1H4-NFC-01a-3

電子掲示板からの文脈を考慮した誹謗中傷コメントの抽出

Harmful Comments Extraction from a Bulletin Board System
- Word Harmfulness Varies depending on Thread Context -

西原陽子 Yoko Nishihara　岩佐一樹 Kazuki Iwasa　福本淳一 Junichi Fukumoto　山西良典 Ryosuke Yamanishi

立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University

Harmful documents on the Web make readers unpleasant. Such documents have been filtered by machine learning methods which learn words used in harmful documents frequently (i.e. black words). The previous methods often fail to filter documents if documents include the black words, but are not harmful. Whether a document is harmful or not varies depending on the context of a document. The context is necessary for filtering harmful documents precisely. This paper proposes a new extraction method of harmful comments from a Bulletin Board System by using the context of a thread. The method extracts harmful comments by two ways. (1) If a black word is included in a comment, the method extracts a comment as harmful. (2) If a word in a comment and a black word appear in previous posted comments frequently, the method adds a word to a black word list and filters a comment as harmful. We evaluated the proposed method. Comments used in the evaluation were those from four threads in Japanese BBS "2-channel." The average of precisions in extraction was 0.47, and the average of recalls was 0.68.

1. はじめに

ウェブ上には日々多数の情報が投稿される．投稿される情報の中には有用なものが多数存在するが，一方で他者を誹謗中傷する情報も存在する．他者を誹謗中傷する情報を放置しておくと，別の人が情報の発信者に対して誹謗中傷する情報を投稿し，更にまた別の人が誹謗中傷の情報を投稿をするなど，誹謗中傷をする多数の情報が短時間の間に連続して投稿されてしまうことがあり，有用な情報の獲得を妨げてしまう．このような争いを避けるためには，他者を誹謗中傷する情報が投稿されたら，できるだけ早期に取り除いてしまうことが望ましい．

誹謗中傷を含む有害な情報を抽出する従来手法にコンテンツをチェックする手法がある．この手法では単語を元にして有害情報の抽出を行うが，ある単語の意味が使用されている文脈で異なる場合，抽出に失敗する恐れがある．「小学生」という単語を例として挙げる．小学生が見る番組について好意的な意見が多数出されている文脈において，「小学生でも楽しめる番組だよね」という文があるとき，「小学生」は本来の意味で使用されており，文も他者を誹謗中傷するものではない．しかし，ある番組に対して誹謗中傷をする意見が多数出されている文脈において，「誉めてるのは信者の脳内に住んでる小学生くらいだもんな w」という文があるとき，「小学生」はファンの精神的な年齢が幼いという意味で使用されており，文は他者を誹謗中傷するものとなる．単語の意味が文脈によって変化することを考慮して，有害情報を抽出することが望ましい．

本研究では電子掲示板に投稿されるコメントの中から，他者を誹謗中傷する文を含むコメントを抽出する手法を提案する．本研究で抽出したいコメントは，スレッドのコメントを読み書きする人や，コメントの中で話題に挙げられている人を誹謗中傷するコメントとする．提案する手法では他者を誹謗中傷する際に使用される単語（バッドワード）と，スレッドの文脈に応じて誹謗中傷する際に使用されることがある単語（スレッドバッドワード）をリストとして用意する．ある文がバッドワードかスレッドバッドワードを含むならば，誹謗中傷をする文と評価して，文を含むコメントを抽出する．時系列順に文を評価する中で，ある単語がバッドワードと共に使われることが多くなってきたらスレッドバッドワードとしてリストへ追加され，少なくなってきたらリストから削除される．

連絡先: 西原陽子，立命館大学情報理工学部，525-8577 滋賀県草津市野路東 1-1-1，nisihara@fc.ritsumei.ac.jp

2. 従来研究：有害情報の抽出

有害情報の抽出方式は大きく 2 つに分けられ，URL を利用する方式とコンテンツをチェックする方式になる．URL を利用する方式では有害な情報を含む Web ページの URL をブラックリストに登録しておき，ブラックリストに載っている Web ページを非表示とする．有害な情報が定常的に掲載されている Web ページに対して有効に働く方式であるが，新しく立ち上がったばかりの Web ページや，ブログのように同一ドメインの下に有害な情報と無害な情報が混在している場合には適用が難しい．本研究で抽出の対象とする情報は，電子掲示板のスレッドの中にあるコメントである．1 つのスレッドは 1 つの URL を持っていることが多いが，スレッドの中に有害なコメントと無害なコメントが混在しているため，URL を利用する方式は使えない．そのため本研究ではコンテンツをチェックする方式を利用する．

コンテンツをチェックする方式では，コンテンツの中に不適切な単語や語句が含まれているかをチェックし，含まれていれば抽出する．Grailheres らはベイジアンフィルタを利用したスパムメールの抽出手法を提案している [Grailheres 04]. Grailheres らの手法は，スパムメールと非スパムメールに出現する文字列の確率を学習し，スパムメールを抽出する．コンテンツと URL の情報を両方を用いて抽出する方式 [井ノ上 01] や，コンテンツを記述する際の HTML タグ内の文字列とコンテンツの両方を用いて抽出する方式 [池田 11] なども提案されている．

コンテンツをチェックする方式では，多くの手法が抽出を行う前に正例と負例を学習する．学習には大量のデータが必要となり，人手で用意することには大きなコストがかかる．学習用データの自動生成を支援する手法も提案されているが [吉川 10]，どこかで人の手を加えねばならない．Web 上の情報は日々新

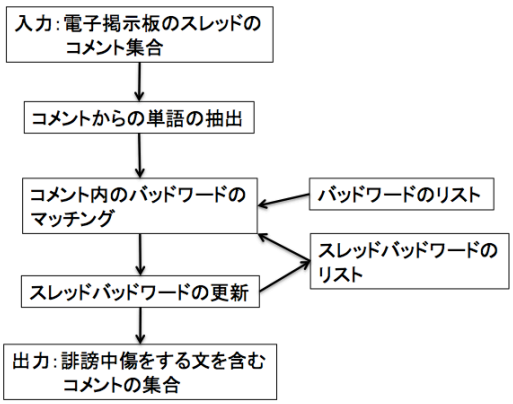

図 1: 提案手法の処理の概要.

しくなり，使われる単語やその組み合わせは変化している．過去に用意した正例が未来においても正例として利用可能とは限らない．人手を介して学習用のデータを日々作って行くことは大きな負担となる．本研究では学習用のデータは用いず，65 個のバッドワードとコメントに含まれる単語を用いて抽出する手法を提案する．65 個のバッドワードは文脈によらず，使用されることにより直ちに誹謗中傷に使われる単語である．これに加えて，どのようなスレッドが与えられても，バッドワードを元にしてスレッドバッドワードのリストを順次作成していくことにより，文脈に応じた抽出を実現する．

3. 提案手法：文脈を考慮した誹謗中傷コメントの抽出

提案手法の処理の概要を図 1 に示す．提案手法に電子掲示板の 1 つのスレッドに投稿されたコメントの集合が入力されると，提案手法はコメント内の各文を形態素解析にかけ，文から単語を抽出する．文から抽出された単語とバッドワード，スレッドバッドワードのマッチングをとり，一致するものがあれば，提案手法は誹謗中傷をする文であると評価する．その後，スレッドバッドワードのリストを更新し，最後に誹謗中傷をする文を含むコメントの集合を出力する．

3.1 入力：電子掲示板のスレッドのコメント集合

提案システムに入力するコメント集合の例を表 1 に示す．電子掲示板のスレッドでは 1 つ目のコメントに話題となる文章が記載され，2 つ目以降のコメントに話題に対する意見が記載されることが多い．本研究では 1 つ目のコメントを話題が記述されたコメント（話題コメント）とし，2 つ目以降のコメントを意見が記述されたコメント（意見コメント）と区別し，意見コメントのみを抽出の対象とする．

3.2 コメントからの単語の抽出，およびバッドワードとのマッチング

本研究で抽出する単語は名詞のみとし，抽出には形態素解析器の茶筌 [松本 00] を用いる．

意見コメントを c_i（$0 \leq i \leq N$，ただし，N は意見コメントの数）とし，c_i に含まれる単語の集合を W_i とする．W_i の各単語 w_j（$0 \leq j \leq M$，ただし，M は意見コメント c_i 中の単語の数）とバッドワードのリストにある単語のマッチングをとる．少なくとも 1 つの単語 w_i がバッドワードのリストにあれば，提案手法は意見コメント c_i を誹謗中傷をするコメントと評価する．

表 1: 提案システムへの入力の例．0 番目のコメントは話題を表すコメント，1 番目以降のコメントは意見を表すコメントとし，本研究では 1 番目以降のコメントを抽出の対象とした．コメントは 2 ちゃんねるのスレッド「仮面ライダー鎧武アンチスレ 23」より抽出された．

番号	コメント
0	テレビ朝日公式サイト ttp://www.tv-asahi.co.jp/gaimu/ 東映公式サイト ttp://www.toei.co.jp/tv/gaimu/ 前スレ 仮面ライダー鎧武アンチスレ　22 http://toro.2ch.net/test/read.cgi/sfx/1384140649/ ※ 原則として＞＞ 950 を取った人が次スレを立てて下さい． ただし，放送の前後はスレの進行が速いため，＞＞ 900 を取った人が立ててください． ＞＞ 950 を取った人は，スレ立て宣言，もしくは不可宣言をしてください． ＞＞ 950 が不可の場合，以降にスレ立てをする際には有志の方が宣言をしてからでお願いします．
1	関連スレ BL@DRAMAtical Murder 145【Nitro+CHiRAL】 http://kilauea.bbspink.com/test/read.cgi/gagame/1381930248/
2	＞＞ 1 乙 バナナで転倒をやってるゴーバスと バナナをちっとも活かせてない鎧武を比べるのは失礼だな
3	＞＞ 2 それやったら，他のライダーや特撮スレもリンクしなきゃいけなくなるだろｗ

バッドワードのリストは 2 種類ある．1 つはスレッドの内容に関係なく，その単語が使われることにより他者を誹謗中傷する可能性が高い単語を集めたリストである．これを単にバッドワードのリストと呼ぶ．もう 1 つはスレッドの内容に応じて作られて行くリストであり，本研究ではこれをスレッドバッドワードのリストと呼ぶ．

3.2.1 バッドワードのリスト

本研究で使用するバッドワードのリストに登録されている単語を表 2 に示す．これらの単語は次の手順で著者の一人により集められた．初めに電子掲示板に投稿されたコメントの中から，他者を誹謗中傷するコメントを選択した．続いて，コメントの中で使われている単語の中から，誹謗中傷に関連すると思われるものを選択した．最後に選択された単語の中から，使われることによって他者を誹謗中傷する可能性が高いものに絞り込んだ．最終的に残った単語は 65 個であった．本研究ではこれらの単語をバッドワードとして用いる．

3.2.2 スレッドバッドワードのリスト

スレッドバッドワードはコメントに含まれる単語の中から選択される．ある単語がコメントの中でバッドワードと共起する割合が高くなってきたらリストに追加され，割合が低くなってきたらリストから削除される．スレッドバッドワードをリストへ追加，リストから削除する 2 通りの方法を説明する．意見コメント c_i よりも前に投稿された意見コメントの集合を PM とする．

1. 追加のみ：単語 w_i が話題コメントに含まれ，かつ W_i 内にバッドワードまたはスレッドバッドワードが少なくと

表 2: 本研究で作成したバッドワードのリスト．65 個の単語が登録されている．内 10 個の単語は性に関わるものであったので，表示からは外した．

ブス，ホモゲ，クソ，糞，自殺，アホ，アフォ，ドアホ，阿呆，バカ，馬鹿，ボケ，クズ，屑，カス，キチガイ，マジキチ，基地外，キモ，キモイ，ウザ，ウザイ，老害，パクリ，パク，フルボッコ，グロ，無能，DQN，ブサイク，不細工，ブサ，駄作，愚作，ババア，ババァ，ブチギレ，イライラ，NG，ショタ，ショボい，しょぼい，基地外，基地害，鬱，ブヒ，ブヒブヒ，アンチ，ワロタ，ワロス，ダサい，ダサイ，イラネ

も 1 つ存在する場合，単語 w_i をスレッドバッドワードとしてリストに追加する．

2. 追加と削除：コメントの集合 PM において，1 つのコメントの中に単語 w_i とバッドワード，またはスレッドバッドワードが共に含まれる割合 r_i を算出する．割合 r_i が閾値 K 以上であれば，単語 w_i をスレッドバッドワードのリストに追加する．反対に割合 r_i が閾値 K 未満であれば，単語 w_i をリストから削除する．

1. の処理では単語が話題コメントに含まれ，かつバッドワードと共に使われているならば，直ちにスレッドバッドワードのリストへ追加する．ある単語が話題コメントに含まれるならば，その単語はスレッドの話題に強く関係すると考えられる．強く関係する単語がバッドワードと共に使われている場合，その単語自身が誹謗中傷を意味するものとなる可能性が高い．この理由により 1. の処理がある．

2. の処理では単語が話題コメントに含まれないが，バッドワードと共に使われているならば，それまでの投稿コメントにおいてバッドワードと共起する割合を評価して，追加，削除を行う．スレッドの文脈に応じて同じ単語でも誹謗中傷に使われる場合とそうでない場合がある．本研究ではコメントの中でバッドワードと共起する割合が高ければ追加し，割合が低くなればリストから削除することにより，スレッドバッドワードのリストを更新していく．

スレッドバッドワードのリストに追加する単語は，その頻度がコメント集合 PM における単語の頻度の平均よりも高いものに限定する．これは，頻度が低い単語はスレッドの話題と関係が弱く，単独で使われた場合，他者を誹謗中傷する意味になることが少ないと考えられるためである．

3.3 出力：誹謗中傷をする文を含むコメントの集合

提案手法の出力例を表 3 に示す．表 3 には抽出された意見コメントの例と，抽出の根拠となったバッドワードが記載されている．バッドワードだけでなく，スレッドバッドワードによっても意見コメントが抽出されている．

4. 提案手法の評価実験

提案手法を用いて誹謗中傷をする文を含むコメントを抽出する実験を行い，手法の評価を行った．

4.1 実験手順

実験者は以下の手順により実験を行った．

1. 電子掲示板のスレッドからコメントを抽出した．

2. 提案手法にコメントを入力し，誹謗中傷する文を含むコメントを抽出した．

表 4: 評価実験に使用した電子掲示板のスレッドのタイトル，コメントの数，正例の数．4 つのスレッドは電子掲示板 2 ちゃんねるから取得された．

スレッドのタイトル	コメント数	正例数
日常のアンチスレ 30	923	495
◆週刊少年ジャンプ総合スレッド◆ Part490	998	221
仮面ライダー鎧武アンチスレ 23	998	333
食戟のソーマアンチスレ 10	997	272

表 5: 提案手法による抽出の適合率，再現率．

スレッドのタイトル	適合率	再現率
日常のアンチスレ 30	0.54	0.86
◆週刊少年ジャンプ総合スレッド◆ Part490	0.68	0.23
仮面ライダー鎧武アンチスレ 23	0.36	0.74
食戟のソーマアンチスレ 10	0.28	0.88
4 つのスレッドの平均	0.47	0.68

3. 被験者にコメントを提示し，他者を誹謗中傷する文を含むコメントを選択する旨を依頼した．

4. 1 人以上の被験者が選択したコメントを正例，それ以外を負例とした．

5. 提案手法が正例を抽出する適合率と再現率を算出した．

実験手順の 1. で用意したスレッドのタイトルを表 4 に示す．表 4 のスレッドはいずれも電子掲示板 2 ちゃんねるにあったものである．他者を誹謗中傷する文が多そうなスレッドを選択した．2 ちゃんねるは最大 1000 件までコメントを書き込むことができるが，1000 件まで書き込まれたスレッドのデータにはアクセスできないことが多かったため，実験時に取れる内で最大の件数を取得した．このため，入力したコメントの数はスレッドごとに異なっていた．

実験手順の 3. でコメントの選択を依頼した被験者は，情報理工学部に所属する大学生 7 名（男性 6 名，女性 1 名）であった．各被験者は 4 つのスレッドの全てのコメントを読み，他者を誹謗中傷していると思った文を含む意見コメントを選択した．

実験手順の 5. の適合率と再現率は式 (1) と式 (2) により算出された．

$$\text{適合率} = \frac{\text{出力に含まれた正例の数}}{\text{提案手法が出力したコメントの数}} \quad (1)$$

$$\text{再現率} = \frac{\text{出力に含まれた正例の数}}{\text{正例の数}} \quad (2)$$

4.2 実験結果

提案手法による抽出の適合率，再現率を表 5 に示す．適合率は平均 0.47，再現率は平均 0.68 であった．

4.3 考察

はじめに，提案システムの適合率と再現率について考察する．適合率の平均は 0.47 であった．適合率が低かった原因としては，スレッドバッドワードを元にして抽出されたコメントの中に負例が多く含まれたことがあげられる．表 6 にスレッドバッドワードにより抽出されたコメントの数，適合率，再現

表 3: 提案システムが出力したコメント，および抽出の根拠となったバッドワードの例．この例は 2 ちゃんねるのスレッド「仮面ライダー鎧武アンチスレ 23」のデータを入力して得られたものの一部である．

番号	コメント	バッドワード	スレッドバッドワード
23	陰謀論は嫌いだが，こんなパクリ作品を社長が OK している段階で，会社ぐるみの詐欺会社というのは明確なんだけどな．	パクリ	（なし）
143	無職が暴れるってまさに鎧武だよなあ，紘汰とか戒斗とか	（なし）	無職
673	虚淵のコピー人間が出来上がる	（なし）	虚，淵
800	つーか何であの無能おばちゃん未だに切られないんだろ？	無能	（なし）

表 6: スレッドバッドワードにより抽出されたコメントの数，適合率，再現率．

スレッドのタイトル	コメント数	適合率	再現率
日常のアンチスレ 30	674	0.52	0.71
◆週刊少年ジャンプ総合スレッド◆ Part490	10	0.30	0.01
仮面ライダー鎧武アンチスレ 23	561	0.27	0.46
食戟のソーマアンチスレ 10	709	0.23	0.60
4 つのスレッドの平均	489	0.34	0.45

表 7: 「◆週刊少年ジャンプ総合スレッド◆ Part490」を入力した時に提案手法が抽出できなかった正例の例．

番号	コメント
70	無知なコンビニ店長が入荷して大惨事になるんだろうな．
94	女装男子キター
169	ワンピースっていつまでトリコに寄生してんの？
198	中二病なんだろ
261	壊れたレコーダーみたいなのが堪えないな 変な宗教でもやってるのか，精神病でも患ってるのか
321	悪意ある第三者が偽造し放題だな

率を示す．スレッドバッドワードにより抽出された文の数は，バッドワードにより抽出された文の数よりも多く，ほとんどは負例であった．このため適合率の平均も 0.34 と，全体平均の 0.47 よりも低くなっている．適合率を向上するためには，スレッドバッドワードにより抽出してしまう負例の数を減らす必要があることが分かった．

再現率の平均は 0.68 であった．他者を誹謗中傷するコメントをできるだけ抽出して，読み手の有用な情報の獲得を支援するという目的においては再現率が高いことが望ましい．このことから，提案手法による誹謗中傷コメントの抽出は一定の目的を達成したと考えられる．

「◆週刊少年ジャンプ総合スレッド◆ Part490」（「ジャンプ」）を入力した際に得られた再現率は 0.23 であり，他の 3 つのスレッドの再現率よりも低かった．「ジャンプ」を入力した際の再現率が低くなった原因としては，バッドワードのリストに含まれる単語が正例の中で使われることが少なかったことがあげられる．表 7 に「ジャンプ」を入力した際に抽出できなかった正例の例を示す．表 7 の 94 番のコメントに含まれる単語は「女装，男子」の 2 つである．94 番のコメントには表 2 のバッドワードが含まれておらず，「女装」「男子」の 2 つの単語もスレッドバッドワードのリストに追加されていなかった．このため提案手法は 94 番のコメントを抽出しなかった．表 7 の他のコメントが抽出されなかったことも同様の理由による．表 7 に示したコメントを抽出する方法としては，タイトルや内容が類似したスレッドを自動的に取得し，それらのスレッドに含まれるコメントも利用してスレッドバッドワードの追加，削除を行う方法が考えられる．

5. おわりに

本研究では電子掲示板に投稿されるコメントの中から，他者を誹謗中傷する文を含むコメントを抽出する手法を提案した．提案した手法は使用することで他者を誹謗中傷する可能性が高い単語をバッドワード，スレッドの文脈に応じて他者を誹謗中傷する可能性がある単語をスレッドバッドワードとし，いずれかのバッドワードを含む文を他者を誹謗中傷する文と評価して，その文を含むコメントを抽出する．電子掲示板 2 ちゃんねるから 4 つのスレッドを選び，スレッド内のコメントに対して提案手法を適用したところ，平均して適合率 0.47，再現率 0.68 により誹謗中傷をする文を含むコメントが抽出できることを確認し，一定の目標を達成できた．今後は適合率の向上を目指して，スレッドバッドワードの追加と削除のアルゴリズムを改善していく．

参考文献

[Grailheres 04] B. Grailheres, S. Brunessaux, P. Leray, Combining Classifiers for Harmful Document Filtering, RIAO 2004, pp.173–185 (2004).

[池田 11] 池田和史, 柳原正, 服部元, 松本一則, 小野智弘, 滝嶋康弘, HTML 要素に基づく有害サイト検出手法, 情報処理学会論文誌, Vol.52, No.8, pp.2474–2483 (2011).

[井ノ上 01] 井ノ上直己, 帆足啓一郎, 橋本和夫, 文書自動分類手法を用いた有害情報フィルタリングソフトの開発, 電子情報通信学会論文誌 D-II, Vol. J84, No.6, pp.1158–1166 (2001).

[松本 00] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬, 浅原 正幸, 日本語形態素解析システム『茶筌』version 2.2.1 使用説明書 (2000).

[吉川 10] 吉川幹人, 佐藤翔平, 関和広, 上原邦昭, リンク構造とコンテンツを複合的に用いた極小訓練事例によるスプログ検出, 情報処理学会論文誌 データベース, Vol.3, No.1, pp.29–37 (2010).