学校非公式サイトにおける有害情報検出

松葉　達明†　桝井　文人‡　河合　敦夫†　井須　尚紀†

†三重大学大学院工学研究科　‡北見工業大学情報システム工学科

{matsuba, kawai, isu}@ai.info.mie-u.ac.jp, f-masui@mail.kitami-it.ac.jp

1　はじめに

「ネット上のいじめ」が新しい「いじめ」の形態として問題となっている.「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上のいわゆる学校非公式サイトの掲示板などに特定の子どもの悪口や誹謗・中傷を書込んだり,メールを送信するなどして,有害情報によるいじめを行うもの[1] である.このような「ネット上のいじめ」では,短期間で深刻化するケースも多い上に,当事者は容易に被害者にも加害者にも成り得る.そのため,エスカレートを見過ごすと事件にまで発展する危険性があり,早期発見・早期対応に向けた取組みが急務である.

文部科学省は,「ネット上のいじめ」を手段や内容に着目して図１のように類型化している.このうち,学校非公式サイトにおける「ネット上のいじめ」に注目してみる.学校非公式サイトの掲示板は,複数のユーザが相互に発言を行い情報交換をする場である.このようなサイトでは,議論の食い違いや学校での諍いなどが発端となり,他のユーザが不快と感じる発言や特定個人を誹謗・中傷する発言が書込まれるケースも頻繁に発生する.[2]

これらの有害情報は,ネットパトロールに基づいて対応されている.ネットパトロールとは,文字通り学校非公式サイトなどを人手でつぶさに書込み内容のチェックを行うことである.ネットパトロールによって有害であると判断された書込みについては,当該掲示板の管理人あるいはプロバイダに削除依頼がなされる.

しかしながら,これらの活動は,教育委員会や学校の教職員や外部委託の情報教育アドバイザーがボランティアベースで行っている場合がほとんどである.掲示板などの書込みについても,書込み内容を印刷や携帯電話のカメラで画面を撮影するなどの処理を経た後に詳細な内容チェックを行うなど,気の遠くなるような作業を行っているのが現状である.また,ネットパトロール支援ツールとして,単語レベルで一致したものを検出する技術もあるが,単語レベルでの検出技術では,有害情報には該当しない多くの情報をも検出するなど,検出精度に問題がある.このような状況では,増大し続ける学校非公式サイト全てを監視し続けることは次第に困難となるであろうし,活動に取り組む人の健康や生活への影響も大きなものとなろう.

そこで本研究では,学校非公式サイトの掲示板に書込まれる有害情報を検出するシステム構築を目指す.これにより,ネットパトロール活動の一部を自動化し,担当者の負担を軽減することができ,有害情報の早期発見・早期対応の支援にもつながる.

本論文は,全5章で構成されている.第1章に続き,第2章では有害情報分類の基本的な考え方について説明する.第3章では,

1. 掲示板・ブログ・プロフでの「ネット上のいじめ」
(a)掲示板・ブログ・プロフへの誹謗・中傷の書込み
(b)掲示板・ブログ・プロフへの個人情報無断掲載
(c)特定個人になりすましたインターネット活動
2. メールでの「ネット上のいじめ」
(a)メールによる特定の子どもに対する誹謗・中傷
(b)「チェーンメール」での悪口や誹謗・中傷
(c)「なりすましメール」での誹謗・中傷
3. その他

図1　文部科学省によるネット上のいじめの類型化

有害情報の定義ついて説明する.第4章では,SVMによる有害情報の分類について説明する.第5章では,本研究の結論と今後の課題について記述する.

2　基本的な考え方

ネットパトロールは掲示板の書込みを一つ一つ確認して,書込み内容の保存,削除依頼をしている.現状では,無数に存在する書込みの確認が最も負担が大きく問題となっている.また,量の問題だけでなく,書込みによる被害者が存在するか否か,第三者が判断するのは困難であり,疑わしい書込みは全てチェックする必要がある.以上を踏まえ,本研究ではネットパトロールにおける有害情報検出作業を支援する手法の構築を目指している.

現在,提案しようとする手法（図2）の概要について説明する.本手法は,有害情報検出,検出した書込みの有害度によるラン

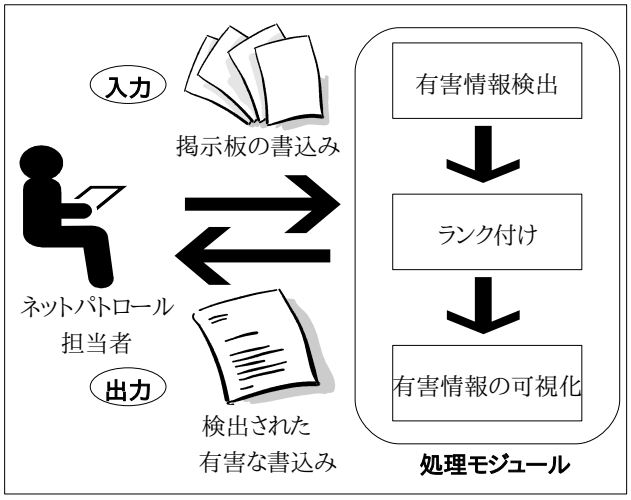


図2　提案手法の概要

キング,有害情報可視化からなる.本紙では,最も基本となる有害情報検出について記述する.有害情報検出では,ネットパトロール担当者に出力すべき,「ネット上のいじめ」に関与する有害情報を検出する.有害情報の定義は第3章で,有害情報検出の具体的な方法は第4章に記述する.

3　有害情報の定義

本章では,有害情報検出処理で検出すべき有害情報の定義について説明する.本研究における有害情報とは,ネット上の掲示板やブログ等の自由に書き込めるスペースにおいて,特定個人の情報を流出したり誹謗中傷等を行い,被害者の実生活に悪影響を及ぼす情報を指す.

実用性を重視するため,実際にネットパトロールに携わっている担当者の助言を受けて有害情報を定義した.作成した有害情報の定義を図3に示す.

有害情報は図3のように3種類に分類できる.harmfulは「松葉達明死ね」のような早急に対処すべき書込みを指す.doubtfulは「大学でうざい奴」のような付近にharmfulな書込みが現れる,またはそれ自体が削除対象となる可能性を含む,チェックしておきたい書込みを指す.normalは無害な書込みである.

・定義の評価実験

作成した定義が,個々の解釈の仕方によって分類判定が分かれるような曖昧な定義になっていないかを評価するため,分類判定の予備実験を行った.

・実験手法

ネットパトロール当事者から頂いた有害・無害情報を含む書込みデータから,無作為に500件抽出し,その書込みを検者に定義を見ながら三つの分類に判別してもらう(書込みデータは三重県域に限定されたサイトから採取したものである).そして判別してもらったデータの一致度(Cohenの Kappa値*,(1)式)を算出し,分類結果にどの程度の差異があるのかを見て,評価を行った.

$$\textit{CohenのKappa値} = \frac{P_o - P_c}{1 - P_c} \quad (1)$$

P_o = *実際に一致した比率*

P_c = *偶然に一致した比率*

・実験結果

実験は成人男性5人,女性1人の計6人に行ってもらった.その結果,一致度は0.67となり,かなりの一致を示した.

・考察

実験結果は強い一致度を示した.この結果から個人の判断に依存する曖昧性はある程度排除されていると考えられる.

本定義によって対象書込みは三つ(harmful(H),doubtful(D),normal(N))のうちどれかに判断される.

・harmful(H):有害(削除)・・・削除依頼を出して対処すべき有害情報である.
・doubtful(D):有害(審議)・・・削除対象となる可能性を含んでいる有害情報候補である.
・normal(N):無害・・・有害情報ではない.

以下,ネットパトロール当事者の助言に基づいた有害情報の定義を示す.

1.個人名の扱い
・個人名そのものの記述　→　H
(例)「松葉　達明」、「松葉」
(メモ)「松葉って凄い人」のように肯定的な表現であっても対象とする.

・イニシャル、ニックネームが記述されている
(個人名に準ずるものと判断)
(例)「松○○明」「エース」
・記述された当事者が特定できる記述　→　H
・記述された当事者が特定できない記述　→　D

・個人の所属や所有物などが記述されている
(例)「北見工大のテキスト情報処理研究室の准教授」,「三重大近くのローソンの女の子」
・記述された当事者が特定できる記述　→　H
・記述された当事者が特定できない記述　→　D

2.個人情報の扱い
(例)「三重県津市一身田一二三」,「090-7112-1234」
(メモ)伏せ字や当字による記述もある.
・一般個人に関する個人情報の記述　→　H
・公共性の高い情報や公開情報の記述　→　D

・個人名や個人情報の書込みを誘導する記述　→　H
(例)「うちの高校のヤリマンって誰?」
(メモ)「～のイケメンは誰?」のような肯定的な表現も対象とする.

・個人に関する情報を提供する記述
(例)「あいつ三重大で○○講義担当してる」,「そいつ確か三重大出身者」
・記述された当事者が特定できる記述　→　H
・記述された当事者が特定できない記述　→　D

3.有害表現の扱い
・誹謗中傷語・暴力誘発語・猥褻語の記述　→　D
(例)「うざい」、「死ね」等の書込みを含む
(メモ)対象者が特定できない場合も含む

・個人間の相互書込みの応酬(匿名も含む)　→　D
(メモ)削除対象に発展する可能性がある

4.以上の条件に合致しない場合　→　N

図3　有害情報の定義

* Cohenの Kappa値とは,単純に何パーセント一致し,何パーセント一致しなかったかという一致度に対し,偶然一致を考慮に入れたものである.Cohenの Kappa値は0～1を推移し,一般的に,値が0.41～0.60の間ならば中等度の一致,0.61～0.80の間ならば強い一致を示し,0.80を超える値をとる場合はほぼ一致していると考えられている.

しかし,「ニート」,「派遣」,「暴走族」等の誹謗中傷として扱うか否か,決めかねる単語で判定に差異が生まれたり,「○○もりまえ○　○か○○」等の記述された当事者が特定できるか否か,現状の定義では非常に判断しづらい書込みで判定に差異が現れていた.よって,有害情報を定義付けるには不十分な部分があるので,今後さらに厳密に定義する必要がある.

4　SVMによる有害情報の分類

掲示板の書込みの有害・無害分類を行うために,二値分類のための機械学習モデルであるSupport Vector Machine (SVM)を用いた[3].SVMは与えられたデータを超空間上で正例集合と負例集合へと分離する際,マージンを最大にすることによって最適な分離超平面を得る学習手法である.SVMの優れた汎化能力と分類能力からテキスト分類の分野でも使われており[4],本研究でも使用する事にした.[**]

・分類実験

SVMによる分類で,どの程度の分類性能を確保できるのか予備実験を行った.

・実験手法

まず,分類実験に使用する書込みデータの説明をする.書込みデータは,第3章の分類判定実験に使用したものと同じで,ネットパトロール当事者から頂いた有害・無害情報を含む書込みデータ2998件の書込みを使用した(書込みデータは三重県域に限定されたサイトから採取したものである).2998件の書込みを図3の定義に従って一つの書込みづつ判別,分類する,そして正例をharmful,doubtfulに該当する書込み,負例をnormalに該当する書込みとしてデータを構築した.その結果,正例が1495件,負例が1503件となった.第3章において,書込みを厳密にはharmful,doubtful,normalの3つに分類している.しかし,有害情報検出処理においてはネットパトロール担当者に確認して欲しい書込み(harmful,doubtfulに該当する書込み)を検出するだけよく,チェック対象の取りこぼしを無いように留意すればよい.その後,ランク付け処理においてharmful,doubtfulによる重み付けで優先して確認して欲しい書込みのランク付けを行う予定である.

次にSVMに利用する素性と特徴量について説明する.まず掲示板の書込みについて形態素解析を行い,分割された形態素ごとに品詞と文字列を特定する.今回は,その特定した品詞と文字列を素性とした.品詞は「連体詞,接頭詞,名詞,動詞,形容詞,副詞,接続詞,助詞,助動詞,感動詞,記号,その他,未知語」の14種類とした.また,素性による分類性能の影響を調べる為,素性を文字列のみ,品詞のみに限定して比較実験を行った.素性の特徴量には,用いたデータ中の出現頻度(2),相対頻度(3),IDF値(4),TF-IDF値(5)の4種類を採用し,比較実験を行った.

$$出現頻度 = 一つの書込みに出現した,ある形態素の頻度 \quad (2)$$

$$相対頻度 = \frac{出現頻度}{全書込みデータ中のその形態素頻度} \quad (3)$$

$$IDF値 = \log\left(\frac{総書込み数}{ある形態素が含まれる書込み数} + 1\right) \quad (4)$$

$$TF-IDF値 = 出現頻度 * IDF値 \quad (5)$$

以上の条件に従って有害情報分類の予備実験を試みた.分類精度評価方法として10分割交差法による適合率(6)と再現率(7)を用いた.10分割交差法とはデータを均等に10分割し,9割で学習を行い,残りの1割を分類するというのを10回繰り返した平均で評価するものである.

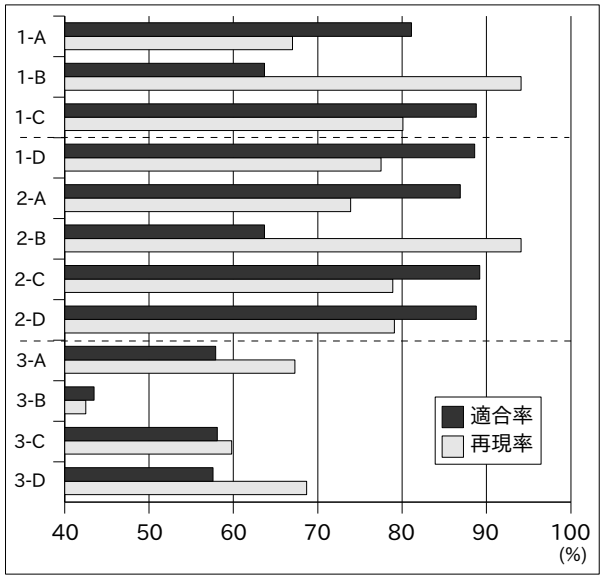
$$適合率 = \frac{s}{n} \quad (6)$$

$$再現率 = \frac{n}{c} \quad (7)$$

s = システムが有害情報と判定した書込みで実際に有害情報だった書込みの総数

n = システムが有害情報と判定した書込みの総数

c = 有害情報書込みの総数



- 1-A：素性(文字列・品詞),特徴量(出現頻度)
- 1-B：素性(文字列・品詞),特徴量(相対頻度)
- 1-C：素性(文字列・品詞),特徴量(IDF値)
- 1-D：素性(文字列・品詞),特徴量(TF-IDF値)
- 2-A：素性(文字列),特徴量(出現頻度)
- 2-B：素性(文字列),特徴量(相対頻度)
- 2-C：素性(文字列),特徴量(IDF値)
- 2-D：素性(文字列),特徴量(TF-IDF値)
- 3-A：素性(品詞),特徴量(出現頻度)
- 3-B：素性(品詞),特徴量(相対頻度)
- 3-C：素性(品詞),特徴量(IDF値)
- 3-D：素性(品詞),特徴量(TF-IDF値)

図4　実験結果

[**] 分類器にはSVM_light(ver6.02)を利用した(http://svmlight.joachims.org/)

・実験結果

実験結果を図4に示す.図4のY軸のラベルは,数値が素性の種類(1:文字列・品詞,2:文字列,3:品詞)を表し,英文字は特徴量の種類(A:出現頻度,B:相対頻度,C:IDF値,D:TF-IDF値)を表している.図3の結果のように,素性が品詞の時を除いて,特徴量を出現頻度にした場合は適合率が80～90%を推移し,再現率は65～75%を推移した.相対頻度の場合は適合率が65%前後,再現率は95%前後だった.IDF値,TF-IDF値の場合は適合率が90%近く,再現率は80%近くになった.

また,素性を文字列に限定すると,特徴量が出現頻度の場合を除いて分類性能に大きな変化は見られず,品詞に限定すると著しく分類性能が低下した.

・考察

まず,特徴量が出現頻度とIDF値,TF-IDF値による場合を見比べるとIDF値,TF-IDF値の方が分類性能が高い事が分かる.そもそもIDF値は,一種の一般語フィルタとして働き,多くの書込みに出現する形態素は重要度が下がり、特定の書込みにしか出現しない形態素の重要度を上げる役割を果たす.よって,IDF値をそのまま使用したり,重み付けに適用すると,有害情報が含む形態素をより特徴付け,分類性能に寄与したと考えられる.

特徴量が相対頻度の場合は再現率が非常に高くなり,適合率は他の特徴量の時に比べて大きく下がった.今回,相対頻度はデータ中の全形態素数で割るのでは無く,個々の形態素総数で算出している.よって,多用される形態素は特徴量は低くなるため,出現頻度の低い形態素によって分類性能が決まる.つまり出現頻度の低い形態素は再現率を上げ,適合率を下げる傾向があると考えられる.

素性については品詞のみに限定した場合,著しく分類性能が低下しているのが分かる.これは品詞の場合,有害情報に出現する特徴的な形態素の品詞が,他の一般的な形態素の品詞と吸収されてしまうため上手く分類できなかったと考えられる.つまり,品詞においては,有害情報に頻出する形態素「死ね」,「殺す」等が,無害な表現である「飛ぶ」,「喋る」等と,動詞として同一視される.よって,品詞は有害・無害情報において出現分布に相関性があり,ランダムに出現していると考えられる.さらに,素性が文字列・品詞,文字列のみの場合を比べて考察してみる.出現頻度が特徴量の場合は,素性を文字列のみにすると適合率,再現率ともに上がり,IDF値,TF-IDF値の場合は大きな差は見られない.つまり,この比較からも出現頻度の場合は有害・無害情報の品詞による普遍化を除去した事が性能の改善に寄与し,IDF値,TF-IDF値は品詞の普遍性をフィルタリングした事が結果に繋がっていると考えられる.よって,品詞を素性として利用するには,品詞単体で使うのではなく,他の品詞との共起を素性にするなどの工夫が必要だと考えられる.

5　おわりに

「ネット上のいじめ」での,誹謗・中傷等を含む有害情報と無害情報の分類を試みた.

まず, 実際にネットパトロールに携わっている担当者に助言を頂き,有害情報の定義を定めた.定義が個々の解釈の仕方によって分類判定が分かれるような曖昧な定義になっていないかを評価するために分類判定実験を行った.Cohenの Kappa値による一致度で評価した所,0.67という強い一致性が見られた.

そしてSVMによる有害情報の分類実験を行った.素性を文字列・品詞,文字列のみ,品詞のみの3種類と,特徴量を出現頻度,相対頻度,IDF値,TF-IDF値の4種類による比較実験を行った.結果,品詞を素性に使うと,有害情報に出現する特徴的な形態素の品詞が,他の一般的な形態素の品詞に同一視されてしまうため上手く分類できない事が分かった.また,特徴量の算出方法による違いでも相対頻度は再現率を上げ,適合率を下げる傾向があり,IDF値による一般的な形態素のフィルタリングで特徴付けて分類性能の改善に寄与する事が分かった.

今後の課題として,より適切な素性候補の発見,特徴量の算出方法等が挙げられる.素性の候補としては,固有名詞・数値表現からなる固有表現が挙げられる.掲示板の書込みにおいて,固有表現は,それだけで有害・無害を判断するような重要な要素となる.固有表現抽出はMUC[5]やIREX[6]でタスクが設定され,盛んに研究が行われており,パターン駆動型[7]やSVM[8]等による抽出手法が提案された.その他にも文脈(対話)を考慮した素性も考えられる.

謝辞

本研究を進めるにあたり,学校非公式サイト情報を提供頂いた財団法人反差別人権研究所みえの松村元樹研究員に感謝致します.

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).文部科学省,2008.

[2] 渡辺凡,砂山渡:電子掲示板におけるユーザの性質の評価.電子情報通信学会技術研究報告,No.652 in 2006-KBSE,pp.25-30,2006.

[3] Vapnik,V.:StatisticalLearnig Theory,Springer(1998).

[4] 平 博順,向内隆文,春野雅彦:Support Vector Machineによるテキスト分類,情報処理学会研究報告.自然言語処理研究会報告,IPSJ SIG Notes 98(99) pp.173-180

[5] DARPA. Proceedings of the Tipstar Text Program PhaseⅢ 18 month workshop. DARPA,1998.

[6] IREX 実行委員会(編).IREX ワークショップ予稿集

[7] 桝井文人,鈴木伸哉,福本淳一: "テキスト処理のための固有表現抽出ツールNExT の開発",第8 回言語処理学会年次大会発表論文集,pp.176-179, 2002.

[8] 山田寛康, 工藤拓, 松本裕治: Support Vector Machineを用いた日本語固有表現抽出,情報処理学会誌, Vol.43, No.1, pp.43-53, 2002