラインでは有害書込み混合率 50%のテストデータと比較して，12%のテストデータでは判定性能が大きく低下していることがわかる．しかし，提案手法では有害書込み混合率 50%のテストデータと 12%のテストデータにおける相関曲線の形に大きな変化がないため，ベースラインよりも判定性能が安定していることを示唆している．

また，有害書込み混合率 12%のテストデータに対してはベースラインの方が高い再現率を示している箇所が見られた．これは，ベースラインでは高い有害極性値を算出できなかった有害書込みが同じ閾値に集まっているためである．これに対し，提

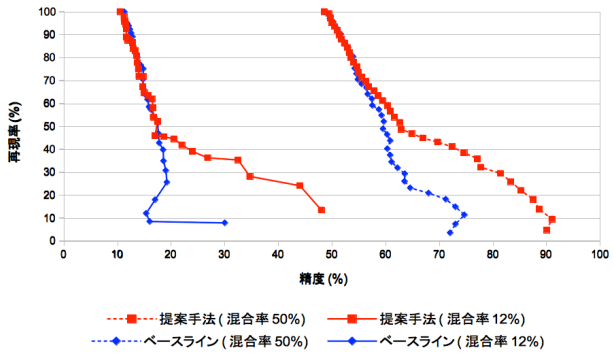

図 1: 有害書込み混合率 50%，12%のテストデータにおける提案手法とベースラインの精度及び再現率

案手法は有害書込みに高い有害極性値を算出することができ，そのような有害書込みを低い閾値に集めることができたため，大きい閾値ではベースラインの方が高い再現率を示していると考えられる．よって，提案手法はベースラインよりも高い有害書込み判定性能を得たといえる．

次に，有害書込み混合率 12%のテストデータに対する処理結果の内容について考察する．精度が 48%となった閾値付近の書込みを調査すると,「ウザイキモイぶす」や「アトピーのやつ死ねよ」という有害書込みが見られた．これらの書込みから抽出された「ウザイ-ぶす」や「アトピー-死ね」というフレーズに対して高い関連度が算出されていたため，有害書込みの有害極性値を高くすることができた．

一方で，有害書込みと同等，もしくはそれよりも高い有害極性値となった無害書込みが多かった．このような無害書込みの例として,「県外に住んでいる」という書込みが挙げられる．この書込みからは「外-住ん」というフレーズが抽出されているが，これらは無害書込みだけでなく，誰がどこに住んでいるなどといった個人情報を晒すような有害書込みにも出現しやすいニュートラルなフレーズである．そのため，有害書込みとの関連度が高くなり，結果として無害書込みの有害極性値が高くなってしまったことが影響として現れた．

対策として,「素晴らしい」などの無害書込みにのみ現れやすい単語を無害極性単語として登録し，無害書込みとの関連度を算出するなどが考えられる．具体的には，まず無害書込みを調査し無害極性単語を辞書に登録し，次に有害極性単語とフレーズの関連度及び無害極性単語とフレーズの関連度を算出する．この際に，無害極性単語との関連度が有害極性単語との関連度より高かった場合は無害なフレーズとする．これにより，ニュートラルなフレーズによる判定性能への影響を低減できると考えられる．

再現率が 100%となった閾値付近の書込みを調査すると,「北見工業大学 4 年の新田」のような人名や所属している学校名など，個人情報のみで構成されている有害書込みが見られた．これは，現在登録している有害極性単語が個人情報との関連度が低く，個人情報のみの書込みに対する有害極性値が低くなったことが影響したと思われる．

この問題に対しては個人情報との関連度が高い単語を有害極性単語として登録し，個人情報を含む書込みを有害と判定するヒューリスティクスを用いる手法が有効と思われる．

6. おわりに

本研究では，ネットパトロール担当者の負担を軽減させるため，有害書込みを検出するカテゴリ別関連度最大化手法を提案した．また，提案手法の実用性を検証するため，現実の有害書込み混合率に即したテストデータに対する判定性能を評価した．まず，現実の学校非公式サイトにおける有害書込み混合率を明らかにした．そして，得られた結果をもとに現実の学校非公式サイトを想定したテストデータを 5 セット作成し，提案手法を適用して評価実験を行なった．また，ベースラインを再現し，提案手法と判定性能を比較した．

評価実験の結果，提案手法はベースラインよりも高い判定性能を得た．しかし，無害書込みに含まれているニュートラルなフレーズの有害極性値が高く，また個人情報で構成される有害書込みの有害極性値が低かったため，判定性能に影響を与えた．

今後は，無害書込みと関連度が高い単語を無害極性単語として登録し，ニュートラルなフレーズを含む無害書込みの有害極性値を下げるような処理を行なう．さらに，個人情報を含む書込みを有害と判定する手法について検討する．

参考文献

[文部科学省 2008] 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)”，文部科学省，(2008)

[松葉 2011] 松葉達明，桝井文人，河合敦夫，井須尚紀：”学校非公式サイトにおける有害情報検出を目的とした極性判定モデルに関する研究”，言語処理学会第 17 回年次大会発表論文集，P2-26(2011.3)

[Turney 2002] Peter D. Turney：”’Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”，Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics，Philadelphia， pp.417-424(2002.7)

[石坂 2010] 石坂達也，山本和英:”2 ちゃんねるを対象とした悪口表現の抽出”，言語処理学会第 16 回年次大会，pp.178-181 (2010.3)

[池田 2010] 池田和史，柳原正ら” 格要素の抽象化に基づく違法・有害文書検出手法の提案と評価”，情報処理学会第 72 回全国大会，pp.71-72（2010.3）

[藤井 2010] 藤井雄太郎，安藤哲志，伊藤孝行：” 有害情報フィルタリングのための 2 語幹の距離及び共起情報による文章分類手法の提案”， 第 24 回人工知能学会全国大会発表論文集，3D2-4，pp.1-4(2010)

[橋本 2010] 橋本広美，木下嵩基，原田実：”フィルタリングのための隠語の有害語意検出機能の意味解析システム SAGE への組み込み”，情報処理学会研究報告，SLP，音声言語情報処理 2010-SLP-81(14)，1-6，2010-05-20