2007－ＥＩＰ－36（４）
2007／5／25

BBSにおける発言の相関に基づいたコミュニケーション阻害行為の発見手法の提案

一藤 裕† 今野 将† 曽根 秀昭††

† 東北大学大学院情報科学研究科 〒980-8578 宮城県仙台市青葉区荒巻字青葉6-3
†† 東北大学情報シナジーセンター 〒980-8578 宮城県仙台市青葉区荒巻字青葉6-3
E-mail: †ichifuji@mail.tains.tohoku.ac.jp, ††{skonno,sone}@isc.tohoku.ac.jp

あらまし 現在，インターネット上の電子掲示板において，故意・過失にかかわらず他の参加者を不快にする発言がたびたび出現している．このような発言はコミュニケーション阻害行為と呼ばれ，繰り返されると，掲示板は荒れ，利用者の減少などを招く．そのため，このような行為を発見する手法が必要とされている．そこで本論文では，各発言が荒らし行為である確率を単語の出現確率からベイズの定理を用いて算出する．算出された結果はその発言が荒らし行為の一部であるかどうかの確率であるため，その確率の評価方法が必要である．我々は，評価手法として単純に評価する場合，複数の発言を結合して評価する場合と文脈に沿った並べ替えを行い評価する場合の3種類の評価方法を提案し，検証実験を行い，有効性の検証を行う．

キーワード 電子掲示板，荒らし行為，ベイズの定理

A detecting method of BBS vandalism based on correlation of comments

Yu ICHIFUJI†, Susumu KONNO†, and Hideaki SONE††

† Graduate School of Information Sciences, Tohoku University
6-3,Aoba,Aramaki-aza,Aoba-ku,Sendai,Miyagi,980-8578,Japan
†† Information Synergy Center, Tohoku University
6-3,Aoba,Aramaki-aza,Aoba-ku,Sendai,Miyagi,980-8578,Japan
E-mail: †ichifuji@mail.tains.tohoku.ac.jp, ††{skonno,sone}@isc.tohoku.ac.jp

Abstract Electronic bulletin board system (BBS) has problems of vandalism. It is necessary for an operator to find such problems quickly when happened. For detecting such vandalism in BBS, we propose three methods to evaluate the probability based on bayesan inference. One is evaluation of the probability simply. Second employs combination of comments, and calculates its probability. Third introduces arrangement of comments as context. The result of evaluation of three methods are shown, and efficiency to find such vandalism is discussed.

Key words BBS, vandalism, bayesian inference

1. まえがき

電子掲示板には，一日に数百万のアクセスがある２ちゃんねる[1]といった大規模なものから，友人同士が情報交換のためだけに利用する小規模なものまで，多種多様なものがある．そこでは，匿名性を保ちつつ文字や記号を用いて情報交換や意見交換等のコミュニケーションを取ることができる．しかし，この匿名性のため，悪意を持った人間が他者のプライバシー情報を公開する，故意にコミュニケーションを阻害する，参加者を挑発するなどといった問題が発生する．そのため，インターネットの健全且つ円滑な利用を促進するために，プロバイダ責任制限法（別名：特定電気通新役務提供者の損害賠償責任の制限及び発信者情報の開示に関する法律）が施行された．本法律は，プライバシーの侵害や著作権侵害に対して，被害者が発言の削除や情報開示などを求めることができることを定めている．この法律により，法律で対象とされた荒らし行為に対しては，要求があってから対処すればよく，管理者やプロバイダは保護されている．しかし，電子掲示板で発生する問題行為は，法律で対象とした以外にも多種多様存在し，掲示板管理者にとっては，利用率を維持するためにも，また，参加者に議論や情報交換のための正常なコミュニケーションの場を提供するためにも，様々な問題行為に対処すべきである．

現在，電子掲示板で発生する問題行為 (以下，荒らし行為と呼び，プロバイダ責任制限法が対象とする行為は含まないこととする) には，故意によるものと過失によるものが存在する．過失で発生する荒らし行為とは，掲示板においてコミュニケーションを行っている最中，お互いの意見や主張が食い違う，または，誤解から言い争いとなり，他の参加者にとって不快に感じる発言の羅列となってしまうような場合のことを指す．故意・過失で起こる荒らし行為には，以下のものが挙げられる．

（1） 多人数 (2 人，自作自演を含む) による罵り合い
（2） 閲覧者を不快にさせる書き込みの連続 (煽り)
（3） 荒らし行為を誘発させるような挑発的な書き込み (釣り)
（4） 無用なコピー&ペーストの繰り返しによる閲覧の阻害

これらの荒らし行為の発見には，主に管理者の巡回や他の閲覧者による管理者への通報がある．管理者の巡回は，管理対象コンテンツを管理者自身が閲覧し，荒らし行為があるかどうかを直接調べる方法である．しかし，管理者が常に管理対象コンテンツを監視することは不可能であり，また，管理者への通報においても，管理者が対処するまで閲覧可能状態が続くため，その間に掲示板が激しく荒れる可能性もある．したがって，問題発言を “可能な限り早く削除する” または “掲載しない” などの，掲示板管理者による適切な対処が必要である．そのためには，いかに早く荒らし行為を発見できるかが重要となる．よって，これら荒らし行為を素早く簡単に発見する方法の確立が必要である．

このような問題を解決するために，出現すると荒らし行為に発展するであろうと思われる単語 (以下，“NG ワード” と呼ぶ) をあらかじめ登録しておき，出現するたびに警報を出力する，または，掲載する前に削除する方法がある．しかし，このように NG ワードのフィルタリングによって，問題発言を全て排除してしまうと，だれでも気軽に利用することができる掲示板の特性を失わせることとなる．また，本音での討論には，多少汚い言葉が混じるものであり，このような言い争い (以下，“フレーミング”) をすべて規制してしまうと，掲示板の存在意義すら失わせることになりうるという観点から，すべてのフレーミングを規制する必要はないという意見も存在する [5] [6] [7]．したがって，NG ワードによるフィルタリングだけでは掲示板の監視は十分ではないと言える．

そこで，いくつかの企業において，この問題に着目し掲示板の発言を監視するサービスが提供された．これらのサービスは，NG ワードをあらかじめ登録しておき，出現するたびに発言を人間が直接チェックし，掲載か非掲載かを管理者のガイドラインに従い判断する．よって，NG ワードを含んだ発言がすべて削除されるという問題は解決するが，NG ワードが出現するたびにチェックするという作業を繰り返すため，監視に多くの人手を使う．そのため，維持費用は高くなり，掲示板の規模が大きくなればなるほど，管理者の費用負担も増大するため，現実的な解決法とは言えない．

この問題を解決するために，我々は，以前，問題発言に含まれる NG ワードだけでなく，相手に好感を与える単語と発言の連鎖に着目した，掲示板の雰囲気を示す指標 “荒み度” を利用した監視支援手法 [3] [4] を提案した．具体的には，単語はあらかじめ辞書を人手を使い用意し，単語の重み付けを行う．その重みと連鎖数から，荒み度を算出し，荒らし行為とそうでない部分の明確な差別化を目指した．[4] では単語の重みを出現率に応じて変化させたが，基準を単語単体で見た場合に不快に感じるかそうでないかで決定していたため，荒らし行為とそうでない部分に曖昧さを残す結果となっている．

曖昧さを解消するために，我々は，単語単体の意味から好感・嫌悪感の重みを決定するのではなく，別の手法を導入することを目指す．掲示板における発言は，複数の単語の組み合わせで構成されている．そのため，各発言に出現する単語に着目することは従来と変わらない．ここで，ある単語と別のある単語が同じ発言内に出現した場合，荒らし行為となる可能性が高くなることに着目する．つまり，あらかじめ荒らし行為によく含まれる単語とそうでない単語の出現確率をあらかじめ学習しておき，そこからある発言が荒らし行為となる可能性を評価し，荒らし行為抽出の曖昧さの解消を実現する．本論文では，荒らし行為となる可能性の評価にベイズの定理を利用する．掲示板には，話の流れが存在しているため，ベイズの定理を用いて算出される各発言の荒らし行為である確率をそのまま利用するだけでは，効率よく荒らし行為を発見できないと思われる．そこで，算出した荒らし行為である確率をそのまま評価する方法，システムによって与えられた番号によって評価する方法，会話の流れによって評価する方法の 3 通りを提案し，どの方法が曖昧さの解消に役立つかを調査する．

2. 荒らし行為と思われる発言確率の複合評価方法の提案

本論文では，掲示板の各発言が，荒らし行為である可能性を算出する方法として，ベイズの定理に着目している．そこで本章では，ベイズの定理について述べ，その後，算出された可能性をどのように評価するかを提案する．

2.1 ベイズの定理

ベイズの定理とは，近年のスパムメールの判別に利用され，件名やメールの中身の単語によるマッチングなどの手法よりもスパム判定率が非常に高く，また，様々な分野で注目を集めている定理である．

ベイズの定理の概要は，事象 A が生起した時の事象 B が生起する条件付確率 (事後確率：$P(B|A)$) と事象 A と事象 B の生起頻度の 2 つの確率 ($P(A)$，$P(B)$) から，事象 B が生起した時の事象 A が生起する条件付確率 (事後確率：$P(B|A)$) を導き出す定理である．$P(A) > 0$ とするとき，以下の式が成り立つ．

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)} \quad (1)$$

2.2 対象とする荒らし行為

荒らし行為には，流し読み程度で用意に発見できるものとそうでないものが存在する．そこで，一見しただけで荒らし行為 (閲覧を阻害するアスキーアートの連続投稿など) と判明できる

ものは対象外とする．また，1章で述べた通り，我々は，ある程度のフレーミングは許容すべきであると考えているため，釣り・煽りなどの行為が発生しても，他の参加者に無視された場合 (以下，“スルー” と呼ぶ)，それは荒らし行為ではないとする．以上より，本論文が対象とする荒らし行為を発見が困難な以下の3つにする．

◎ 多人数 (2人，自作自演を含む) による罵り合い
◎ 閲覧者を不快にさせる書き込みの連続
◎ 不快な書き込みを誘発させるような挑発的な書き込み

但し，出現した発言がどの荒らし行為に属するかの判別は行わず，荒らし行為かそうでないかの判別だけを行うものとする．

2.3 ベイズの定理を利用するための準備

本論文では，対象とする掲示板の発言が，荒らし行為の発言かそうでない発言かの2クラスに分類する必要がある．そこで，クラス分類の代表的な手法の1つである，単純ベイズ分類器の学習アルゴリズムを利用する．これは，訓練データから各トークン (ここでは “意味を持つ最小単位の単語” を指す) が荒らし行為に含まれている確率を推定する．その後，これを基に，対象の発言が荒らし行為である確率を推定するものである．

英語を始めとする多くの言語では，トークンに分解することは，単語に着目するなどし容易にできるが，日本語のような明確な区切りがない場合には難しい．現在，日本語をトークンに分解する手法には，n-gram，形態素解析，漢字・かな・カナで区切るなどの手法が存在する．これらの方法の間での定量的比較はあまり行われていないため，今回，トークン化には形態素解析 (茶筌 [2]) を利用することにする．

2.4 学習データの収集

本手法では，学習用のデータが必要である．そこで，学習データとして，ゲーム関連の話題・雑談・相談関連の話題・パソコン関連の話題の掲示板を用意した．これらを，6人の20歳以上でかつインターネット暦が3年以上の男性にあらかじめ10の掲示板を読んでもらい，荒らし行為である発言の抽出を行ってもらう．その後，3人以上が荒らし行為と認めた発言を，学習用データとして利用する．また，荒らし行為であるとだれも判断しなかった発言を通常発言用の学習データとして利用する．

得られた学習用データ数はそれぞれ，荒らし行為である発言の総数が173個，荒らし行為でない発言の総数が10000個である．次章にて，この学習データで学習したシステムの検証実験を行う．

2.5 算出確率の評価方法

掲示板には，話の流れが存在している場合があるため，単純に算出した確率を評価しても荒らし行為の発見ができない可能性がある．そこで，本論文では，以下の3つの評価方法を提案する．

単純評価 発言時にシステムから与えられるナンバーに沿って算出した確率を発言順に並べグラフ出力し，荒らし行為の範囲を抽出する．

複数評価 掲示板には話の流れが存在するため，複数の発言を一まとめにし荒らし行為である確率を算出する．例えば，4，5，6，7，8の数を与えられた発言があった場合，発言ごとに算出するのではなく，4-6の発言を一つのテキストデータとし，荒らし行為である確率を算出する．その後，5-7，6-8というように算出し，発言順に並べグラフ出力し，荒らし行為の範囲を抽出する．

並べ替え評価 掲示板には話の流れが存在するが，それがいつも番号順に並んでいるわけではない．そこで，話の流れに沿うように並び替えを行う．並び替えを行い，細分化した後，算出した荒らし行為である確率が高いものが集まっているかを調査し，荒らし行為の範囲の抽出をする．ここで，並び替えには，特定の発言に意見を述べるときなどに使われる “＞＞”(“アンカー” と言う) を利用して行う．但し，対象とするアンカーは “＞＞番号” のスタイルのものだけとし，引用文などは対象としない．

以上の3つの評価方法について，次章にて検証実験を行い結果を示す．

3. 検証実験

本章では，2章で算出した荒らし行為である可能性の評価方法の有用性を調査する．まず，実験対象掲示板を示し，その後，それぞれの実験結果を示す．

3.1 対象掲示板

ベイズの定理を用いて算出した荒らし行為であるという確率の評価方法の実用性を示すための実験対象として，インターネット上の巨大掲示板2ちゃんねるのような，発言の順番に表示され，文脈などによりスレッディングされない掲示板をサンプルとして選択した．このような掲示板の特徴として，発言した時間が早い順番に表示されていくことがあげられる．このため，2つの話題が同時に進行する場合や，複数の人間が参加しほぼ同時に発言する場合において，どの発言に対する発言なのか分からないものになってしまうこともある．現在，これを回避するために，アンカーを用いてどの発言に対する発言なのかを明示することも多い．

3.2 評価・比較実験

本実験のサンプルとして，2ちゃんねる [1] より，“いい加減，少年犯罪とTVゲームを関連づけるの止めろ パート3” という題名の掲示板の適用結果を示す．この掲示板では，多少の口汚い言葉が含まれるが，議論として成り立っている部分と，ただの煽りとなっている部分の両方が存在している．本手法では，議論として成り立っている部分は抽出せず，ただの煽りとなっている部分の抽出ができれば成功と言える．

単純評価

確率を個別に評価した結果は，図1のようになった．この出力方法では，○で囲まれた範囲の発言が荒らし行為の可能性が高いものとして出力されている．実際の掲示板の内容と照らし合わせて見る．

I の範囲では，URLのみの発言と煽りと取られるような発言1つが存在している．*II* の範囲では，特に荒らし行為というわけではない発言が存在している．*III* の範囲では，フレーミングが行われており，荒らし行為とも取ることができる発言が多

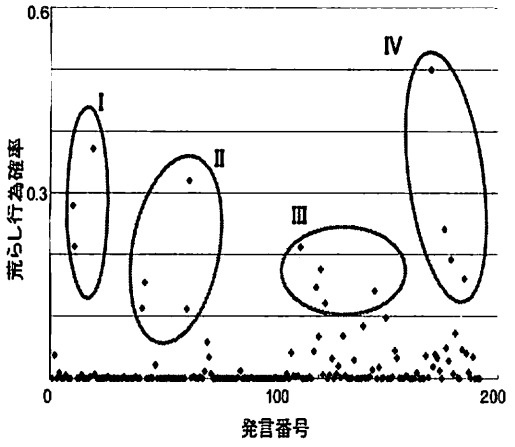

図 1　確率を個別に評価した出力結果

数存在している．*IV* の範囲では，明らかな煽りと取れる発言とそれに追随した発言が存在している．

I と *II* に関しては，今回行った実験における学習データの量に影響を受けているように思われる．よって，より多くのデータを学習させることにより，誤った分類を減らすことが期待できると思われる．

III に関しては，我々はある程度のフレーミングを容認するという立場から，低確率だが荒らし行為になりうるという算出評価は有益であると言える．*IV* に関しては，完全に対象とする荒らし行為であるため，発見できることが分かる．

このように，発言単体の確率をグラフ評価した結果だけでは，*I* のようなスルーされた場合でも高い確率で算出されてしまう可能性があるといった問題が残ることが判明した．

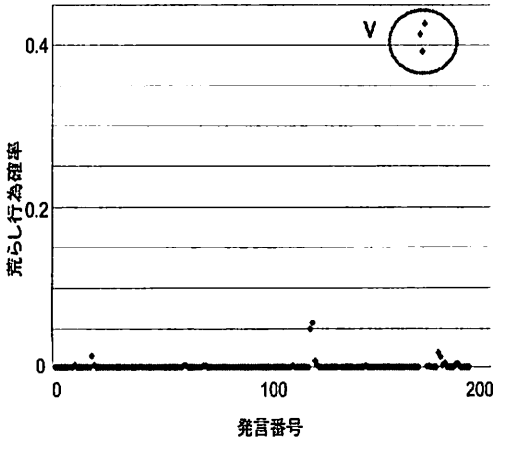

図 2　複数の発言をまとめて確率を算出し評価した出力結果

複数評価

複数の発言をまとめて確率を算出し評価した結果は，図 2 のようになった．今回は，3 つの発言が出現したときの荒らし行為である確率を算出している．図 1 と異なり，荒らし行為の可能性となっている箇所は *V* の 1 箇所のみを示している．ここは，明らかな煽りと取れる発言とそれに追随した発言が連なっている箇所であり発見すべき対象である．また，単純評価で出力された *I* と *II* の部分が連続評価では出力されていない．よって，この結果は本評価手法が機能していることを示している．

しかし，*III* の範囲は飛び飛びでフレーミングが出現しており，単純に発言順にまとめて評価したため，フレーミングの箇所が打ち消されてしまっている．以上より，連続評価は，連続で発生した荒らし行為には有益であるが，飛び飛びで出現した場合には，対応できないことが判明した．

並べ替え評価

アンカーに従って並べ替えをした結果，フレーミングの範囲で，且つ，アンカーによる連鎖が発生している箇所は表 1 となった．

表 1　並び替え評価の確率

発言番号	算出確率
108	0.041386
109	0.004528
110	0.003333
112	0.21557
114	0.000186
116	0.002183
118	0.04427
119	0.146583
120	0.067793
122	0.005987

表 1 の中の 108 から 122 まで，フレーミングが発生している．しかし，発言番号 115 と 117 は別の話題で会話をしており，算出確率はともに 0 である．したがって，並び替えることによって，無関係な発言を考慮する必要がなくなる場合が存在する．

このようにアンカーのついた発言の連鎖が多い話題は稀な部類に入る．そのため，アンカーにのみ頼った並べ替えではなく，引用や前後関係から並び替えることのできる手法が必要である．

以上の評価・比較実験より，発言単体が荒らし行為の可能性である確率を算出し，その評価方法を変化させることにより荒らし行為の範囲の抽出を実現できる可能性が明らかとなった．今後，各評価方法を組み合わせる，または，導入する方法の提案が必要である．

4. ま と め

本論文では，荒らし行為発見支援のために，あらかじめ荒らし行為によく出現する単語とそうでない単語を学習データとして用意し，そこから，ベイズの定理を利用し，表示された発言が荒らし行為である確率を算出・評価することにより荒らし行為の発見の曖昧さの解消を目指した．その評価方法として，単純評価，複合評価，並び替え評価の 3 通りの評価方法を提案し，同一掲示板に試行した．その結果，単純評価は掲示板の特徴である話題の通りに並んでいないということから，荒らし行為の

範囲というものを出力するには向いていないということが明らかとなった．次に，複合評価では，話の流れの通りに並んでいる範囲を抽出することができることが明らかとなった．最後に，並べ替え評価では，アンカーを利用した並び替えを行い，確率を評価することを試みたが，アンカーの数が少ないものが多く，期待するほどよい結果は得られなかった．従って，複合評価を軸に，アンカー利用した発言をどのように評価していくかが今後の課題となる．

文　献

[1] 2 ちゃんねる, http://www.2ch.net/

[2] 形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/

[3] Yu Ichifuji, Susumu Konno, Hideaki Sone, "A method to monitor a BBS using feature extraction of text data", International Conference on Human.Society@Internet, (2005) pp.349-352

[4] 一藤 裕, 今野 将, 曽根 秀昭, "テキストベースコミュニケーションにおける阻害行為に関する教科手法の提案", 電子情報通信学会技術研究報告 IA, Vol.206, No.62, pp. 43-47, 2006.

[5] 大澤幸生, 松村真宏, 中村洋, "フレーミングは議論を阻害するか -2 ちゃんねるは何故面白い？", 第 11 回 ITRC 研究会, 2002.

[6] 柴内康文, "言い争う—「フレーミング論争の検証」", 現代のエスプリ, 川浦康至 (編), vol.370, 至文堂, 1998.

[7] 松村真宏, 三浦麻子, 柴内康文, 大澤幸生, 石塚満, "2 ちゃんねるが盛り上がるダイナミズム", 情報処理学会誌, vol.45, no.3, pp.1053-1061, 2004.