計測自動制御学会東北支部 第 254 回研究集会 (2009.12.17)
資料番号 254-7

電子掲示板における不適切発言の評価方法に関する研究

Evaluation for vandalism comment in BBS

○一藤 裕*，今野 将**，曽根 秀昭*

○ Yu Ichifuji*, Susumu Konno**, Hideaki Sone***

*東北大学, **千葉工業大学

*Tohoku University of Technology, **Chiba Institute of Technology

キーワード： 電子掲示板 (Electronic bulletin board system) ベイズ理論 (Bayesian theorem)，情報倫理 (information ethics), 利用者補助 (User assistance)

連絡先： 〒 980-8578 仙台市青葉区荒巻字青葉 6–3 東北大学 サイバーサイエンスセンター 4F 曽根・水木研究室

一藤 裕，Tel.: (022)795-6094，Fax.: (022)795-6096，E-mail: ichifuji@mail.tains.tohoku.ac.jp

---

1. はじめに

電子掲示板は、インターネット上のコミュニケーションツールの 1 つである。電子掲示板を利用することにより、不特定多数のユーザと非リアルタイムでコミュニケーションをとることや情報交換を行うことができる。特に、匿名性を有することが多いため、相手の立場に関係なく本音で議論を行うことや、普段言うことが難しいことも気軽に発言することができる。そのため、電子掲示板の発言には、通常のコミュニケーションを目的とした発言（以下、通常発言）のほかに、匿名性を悪用した他人を不愉快にすることを意図した発言（以下、不適切発言）がある。また、不快な単語を含むがコミュニケーションを意図した通常発言とも不適切発言ともとれる発言（以下、曖昧発言）も存在する。

不適切発言は、通常コミュニケーションを阻害し利用者を遠ざけるため、通常コミュニケーションを維持するためには、不適切発言を減少させる必要がある。

不適切発言には、2 種類に分類することができる。一つは、意図的に不適切発言を行う場合である。もう一つは、過失で行う場合である。意図的に不適切発言を行うユーザは、注意しても不適切発言をやめることはない。しかし、過失の場合、ユーザの経験不足など理由により、その発言が相手にどう受け取られるかを把握できていないために生じる。したがって、掲載前にその発言が不適切発言であることをユーザへ提示できれば、その発言を推敲、もしくは、とりやめる可能性がある。つまり、不適切発言を減少させることが可能となる。

以上より、電子掲示板の通常コミュニケーションを維持するためには、不適切発言を投稿するまえにユーザに対し注意喚起をすることが有効な手段の一つであると言える。そのためには、投稿発言が通常発言か不適切発言か分類する手法が必要である。そこで、本稿では、ユーザが発言を行う前に、その発言が不適切発言か通常

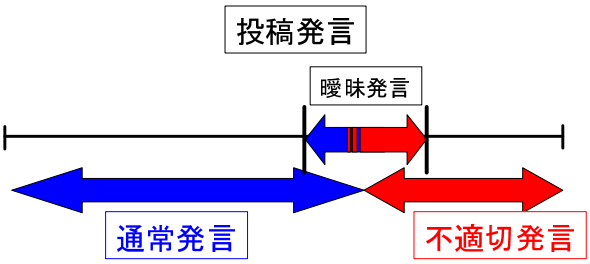

Fig. 1　3 種類の発言の関係

発言かを自動的に分類する手法を提案する。

2.　発言分類のための着眼点

電子掲示板には、通常発言・不適切発言とそのどちらにもなりえる曖昧発言の 3 種類があり、Fig.1 のような関係となっている。通常コミュニケーションを維持するため本稿では、投稿された発言が不適切発言と通常発言の 2 種類に分類することを目指す。

2.1　投稿発言の分類方法

電子掲示板での発言に関する特徴として、文字や記号で構成されている点が挙げられる。そのため、発言を分類する手掛かりとして、発言に使われている単語が非常に重要であると言える。しかし、単純に単語のみを基準とするだけでは、分類することは難しい。なぜなら、不適切発言を引用した発言や、不適切な単語をあえて褒め言葉として使うなど文法や本来の意味にとらわれない使われ方をするからである。

また、掲示板の発言間隔に関する特徴として、チャットと同様にほぼリアルタイムコミュニケーションをがなされる点が挙げられる。また、事件などの最新ニュースが掲示板のトピックの場合、数分程度で 1000 発言以上書きこまれることもある。そのため、投稿してから掲載されるまで時間がかかると、通常のコミュニケーションを阻害する結果となり、ユーザ離れを引き起こすことにもなりかねない。

以上より、なるべく軽い処理で発言を分類する方法が望ましい。そこで本稿では、電子メールにおける Spam メールフィルタに着目する。そのなかでも特に、軽い処理で判別精度が高いベイジアンフィルタ [1)2)] の応用を考える。ベイジアンフィルタはあらかじめ Spam メールに使われる頻度の高い単語を学習し、そのデータを用いて判別を行う。電子掲示板の発言も電子メールと同様に文字や記号で構成されているため、不適切発言を電子メールにおける Spam メールと同じように分類することが可能ではないかと考えたからである。

しかし、Spam メールは広告・勧誘が目的であり、ある程度出現する単語が限定されるが、不適切発言は状況により変化する。したがって、ベイジアンフィルタをそのまま適応させるだけでは、分類精度が不十分となる可能性が高い。そこで、単語だけでなく、単語のペアを利用することを考える。なぜなら、電子掲示板の発言には他の発言を引用し、その発言はよくないという意思を伝える場合などがあるからである。つまり、そのような発言を単語のみで分類すると、不適切な単語が多く占める発言のため、注意する発言を不適切発言として分類することになってしまうからである。しかし、単語のペアを作成することにより、不適切単語とそれを否定する単語の組み合わせが出現することにより、不適切発言として分類されなくなるのである。

以上より、学習データとして、不適切発言に出やすい単語および単語のペアを用意し、2 通りの判別を行う。その後、判別結果の組み合わせから、投稿発言を Table 1 の通り、それぞれ White 発言・Gray 発言・Black 発言の 3 種類に分類する。White 発言は通常発言に、Gray 発言は曖昧発言に、Black 発言は不適切発言に該当する。

Table 1　投稿発言の分類

単語	ペア	分類結果	該当発言
○	any	White	通常発言
×	○	Gray	曖昧発言
×	×	Black	不適切発言

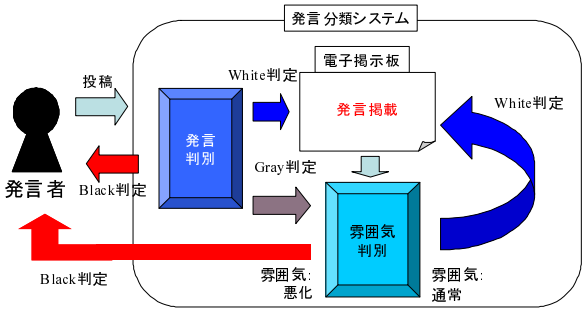

Fig. 2　発言分類方法の概要

2.2　曖昧発言の評価方法

曖昧発言は、不適切発言にも通常発言にも受け取られる発言であると述べた。そこで、Gray 発言を White 発言、Black 発言に分類することを考える。

3.　発言分類手法の提案

発言分類手法は、Fig.2 のように、発言を White・Gray・Black 発言の 3 種類に分類する発言判別手法と、電子掲示板の雰囲気判別手法の 2 つで構成されている。

3.1　発言判別手法

判別手法は以前、我々が提案した手法を使用する [3]。これは、Fig.3 のように、判別フェーズと学習フェーズで構成されている。学習フェーズでは、主観評価によって、あらかじめ通常発言と不適切発言を用意し、それぞれに出現しやすい単語および単語のペアを学習させる。

判別フェーズでは、学習フェーズで学習したデータを用いて、投稿発言を White 発言・Gray 発言・Black 発言の 3 種類に分類する。

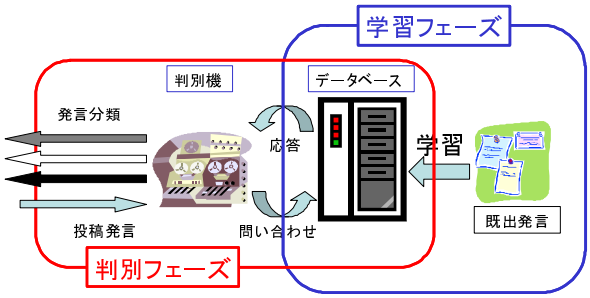

Fig. 3　発言分類方法の概要

3.2　雰囲気判別手法

電子掲示板の雰囲気の判別方法は、以前我々が提案した手法を使う [4]。

電子掲示板の雰囲気は、発言の蓄積によって形作られる。たとえば、不適切発言が集中する箇所は、その掲示板の雰囲気が悪化していると言える。そこで、発言中に使われる単語に着目し、発言の雰囲気を判別する。具体的には、雰囲気を悪くする単語（例：“死ね”、“消えろ”）と雰囲気を良くする単語（例：“ありがとう”、“いいね”）を用いて、雰囲気の判別を行う。

3.3　曖昧発言判別手法

発言判別手法によって Gray 発言と分類された発言を、White 発言と Black 発言のどちらかに電子掲示板の雰囲気を判断基準として分類する。雰囲気が悪化している場合、Gray 発言は Black 発言に分類する。反対に、雰囲気が通常または良い場合、Gray 発言は White 発言に分類する。

このように発言判別手法、雰囲気判別手法および曖昧発言判別手法の 3 種類を用いて投稿発言を White 発言と Black 発言に分類する。

4.　検証実験

提案した発言分類手法の妥当性を示すために、実際の掲示板を利用して発言の分類を行い、White 発言が主観評価による通常発言とどれだけ一致

Table 2　発言判別手法の結果

分類結果	通常	不適切	検証発言数
White 発言	646	18	664
Gray 発言	248	54	302
Black 発言	102	28	130
Total	996	100	1096

するか、Black 発言が主観評価による不適切発言とどれだけ一致するかの実験を行った。

4.1　実験対象掲示板

検証実験を行うために、学校裏サイトと呼ばれる電子掲示板の中から無作為に抽出した。本稿では、1 つの掲示板を例に挙げる。対象とした発言は、2007 年 5 月から 2008 年 11 月までの全 1096 発言である。この掲示板では、挑発・誹謗中傷のような不適切発言の連鎖と通常コミュニケーションが繰り返し発生していた。すべての発言は、あらかじめ複数人の成人男性に読んでもらい主観評価を行っている。

4.2　実験結果

1096 発言に対し、発言判別を行い 3 種類の発言に分類した。結果は、Table 2 となった。表中の “通常” は、主観評価で通常発言と評価された数を示しており、また、表中の “不適切” は、主観評価で不適切発言と評価された数を示している。

その後、Gray 発言を分類するために、雰囲気の判別を行った。雰囲気の変化を視覚的にとらえるために、Fig.4 を示す。Gray 発言と雰囲気の対応は、Table 3 のようになった。表中の “通常” は、主観評価によって通常発言と評価された発言数を示しており、また、表中の “不適切” は、主観評価によって不適切発言と評価された発言数を示している。“雰囲気’ ’は雰囲気判別手法により判別された結果を示している。

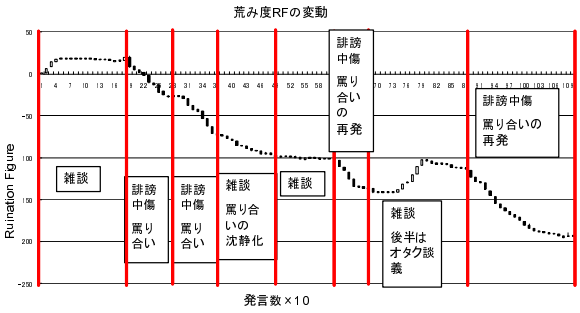

Fig. 4　掲示板の雰囲気の変化

Table 3　雰囲気の変化による Gray 発言と主観評価の比較

範囲	雰囲気	Gray	通常	不適切
1〜180	通常	39	39	0
181〜370	悪化	62	33	29
371〜610	通常	58	56	2
611〜680	悪化	26	21	5
681〜870	通常	40	39	1
871〜1097	悪化	77	60	17

Table 3 より、雰囲気が通常であった 1〜180、371〜610、681〜870 の範囲における Gray 発言は、主観評価においても、137 発言中 134 発言が通常発言と評価される結果となった。また、雰囲気が悪化となった 181〜370、611〜680、871〜1097 の範囲における Gray 発言は、不適切発言の割合が雰囲気が通常の場合と比べて増加する結果となった。

電子掲示板の発言を White 発言と Black 発言に分類した結果と主観評価の比較を Table 4 に示す。Table 4 より、主観評価により通常発言と評価された 996 発言中 780 発言 (78.3%) を分類することができた。また、主観評価により不適切発言と評価された 100 発言中 79 発言 (79.0%)

Table 4　発言分類手法と主観評価の比較

分類結果	主観 [通常]	主観 [不適切]
White 発言	780	21
Black 発言	216	79
一致率	78.3%	79.0%

を分類することができた。

以上より、提案手法はこの掲示板の発言を約80%の精度で発言を分類できる結果となった。

5. まとめ

電子掲示板では、不適切発言によって通常コミュニケーションを阻害する問題がある。通常コミュニケーションを維持するためには、不適切発言を減らす必要がある。不適切発言には、過失によるものがあり、そのような発言をするユーザに対し、不適切発言であることを示すことができれば、不適切発言を減少させることにつながると考えた。したがって、投稿発言が不適切発言と通常発言に分類する手法が必要であった。本稿では、不適切発言と通常発言に分類するために、発言判別手法と雰囲気判別手法を組み合わせた発言分類手法を提案した。発言判別手法では発言を White 発言、Gray 発言、Black 発言の 3 種類に分類し、雰囲気判別手法を利用して Gray 発言を White 発言と Black 発言に分類した。その結果、投稿発言を通常発言を 78.3%、不適切発言を 79.0%の精度で分類することが可能であることを示した。

参考文献

1) P. Graham: A Plan for SPAM, http://paulgraham.com/spam.html
2) 田端 利宏: SPAM メールフィルタリング：ベイジアンフィルタの解説，情報の科学と技術，56(10)，464-468 (2006)
3) 一藤 裕，今野 将，曽根 秀昭: 掲示板の発言に対する自動判別を用いたユーザ教育支援手法の改良，電子化知的財産・社会基盤研究会，Vol.108 No.459，pp.219-224 (2009)
4) Yu Ichifuji, Susumu Konno and Hideaki Sone: A Method to Monitor a BBS Using Feature Extraction of Text Data, Proceedings of 3rd International Conference on Human.Society@Internet, LNCS 3597, pp.349-352 , Japan, (2005)