社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2006－ＱＡＩ－19（９）
2006／5／24

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評価手法の提案

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　近年，プロバイダ責任制限法が施行され，コンテンツの管理者や運営者の責任が明確となり，問題行為に対する適切な処置を求められるようになった．その結果，管理者や運営者の管理負担が増大することとなった．よって，その管理負担を軽減するために，問題行為を発見を支援する手法が必要であるといえる．今回，我々は，コンテンツの中から，荒らし行為と呼ばれるコミュニケーションを阻害する問題行為が発生する電子掲示板を選択し，その荒らし行為の発見支援を目指す．具体的には，掲示板における各発言に含まれる閲覧者に心理的影響を与える単語と，他の閲覧者が興味を持った証拠であるアンカーと呼ばれる記号に着目し，それらを利用し，掲示板の雰囲気を数値化することにより評価するシステムの提案を行う．
キーワード　電子掲示板，荒らし行為，tf-idf 法

Proposal of an evaluation method for obstruction in text-base communication

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 a problem of abuse on communications. The authors had proposed a method that aids an operator to find such abuse. The method employs "Ruination Figure [RF]" which indicates influence of the BBS. The evaluation of RF is based on two types of feature extraction. One is presence of "words" which causes both positive and negative psychological effects on readers. The other is number of "responses" which is a reaction of the reader to a certain comment. Each influence of words is treated as a constant in this method and this validity is discussed. Another discussion is development of a way to detect a critical abuse from trend of the RF.
Key words　BBS, Obstruction in Communication, tf-idf

1. まえがき

インターネットの健全か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：特定電気通信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．本法律の対象は，プロバイダ（以下，"ISP"）だけでなく，電子掲示板などの管理者や運営者も含まれるため，本法律の施行により，各々が果たす管理責任が明確となり，適切な対処を求められることになる．しかし，電子掲示板等の企業や個人で運営しているコンテンツを常時監視することは，管理側にとって非常に負担がかかる．従って，コンテンツの内容を効率的に把握し，発生している問題に即時対処できる方法の確立が必要であると言える．今回我々は，コンテンツの中でも，悪意を持てばだれでも簡単にプライバシーの侵害や著作権侵害

を行うことができる電子掲示板に着目し，問題行為の発見を支援する手法を提案する．

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故意にコミュニケーションを阻害するという問題が発生する．また，悪意を持っていなくても，過失により知らずに荒らし行為を行ってしまう場合もある．故意・過失で起こる荒らし行為には，以下のものがあげられる．

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

プロバイダ責任制限法では (1) の犯罪予告・個人を特定できる情報の書き込みが対象であるが，(2)~(5) の荒らし行為が (1) の行為に変化することがあるため，荒らし行為全般を発見することが，円滑かつ健全なコミュニケーションの維持に必要となる．

荒らし行為の発見は，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[3][4][5]．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．

そこで，我々は，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[2]を改良することを提案する．具体的には，単語はあらかじめ辞書を人手を使い用意し，[2] では同一としていたそれぞれの単語が与える影響力を，tf-idf 法を利用して重み付けを行う．その重みと連鎖数から，荒み度を算出し，従来よりも荒らし行為とそうでない部分の明確な差別化を行い，最終目標である機械判断への足がかりとする．

2. 荒み度の算出

荒み度とは，掲示板の雰囲気を数値化したものである．これは，各発言に出現する単語と発言の連鎖から算出される．まず最初に，我々が提案した従来の算出方法[2]の述べ，その問題点を明らかにする．その後，問題点を解決するために，新たな算出方法を提案する．

2.1 対象となる荒らし行為

荒らし行為には，流し読み程度で容易に発見できるものとそうでないものが存在する．そこで本論文では，対象とする荒らし行為を，発見が困難であり，また，犯罪を助長する行為や個人情報の公開につながる行為として，以下の３つを対象とする．

◎ 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

これらの荒らし行為を，掲示板管理者がすべての発言を読むことなく発見することができる支援手法を実現する．

2.2 従来の荒み度の算出方法

従来の手法では，荒み度 $[RF]$ を算出するために 7 個の式と 2 個の集合を定義する．Positive Word $[pw]$ は，閲覧者に単体で好感を与える単語（ありがとう，ガンバレ等）の集合を表し，pw の各要素に与える重みを Positive Word Weight $[pww](pww > 0)$ とする．その逆に，閲覧者に単体で不快を与える単語（死ね，黙れ等）の集合は Negative Word $[nw]$ で表し，nw の各要素に与える重みを Negative Word Weight $[nww](nww < 0)$ とする．

各発言中に出現した $pw \cdot nw$ に一致した数を Concord Number $[cn]$ で表し，それぞれ $cn(pw)$，$cn(nw)$ で表す．単語から算出できる各発言の影響力を Word Score $[Ws]$ で表す．ある発言の番号を “t” とすると，$Ws(t)$ は $pw \cdot nw \cdot cn$ を使って，以下のように表す．

$$Ws(t) = pww * cn(pw) + nww * cn(nw) \quad (1)$$

従来の手法では，$pww = 1$，$nww = -1$ と設定したため，実際の Ws(t) は

$$Ws(t) = cn(pw) - cn(nw) \quad (2)$$

となる．

しかし，これだけでは特定の単語が出現した発言すべてが反応するため，効率よく荒らし行為を発見することは難しい．そこで，単語以外に閲覧者に影響を与えるものとして，松村らの発言の連鎖を用いて他者に最も影響を与えた起点となる発言を探す研究[6][7]に着目した．これより，発言の連鎖は掲示板の流れや傾向を表現するに足ると考えた．

例えば，発言 8 の中で ">>3" とあった場合，発言 8 は発言 3 に対して意見を述べているので，発言の連鎖が発生していると捉えている．さらに，発言 8 に対して，発言 12 から意見があった場合，3 の発言によって 8 以下の連鎖を生んだと考え，間接的に発言 3 の意見が影響したと考え，発言の連鎖数 (Res と定義する) は $Res(3) = 2$ とする．したがって，Res を用いて，Weber-Fechner's law(ウェーバー・フェヒナーの法則)(注1)から，発言の連鎖が閲覧者に与える影響を Comment Chain Score [ccs] とし，次のように定義する．

$$\begin{aligned} &Ws(t) \geqq 0 \text{ のとき} \quad ccs(t) = log_2 Res(t) \\ &Ws(t) < 0 \text{ のとき} \quad ccs(t) = -log_2 Res(t) \end{aligned} \quad (3)$$

ただし，$Res(t) < 2$ の場合は $ccs(t) = 0$ とする．なぜなら，この話題は 2 人目で完結したとみなせ，その他にまったく影響を与えないとみなすことができるからである．

式 1・3 から算出する各発言の影響力を Statement Score [Ss] と定義し，t 番目の発言の Ss(t) は，

$$Ss(t) = Ws(t) + ccs(t) \quad (4)$$

となる．$Ss(t)$ が正であればその発言は好感を与えているとみなす．逆に，$Ss(t)$ が負であれば，その発言は不快感を与えているとみなす．

荒み度 "Ruination Figure" [RF] は掲示板全体の指標であるので，各発言の影響力 Ss を順に加算したものとなる．よって，最後に書き込まれた発言が t 個目時，荒み度 $RF(t)$ は，次のようになる．

$$RF(t) = \sum_{1}^{t} Ss(i) \quad (5)$$

2.3 従来手法の問題点

前節で述べた算出方法に基づき，荒らし行為の発見を行った結果，全体の約 70 から 80%を発見することができた．しかし，発見すべき荒らし行為の範囲を，指摘できないものが存在することや，荒らし行為とそうでない範囲の区別が難しい箇所が存在するといった問題点もある．それらの原因として，1 つの pw は 1 つの nw の出現で打ち消すことができるため，nw の出現を打ち消すほどたくさんの pw が出現したことや，$ccs(t)$ が大きくなりすぎたことが考えられる．

(注1)：人間にある刺激を与えた場合，その絶対値に比例した感覚を生ずるのではなくて，与えた刺激の対数に比例する感覚を生ずるという法則

そこで，本論文では，pww，nww の各単語の持つ重みを出現率によって変化させ，荒み度 [RF] に反映さえることを提案する．

2.4 単語の重みを考慮した荒み度の算出方法

前節で述べた問題点を解決するために，各単語が与える心理的影響は，それぞれで異なっているという観点から，単語の出現率に着目し，単語の出現率に基づき重みを変化させるという，単語の重みを考慮した荒み度の算出方法を提案する．具体的には，各々の重みを，単語の出現率に着目し，多く出現する単語の重みは小さく，ほとんど出現しない単語の重みは大きくするように設定する．ある単語 x が pw に属する場合，その重み $pww(x)$ は式 6 で表す．

$$pww(x) = n * log(s/y) \quad (6)$$

式 6 中の n は，ある発言中に単語 x が出現した回数を表し，s はその掲示板の全ての発言数を表し，y は単語 x が出現した発言の数を表している．nww の場合も同様である．

単語の重みを変化させるため，式 1 は，以下のようになる．

$$Ws(t) = \sum pww(i) * cn(pw(i)) + \sum nww(j) * cn(nw(j)) \quad (7)$$

また，式 7 の Ws(t) と式 3 の ccs(t) とのバランスを考慮するため，

$$Max(|Ws|) \cong Max(|ccs|) \quad (8)$$

となるよう調整し，単語と連鎖の力関係を等しくする．

その結果，式 4 は，次のようになる．

$$Ss(t) = Ws(t) + \{ccs(t) \times Max(|Ws|)\}/Max(|ccs|) \quad (9)$$

以上のように，単語の重みを出現率によって変化させ，また，単語による影響と連鎖による影響のバランス考慮した荒み度の算出方法を用いて，従来手法の問題点の解決を目指す．

次章にて荒み度 RF の評価方法について詳しく述べる．

3. 荒み度の評価方法

荒み度 RF は各発言の影響力を蓄積したものであり，そこから荒らし行為を発見するためには，管理者がその変動を見る必要がある．そこで著者らは，荒み度から得られた情報を損なうことなく効率的に表現することを考え，管理者が判断しやすいように，株価の変動を表現するために使われるローソク足チャート [8][9] を用いて荒み度の変動を表現することを提案する．

荒み度の変動をローソク足チャートで表現するために，株価の始値・高値・安値・終値を荒み度に対応させる必要がある．そのため，掲示板の各発言の荒み度 RF を 10 刻みで，各集合ごとに最初の値・極大値・極小値・最後の値を始値・高値・安値・終値に見立てて，ローソク足チャートを出力する．

本論文では，このローソク足チャートの変化の傾向と程度から荒れていると思われる範囲を掲示板管理者の目視により抽出

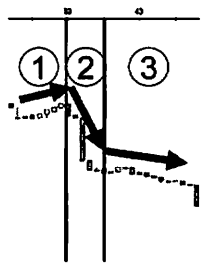

図 1　ローソク足チャートの抽出例

する．例えば，図 1 が出力されたとする．1 では緩やかだった変化が 2 の範囲では，急激に変化していることがわかる．その後，3 の範囲ではまた変化が緩やかになっている．よって，2 の範囲でなんらかの悪意のある発言が出現したのではないかということが予想でき，2 の範囲が荒らし行為が発生した可能性があると判断できる．このようにして，目視により荒れている範囲を抽出する．

4. 実　　験

本章では，本提案手法が従来手法に比べどれだけ優れているかを，実際の掲示板に対して実験を行うことにより，明らかにする．まず，実験対象掲示板を示したあと，評価方法について述べ，実験結果を示す．

4.1　対象掲示板

荒み度を用いた監視支援の有効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 ちゃんねるから，テーマの異なった 12 個の掲示板をサンプルとして選択した．各掲示板には番号を付与し，それぞれの特徴を述べる．1 番から 9 番の掲示板は，問題となるような単語の出現率が低い傾向を持っており，10 番目から 12 番目の掲示板は，逆に高い傾向を持っている．1，5 番の掲示板はパソコン関連の話題，2，3，6，9，11 番の掲示板はゲーム関連の話題，残りは雑談・相談関連の話題の掲示板である．

4.2　評 価 方 法

これらの掲示板の荒れている範囲の抽出は，提案手法を利用する人間の主観が入らないようにするために，インターネット暦 3 年以上の 20 代の男性 6 人に，サンプル掲示板を読み，荒れていると判断した範囲を抽出する主観評価を実施した．

本手法を実験するために，荒み度を算出するために pw と nw のそれぞれの辞書を作成する必要がある．今回，実験対象とは違う 3 つの掲示板から pw に属する単語を nw に属する単語を人手を用いて抽出し，pw には 121 個，nw には 308 個の該当単語をそれぞれ登録し，それぞれの辞書を作成した．このようにして作成した辞書を，登録単語を変えずに全ての対象掲示板に対して使用した．そこから得られた提案手法による結果と全ての単語に同一の重みを与えた従来手法の場合の結果と主観評価の結果との比較を行った．

本手法の目的は，発生した荒らし行為を漏れなく発見することである．従って，今回は従来の手法よりも多く荒らし行為の範囲を指摘できることで成功とする．

実験結果の一例として，11 番の "結局 EQ2 が WoW にまるで歯がたたなかった件について" という題名の掲示板の旧手法の結果を図 2 に，新手法の結果を図 3 に示す．主観評価による荒らし行為の範囲は，61-160，281-300，631-670，751-790 となっている．

4.3　比 較 検 証

従来の手法による出力グラフは図 2 に，本提案手法による出力グラフは図 3 となった．各グラフの直線で囲まれた範囲が主観評価と評価手法が一致した範囲を示している．また，点線部分は荒らし行為でないのに荒らし行為と判断された範囲を示している．

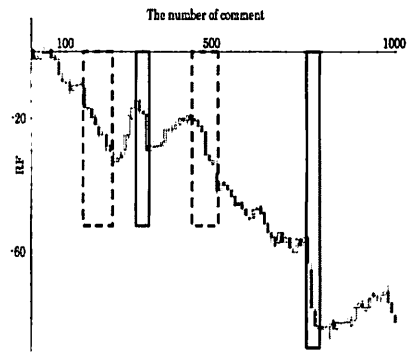

図 2　The result of the experiment by the former method

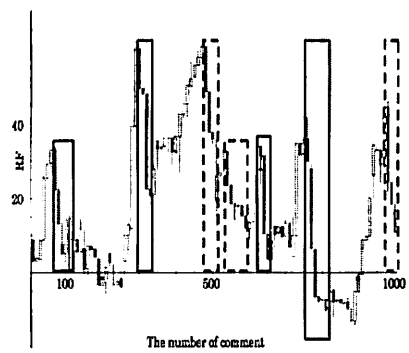

図 3　The result of the experiment by the new method

図 2，3 より，従来の手法では見つけることのできなかった 2 箇所を提案手法では発見できていることが分かる．提案手法の優位性を示すため，主観評価と従来・提案手法の一致の一覧表を表 1 に示す．表 1 より，単語の重みを考慮した本手法が，従来手法よりも確実に荒らし行為の範囲を捉えていることがわかる．ただ，グラフより荒らし行為の始点や終点が実際より長

表 1　主観評価と従来・提案手法の比較

主観評価	従来の手法	提案手法
61-160	×	○
281-300	○	○
631-670	×	○
751-790	○	○

い場合や短い場合がある．これは，今回の主観評価の仕方と荒み度の表示方法に問題があると思われる．例として，299 から 312 まで荒れているとしたとき，評価者には，291 から 320 まで荒れていると表記するようにしたことと，荒み度を 10 区切りで表現していることの両方が，今回のようなあいまいさを残す原因として考えられる．このことは今後の課題として解消すべき点である．

実験で使用した 12 の掲示板に対する結果を表 2 に示す．表中の分母は，各手法によって荒れていると判断できる箇所を示しており，分子は主観評価により荒れていると判断された箇所を示している．また，漏れは主観評価で荒れていると判断された箇所を各手法が見つけられなかった数を示している．

表 2　従来手法と提案手法の比較

種類	従来手法	漏れ	発見率 [%]	新手法	漏れ	発見率 [%]
1	0/2	0	100	0/0	0	100
2	1/2	0	100	1/2	0	100
3	0/0	1	0	0/0	1	0
4	2/2	1	66.7	3/4	0	100
5	1/4	0	100	1/1	0	100
6	0/0	0	100	0/0	0	100
7	2/4	1	66.7	3/6	0	100
8	2/3	1	66.7	2/3	1	66.7
9	0/0	0	100	0/0	0	100
10	2/2	1	66.7	3/5	0	100
11	3/4	1	75	4/6	0	100
12	2/2	0	100	2/2	0	100
Total	16/25	5	76.2	19/29	2	90.5

表 2 の発見率は，次の式で表す．

$$\text{発見率}[\%] = \frac{\text{手法によって発見できた荒らし行為の数}}{\text{掲示板に出現するすべての荒らし行為の数}} \times 100 \quad (10)$$

実験の結果，従来の手法では見つけることのできなかった箇所が 5 個，発見率が約 76.2%に対し，新手法では 2 個に減少し，発見率約 90.5%に上昇している．また，それと同時に荒れていると判断できる箇所も増えている．しかし，本手法の目的は，荒れている範囲を漏れなく指摘することであり，指摘できなかった荒れている範囲を従来の手法よりも多く指摘できているため，本手法は従来の手法よりも優れているといえる．

5. ま と め

本論文では，荒らし行為に対する掲示板管理者の監視負担を軽減するために，単語がもつ閲覧者への心理的影響を考慮し“荒み度”を算出する方法を提案し，各々の場合で，登録単語を変えずに実験を行った．12 個の掲示板に対して実験を行った結果，従来の方法で，見つけられなかった範囲を指摘することができ，発見率を約 14%上昇させることができた．これにより，従来の手法よりも，精度の高い監視支援手法を実現したといえる．

これらの発見方法をより効率的に行うために，機械的な判断の導入や，あいまいさを解消させるために発言の区切りかたの改善が今後必要である．

文　献

[1] 2 ちゃんねる, http://www.2ch.net/

[2] 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

[3] 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.

[4] 柴内康文, “荒い争う—「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.

[5] 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.

[6] 松村真宏, 大澤幸生, 石塚満, “テキストによるコミュニケーションにおける影響の普及モデル,” 人工知能学会論文誌, Vol.17, no.3, pp.259-267, 2002.

[7] 松村真宏, 大澤幸生, 石塚満, “影響の普及モデルに基づくオンラインコミュニティ参加者のプロファイリング”, 人口知能学会論文誌, vol18, non.4, pp.165-172, 2003.

[8] 伊藤智洋, “儲かる！株の教科書 テクニカル指標の読み方・使い方”, 日本実業出版社, 2004.

[9] 阿部達郎, 柳谷雅之, 野村光紀, 葭部音士, 野村 光紀, “株はチャートでわかる！-テクニカル分析がチャートギャラリーでわかる！できる！ パンローリング相場読本シリーズ”, パンローリング, 2000.