ネットワーク上の不適切発言の評価と利用者教育への応用

一藤 裕

東北大学大学院情報科学研究科
応用情報科学専攻
博士課程後期3年の課程
A5ID4001

2010年1月提出

目 次

図 目 次

第1章　序論

1.1　背景

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一つである.「2ちゃんねる」[1]や「Yahoo 掲示板」[2],「mixi」[3] などを代表とする巨大なものから，個人のホームページに設置された少数で利用するためのものまで，大小様々な規模が存在しており，一般的なインターネット上のコミュニケーションツールとなっている.

電子掲示板では，電子掲示板利用者（以下,「ユーザ」と呼ぶ）が文字や記号を使って発言したい内容を書き込み，やり取りを行う．書き込まれた内容は，電子掲示板に設定が許す限り保存され，インターネットを通じていつでも閲覧可能となる．これにより，相手がインターネットに接続していなくとも発言を投稿しておくことにより，非リアルタイムでの会話が成り立つ.つまり，非リアルタイムでチャットのようにコミュニケーションを取ることが可能である．また，電子掲示板の多くは，匿名で相手の顔・年齢・身分も公開することなしに利用可能であるため，誰とでも気軽にコミュニケーションを取ることも可能である．話題も多種多様であり，世間話や特定のトピックに関する議論の場としても使われる．このような会話や議論以外にも，情報収集や質疑応答などの様々な使い方がされている.

このような電子掲示板において，“他の閲覧者を挑発し不快にする書き込み”，“誹謗中傷する書き込み”，“他人の個人情報を不特定多数に公開する書き込み” が発生し対処に追われている.このような行為（以下,「不適切行為」と呼ぶ）を放置してしまうと，通常コミュニケーションを阻害するだけでなく，現実社会でのイザコザへの発展やプライバシー侵害による発言者への訴訟などの問題へと発展する恐れがある．2004 年 6 月には，長崎・佐世保の小学六年生の女児がインターネットの掲示板に悪口を書き込まれ，書き込みを行った同級生を殺害する事件が発生している [4]．また，2006 年 10 月 6 日には，同級生から性格を中傷する内容が断続的にブロ

グに書き込まれたり，直接メールを送られるなどし，被害者が自殺未遂をする事件が発生している [5]．さらに 2009 年には，文部科学省 [6] が学校裏サイトの実態調査を行い，掲載されている発言の約 27%に「死ね」「殺す」といった暴力的な言葉が存在すると報告された．

このような不適切行為が発生した場合，迅速に処理すると同時に，このような不適切行為をしないようにユーザに対し情報倫理教育を施すことが社会問題を発生させないためにも必要不可欠である．ここで，不適切行為が発生する原因に着目する．このような不適切行為には，意図的に行ったものと過失によって発生したものの 2 つが考えられる．意図的に行うユーザに対しては，不適切行為発生後に対処すればよい．しかし，過失によって発生したものは，ユーザの情報モラルが未熟であることや書き込んだ発言が与える印象を正しく理解していないなどユーザの経験不足に起因するものと考えられる．よって，投稿する発言が不適切行為に該当することをユーザに認識させることができれば，過失によって発生する不適切行為を事前に防ぐことが可能であると考えられる．そこで，ユーザへ投稿発言が不適切行為になりうる可能性を示唆することにより，どのような発言が不適切行為となるかを教育できれば，過失によって発生する不適切行為を防ぐことが可能となる．しかし，現在，上記の不適切行為に対し，主に発生後の対処及びそのような不適切行為をブロックすることを主眼においた取り組みがされている．次節にて，対処方法とその不適切点について述べる．

1.2　既存研究による取り組み

1.1 節では，ユーザの経験不足などにより，不適切行為が電子掲示板上で発生し現実社会でのいじめや誹謗中傷・プライバシーの侵害といった社会問題に発展する現状について述べた．1.2 節では，電子掲示板における名誉毀損やプライバシー侵害への対処法として，政府による対策を述べる．また，既存の対処方法として，電子掲示板における不適切行為を防ぐためのフィルタリング手法の特許や監視サービスについて述べる．

1.2.1　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

インターネットが急速に広まり，法整備が追いつかずインターネット上で様々な問題が発生した．特に，電子掲示板上において，特定の個人への誹謗中傷や特定商品を貶める内容，流出した顧客情報やプライベート情報までが投稿，公開され，誰もが閲覧可能な状態となった．これに対し，被害者達は，情報の削除などを電子掲示板の管理者へ依頼し，それらの情報を書き込んだユーザの IP アドレスなどの加害者情報の公開を依頼した．しかし，通信の秘密の観点より，どこまで守秘義務を負うのかが明確となっておらず，無制限にプロバイダなどが責任を負う可能性があった．そこで,「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」[7]（通称,“ プロバイダ責任制限法 ”）が制定・施行された．

プロバイダ責任制限法は，電子掲示板の匿名性を悪用し，他者の個人情報を不特定多数の人間に晒すなどのプライバシーを侵害する行為が発生した場合，プロバイダ・サーバ管理運営者に対し，発信者の情報開示ができることや，損害賠償責任が制限されることが明記された．この法律は，平成 13 年 11 月 30 日に制定・公布され，平成 14 年 5 月 27 日に施行された．これにより，インターネット上でプライバシー侵害が発生した場合，プロバイダ等のとるべき行動基準が明確化され，迅速かつ適切な対応を促進し，また，被害者が加害者情報を請求できる権利が保障され，インターネットの円滑かつ健全な利用を促進することが期待された．但し，対象となる不適切行為は，プライバシーの侵害などの犯罪行為であり，すべての不適切行為が対象となっているわけではない．また，プライバシー情報などの公開によりダメージを受けた被害者以外，プロバイダに対し加害者情報を請求できない問題がある．

1.2.2　NG ワードを利用したフィルタリング手法

電子掲示板では，名誉毀損にはならないが，他者を不快にする不適切な書き込みが多く投稿されている．このような不適切な書き込みが投稿されると，電子掲示板の通常コミュニケーションを阻害することとなるため，投稿を防ぐことが必要である．これら不適切な書き込みを防ぐために，複数の電子掲示板では NG ワードフィルタリング手法が導入されている．これは，特

定の単語をあらかじめ NG ワードとして登録しておき，その単語が含まれる発言が投稿された場合，その発言をブロックする，または，管理者へ通知するフィルタリング手法である．また，これを発展させ電子掲示板システム [8] という発言フィルタリングシステムも提案されて特許も取得されている．これは，掲示板ごとに話題が異なるため，話題ごとの単語データベースと，どの話題の掲示板にも出やすい共通単語データベースを作成し，図 1.1 のように複数のフィルタを用意しフィルタリングを行うシステムである．各データベースには，掲載してはならない単語が格納されており，その単語が含まれた発言が書き込まれると，発言者にはブロックされたことが通知され，掲示板管理者には不適切な発言が投稿されたことが通知される．つまり，NG ワードフィルタリングを発展させたシステムである．

問題として，NG ワードによるフィルタリングは，データベースにない単語を使うことでフィルタをすり抜けることができてしまう．また，特定の単語をブロックするために，表現の幅を狭める問題や通常のコミュニケーションを阻害する可能性がある．

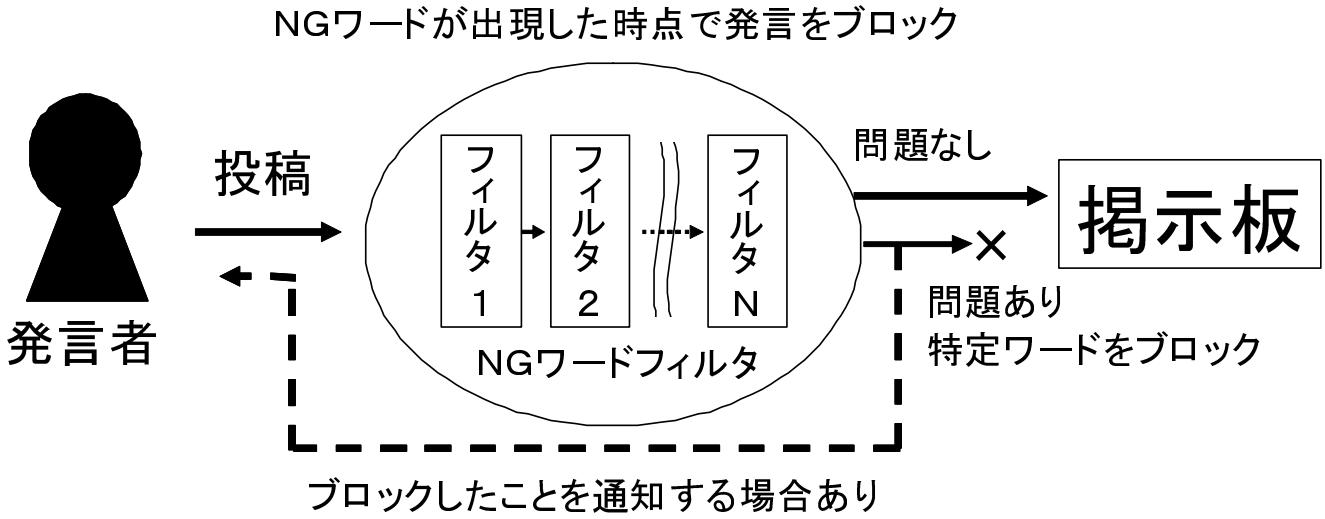

図 1.1: NG ワードフィルタを利用した掲示板概略図

1.2.3　人間による掲示板監視サービス

電子掲示板の発言には誤字脱字などが非常に多く，複数の文字を組み合わせて漢字を作り表現する場合がある．また，チャットのように使われることが多い．そのため，発言単体では不

適切な発言かどうか判断できない発言が存在し，誤字脱字や文脈を考慮する必要があり，NG ワードフィルタリングでは防ぐことができないものがある．

そこで，人手を使って実際に投稿された発言の判断を行う監視サービスが提供されている [9][10]．これは，書き込まれた発言を，電子掲示板に掲載前に確認を行い，不適切行為かどうかを確認するものである．人手を使ったサービスのため，発言ごとの関係や文脈を考慮し確実に発言を判別することが可能である利点を持つ．

しかし，不適切行為かどうか判断が難しい発言に対しては，図 1.2 のように監視者の感情により評価が変化する可能性がある．また，1 つ 1 つ確認するために，掲載までの時間を要する欠点を持つ．さらに，人件費などのコストが非常に高く，すべての掲示板に導入することは不可能であるという欠点も持つ．

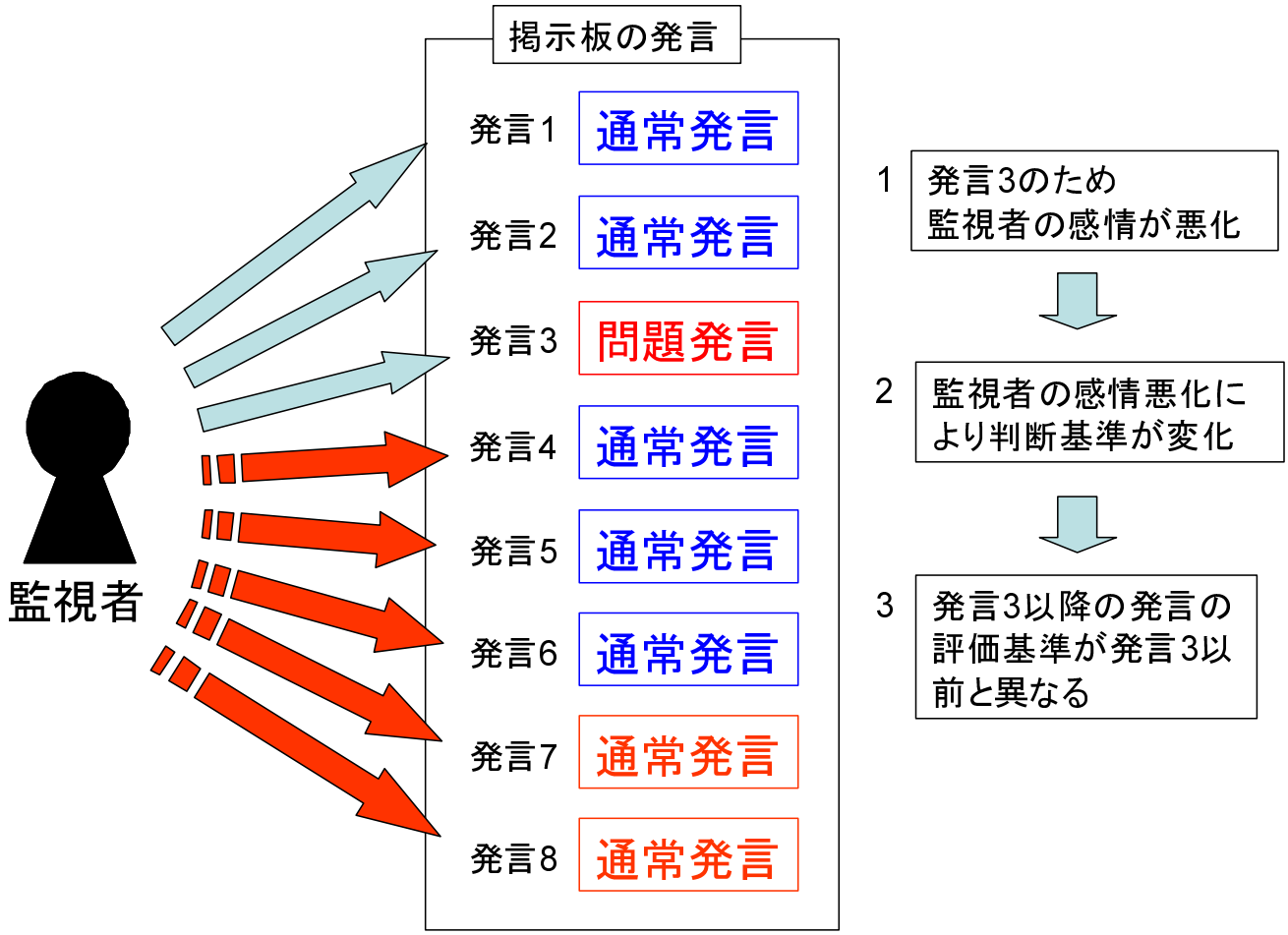

図 1.2: 主観評価基準の変化の例

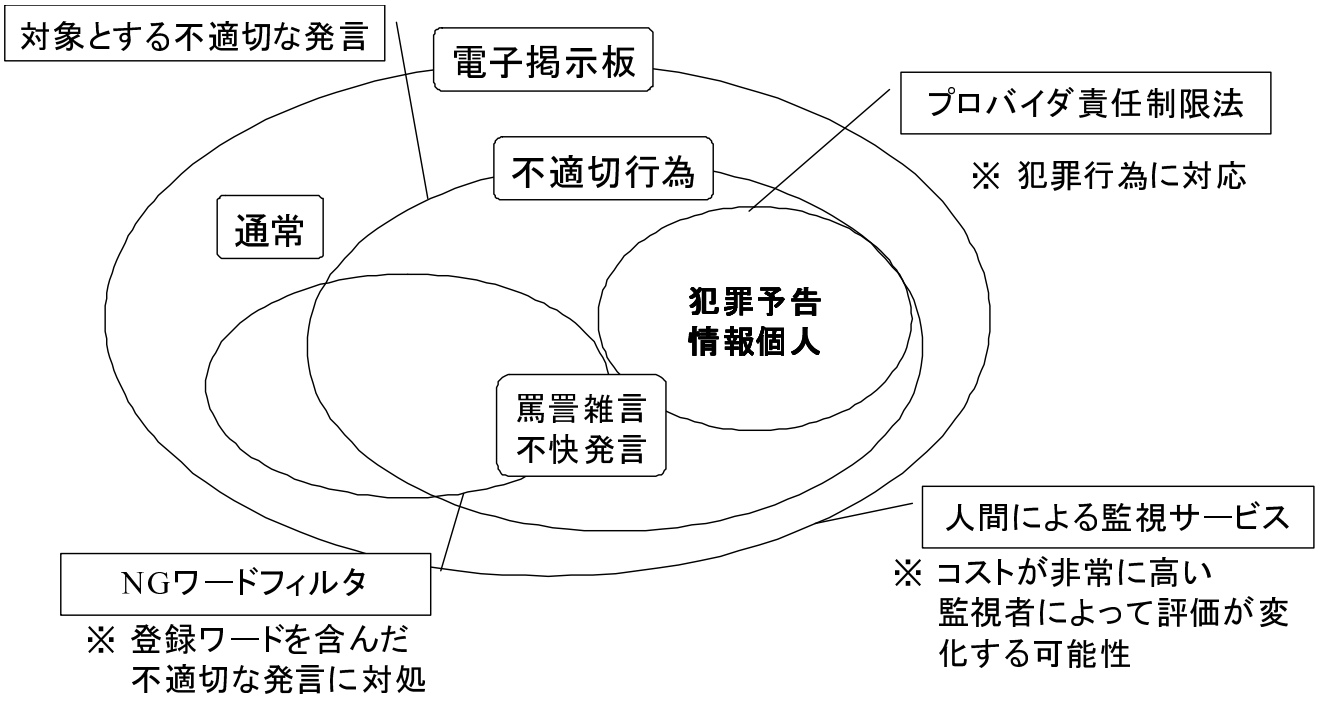

図 1.3: 現在の対策が取り扱う問題範囲

1.2.4　自治体レベルの取り組み

学校裏サイトやプロフがいじめの温床となっているが，学校だけの対策では限界があるため，自治体レベルで監視や対処に乗り出すケースが増加している [11]．石川県では，2009 年 4 月から，インターネットに詳しい県立高の教員 8 人が「監視役」を兼務し，週 2 回，弁護士や県警職員のアドバイスを受けながら，ネット上にある県内の小中高校の裏サイトについてチェックを始めたとある．また，東京都江東区では，民間業者へ監視業務を委託したともある．これは，現場任せでは対策が追い付かず，また，毎日のように更新される書き込みを監視することは教職員には無理に近いという現状があるからである．

また，学校裏サイトの事情を把握するために，学校裏サイトチェッカー [12] が立ち上げられた．このサイトは，学校裏サイトと思われるサイトをユーザが登録し，相互に情報を交換し，被害を最小限に抑えるために協力し合うことを目的としている．ただし，学校裏サイトチェッカーに登録されたサイトの多くは，すぐ閉鎖しまた別のサイトを立ち上げるなどの対応をしており，根本的な解決には至っていないと思われる．したがって，監視し削除するのと同時に，このよ

うなサイトにおいて不適切な行為をしないように教育することが必要不可欠であると言える．

1.3　不適切行為を減少させるためのアプローチ

NG ワードフィルタでは，特定の言葉や単語をあらかじめ NG ワードとして登録しておき，NG ワードが出現した場合，その発言をブロックすることで不適切な行為へ対処している．しかし，特定ワードをブロックしてしまうと，表現の幅を狭めることとなり，円滑なコミュニケーションの阻害となる可能性がある．また，特定ワードを同じ意味の別のワードに置き換えることや他の言い回しをすることにより，フィルタリングを回避されてしまう可能性もある．

電子掲示板監視システムは，NG ワードフィルタの発展形であり，目的は監視である．これは，不適切な発言が投稿されたときに管理者へ自動的に通報され，管理者がどう対処を行うかを決定するための支援を行うシステムである．この場合，ユーザがなぜ不適切な発言として処理されたかが分からず，不適切な発言を投稿するユーザの数を減少させることは難しい．また，NG ワードフィルタを発展させたものであり，先ほどと同様に別の言い回しなどで回避される可能性もある．

人間による電子掲示板の監視では，人件費などコスト高が問題であり，大小様々ある電子掲示板のすべてに導入することは不可能である．また，監視者の気分によって判断の難しい発言の評価が変わる可能性がある．

以上より，図 1.3 のプロバイダ責任制限法が対象外で，かつ，相手を挑発・誹謗中傷する不適切な行為（以下,「不適切発言」と呼ぶ）を対象とする．電子掲示板に投稿される発言が不適切発言であった場合，投稿したユーザに対して不適切発言であることを提示することができれば，ユーザは不適切発言がどういうものかを学習する機会が与えられることになる．つまり，ユーザに不適切発言がどういうものかを教育支援することが，不適切発言を投稿するユーザの数を減らすために必要不可欠である．教育支援が実現できれば，過失で不適切発言を投稿していたユーザは，提示される度に不適切発言がどういうものかを学習し，最終的に不適切発言がどういうものかを理解し，過失による不適切発言の投稿をしなくなると期待できる．

1.4　研究目的

ユーザに不適切発言を投稿しないように，不適切発言投稿時にユーザに対し不適切発言であることを提示する教育支援を実現することを目指す．ユーザに不適切発言であることを提示するためには，発言を不適切発言とそれ以外（以下,「通常発言」と呼ぶ）に分類する手法が必要不可欠である．そこで本稿では，特定の言葉をブロックせずに発言全体を評価し，電子掲示板の話題や場にあった評価を行うために必要な要素を明らかにし，発言を通常発言と不適切発言の 2 種類に自動分類する手法を確立することを目的とする．電子掲示板の発言の自動分類が可能になれば，特定ワードを禁止することなく不適切行為を発見し提示することが可能となるため，円滑なコミュニケーションを維持することができる．

本研究が想定する教育支援のユーザモデルと状態遷移を図 1.4 に示す．本稿では，意図せず不適切発言を行ったユーザに対し不適切発言となることを提示することで支援を完了とする．不適切発言を投稿する度に，不適切発言であることをユーザに提示する．最終的にユーザは，不適切発言がどういうものかを理解し経験を積んだ状態になり，過失で不適切発言をする確率が低くなる．ただし，提示された結果によって，ユーザが不適切発言を推敲するのか取りやめるのかについてはここでは取り扱わない．

1.5　本論文の構成

本論文の構成は以下の通りである．

2 章では，電子掲示板の発言を自動分類するために，電子掲示板の特徴を解析する．まず取り扱う発言の種類を定義し，電子掲示板の構造や特徴を解析することにより，発言を自動分類するために必要な要素を明らかにする．

3 章では，電子掲示板の発言単体に着目し，分類をするために必要な要素を明らかにする．そのために，電子掲示板の発言の特徴から，既出発言から不適切発言と通常発言がどういうものか学習させ，学習データを用いた処理を行うことにより，発言の自動分類を目指す．

4 章では，電子掲示板の発言の評価の基準として各電子掲示板の雰囲気の数値化を行い，雰囲気評価を実現する．そのため，電子掲示板の雰囲気を数値化するために必要な要素を明らかに

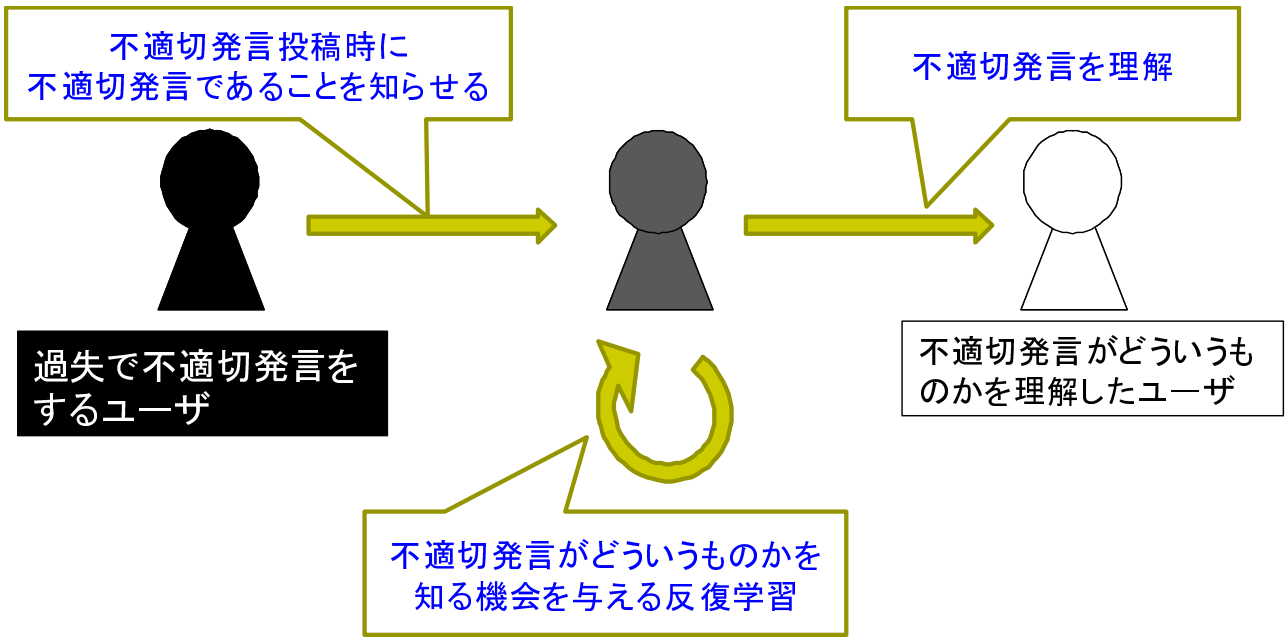

図 1.4: 対象とするユーザモデルと状態遷移

する．電子掲示板の雰囲気はそれぞれの発言の蓄積によって構成されると考えられるため，電子掲示板の雰囲気を数値化するために，発言中の単語に着目する．また，掲示板が盛り上がったことを示す発言の連鎖数にも着目し，電子掲示板の雰囲気評価を行う．

5 章では，利用者教育支援を実現するために必要な発言分類手法を提案し検証を行う．不適切発言と通常発言の狭間の発言が電子掲示板の雰囲気に依存すると推測し，発言ごとの評価に加え発言の雰囲気の変化に着目した電子掲示板の発言の分類評価を実現する．

6 章では，まとめを行う．

第2章　不適切発言の自動分類のための特徴解析

2.1　緒言

1 章では，電子掲示板に書き込まれる不適切発言による問題について述べた．その問題を解決するために，経験不足から不適切発言を投稿するユーザに対し不適切発言がどういうものかを教育することが必要不可欠であると述べた．そこで，投稿発言が不適切発言であった場合，ユーザへ不適切発言であることを提示する教育支援を実現するために，投稿発言を自動的に通常発言と不適切発言に分類する手法が必要であると述べた．本章では，電子掲示板の発言を自動分類するために，電子掲示板システムや発言の閲覧方法および使用方法に着目し，電子掲示板の発言を自動分類するために必要な要素を明らかにする．

2.2　電子掲示板の構造

電子掲示板は，参加者が文字や記号を使って自由に発言を投稿しあうことでコミュニケーションを図ることができる Web 上のアプリケーションの一つである．電子掲示板に投稿された発言は，システムが許す限りいつでも閲覧可能状態となっている．そのため，このことを利用して，非リアルタイムで他のユーザとチャットのようコミュニケーションを取ることができる．

電子掲示板は，個人用 Web サイトに開設する小規模なものから，2 ちゃんねるや Yahoo 掲示板のような大規模なものまで多種多様であり，また，携帯電話の高性能化と普及により，若年層の利用も増加し，いじめの温床とも言われる学校裏サイトと呼ばれる掲示板も存在していると述べた．

電子掲示板の利用方法には，不特定多数が自由に利用できるものや会員制のもの，特定のユーザのみ利用可能なものなどがある．個人が開設した掲示板や 2 ちゃんねるのような電子掲示板

は不特定多数のユーザが利用可能である．これらの電子掲示板では，発言投稿時に発言を識別するために，番号・名前・投稿した時間・ユーザ ID などが同時に表示されることもある．これらの情報の中では，ユーザは発言者の名前のみ自由に決定することができる．そのため，名前を変えて発言を書き込み，あたかも複数の参加者が不適切発言の応酬を繰り広げていると見せる（「自作自演」と呼ばれる）ことが可能である．

会員制のものには，個人を確実に特定できるよう個人情報の記入を求められるものから，メールアドレスのみで登録可能なものがある．会員制の利点としては，不適切行為を行った場合，管理者側が不適切行為を行ったユーザの参加資格の停止処分などの対処ができることが挙げられる．しかし，現在，メールアドレスは用意に取得できるため，1 人が複数のアカウントを取得し，電子掲示板で不適切行為をすることも可能である．したがって，不適切発言を行った発言者の情報をブラックリスト化し取り締まることは，会員登録条件の厳しくない会員制電子掲示板ではあまり現実的ではないと言える．

電子掲示板の表示形態は，図 2.1 のような時系列に沿って発言が一覧表示される単純なもの（「伝言板型」と呼ばれる）や，図 2.2 のような各々の発言の参照関係ごとに表示されるもの（「ツリー型」と呼ばれる）の 2 種類が一般的である．表示形態の違いのため，本稿では，主に図 2.1 の時系列に沿って表示される電子掲示板を取り扱う．

以上より，電子掲示板の発言を自動分類するためには，発言者 ID や投稿時間などよりも発言内容に着目することが重要であると言える．なぜなら，一度不適切発言を投稿したユーザであっても，次に必ず不適切発言を投稿するとは限らないからである．

2.3　取り扱う発言の定義

電子掲示板では，相手を直接見て対話するわけではないため，議論などが白熱すると，ある程度口汚い発言になる場合がある．特に本音の混じる議論では，このような発言（以下,“ フレーミング ”と呼ぶ）は当然であり，結果としてよりよい結論が出ることも多い．したがって，フレーミングをすべて規制してしまうと，議論を妨げる結果となり，電子掲示板の存在意義すら失いかねないとも言われている [13][14]．しかし，行き過ぎてしまうと不適切発言となってし

図 2.1: 伝言板形の電子掲示板

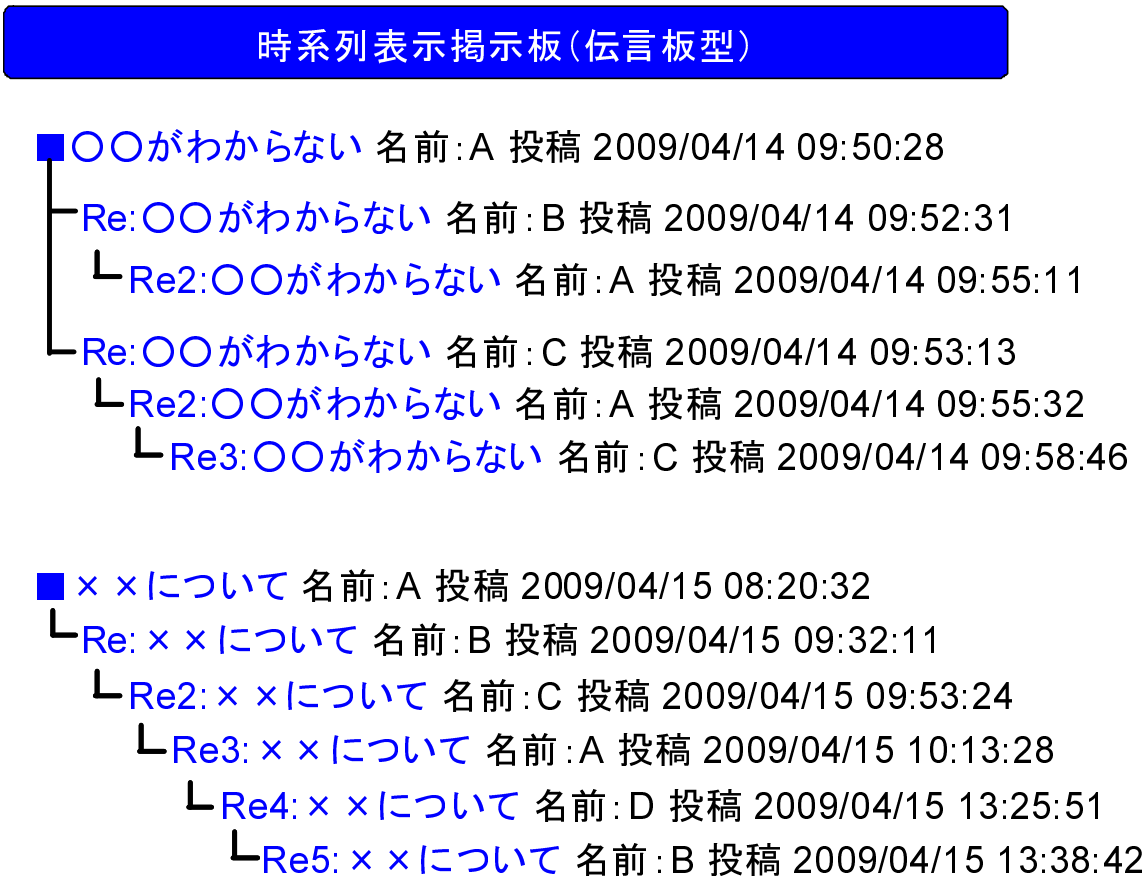

図 2.2: ツリー型の電子掲示板

まう．つまり，不適切発言と通常発言の間には，どちらにも取ることのできるフレーミングのような発言（以下,“ 曖昧発言 ”）が存在していると考えられる．単語単体で見ると不快なものも，通常発言として捉えることもできる発言もあるため，単純に NG ワードフィルタでブロックすると円滑なコミュニケーションを阻害する恐れも生じる．したがって，これらの発言を分類し発言者に対して教育することは，従来の NG ワードフィルタでは不可能である．

そこで本稿では，不適切発言と通常発言と曖昧発言を以下のように定義する．

不適切発言の定義

- 閲覧者をただ不快にする発言
- 閲覧者を挑発する発言
- 閲覧者を煽る発言

通常発言の定義

- 通常コミュニケーションを意図した発言

対象外の発言

- アスキーアートではない発言
- 繰り返し投稿される無意味な発言

アスキーアートとは，文字や記号を用いて絵を描くもので，閲覧を邪魔するものや皮肉などに使われる．今回は文章を評価対象としているため，アスキーアートや繰り返し投稿される無意味な発言は対象外とする．

曖昧発言の定義

- 通常発言とも不適切発言ともとれる発言
- 不適切な単語が含まれているが通常コミュニケーションを意図した発言

不適切発言，通常発言，曖昧発言の 3 種類の発言の関係は，図 2.3 となっている．曖昧発言は，通常発言と不適切発言の境界付近に存在している．教育支援を実現するためには，投稿発言を通常発言と不適切発言の 2 種類に分類することを目指す．そのため，曖昧発言も通常発言と不適切発言の 2 種類に分類することを目指す．

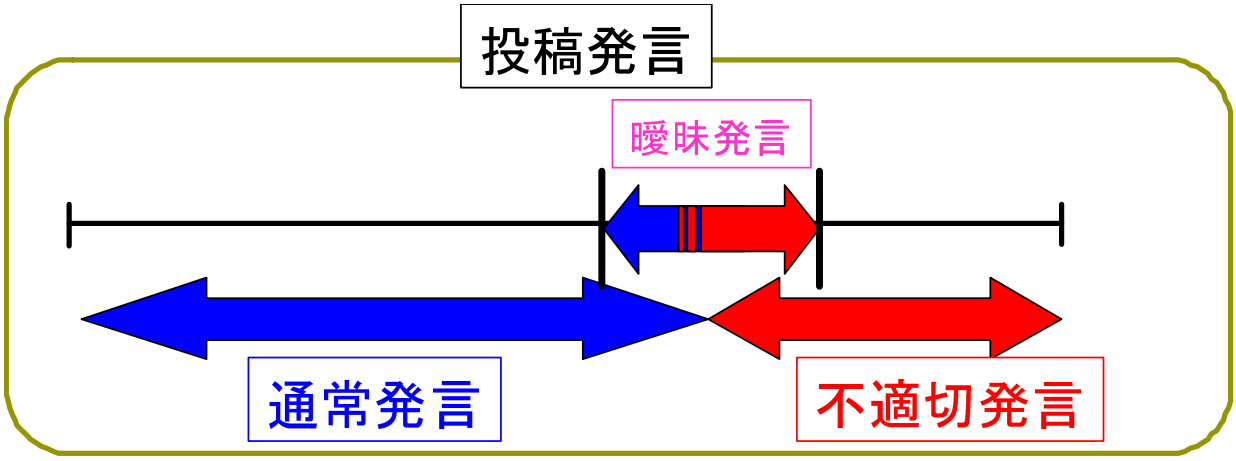

図 2.3: 電子掲示板の発言の関係

2.4　電子掲示板システムの利用方法及び特徴

電子掲示板中の発言は，文字と記号のみで書かれている．それぞれの発言は，発言順に保存され，システムが許容する限り，過去の発言を読むことが可能となっている．そのため，特定の発言に対し，意見を述べたい場合，アンカーと呼ばれる記号（‘>’，‘>>’）や矢印を用いることで可能となる．例として，図 2.4 の電子掲示板を挙げる．223 番の発言は，“>> 212” があるので，212 番の発言に対する発言であることを示している．また，224 番の発言は，“> 引用文 ” があるので，引用文に対する発言であることを示している．225 番の発言は “↑” という矢印を使い，直前の発言に対する発言であることを強調している．このようにアンカーの後ろに指定したい発言番号を用いる方法や，引用文を持ってくる方法や矢印を使い強調する方法がある．これらアンカーや矢印を見ることにより，発言がどのように繋がっているかを確認することが可能である．

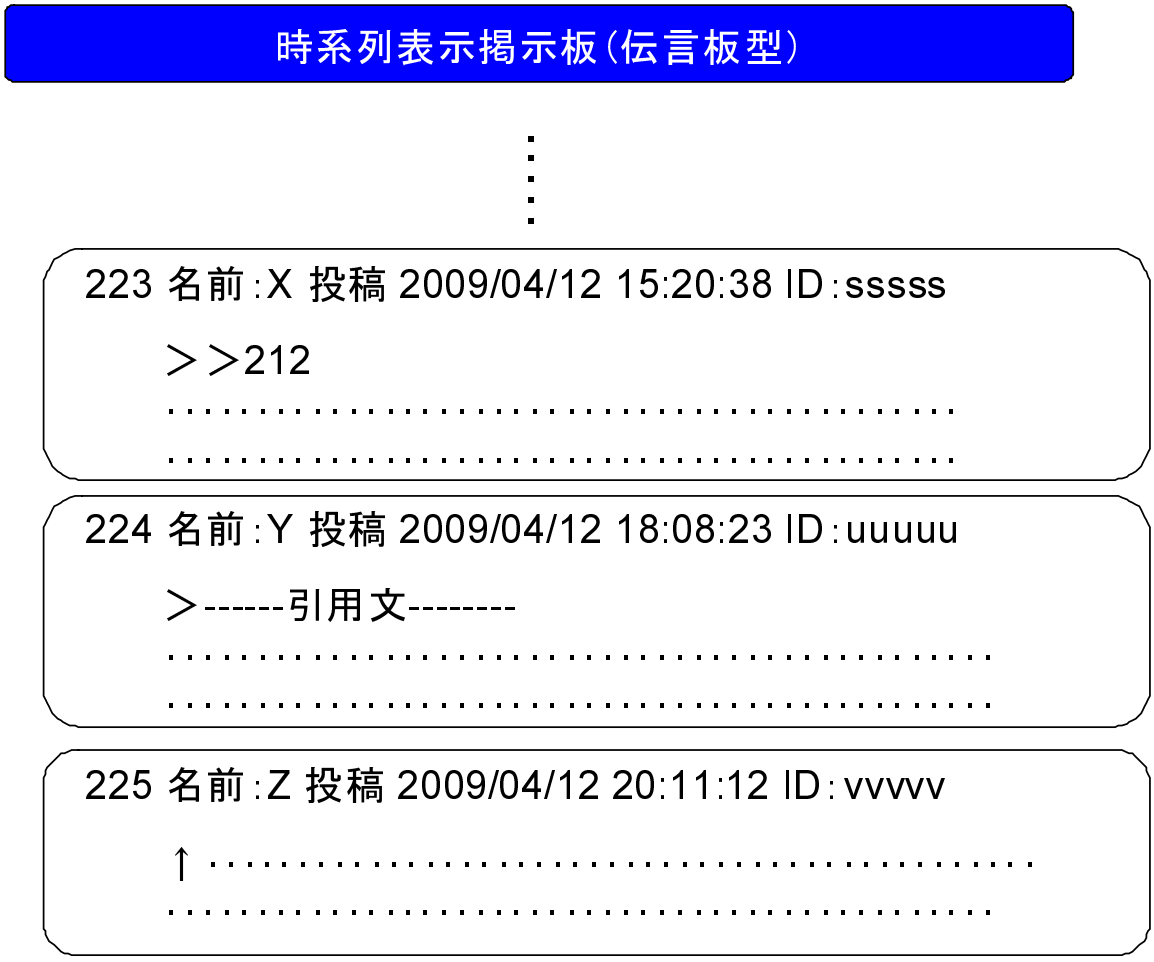

図 2.4: 発言の引用例

電子掲示板は主に話題やトピックごとに細分化されている．そのため，話題やトピックに依存した用語なども存在する．特に，特定の話題やトピックの掲示板では適切な発言も，違う話題やトピックの掲示板では不適切な発言となってしまうこともある．また，同一掲示板であっても会話の方向性により，その場の雰囲気も異なる

以上の特徴より，掲示板ユーザは，掲示板に参加するために，過去の発言を読み，どのような話の流れで，現在に至っているかを理解し発言を書き込むことが一般的なマナーとされている．つまり，ユーザは現状の電子掲示板の雰囲気を読み取り，その場の雰囲気にあった発言を書き込んでいるのである．このことから，電子掲示板の発言を分類するためには，まず，投稿発言がどういうものかを評価し，その後，過去の発言から電子掲示板の雰囲気を評価し，その雰囲気に合った発言かを評価することが必要である．

以上より，電子掲示板の発言を自動分類するためには，以下の 3 点が必要である．

1. 電子掲示板の各発言の分類
2. 電子掲示板の雰囲気の評価
3. 雰囲気を考慮した発言の分類

電子掲示板は文字と記号で書かれ，各々の発言は電子掲示板に蓄積されている．したがって，以上の 3 点を評価するために，電子掲示板の発言および蓄積されている過去の発言を評価要素として分類を行う．

2.5　結言

電子掲示板は，話題が細分化され，その話題専用の掲示板が確立されている．そのため，ある話題の掲示板では不適切でないとされる発言であっても，別の掲示板では不適切となる発言が存在することを述べた．また，1 つの話題の中であっても，やり取りの積み重ねを経て与える印象も変わることも述べた．したがって，電子掲示板の発言を自動分類するためには，各投稿発言を評価分類すること，電子掲示板の雰囲気を評価すること，そして，雰囲気を考慮した分類を行うことが必要である．以上より，図 2.5 のような構成で話を進める．3 章では，電子掲示板の各発言を評価分類するために必要な要素を明らかにし，発言ごとの評価分類が可能であることを示す．4 章では，電子掲示板の雰囲気を評価するために必要な要素を明らかにし，雰囲気の評価が可能であることを示す．5 章では，教育支援を実現するために，雰囲気を考慮し発言を分類することで，対象電子掲示板に合った自動分類が可能であることを示す．

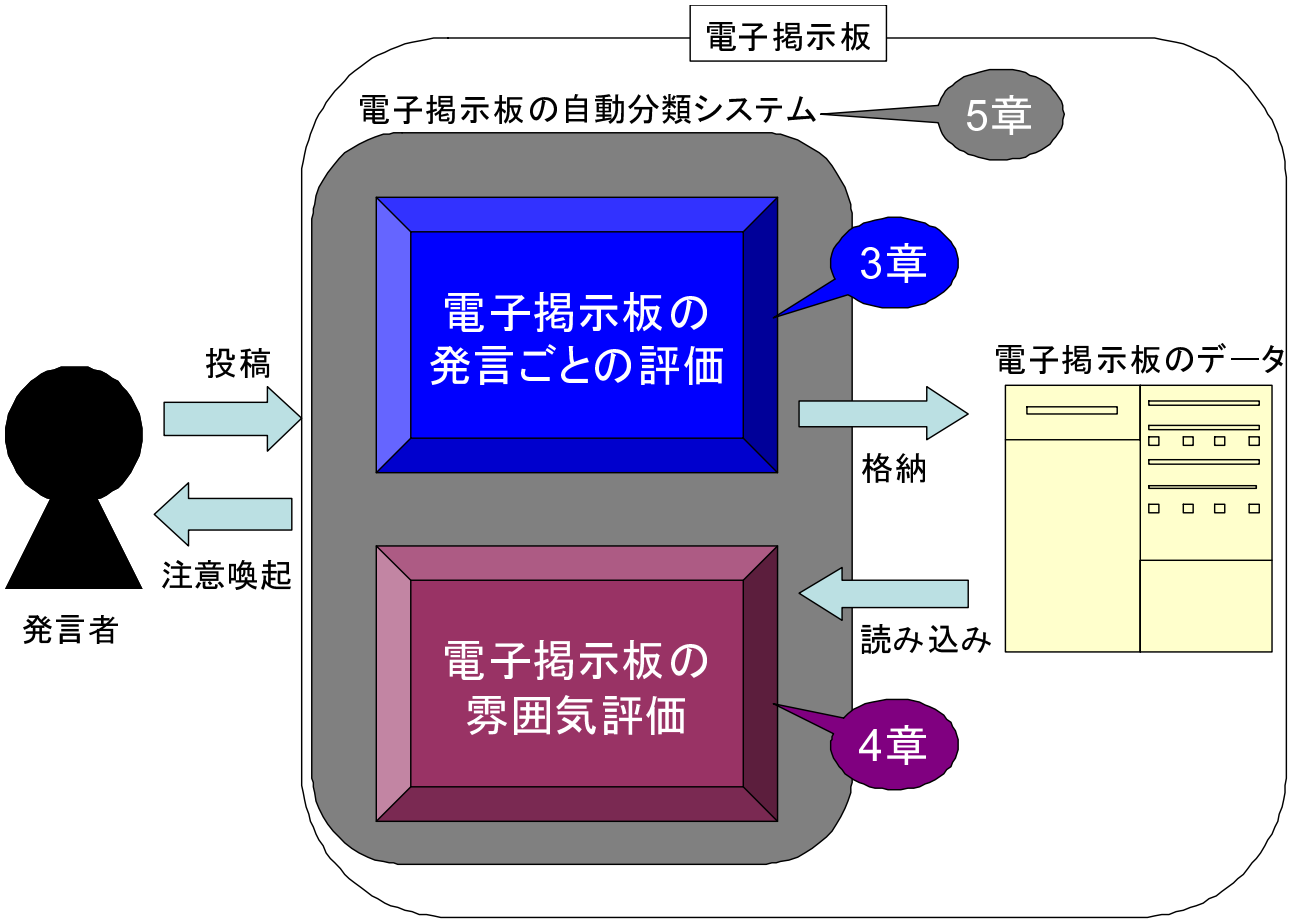

図 2.5: 発言分類システムの概要

第3章　電子掲示板の発言分類について

3.1　緒言

電子掲示板の発言の自動分類手法の確立のために，電子掲示板の発言単位の評価が必要であることを 2 章で述べた．本章では，図 3.1 の発言単位の分類を行うために必要な要素を明らかにする．つまり，電子掲示板の各発言を不適切発言と通常発言の 2 種類に分類するために必要な要素を明らかにし，発言単位の自動分類手法の提案を行う．

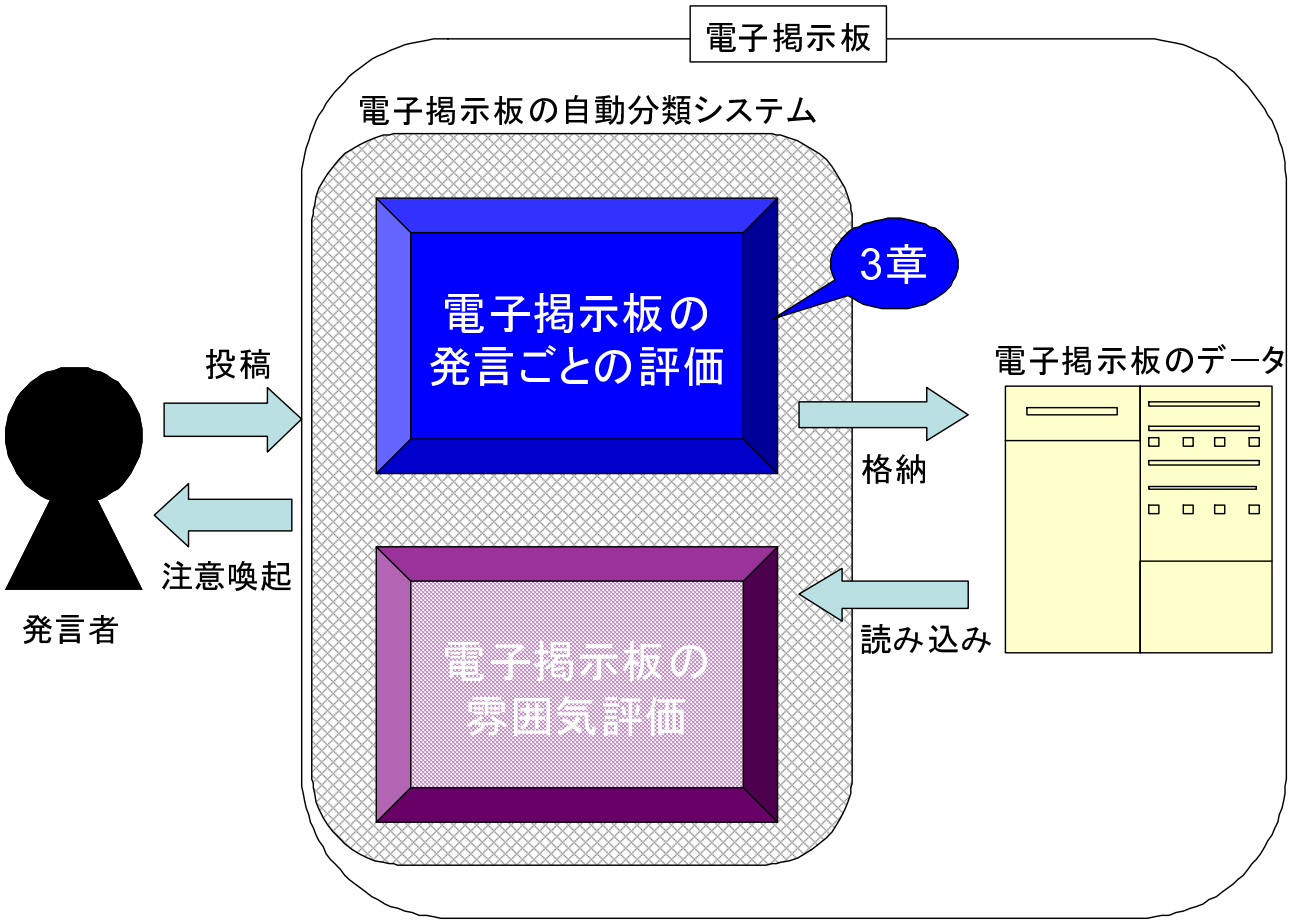

図 3.1: 3 章で提案する発言分類手法

発言単位の分類には，NG ワードフィルタは使用しない．なぜなら，NG ワードフィルタでは，1 章で述べた通り，不適切発言以外の発言も不適切単語が出現したために不適切発言と分類されてしまい，円滑なコミュニケーションを阻害する可能性があるからである．そこで電子掲示板に投稿された発言が，不適切発言となるかどうかを分類評価するために，発言の文章を評価することを考える．

3.2　文章を評価する研究

文章を評価する手法として，N-gram 法や形態素解析法などが一般的であり，これらを用いて全文検索の効率化や Spam メールの分類などの研究が盛んに行われている．

3.2.1　N-gram 法

文書から特定の文や箇所を検索する手法として N-gram 法がある．N-gram 法では，文章の頭から N 文字分切り出し，次に文章の頭から 1 文字目から N+1 文字目まで切り出すことを最後の文字まで繰り返す．例えば $N = 2$ の場合,「文章を評価する」という文章は,「文章」「章を」「を評」「評価」「価す」「する」に分解される．このように 1 文字ずつずらして切り出すことにより，特定の並びの文字列の出現頻度を求める手法である．問題点として，$N = 2$ における「を評」や「価す」など意味のわからないインデックスが大量に作成されるため，データ量の肥大化が挙げられる．

N-gram の出現頻度を用いることにより入力された文書の感情を類似度から推定する研究 [15] や，N-gram 法の欠点となるインデックスの肥大化を解消することを目的とした新たな N-gram 法の提案を行う研究 [16] など N-gram を改良する手法や検索などの研究が盛んに行われている．N-gram 法は定型文には有効であるが，電子掲示板の発言のように不定型文かつ語順も独特の場合が多く，投稿発言が不適切発言と通常発言の 2 種類に分類するためには向かないと思われる．

3.2.2　形態素解析法

N-gram 法とは別に，単語や品詞ごとに分解する手法として，形態素解析がある [17][18]．英文では，単語間にスペース，コンマ，ピリオドがあるため単語ごとに分解することが容易である．しかし，日本語では，単語間にスペースが存在せず，また，語形変化が非常に多いため，分かち書きが非常に難しい．そのため，解析用の辞書を用意し利用することにより，文章を単語に分解する．

形態素解析を用いた電子メールにおける受信者にとって迷惑となる Spam メールの対処が存在する．電子メールは，Web コミュニケーションツールの一つで，電子掲示板と同様に文字や記号でやり取りを行う．そのため，Spam メールと同じように電子掲示板における不適切発言を取り扱うことができるのではないかと考えた．そこで，Spam メールフィルタリング手法に着目し，形態素解析を利用した発言分類を実現することを目指す．

3.3　電子掲示板の発言判別要素

3.3.1　電子メールにおける Spam メール対処方法

電子掲示板の発言は，文字と記号を使って文章で書かれている．本稿では，NG ワードフィルタリングのように特定の単語やフレーズの出現の有無によって発言分類を行うのではなく，文章を評価することにより，不適切発言かどうかを評価したい．そこで，電子掲示板の発言と同様に，文字と記号でやりとりを行う電子メールに着目する．

電子メールには，受信者が不快に思う Spam メール（受信者が事前に承諾しない広告メール）がある．Spam メールが大量に送信されると，重要なメールが埋もれてしまい，正常な利用を妨げる問題が発生する．現在，この問題に対し，主に 2 つの対処方法がある．

1. 送信者情報を利用した対処
2. 内容を利用した対処

以下に，それぞれについて詳細を説明する．

1. 送信者情報による対処

   特定の送信者のメールアドレス，または，特定のドメインからのメールを受信しないことにより，Spam メールを回避する対処方法である．代表例として，ブラックリストが挙げられる．これには，受信者が設定する場合とサーバ側で設定される場合の 2 通りが存在する．受信者側で設定する方法として，メーラーソフトウェア [19] が挙げられる．ブラックリストのデータは更新されるため，分類精度を高く維持できる．

2. 内容による対処

   Spam メールは勧誘や広告を目的としたメールである．そのため，電子メールの文中には，誘導先の URL や勧誘を目的とする言葉が多く使われている．そこで，事前に複数の Spam メールを用意し，Spam メールに出やすい単語を学習させ，各々の単語が出現したとき，そのメールが Spam メールである確率を算出する．そのデータを用いて，新たな電子メールが Spam メールである確率を算出し，あらかじめ設定した閾値を超えた場合，Spam メールと判定する対処方法である．代表例として，ベイジアンフィルタ [20] が挙げられる．ベイジアンフィルタは，学習データを増やすことにより，精度を上昇させることが可能である．ベイジアンフィルタの精度の上昇を目的とした研究が数多くなされている [21][22].

ここで，電子掲示板の不適切発言を電子メールの Spam メールとみなし，同様に確率を算出することにより，不適切発言を特定の単語の有無で判断せず，分類できるのではないかと考えた．ただし，電子掲示板の発言と電子メールには多くの違いがあるため，そのまま適用することは難しい．そこで，電子掲示板の発言と電子メールの相違点を挙げ，不適切発言を自動分類するために必要な要素を明らかにする．

3.3.2　電子掲示板の発言と電子メールの相違点

電子メールの送信者情報に対応するものとして，電子掲示板の発言者の名前や ID がある．しかし，2 章で述べた通り，発言者の名前はユーザが自由に決定でき，いつでも自由に変更可能である．また，ユーザの接続 IP アドレス情報などから自動的に決定される ID も，容易に変更

可能である．そのため，特定の発言者名や ID によって，不適切発言かどうか評価することは難しい．さらに，一度，不適切発言を行ったユーザであっても，次に必ず不適切発言を投稿するとは限らない．特に，過失で不適切発言を行った可能性もあるため，特定の発言者名や ID ですべての発言を不適切発言と判別することは，活発な議論の妨げやユーザ離れにつながる可能性がある．

また，Spam メールと不適切発言には文体にも違いがある．Spam メールは広告が主目的であり，Spam メールに出現しやすい単語の傾向が捕らえやすく，ある程度文体も固定される．しかし，電子掲示板の発言は，閲覧者，発言したときの状況およびトピックによって不適切発言となる発言が変わるため，出現しやすい単語の傾向が捕らえにくい．例として，曖昧発言が挙げられる．曖昧発言は，不適切発言でも使われる単語が含まれているが，議論などが白熱している状況や強調する場合でしばしば出現する．そのため，単語単体をみるだけでは，曖昧発言が通常発言と不適切発言のどちらに分類されるかを判断することは難しい．

また，電子掲示板の発言には，引用文を用いて発言を行う場合がある．引用文が不適切発言で，発言趣旨が不適切発言を注意する場合，NG ワードフィルタリングのように不適切単語の有無で分類した場合，不適切発言と分類されることになる．さらに，発言を推敲しないで投稿する場合もあるため，最初と最後で主張が逆転し一貫しない発言も多く存在する．特に，不適切発言によく出現する単語であっても，褒め言葉として使われる場合もあり，Spam メールのように，各単語が出現したとき不適切発言である確率を利用するだけでは不十分であると予測される．したがって，分類要素である単語が発言中の単語の出現状況に応じて値を変化させることができれば，電子掲示板の発言を自動分類することが可能となる．

3.3.3　発言分類要素の提案

発言中の単語の出現状況に応じて，分類要素である単語の値を変化させるために，出現した単語の総当たりによる 2 つの単語の組み合わせ（以下，“単語のペア”）を作成することで表現することを提案する．例として，ある発言中に “単語 A”，“単語 B”，“単語 C”，“単語 D” が出現した場合を考える．“単語 A” が不適切発言に出やすい単語であるが，褒め言葉としても使われるものとする．この場合，単語のみで分類を行った場合，不適切発言と分類される可能性が

高い．しかし，単語のペアを作成することにより，“単語 A-単語 B”，“単語 A-単語 C”，“単語 A-単語 D” と “単語 A” に関連する分類要素が 3 パターン生成される．その結果，組み合わせた単語により複数の値を持つことが可能となり，通常発言にも不適切発言にも使われる単語を適切に分類することが可能となることが期待できる．

3.3.1 項より，電子掲示板の発言内容にのみ着目する．そして，発言中の単語及び単語の組み合わせを用いてベイジアンフィルタの処理方法を用いることにより，発言の自動分類手法を確立する．

3.4　発言単位での自動分類手法

前節で述べた通り，一貫性のない電子掲示板の発言を自動分類するために，単語及び単語のペアを用いることを提案した．そこで本稿では，軽い処理で行える範囲内で，発言中の単語に新たな手がかり加える判別手法を提案する．提案手法は，図 3.2 のように既出発言から不適切発言と通常発言を学習させる学習フェーズと，その学習データを使って新たな発言を分類する分類フェーズの 2 つで構成される．提案手法を実現する 2 つのフェーズである学習フェーズと分類フェーズの詳細について述べる．

学習フェーズ

学習フェーズでは，不適切発言と通常発言から単語と単語のペアの出現数を学習する．このフェーズは，次の 3 つの Step で構成される．

1. 主観評価による発言の分類
2. 発言の分解とペアの作成
3. データベースへの登録

まず，Step1 では，既出発言を実際に読み，主観評価によって不適切発言と通常発言に分類する．このとき，学習させた通常発言と不適切発言の数を登録する．

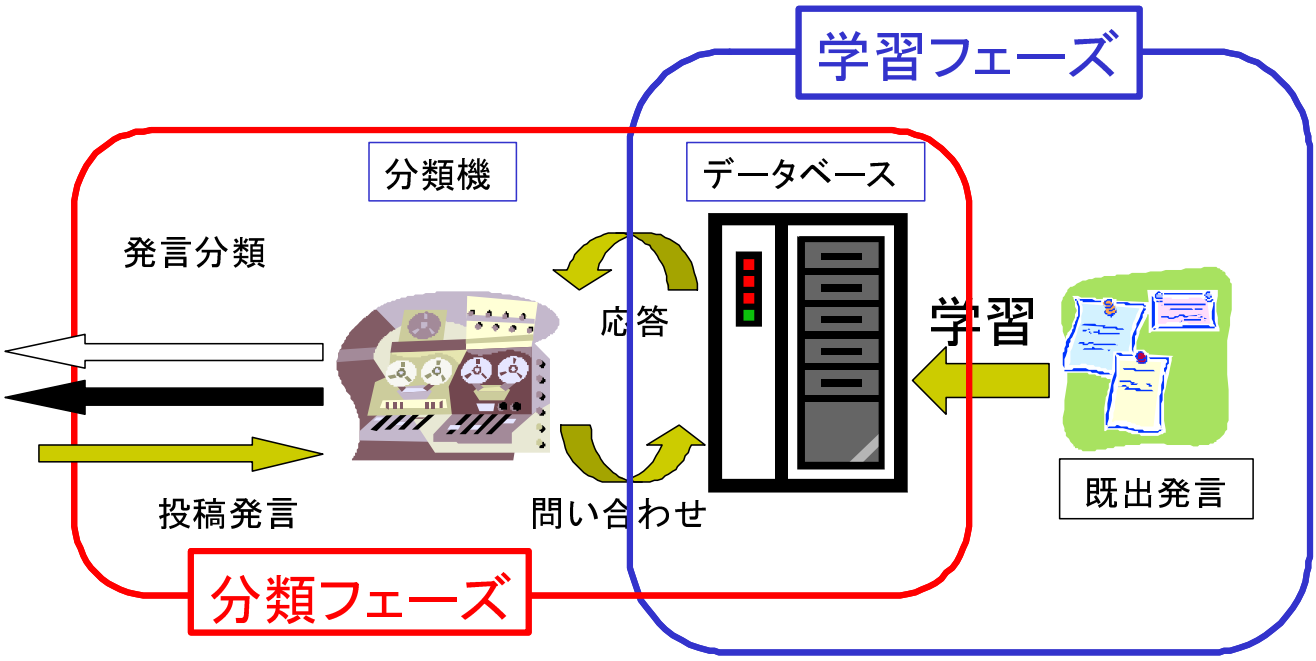

図 3.2: 発言単位の発言分類システムの構成

Step2 では，発言を単語に分解するために形態素解析を行う．その後，助詞・助動詞を除く．これは，事前に掲示板の発言を調査した結果，助詞・助動詞は不適切発言・通常発言のどちらにもよく出現し，出現数にあまり差異が見られなかったためであり，助詞・助動詞を除くことにより，判別フェーズにおける処理をより軽くすることが可能になるからである．その後，残った単語から，単語のペアを作成する．作成方法は，図 3.3 のように総当りで作成する．これにより，複数の意味を持つ単語や使われ方によって複数の意味を持つ単語に対応する．この処理を発言単位で行う．

Step3 では，Step2 で作成した単語と単語のペアをそれぞれのデータベースに登録する．このとき登録されるデータは，単語のデータベースには，単語とその単語が不適切発言に出現した数，その単語が通常発言に出現した数，学習させた通常発言の数，学習させた不適切発言の数の 5 項目である．単語のペアの場合も同様である．Step2・Step3 の具体的な流れを図 3.3 に示す．

分類フェーズ

分類フェーズでは，学習フェーズで登録したデータを用いて，新たに投稿された発言を不適切発言と通常発言の 2 種類に分類する．このフェーズは，次の 7 つの Step で構成される．

1. 発言の分解とペアの作成
2. データベース問い合わせ
3. 各要素が不適切発言に出現する確率の算出
4. 発言が不適切発言である確率の算出
5. 単語と単語のペアの結果の複合
6. 単語による結果と単語のペアによる結果が異なる場合の再評価
7. 発言者への提示

まず，Step1 では，投稿発言を形態素解析し単語に分解する．その後，助詞・助動詞を除き，残った単語から単語のペアを作成する．その後，発言に出現した単語及び単語のペアとそれぞれの出現数を取得する．

Step2 では，分解した単語と単語のペアが過去に不適切発言と通常発言に出現した数および学習させた発言数を取得する．

Step3 では，獲得した出現数から各単語・各単語のペアが不適切発言に出現する確率を算出する．以下に，単語の場合の出現確率の算出式を示す．なお，単語のペアの場合でも算出式は同じである．発言が不適切発言となる確率を算出するために，まず，出現単語が不適切発言に出現する確率をそれぞれ算出する．例として，ある単語の不適切発言に出現する確率を求める．

学習データベースにおいて，単語 x^i が不適切発言に出現した総数を $S_b(x^i)$，通常発言に出現した総数を $S_g(x^i)$，学習させた不適切発言の総数を BN，学習させた通常発言の総数を GN としたとき，求める単語 x^i が不適切発言に出現する確率を $\pi(x^i)$ とすると，式 3.1 となる．

$$\pi(x^i) = \frac{\frac{S_b(x^i)}{BN}}{\frac{S_b(x^i)}{BN} + \frac{S_g(x^i)}{GN}} \tag{3.1}$$

Step4 では，Step3 で算出した各単語の確率を利用して発言が不適切発言である確率を算出する．ここで，発言 y が不適切発言である確率を $p(y)$ とすると，式 3.2 で計算される．

$$p(y) = \frac{(PB)^{1-m}\prod_{i=1}^{m}\pi(x^i)}{(PB)^{1-m}\prod_{i=1}^{m}\pi(x^i) + (1-PB)^{1-m}\prod_{i=1}^{m}(1-\pi(x^i))} \tag{3.2}$$

式 3.2 中の は，式 3.3 で算出する．

$$PB = \frac{BN}{BN + GN} \tag{3.3}$$

算出される値は 0 から 1 の間の値をとり，1 に近いほど不適切発言であることを示す．

Step5 では，単語を用いて算出された不適切発言である確率と単語のペアを用いて算出された不適切発言である確率を用いて，不適切発言か通常発言かの分類を行う．算出された確率が設定した閾値を超えた場合，“不適切発言候補” と分類する．不適切発言候補は主観評価による不適切発言に相当する．逆に，閾値を超えなかった場合，“通常発言候補” であると分類する．通常発言候補は，主観評価による通常発言に相当する．単語による分類結果と単語のペアによる分類結果の 2 種類の組み合わせにより，表 3.1 のように分類する．それぞれのパターンについてどう分類するかについて述べる．

表 3.1: 単語と単語のペアの組み合わせ分類結果

分類パターン	単語による分類	単語のペアによる分類	分類結果
パターン 1	不適切発言候補	不適切発言候補	不適切発言候補
パターン 2	不適切発言候補	通常発言候補	未判定発言
パターン 3	通常発言候補	any	通常発言候補

分類パターン 1 は，単語を用いて算出された不適切発言である確率と単語のペアを用いて算出された不適切発言である確率がともに閾値を超えたパターンである．この発言は，不適切発言に出現しやすい単語が多く，また，単語のペアを作成しても同様の結果となったことから，不適切発言候補に分類する．

分類パターン 2 は，単語を用いて算出された不適切発言である確率は閾値を超えているが，単語のペアを用いて算出された不適切発言である確率は閾値以下であったパターンである．これは，不適切発言によくでる単語が多く出現するが，ペアを作成すると通常発言となる発言である．つまり，フレーミング発言や，首尾一貫しない発言，引用文を用いて逆の意見を述べる発言がこれに該当していると言える．このようなパターンは，不適切発言候補とも通常発言候補とも受け取ることができるため，ここでは未判定発言として分類することにする．したがって，このパターンの場合，通常発言候補なのか不適切発言候補なのか再分類を行う必要がある．Step6 で再分類方法について述べる．

分類パターン 3 は，単語を用いて算出された不適切発言である確率は閾値以下であったパターンである．この場合，不適切な単語がほぼないため，単語のペアによる分類にかかわらず通常発言候補と分類する．単語のペアによる分類が不適切発言候補の場合は，記号を多用する発言に多くみられる．つまり，笑いを意味する “w” や “！”，“？” といった記号がペアを作成すると不適切発言でよく使われる要素になるのである．

Step6 では，Step5 において対象発言が未判定発言の場合について，再分類を行う．未判定発言が不適切発言候補か通常発言候補かの分類要素として，発言中の不適切単語の構成率に着目する．なぜなら，未判定発言には必ずある程度の不適切単語が含まれている．閲覧者は，文中に含まれる不適切単語が多ければ多いほど，その発言を不適切発言と感じるのではないだろうかと推測したからである．そこで，分類に使用した単語のうち，閾値以下の単語を通常単語とし，閾値以上の単語を不適切単語としたとき，通常単語の構成率を式 3.4 で表す．

$$\text{構成率} = \frac{\text{評価に使われた通常単語数}}{\text{評価に使われた通常単語数} + \text{評価に使われた不適切単語数}} \tag{3.4}$$

構成率がある一定値を超えた場合，通常発言候補に分類する．逆に，構成率がある一定値を超えない場合，不適切発言候補に分類する．

最後に Step7 では，Step5，Step6 で出力された結果，不適切発言候補と評価された場合のみ，ユーザへ注意を促す．

Step1 から Step5 までの流れを図 3.4 に，Step6 の流れを図 3.5 に示す．

3.5　分類精度の検証

3.5.1　実験環境

検証実験を行うため，提案した指標を用いた試作システムを構築した．試作システムは，表 3.2 の通り CPU：Core2Duo の 1.86GHz，メモリ：4GB のスペック，OS：FreeBSD7.0 のパーソナルコンピュータに，プログラム言語 Ruby（version1.8.6）で記述し，データベースは Mysql（version5.0.45）を使用し，Web サーバは Apache（version2.2.6）を使用した．また，学習・分類時に行う形態素解析には，形態素解析用ソフト Mecab（version0.96）を使用した．この際，電

表 3.2: 試作システム環境

環境	スペック
CPU	1.86GHz
メモリ	4GB
OS	FreeBSD7.0
プログラム言語	Ruby（version1.8.6)）
データベース	Mysql（version5.0.45）
Web サーバ	Apache（version2.2.6）
形態素解析用ソフト	Mecab（version0.96）

子掲示板で多用される造語などはあらかじめ登録し対応できるようにしている．また，本提案システムは，不適切発言候補に分類された場合，発言者へ注意喚起するために，発言を書き込んだとき，その発言が不適切発言候補であることを提示する機構として，分類フェーズの処理をスクリプト化し html ファイルに追加した．検証実験用データは，いわゆる学校裏サイトと呼称される掲示板の発言を収集し利用した．収集したデータは 2004 年 11 月から 2008 年 4 月までに書き込まれた発言で，複数の人間の主観評価により誹謗中傷発言を 520 発言，通常発言を 1276 発言抽出し，データベースに登録している．

3.5.2　検証内容

本提案手法によって，発言単位の自動分類が可能であり，単語のペアが有効であることを検証する．分類結果として，以下の 4 通りがある．

1. 通常発言を通常発言候補と分類
2. 不適切発言を不適切発言候補と分類
3. 通常発言を不適切発言候補と分類 (False Positve)
4. 不適切発言を通常発言候補と分類 (False Negative)

(1)，(2) の判別がよく，(4) の False Negative が少ない分類ができることが望ましい．但し，(3) の False Positive の数はある程度許容する．なぜなら，発言が掲載されるまえに発言者に注意を促す教育支援を実現するために分類を行うため，False Negative は許容できないが，False Positive が発生しても，他者に悪影響を与える発言がされることはないためである．

単語のペアを作成する理由は，誹謗中傷でも褒め言葉としても使われる単語に対し，複数の値を持たせることであった．実際に，単語のペアを作成することにより，ペアとなった単語により値が変化し，かつ，不適切単語の組み合わせの場合は高い値を，それ以外では，低い値を持つことができているかを検証する．

3.6　検証結果

検証実験の結果として，まず，判別手法 Step5 までの結果を表 3.3 に示す．

表 3.3: 単語と単語のペアを利用した発言の分類結果

分類パターン	システムによる分類	主観評価：通常発言	主観評価：不適切発言
パターン 1	不適切発言候補	102	28
パターン 2	未判定発言	248	54
パターン 3(通常発言候補)	通常発言候補	41	1
パターン 3(不適切発言候補)	通常発言候補	605	17

パターン 1 は，分類システムが不適切発言候補と分類した発言である．その発言のうち，主観評価によって通常発言であると分類された発言，つまり False Positive が 102 発言あったことを示している．また，主観評価によって不適切発言であると分類された発言が 28 発言あったことを示している．

パターン 3(通常発言候補) は，分類システムによると通常発言候補に分類される．その発言のうち，主観評価によって通常発言であると評価されたものは 605 発言あったことを示している．それに対し，主観評価によって不適切発言に分類されたもの，つまり，False Negative が 17 発言あったことを示している．False Negative となった発言の多くは，不適切な単語だが，データベースに登録されていない単語があったことや，誤字脱字や電子掲示板でよく使われる表現によって，正しく形態素解析できずに通常発言と評価されてしまったことが原因であった．また，通常単語のみを使った皮肉の発言も，False Negative となってしまっている．

学習データの不足や，形態素解析用辞書の不足に対しては，学習データベースの充実，及び，形態素解析用辞書の充実で対応できる．ただし，学習させすぎると登録されるデータの数が膨大となり，分類精度を落とすことになりかねない．よって，学習データも出現頻度が少ないものは削除する，または，サブのデータベースを用意し，一定数を超えたときに学習データとして使うなどの対応策が必要であると言える．また，通常単語のみを使った皮肉に対しては今のところ分類不可能である．

パターン 3(不適切発言候補) は，単語による分類は通常発言候補であるが，単語のペアによる分類は不適切発言候補である場合である．この発言は，感情を示す記号が多く使われている発言に多いと予測し，通常発言候補に分類される．実際に，主観評価による分類が通常発言である発言が 41 発言に対し，主観評価による分類が不適切発言である発言，つまり，False Negative は 1 発言のみであった．パターン 3 の結果より，単語による分類が通常発言候補の場合，通常発言候補に分類することが妥当であることが示された．

パターン 2 は，通常発言候補にも不適切発言候補にもなりうる未判定発言であり，再分類が必要であると述べた．再分類には，発言中の分類要素の通常単語構成率によって分類される．事前実験より，不適切発言の通常単語構成率の平均は 66%であり，通常発言の通常単語構成率の平均は 76%であった．そこで，70%を閾値に設定した．通常単語構成率が 70%未満の場合，不

適切発言候補に分類し，70%以上の場合，通常発言候補に分類する．その結果，パターン 2 の未判定発言は，表 3.4 となった．

表 3.4: 単語の構成率に着目した未判定発言の分類結果

分類結果	主観評価：通常発言	主観評価：不適切発言
未判定発言	248	54
誤判定	77	25
正判定	171	29

未判定発言のうち，主観評価によって通常発言であると分類された発言を通常発言候補と正しく分類できたものが，248 発言中 171 発言であり，正答率 69．0%であった．また，主観評価によって不適切発言であると分類された発言を不適切発言候補と正しく分類できたものが，54 発言中 29 発言であり，正答率 53.7%であった．これより，未判定発言が通常発言候補となるのか不適切発言候補となるのかを分類する要素として，発言中の通常単語の構成率はあまり有効ではないという結果が得られた．

パターン 2 の再分類結果を反映した最終的な結果は表 3.5 となった．表 3.5 より，通常発言を

表 3.5: 提案手法による発言単位での分類結果

分類結果	主観評価：通常発言	主観評価：不適切発言
通常発言候補	817	44
不適切発言候補	179	57
正分類率	82.0%	56.4%

82.0%正しく自動分類できたが，不適切発言を正しく自動分類できたのは 56.4%であった．これは，未判定発言を正しく分類できなかったのが原因の一つとして考えられる．学習データは発言を個別に読み，主観評価による分類を行った．しかし，主観評価を行う際に，発言順に読んで分類してもらったため，発言がどういう状況で書かれているのかを無意識的に考慮した可能性が高い．そのため，通常発言候補にも不適切発言候補にも受け取ることができる未判定発言の分類に影響を与えたのではないかと考えられる．そこで，未判定発言が，通常発言候補に分類されるのか不適切発言候補と分類されるのかは，発言中の単語の構成率ではなく，その発言がどういう流れで投稿されたかが分類要素として重要であると考えられる．

例として，不適切発言がまったくない掲示板を考える．この場合，不適切単語を含んだ曖昧発言が出現した場合，閲覧者が冷静である可能性が高くどういう意図をもっての発言かを理解し主観評価では通常発言と分類される可能性が高い．逆に，不適切発言が多く出現している場合，その掲示板で曖昧発言が出現した場合，雰囲気が悪いため不適切発言として受け入れられる可能性がある．つまり，曖昧発言が通常発言なのか不適切発言なのかを決定するのは，電子掲示板のその場の雰囲気に依存すると推測できる．したがって，電子掲示板の雰囲気を考慮した発言分類が必要である．

単語と単語のペアの有効性を示すため，“キモい” という単語を例示する．この単語は,「気持ち悪い」という単語から派生したもので，意味もほぼ同じである．しかし，現状，対象を褒める言葉として使われることも多く，不適切発言でも通常発言でも使われる単語である．この単語と単語のペアが持つ値は，表 3.6 となった．“キモい” は，基本的に「気持ち悪い」の意味な

表 3.6: 単語のペアの有効性

登録データ	不適切発言に出現する確率	使用例
キモい	0.958	基本的には不適切単語．褒め言葉としても使用.
キモい-糞	0.903	どちらも不適切単語．ペアでも高い値を維持
お前-キモい	0.813	やり取りの相手を挑発していると予測
こんな-キモい	0.397	物体に対する評価をしていると予測
この-キモい	0.188	特定の物に対する評価をしていると予測
だけど-キモい	0.279	反対の意見を述べていると予測

ので，この単語に与えられた値も 0.958 と非常に高い．しかし，“この” や “こんな” といった対象を示す単語と組み合わせることにより，値がそれぞれ，0.397，0.188 と低くなっている．この結果は，まさに予測した通りであり，発言が通常発言候補なのか不適切発言候補なのか分類するために，単語のペアは十分有効であることが判明した.

3.7　結言

電子掲示板のユーザ教育支援のために，発言の自動分類手法が必要であった．そこで，まず発言単位での分類手法の提案を行った．発言は，文字と記号でのみ構成されているため，分類

要素として発言中の単語や文章が重要である．そこで，発言を分類するために，電子掲示板の発言と同様に文字と記号で構成されるものとして電子メールに着目した．電子メールにも通常のメールと Spam メールが存在し，内容から分類する Spam メールフィルタリングがある．そこで，電子掲示板の不適切発言を電子メールの Spam メールと同様に分類できるのではないかと考えた．本稿では，Spam メールフィルタリングに使われるベイジアンフィルタに着目し同様の処理方法で発言の分類をすることを考えた．ただし，電子掲示板の発言には，1 つの単語においても使われ方によって間逆の意味を持つものも多い．そこで，単語に複数の意味を持たせるために単語を組み合わせ，複数の意味を持たせることで，この問題を回避した．

以上より，発言中の単語および，発言中の単語の組み合わせを用いて，ベイジアンフィルタと同様の処理を行うことによって，各発言が不適切発言となる確率を算出し設定した閾値より高い値が算出された場合，不適切発言候補であると分類した．単語と単語のペアの出力結果をもとに，発言を通常発言候補・不適切発言候補・未判定発言の 3 種に分類した．その後，未判定発言を発言中の通常単語構成率をもとに通常発言候補か不適切発言候補かに分類することを試みた．しかし，結果は 53%程度であり，未判定発言を評価するのに十分な結果が得られなかった．この結果，未判定発言を分類する要素は，発言中の通常単語の構成率ではないことが明らかとなり，電子掲示板の雰囲気に依存するのではないかという予測が得られた．

本章では，電子掲示板の各発言を通常発言候補・不適切発言候補の 2 種類に分類するパラメタとして，発言中の単語および発言中の単語の組み合わせが有効であることを確認した．

次章にて，電子掲示板の雰囲気を自動評価する手法の提案を行う．

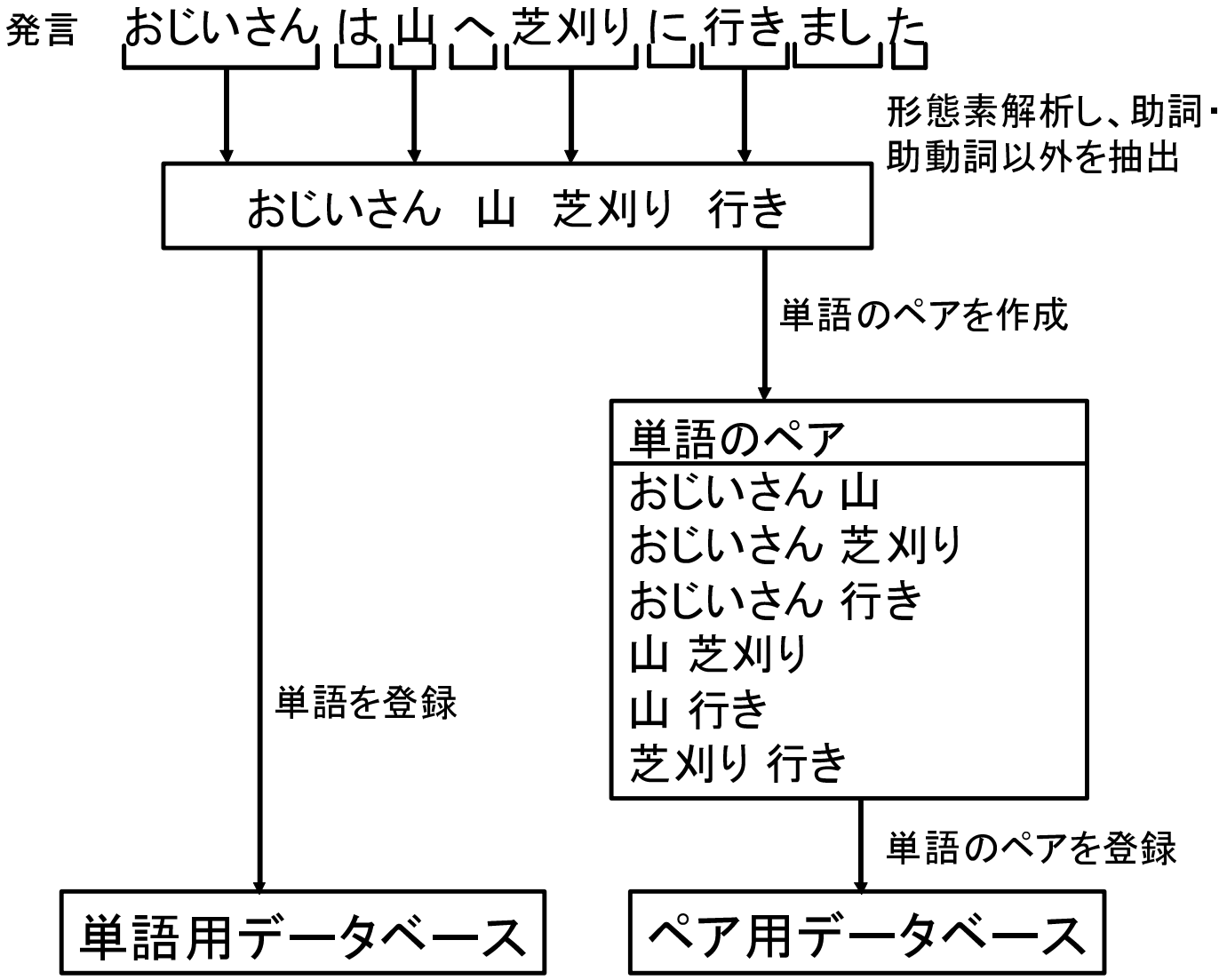

図 3.3: 学習データの登録方法の例

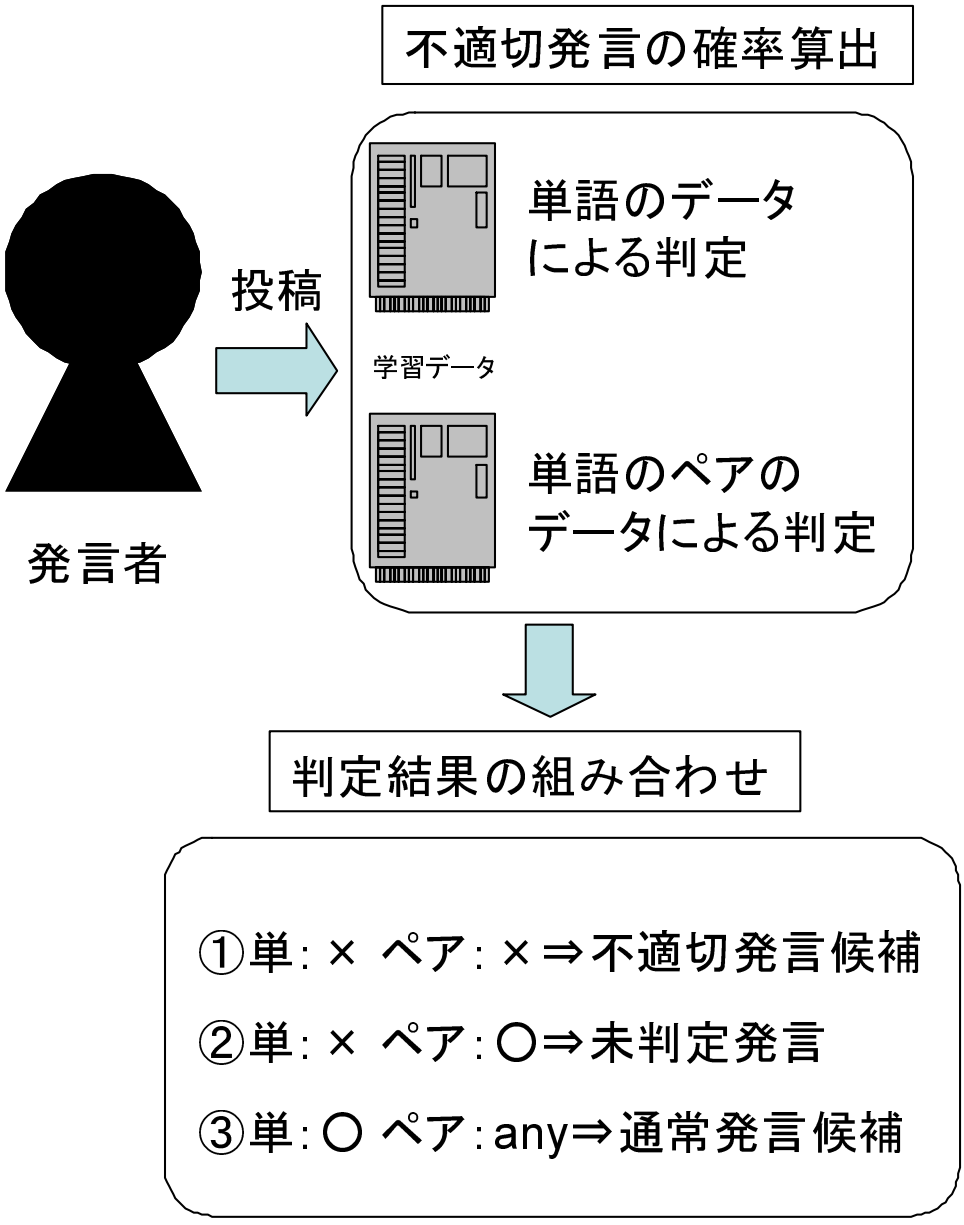

図 3.4: 発言単位の分類の流れ（その１）

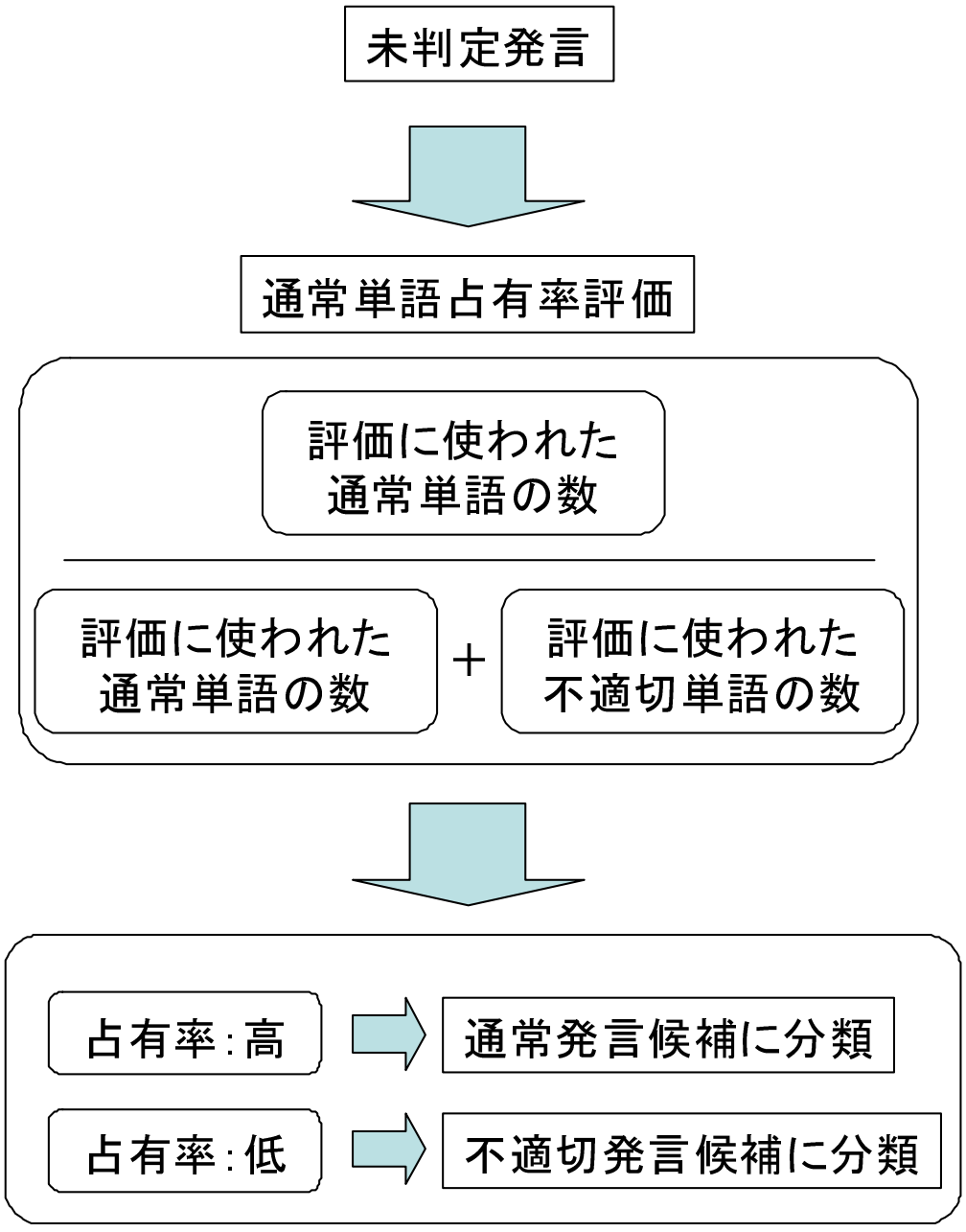

図 3.5: 発言単位の分類の流れ（その 2）

第4章　電子掲示板の雰囲気評価について

4.1　緒言

3章で，電子掲示板の発言を通常発言候補・不適切発言候補・未判定発言の3つに分類した．その結果，未判定発言を分類するためには，電子掲示板の雰囲気が有効であると推測された．そこで，本章では，図4.1の電子掲示板の雰囲気を評価するために，電子掲示板の発言から必要な要素を明確化する．

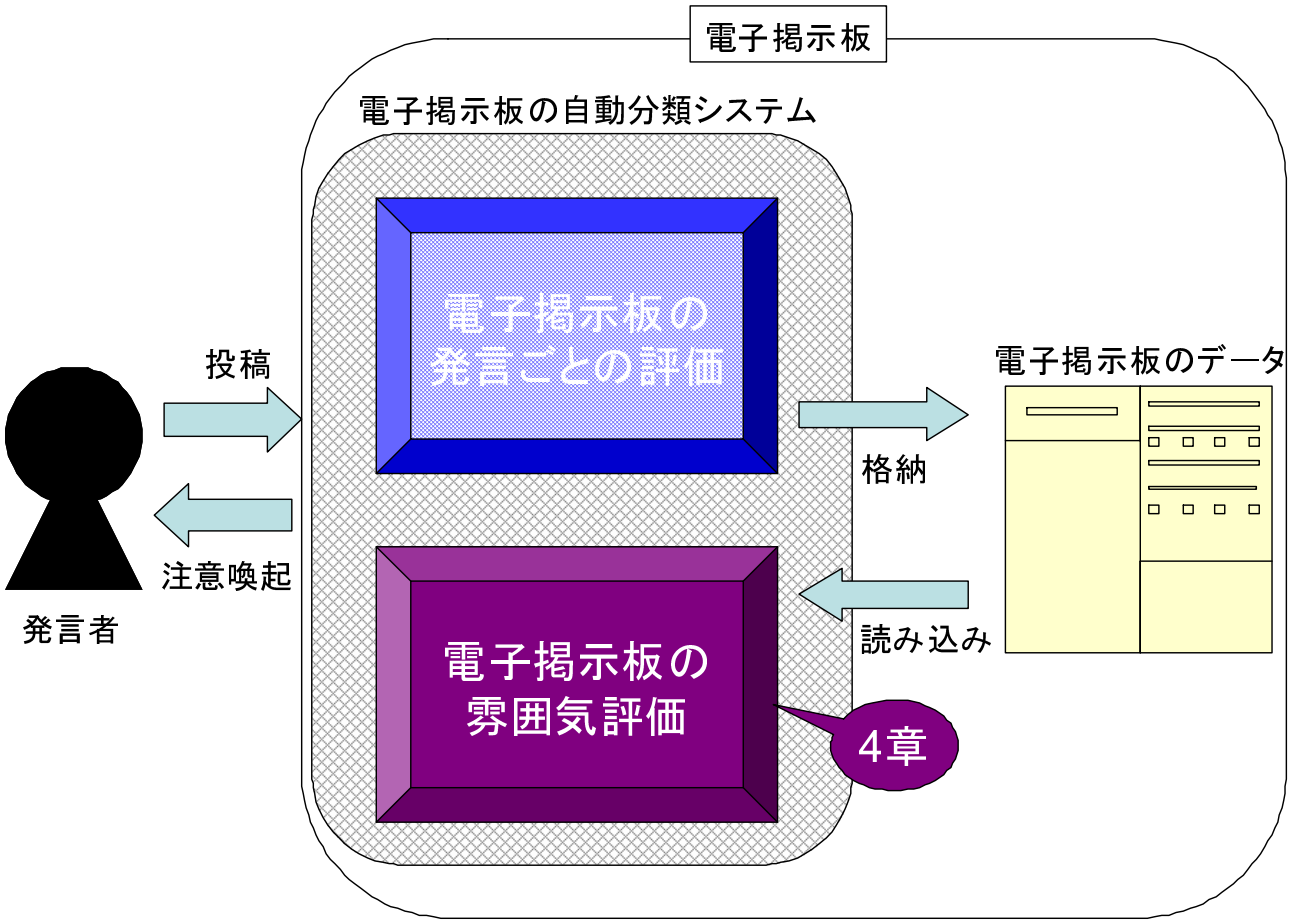

図 4.1: 4章で提案する雰囲気評価手法

電子掲示板は，話題によってそこに書き込まれる発言が与える印象が変化する．たとえば，凶悪な事件を話題とした電子掲示板において，ユーザが犯人に対し「最悪。死ね。」と書いても閲覧者は不快に感じず，むしろ，その意見に同調する．しかし，交流を目的とした電子掲示板では，同じような内容を書いた場合，場にそぐわず，また，不快な内容であるため，非常に嫌悪され不適切発言とみなされる．

また，同じ話題，同じ電子掲示板内であっても，書き込まれた発言により，雰囲気は刻々と変化する．つまり，電子掲示板の雰囲気は，発言内容及び発言の流れによって決まると言える．

そこで，掲示板の雰囲気を知るための手掛かりとして，発言内容と発言のつながりに着目する．発言中の閲覧者に嫌悪感を与える単語に加え，閲覧者に好感を与える単語にも着目する．これら 2 種類の単語と発言のつながりから，掲示板の雰囲気を数値化し評価する．

本章では，掲示板の雰囲気を数値化し評価することにより，対象電子掲示板の雰囲気の良し悪しが判別できることを明らかにすることを示す．

4.2　電子掲示板の雰囲気評価指標の提案

4.2.1　荒み度の提案

電子掲示板は多くの発言が連なって蓄積され構成される．また，電子掲示板の発言は文字と記号で構成されているため，発言者の意図や心理を判断できるものが文章のみである．そこで，電子掲示板に蓄積された発言のつながりと発言中に使われる単語が，電子掲示板の雰囲気を決定していると捉えることができる．これより，掲示板の雰囲気を評価するためには，一つ一つの発言を確認し，また，各々の発言がどのようにつながっているかに着目すればよい．つまり，発言一つ一つは閲覧者に直接与えた印象を示し，発言の積み重ね（以降,「発言の連鎖」と呼ぶ）は，その発言が他の発言者に与えた印象の大きさを示しており，発言の連鎖数が大きければ，閲覧者は内容の善し悪しにかかわらず，その発言が印象を与えたと知ることができる．松村ら [23][24] も発言の影響の普及モデルを作成し，それに基づいた発言者のプロファイリングを実現している．

これら発言の相互の印象と発言単体の印象が積み重なり，電子掲示板の雰囲気を作り出して

いると考えられる．したがって，発言の印象を掲示板の雰囲気とみなし評価する．

電子掲示板の雰囲気を評価するために，閲覧者に直接印象を与える単語と，発言の連鎖数を評価要素として着目する．ここで，単語が閲覧者に与える印象と発言の連鎖数を考慮したものを電子掲示板の雰囲気の評価指標 “荒み度 (Ruination Figure[$RF\{t\}$])” として提案する．

荒み度は実数であり，正であればよい内容の発言であるとみなし，負であれば悪い内容の発言であるとみなす．これらを掲示板に投稿された発言まで算出し，その変動程度から電子掲示板の雰囲気が良いか悪いかを評価する．次節にて，詳細設計を行う．

4.3　雰囲気評価指標の算出方法

本節では，発言番号 t の荒み度を算出するために新たに定義した 2 つの集合 pw，nw，3 つの値 pww，nww，$cn\{t\}$，4 つの式 $Os\{t\}$，$css\{t\}$，$Ss\{t\}$，$RF\{t\}$ について詳細な説明を行う．

4.3.1　発言が直接与える印象値の算出

pw • pww • nw • nww • $cn\{t\}$ • $Os\{t\}$ の定義

掲示板の荒み具合を評価する荒み度を算出するために，次の 2 つの集合と 3 つの値と 4 つの式を定義する．

Positive Word $[pw]$

pw は,「ありがとう」・「サンクス」といった相手に謝意を示す時や,「お願いします」・「すいません」といった相手に謝罪や依頼をする時に使われる単語・言葉，つまり，相手に好感を与えるような単語の集合を表す．これを次のように集合として定義する．

$$pw = \{\ \text{ありがとう}, \text{頑張れ}, \text{お願いします}, \cdots, \text{お疲れさま}\ \} \tag{4.1}$$

Positive Word Weight $[pww]$

pww は Positive word weight の略で，pw の要素に与える重みを表す．各要素間の重みには差は無く，一様に正値の重みを与えると定義する．pww は次に示す．

$$pww = x \qquad (x > 0) \tag{4.2}$$

Negative Word $[nw]$

nw は,「ばか」,「死ね」といった相手を怒らせるような単語・言葉の集合，つまり，pw とは逆に相手に嫌悪感を抱かせるような単語の集合を表す．nw を次のように集合として定義する．

$$nw = \{\ \text{ばか}, \text{死ね}, \text{お前}, \cdots, \text{くたばれ}\ \} \tag{4.3}$$

Negative Word Weight $[nww]$

nww は nw の要素に与える重みを表す．各要素間の重みには差は無く，一様に負値の重みを与えると定義する．nww は次に示す.

$$nww = y \qquad\qquad (y < 0) \tag{4.4}$$

Concord Number $[cn\{t\}]$

$cn\{t\}$ は，発言 t 中に pw，nw が出現した回数を表している．つまり，ある発言において pw の要素に一致したワードがあれば，一致した総数を $cn_{pw}\{t\}$ で表し，また，nw の要素に一致したワードがあれば，一致した総数 $cn_{nw}\{t\}$ で表す.

Opinion Score $[Os\{t\}]$

$Os\{t\}$ は，発言 t がもつ印象値であり，発言中の単語の出現数，つまり，pw・nw・cn_{pw}・cn_{nw} から算出する．したがって，これを次の式で定義する．

$$Os\{t\} = pww * cn_{pw}\{t\} + nww * cn_{nw}\{t\} \tag{4.5}$$

$Os\{t\}$ の算出例を示す．図 4.2 の発言があったとする．

この内容の発言は，2 ちゃんねるなどで電子掲示板に書き込める発言数が一定数を達し，書き込めなくなったために新たに同じ内容の電子掲示板を設置したユーザへ感謝の意を述べている．発言中には，お疲れ様の意を示す「乙」と感謝の意を示す「ありがとう」がある．つまり，pw が 2 つ出現している．それに対し，nw は出現していない．したがって，$cn_{pw}\{2\} = 2$，$cn_{nw}\{2\} = 0$，$pww = 1$，$nww = -1$ となる．以上より，$Os\{2\}$ は次式となる．

$$Os\{t\} = 1 * cn_{pw}\{2\} - 1 * cn_{nw}\{2\} = 1 * 2 - 1 * 0 = 2 \tag{4.6}$$

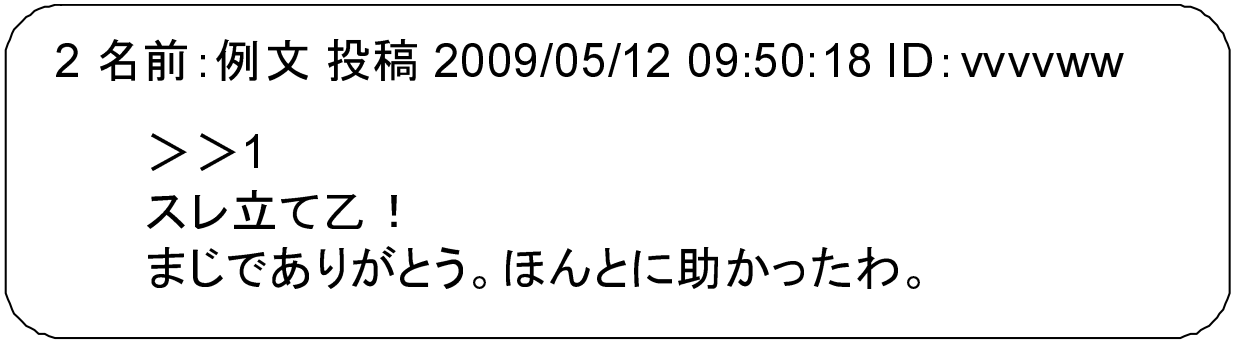

図 4.2: 電子掲示板の発言例

4.3.2　発言の連鎖数が与える印象値の算出

発言のつながりを見るため，取り扱う発言の連鎖を定義する．発言の連鎖には，2 章で述べた通り，次の 3 つの方式が考えられる．

1. 発言番号を指定する
2. 引用文を用いる
3. 矢印を用いる

1 は最も典型的な発言の連鎖である．アンカー（“>”，“>>”）と呼ばれる記号に指定したい発言番号を組み合わせることにより，その発言に意見を述べていることを示している．番号を指定するため，過去のどの発言に対しての発言かが明らかである．

2 は，発言全体ではなく，発言中の特定の部位に対して意見を述べる場合に使われる．直前または極近い発言に対し使われることが多い．

3 は，直前の発言に対し，意見を述べる場合に使われる．

以上より，発言の連鎖として，1 のアンカーを使った場合を取り扱う．

4.3.3　発言の連鎖数が与える印象値の算出

Comment Chain Score$ccs\{t\}$ の定義

$ccs\{t\}$ は，発言 t の連鎖数 $Res\{t\}$ から算出する．発言 t の連鎖数 $Res\{t\}$ の算出方法は次の方法で定義する．図 4.3 のように発言が連鎖している場合，発言 3 は，発言 96, 発言 119, 発言 215 に印象を与えたと見なすことができる．また，発言 119 には印象を与えたと見なせる発言 131 がある．つまり，発言 3 にとっては，発言 119 を経由して発言 131 にも印象を与えたと考えられる．以上より，発言 t における発言の連鎖数を $Res\{t\}$ と定義すると，$Res\{3\} = 4$，$Res\{119\} = 1$ と表すことができる．

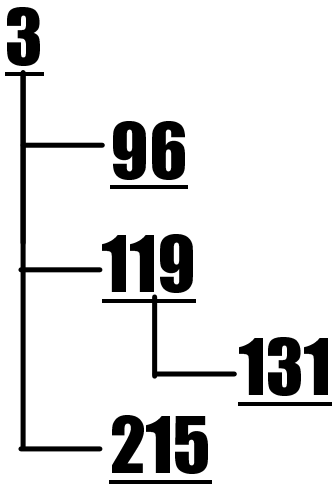

図 4.3: 発言の連鎖の木構造

$Res\{t\}$ を用いて，発言 t の $ccs\{t\}$ は次の式で定義する．

$$ccs\{t\} = i * log_n Res\{t\} \qquad (i = \pm 1 : Os\{t\} \text{ に依存}) \tag{4.7}$$

$Os\{t\}$ が正の場合は $i = 1$ となり，負の場合は $i = -1$ となる．これは，発言の連鎖数 Res が多いということはその数だけ，発言の内容の良し悪しにかかわらず他に印象を与えた発言であるということを示しており，$ccs\{t\}$ を $Os\{t\}$ に加えることにより，その発言の印象値を表現することが可能となる．

ここで $css\{t\}$ に log（Logarithm）を使用した理由を述べる．発言の連鎖数は，その発言が他の閲覧者にどれだけ印象を与えたかを示すひとつの指標であると述べた．つまり，発言の連鎖数が多いという情報は，閲覧者への 1 つの刺激として見なすことができる．人間の感覚は，与えられた刺激の対数に比例するというウェーバーフェフィナーの法則がある．つまり，ある程度の数の連鎖数を持つ発言を見た閲覧者は，深い印象を受けると考えられるが，その発言の連鎖数が 10 の場合と 15 の場合では与える印象がそれほど変わらないということが言える．これより，閲覧者が発言の連鎖数という印象は対数で表現できる．そこで，発言の連鎖数が与える印象値 $ccs\{t\}$ を式 4.7 で定義する．

4.3.4　発言が与える印象値の算出

$Ss\{t\}$・$RF\{t\}$ の定義

Statement Score $[Ss\{t\}]$

$Ss\{t\}$ は，発言 t が掲示板に与えた印象値を表す．これは，発言が直接与えた印象値を算出する式 4.6 と発言の連鎖数が与えた印象値を算出する式 4.7 を加算したもので，次の式で定義する．

$$Ss\{t\} = Os\{t\} + ccs\{t\} \tag{4.8}$$

Ruination Figure $[RF\{t\}]$

$RF\{t\}$ は，発言 t までの印象値を発言順に加算していったもので，t 番目における掲示板の雰囲気を示す荒み度 $RF\{t\}$ は次の式で定義する．

$$RF\{t\} = \sum_{1}^{t} Ss\{t\} \tag{4.9}$$

式 4.9 で算出した荒み度の変動から電子掲示板の雰囲気の善し悪しを評価する．荒み度がマイナス方向へ変化した場合，その範囲では，不適切単語が多く印象深い発言があったと推測されることから，否定的な話題についてのやり取りが行われている，または，掲示板に不適切発言が多く荒れていると推測できる．よって，電子掲示板の雰囲気を “悪化状態” と評価する．逆に，荒み度が変化せず一定の場合やプラス方向へ変化した場合，その範囲では，不適切単語が少な

い発言が多いと推測されることから，肯定的な話題についてのやり取りが行われていると推測される．よって，通常コミュニケーションが行われていると考えられるため，電子掲示板の雰囲気を “通常状態” と評価する．

4.4　雰囲気評価指標 “荒み度” の妥当性の検証

電子掲示板の雰囲気を評価するために，発言中の単語と発言の連鎖数が必要な要素であると述べた．そこで，実際の電子掲示板に対し，発言中の単語と発言の連鎖数を用いて雰囲気を数値化し，荒み度の変化が雰囲気の変化を示すことができることを検証する．

4.4.1　実験準備

荒み度の妥当性を示すために，荒み度 RF を算出する試作システムを作成した．試作システムは，表 4.1 のスペックのパーソナルコンピュータを使用し，表 4.2 の通り，ruby 言語で記述し，pw を 121 個，nw を 308 個人間が選択し用意した．

表 4.1: 使用したパーソナルコンピュータのスペック表

パーツ	スペック
CPU	1.86GHz
メモリ	4GB
OS	FreeBSD7.0

表 4.2: 使用言語およびデータ

使用言語およびデータ	内容
プログラム言語	Ruby（version1.8.6)）
pw	121
nw	308

荒み度の変化を視覚的に捉えやすくするため，ローソク足チャートをつかってグラフに出力する [25][26]．ローソク足チャートは，図 4.4 のように始値，終値，最高値，最小値をローソク

足という 1 つのアイテムで表示するもので，荒み度の変化を視覚的に捉えるのに適したグラフの 1 つである．

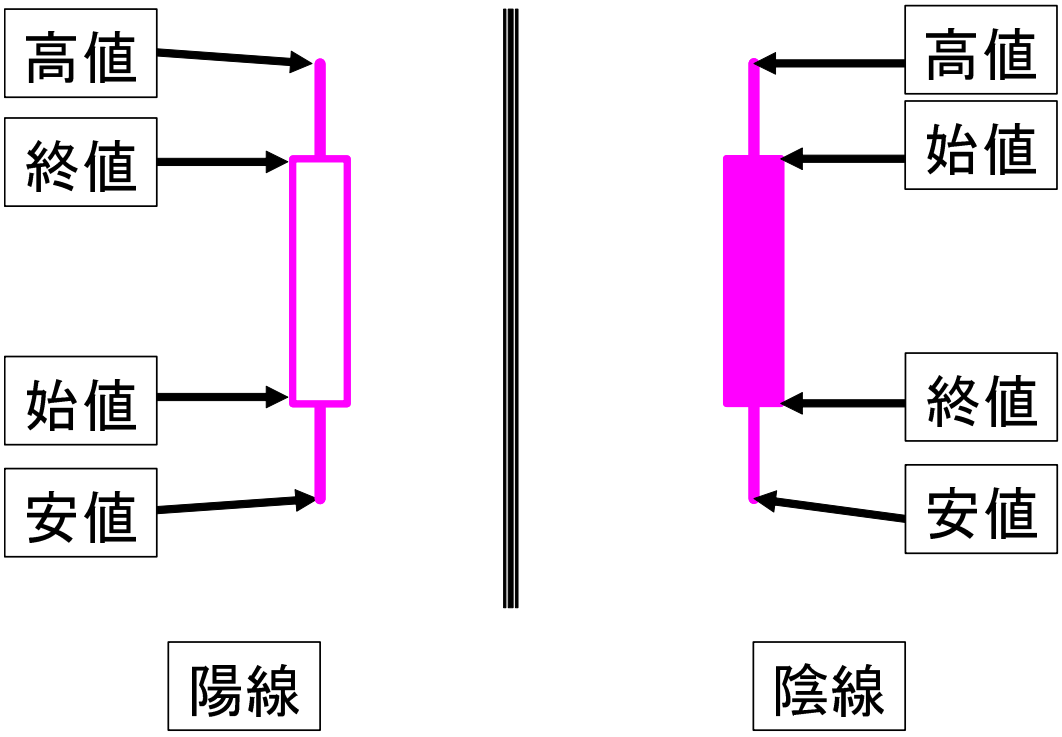

図 4.4: ローソク足

始値と終値で四角を作り，それよりも最高値が大きい場合，四角より上にひげと呼ばれる棒を使って表示される．逆に，四角よりも最小値のほうが小さい場合，下へひげを使って表示される．始値より終値のほうが大きい値の場合，陽線と呼ばれる白抜きのアイテムで表示される．逆に，始値より終値のほうが小さい値の場合，陰線と呼ばれる黒塗りのアイテムで表示される．ローソク足は，図 4.5 のようなパターンが存在する．ローソク足チャートは主に，株価の変動などに用いられている．

4.4.2　実験対象掲示板

実験結果の一例として，学校裏サイトと呼ばれる電子掲示板を挙げる．この掲示板では，挑発・誹謗中傷の連鎖の発生と沈静化が繰り返し発生した．実験に利用した発言は，3 章でも使用した 2007 年 5 月から 2008 年 11 月までの 1097 発言を対象とした．

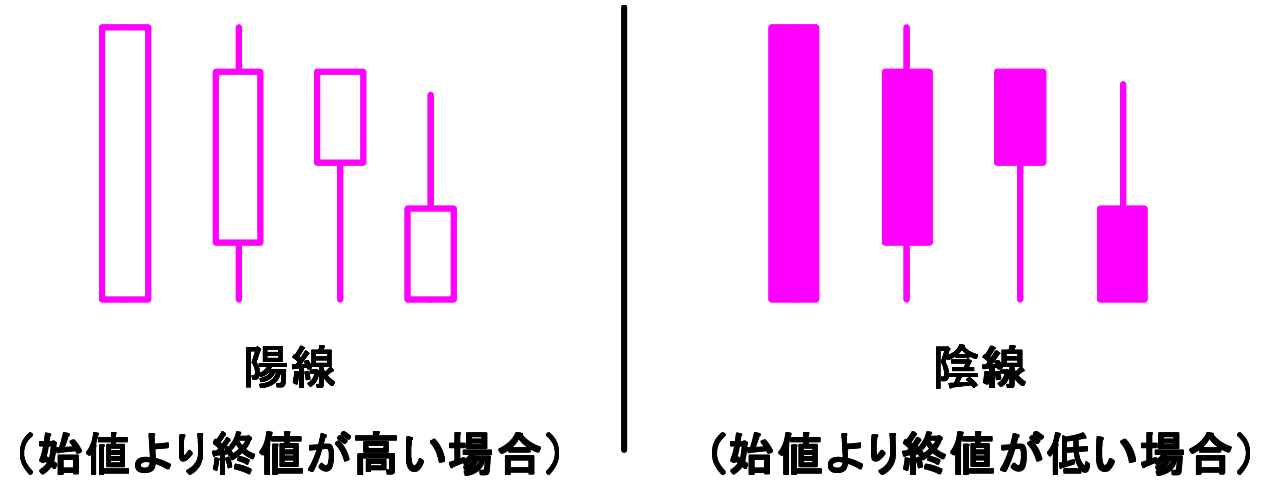

図 4.5: ローソク足の種類

4.4.3　実験結果

実験対象掲示板の荒み度 RF を算出した結果，図 4.6 が得られた．図中の縦線は，荒み度の変化の度合いが著しく変化した箇所を示しており，四角の内容はその範囲で発生した話題を示している．荒み度と内容の比較を表 4.3 に示す．

表 4.3: 荒み度と発言内容の比較

範囲	荒み度による評価	話題（人間による評価）
1-180	通常	雑談
181-280	悪化	誹謗中傷・罵り合い・挑発
281-370	悪化	誹謗中傷・罵り合い・挑発
371-490	通常	挑発行為の沈静化
491-610	通常	雑談
611-680	悪化	誹謗中傷・罵り合いの再発
681-870	通常	雑談・オタク談義
871-1097	悪化	誹謗中傷・罵り合いの再発

範囲 1-180 は，主に自己紹介が行われ，共通する部活の先輩・後輩の話や，4 月始めであったため，クラス替えの話題で占められていた．

範囲 181-280 は，始めに挑発する書き込みがあり，それに呼応したユーザとの挑発・罵り合いが繰り返し行われていた．

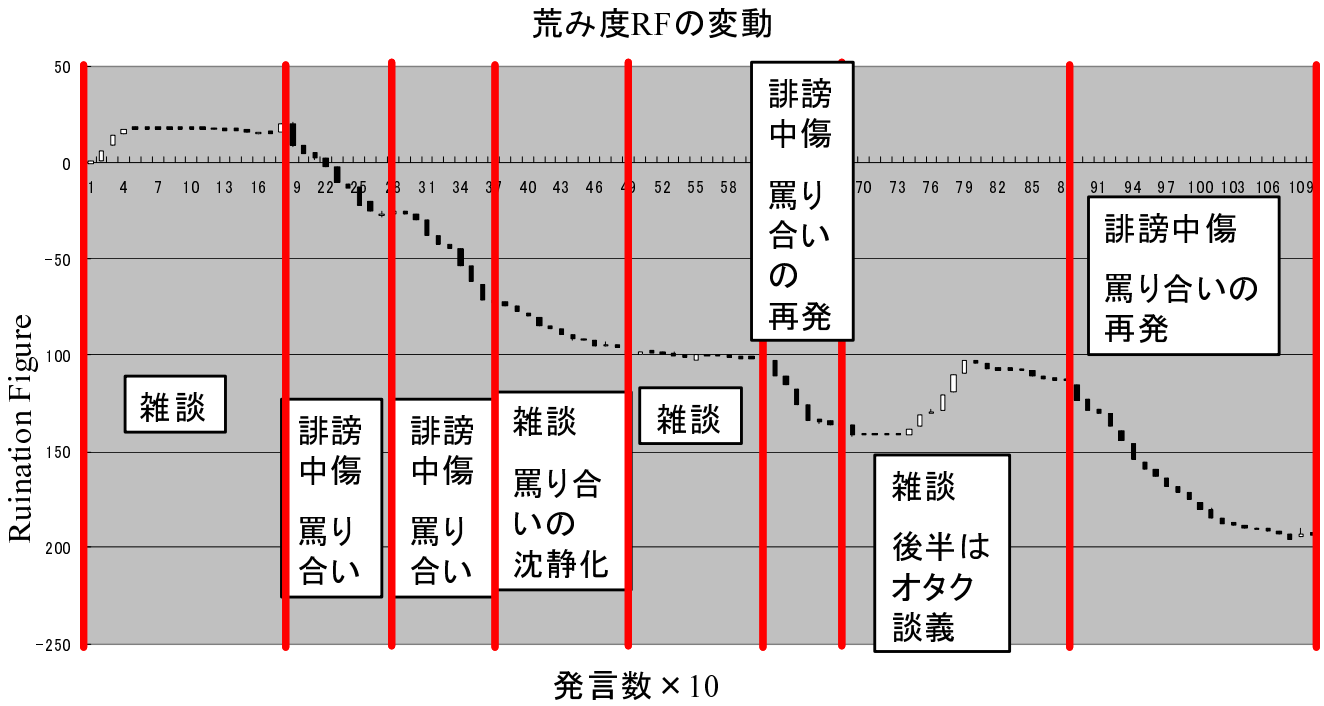

図 4.6: 荒み度の変化と発言内容の比較

範囲 281-370 は，挑発・罵り合いを悪化させるために，さらに別のユーザが参加し，罵り合いや挑発を繰り返し行っていた．

範囲 371-490 は，前半は，罵り合いを行っていたユーザを非難する発言で占められていたが，後半では，ゲームのキャラクタの話に移行した．

範囲 491-610 は，翌日にもう一度掲示板で話をするための集合時刻などの取り決めを行っていた．後半では，実名を書き込まれたため，その発言を消す方法についての会話で占められていた．

範囲 611-680 は，実名を晒したと思われるユーザの挑発発言とそれに反応した数名のユーザの罵り合い・挑発の応酬が繰り広げられていた．

範囲 681-870 は，罵り合いに参加していたユーザがいなくなったため，別のユーザがキャラクタになりきり雑談を行っていた．

範囲 871-1097 は，前半はゲームやアニメのキャラクタになりきった発言が多数書き込まれていた．その後，その発言者をうとましく思ったユーザが挑発的な内容で発言をやめるよう書き

込んだため，誹謗中傷の応酬が再発した．

以上より，荒み度の変化の方向および度合いを見ることにより，その範囲で交わされるコミュニケーションがどういうものか，つまり，雰囲気がどうなっているのかを推定することが可能である結果が得られた．

この掲示板以外においても，同様の結果が得られた．その結果，電子掲示板の雰囲気は，発言中の単語と発言の連鎖から推定することが可能であることが明らかとなった．

4.5　結言

電子掲示板の話題は多種多様であり，話題の数だけ電子掲示板が存在する．そのため，電子掲示板の発言を自動分類するためには，対象となる電子掲示板の雰囲気を評価する必要があった．電子掲示板の雰囲気は，話題や発言の内容，発言中の単語によって決定されると考えた．そこで，各々の発言が好感を与える発言と嫌悪感を与える発言がどのように連なっているかに着目し，電子掲示板の雰囲気を数値化することを目指した．電子掲示板の雰囲気の評価指標として，好感を与える単語と嫌悪感を与える単語および発言の連鎖数を利用し，各発言が電子掲示板に与える印象を数値化した．それらの発言の連なりが雰囲気を決定していると考え，各発言が与える印象値の総和を荒み度 RF として提案した．実際に評価指標 “荒み度” を用いることにより，電子掲示板の雰囲気が “通常” なのか “悪化” なのかを知ることができることを確認した．これより，電子掲示板の雰囲気評価に必要な要素として，発言中の単語と発言の連鎖数が有効であることを確認し，雰囲気を数値化することが可能であることを示した．

第5章　電子掲示板の自動分類評価について

5.1　緒言

経験不足が原因で不適切発言を投稿してしまうユーザに対し，投稿発言が不適切発言であることを提示する利用者教育支援の実現するために，電子掲示板に投稿された発言を通常発言と不適切発言の2種類に分類する自動分類手法を確立する．そこで本章では，投稿された発言を不適切発言と通常発言の2種類に分類する手法の提案と検証を行う．提案する手法は，図5.1のように構成される．

電子掲示板の発言を自動分類するために，まず発言単体に着目し，3章で発言単位での自動分類手法を提案した．提案手法による分類結果から，未判定発言を通常発言候補と不適切発言候補の2種類に分類するためには，発言単体にのみ着目するだけでは不十分であることが明らかとなった．そこで，未判定発言を分類するためには，投稿時の電子掲示板の雰囲気が重要であると着目し，4章で発言中の単語と発言の連鎖に着目し評価する手法を提案した．本章では，3章で分類した未判定発言を4章で提案した雰囲気評価を利用して分類することにより，投稿発言が通常発言候補と不適切発言候補の2種類に分類する手法を提案しその妥当性を示す．

5.2　雰囲気を考慮した発言分類

未判定発言は，不適切な単語を含むが，単語のペアを作った場合，不適切ではない可能性のある発言である．つまり，図2.3のように，通常発言と不適切発言の境界付近の発言であると考えられる．この未判定発言が，通常発言候補なのか不適切発言候補なのかを決定する要素として，3章では，発言中の通常単語の構成率を利用し分類を行った．結果として，通常単語の構成率は分類要素として妥当ではないことが明らかとなった．

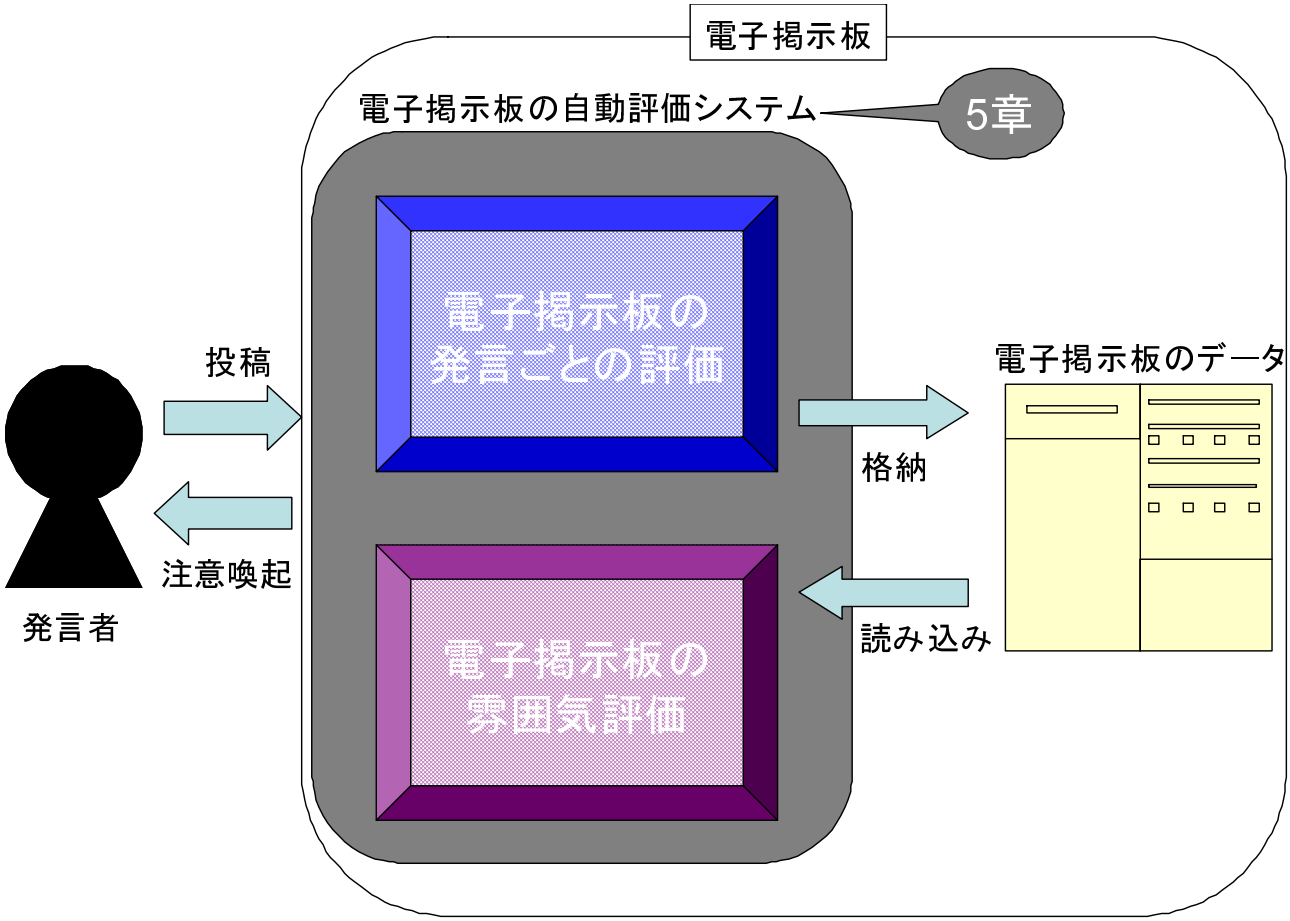

図 5.1: 5 章で提案する発言の自動分類手法

ここで，未判定発言を通常発言候補と不適切発言候補に分類するための条件について考える．未判定発言は，不適切な単語が含まれているが，通常発言に出現する単語のペアによって構成される発言である．そのため，投稿前に書き込まれていた掲示板の発言の内容 (電子掲示板の雰囲気) によって未判定発言の受け取られ方が異なると考えられる．例として，未判定発言に分類される可能性の高い次の 3 つの場合について考える．

1. 皮肉を込めるために多少不適切単語を使った場合
2. ただ褒めるだけでは物足りないため不適切単語を使って強調した場合
3. 悪意のある発言に応対した場合

それぞれの例が出現する掲示板の雰囲気を考える．

電子掲示板の発言は文字と記号で書かれているため，(1)，(2) の未判定発言は「他の発言との差別化を図りたい」「書き込んだ発言を目立たせたい」という発言者の意図が反映された結果の一つとして考えられる．したがって，電子掲示板の雰囲気は円滑にコミュニケーションが行える “通常状態” であると言える．

(3) の発言が出たときは，掲示板の雰囲気が “悪化状態” である場合が多いと考えられる．なぜなら，悪意のある発言に応対した発言であるため，事前に不適切発言があることを示しており，不適切発言が出現する状況は閲覧者を不快にする状況だからである．したがって，雰囲気が悪化し不適切発言が多く出現している状況において，未判定発言が出現すると，状況を悪化させる発言と同様に受け取られ，結果として不適切発言になると予測される．

以上より，雰囲気が通常状態の時は，未判定発言に分類された発言が出現した場合，再分類時には通常発言候補に分類し，雰囲気が悪化状態の時は，未判定発言に分類された発言が出現した場合，再分類時には不適切発言候補に分類することを提案する．

これより，電子掲示板の発言を自動分類し，不適切発言候補であった場合，発言者へ注意を促す利用者教育支援システムは，図 5.2 のようになる．電子掲示板に新たな発言が書き込まれるとき，まず，発言単体で分類を行う．

分類結果が不適切発言候補であった場合，ユーザへ不適切発言となることをユーザへ示し，電子掲示板への掲載は行わない．

分類結果が通常発言候補に分類された場合，電子掲示板にその発言を掲載し，ユーザへの注意喚起は行わない．

分類結果が未判定発言に分類された場合，電子掲示板の既出発言中の単語および発言の連鎖から電子掲示板の雰囲気の状況を算出する．雰囲気が “通常状態” であった場合，未判定発言は通常発言候補に分類される．その結果，発言は電子掲示板に掲載され，ユーザへの注意喚起は行わない．逆に，雰囲気が悪化状態であった場合，発言は 不適切発言候補に分類され，ユーザへ不適切発言候補であることをユーザへ示し，電子掲示板への掲載は行わない．このようにして，投稿された発言が通常発言候補か不適切発言候補かを自動分類し，不適切発言候補に分類された場合ユーザへ注意喚起することにより，過失で書き込まれた不適切発言を減少させることが可能となる．また，ユーザへ注意喚起をすることにより，ユーザは発言を推敲する機会が

与えられるとともに，どういう発言が不適切発言となるかを学習する機会を与えられる．その結果，ユーザは不適切発言がどういうものかを理解し，不適切発言の投稿をしなくなる．つまり，不適切発言を投稿するユーザに対し，利用者教育支援が実現できる．

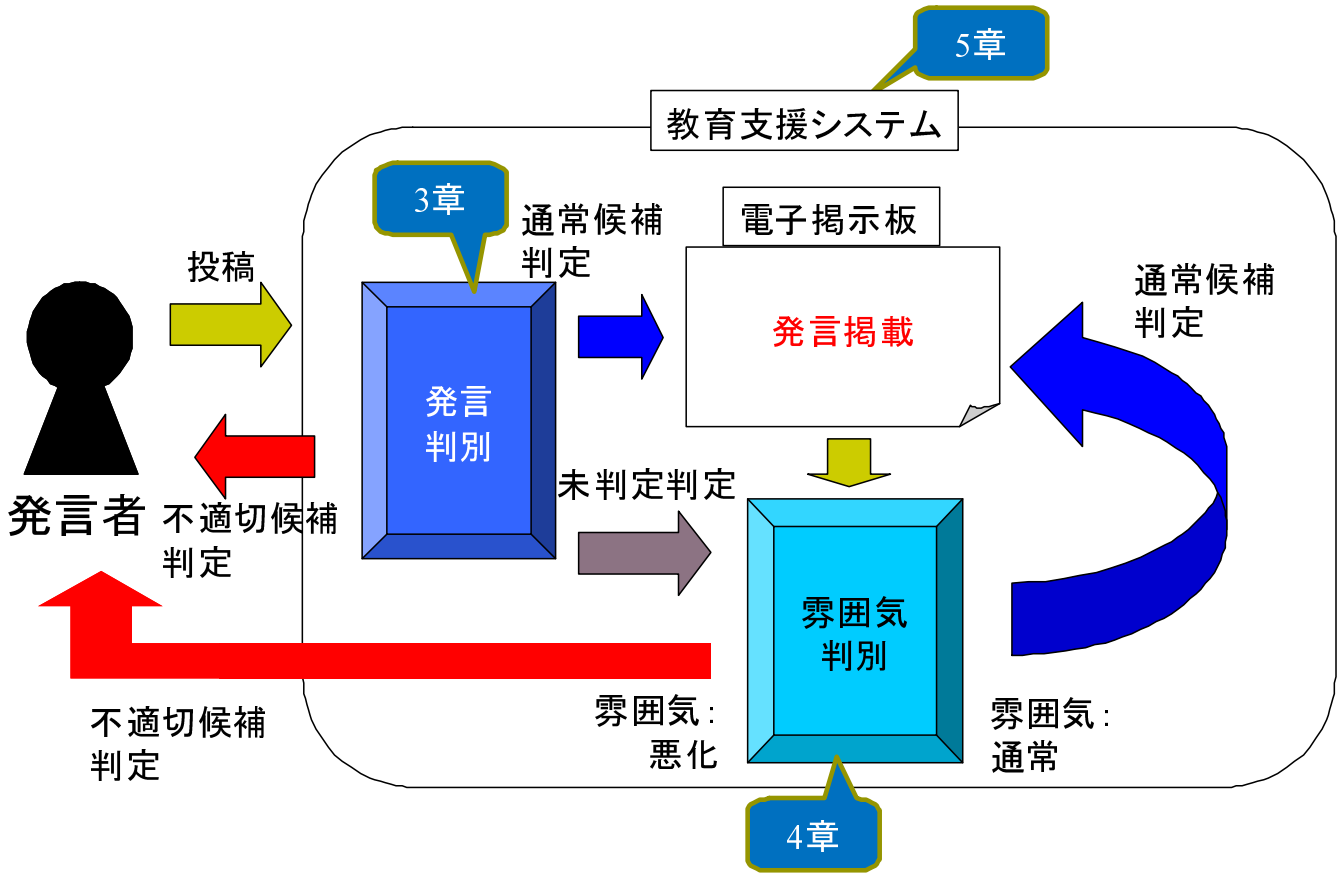

図 5.2: 電子掲示板の雰囲気を考慮した発言分類システム

5.3　検証実験準備

5.3.1　実験環境

検証実験を行うために作成した図 5.2 の試作システムは，CPU：Core2Duo の 1.86GHz，メモリ：4GB のスペック，OS：FreeBSD7.0 のパーソナルコンピュータに，プログラム言語 Ruby（version1.8.6）で記述し，データベースは Mysql（version5.0.45）を使用し，Web サーバは Apache を使用している．また，学習・判別時に行う形態素解析には，形態素解析用ソフト Mecab（ver-

sion0.96）(形態素解析ソフト MeCab 2009) を使用している．この環境は，3 章との比較を行うため，使用した学習データ，パーソナルコンピュータ，プログラム言語，使用ソフトは同一の物としている．表 5.1，表 5.2 に試作システム環境を示す．

表 5.1: 使用したパーソナルコンピュータの性能

環境	スペック
CPU	1.86GHz
メモリ	4GB
OS	FreeBSD7.0

表 5.2: 試作システム環境

環境	ソフトウェア
プログラム言語	Ruby（version1.8.6)）
データベース	Mysql（version5.0.45）
Web サーバ	Apache（version2.2.6）
形態素解析用ソフト	Mecab（version0.96）

5.3.2　検証内容

未判定発言が出現した時，電子掲示板の雰囲気が通常状態であれば，その未判定発言を通常発言候補に分類する．逆に，電子掲示板の雰囲気が悪化状態であれば，その未判定発言を不適切発言候補に分類することを提案した．そこで，発言の自動分類を行うために，未判定発言が出現した場合，その場の雰囲気に応じて通常発言候補または不適切発言候補に分類し，主観評価と一致するか検証を行う．

また，雰囲気に応じて発言を分類することにより，電子掲示板の発言の自動分類の精度が十分であることを確認する．

5.3.3　実験対象掲示板

実験結果の一例として，3 章，4 章で使用した学校裏サイトと呼ばれる電子掲示板を挙げる．これは，3 章の結果との比較を行うためである．この掲示板では，挑発・誹謗中傷の連鎖の発

生と沈静化が繰り返し発生した．実験に利用した発言は，2007 年 5 月から 2008 年 11 月までの 1097 発言を対象とした．

電子掲示板の雰囲気は，4 章の荒み度 RF を利用した雰囲気判別の結果より，表 5.3 となっている．

表 5.3: 発言範囲の雰囲気の状態

発言範囲	雰囲気
1-180	通常状態
181-370	悪化状態
371-610	通常状態
611-680	悪化状態
681-870	通常状態
871-1097	悪化状態

5.4　雰囲気を考慮した分類結果の分析

発言単位での自動分類後，未判定発言に分類された発言とその発言が書き込まれた掲示板の雰囲気及び，その発言の主観評価の比較を表 5.4 に示す．表中の “主観評価：通常発言” の欄は，発言範囲において未判定発言に自動分類された発言を主観評価によって通常発言であると評価した発言数を示している．“主観評価：不適切発言” の欄は，逆に，未判定発言と自動分類された発言を主観評価によって不適切発言であると分類された発言数を示している．発言数の欄は，その発言範囲中にあった未判定発言の総数を示している．

表 5.4: 雰囲気に依存した未判定発言の主観評価による分類

発言範囲	雰囲気	未判定発言数	主観評価：通常発言	主観評価：不適切発言
1-180	通常	39	39(100%)	0（0%）
181-370	悪化	62	33（53.2%）	29（46.8%）
371-610	通常	58	56（96.5%）	2（3.4%）
611-680	悪化	26	21（80.8%）	5（19.2%）
681-870	通常	40	39（97.5%）	1（2.5%）
871-1097	悪化	77	60（77.9%）	17（22.1%）

雰囲気が通常状態のとき，未判定発言の多くは主観評価で通常発言に分類されるという結果が得られた．通常状態である「1-180」[371-610」「681-870」の範囲では，通常発言 134 個に対し不適切発言はわずか 3 個という結果から，多少不適切単語が含まれた発言を投稿しても通常状態である場合，その発言は通常発言候補に分類することが妥当であることが明らかとなった．3 個の不適切発言がパターン 2 の未判定発言に分類された原因は，学習データが不十分のため，単語のペアによる判定が通常発言候補と出力されたことが原因であった．もし，対応したデータがあれば，単語による判定も単語のペアによる判定も不適切発言候補と評価されるため，雰囲気が通常状態である場合は，未判定発言は通常発言候補に分類されることが妥当であるという結果が得られた．

逆に，雰囲気が悪化しているときは，未判定発言は主観評価では通常発言と不適切発言が半数程度という結果が得られた．雰囲気が悪化した時は，曖昧発言が出現すると，雰囲気の悪化に拍車を掛けることとなるため，不適切発言に分類されると考えられる．しかし，雰囲気が悪化状態におけるすべての曖昧発言が不適切発言となるわけではない．この掲示板では，主観評価によると未判定発言の 5 割から 8 割が通常発言であった．したがって，この悪化状態においての未判定発言をすべて不適切発言候補に分類すると False Postive が増加することは明らかである．しかし，本研究の目的は，投稿発言を自動分類し，不適切発言候補に分類された発言を投稿したユーザへ注意を促すことによる不適切発言の減少であり，不適切発言を通常発言候補に誤分類してしまう False Negative を減らすことが False Positive の増加よりも重要である．したがって，雰囲気が悪化状態の場合，未判定発言は不適切発言候補に分類する．

未判定発言を 3 章で提案した発言単位で分類した結果を表 5.5 に，本章で提案した雰囲気を考慮して分類を行った結果を表 5.6 にそれぞれ示す．

表 5.5: 単語の構成率に着目した未判定発言の分類結果

分類	主観評価：通常発言	主観評価：不適切発言
未判定発言	248	54
誤判定	77	25
正判定	171	29

表 5.5 と表 5.6 を比較した結果，雰囲気を考慮することにより不適切発言を通常発言候補に分

表 5.6: 雰囲気を利用した未判定発言の分類結果

分類	主観評価：通常発言	主観評価：不適切発言
未判定発言	248	54
誤判定	104	3
正判定	144	51

類する False Negative を減少させることが可能であることが判明した．それに対し，通常発言を不適切発言候補に分類する False Positive を増加させることになることも判明した．ただし，不適切発言を判別しユーザへ注意を促す教育支援を実現することが目的であり，False Positive の増加よりも False Negative の減少のほうが重要である．以上より，雰囲気を考慮した分類は有効であることが示された．

雰囲気を考慮した発言の自動分類の結果は，表 5.7 となった．表 5.7 中の正答率のカッコ内の

表 5.7: 雰囲気に着目した発言分類結果

分類結果	主観評価：通常発言	主観評価：不適切発言
通常発言候補	749	20
不適切発言候補	206	79
正答率	78.4%（82.0%）	79.8%（56.4%）

数値は，通常単語の構成率を用いて発言の再分類を行った時の正答率を示している．その結果，不適切発言を 56.4%から 79.8%の精度で分類できるように改善された．通常発言の正答率を維持しつつ，不適切発言の分類精度を 20%以上改善することができた．以上より，発言を自動分類するためには，発言の個々を分類し，電子掲示板の雰囲気を利用して再分類する本提案手法が有効であることが示された．

5.5　結言

本章では，ユーザの教育支援を実現するために，3 章 4 章の結果を踏まえた自動発言分類方法を提案した．未判定発言が不適切発言と受け取られるのか通常発言と受け取られるのかは，発言中の通常単語の構成率ではなくその発言が書き込まれた電子掲示板の雰囲気であると考え，

電子掲示板の雰囲気に応じた未判定発言の分類を試みた．具体的には，電子掲示板の雰囲気が通常状態の場合，未判定発言の意図は,「目立ちたい」「意味を強調したい」と捉え通常発言候補に自動分類し，また，電子掲示板の雰囲気が悪化状態の場合，未判定発言の意図は,「不適切発言への不適切な対応」と捉え不適切発言候補と自動分類した．その結果，発言単体で分類を行った場合に比べ，通常発言の分類精度を維持しつつ 20%以上不適切発言の分類精度を改善することができた．発言を誤分類する主な原因は，主観評価によって不適切と評価するが学習データに登録されていない単語や単語のペアがあったためであり，学習データを充実させることにより対応が可能であると考えられる．

以上より，電子掲示板の発言を自動分類するためには，発言単位で分類し，その後，未判定発言に分類された場合，電子掲示板の雰囲気を考慮して再分類することが必要であることを示した．この自動発言分類手法の確立により，過失で不適切発言を書き込んでしまう経験不足のユーザに対し，不適切発言となりうることを示唆する教育支援が実現可能であることを示した．

第6章　結論

本論文の各章のまとめは以下の通りである．

1 章では，電子掲示板で発生する問題，及び，その問題に対する現状の対応方法について述べた．また，対応策は主に問題となる発言をブロックする，または，早期発見早期対処のスタンスであるため，問題となる発言を投稿しないようにユーザに対し教育支援を行うことが重要であることを述べた．そこで，経験不足のために不適切発言を投稿するユーザを対象とし，不適切発言を投稿した場合，ユーザに不適切発言であることを提示し自主的に学習してもらう教育支援の実現を目指した．そのために，発言が不適切発言なのか通常発言なのかを分類する手法が必要不可欠であった．そこで，本稿では，発言を自動分類する手法の確立を目的とした．

2 章では，電子掲示板を自動分類するために，電子掲示板の構造や特徴に着目し，評価すべき対象を明らかにした．電子掲示板は文字と記号で構成され，書き込まれた発言が蓄積することで，その電子掲示板の方向性が決定される．つまり，蓄積された発言が嫌悪感を抱く言葉を多く使う掲示板では，嫌悪感を抱く言葉を多く使用してもそれほど目立たない．しかし，嫌悪感を抱く言葉を多く使わない掲示板において，嫌悪感を抱く言葉を多く使用した発言を書き込むと，非常に目立つ．また，1 つの電子掲示板でも，時間がたつにつれ，話題も変化し使われる言葉も変化する．その結果，書き込まれた場所や時によって，嫌悪感を抱く言葉を多く使用した発言が目立つ場合と目立たない場合が存在する．このことから，電子掲示板に投稿された発言が不適切なのか適切なのかを評価することが必要であることを述べた．また，電子掲示板に書き込まれた過去の発言から，その掲示板の場の雰囲気を評価する必要があることを述べた．

3 章では，電子掲示板の発言単位で通常発言と不適切発言の 2 種類に分類する手法の確立を行った．ここでは，電子掲示板の発言と同じように文字と記号で構成される電子メールの spam メールを不適切発言と同等のものと見なし，spam メールフィルタの一種であるベイジアンフィ

ルタと同様の処理を行うことで発言を分類することを目指した．ただし，掲示板の発言は不定型文であり，単語が持つ意味が使われ方によって間逆になることから，単語のみではなく，単語を組み合わせ，単語のペアを作成することにより 1 つの単語に複数の意味を持たせることを提案した．単語と単語のペアを学習データとし，新たな発言が不適切発言となる確率を算出し，2 種類のデータの結果の組み合わせから発言の分類を行った．その結果，多数が不適切発言と評価する発言を不適切発言候補として分類し，多数が通常発言と評価する発言を通常発言候補に分類することが可能であることを示した．また，未判定発言は，発言単体を評価するだけではどちらに分類されるのかを決定できないことも示した．検証実験より，単語のペアを作成することにより，1 つの単語に複数の値を持たせることを実現でき，また，それぞれの値も妥当であることを示した．

4 章では，電子掲示板の雰囲気を評価することを目指した．そのため，発言中の好感を与える単語と嫌悪感を与える単語と発言の連鎖数に着目し，各発言が電子掲示板に与える印象を数値化し，その蓄積を荒み度という評価指標で表すことを提案した．荒み度の変化の度合いから，電子掲示板の雰囲気が通常状態なのか悪化状態なのかを示すことが可能であることを示した．電子掲示板の発言中の単語と発言の連鎖数が雰囲気を示す要素として有効であることを示した．

5 章では，3 章の発言単位の分類手法と 4 章の雰囲気の評価を組み合わせることにより，教育支援を実現するために必要な電子掲示板の自動分類手法の確立を目指した．3 章で提案した発言単体の分類手法において，発言単体では分類することができなかった未判定発言を，4 章で提案した書き込まれた電子掲示板の雰囲気から再分類する手法を提案した．検証実験より，雰囲気が通常状態では，未判定発言は通常発言と受け取られることが明らかとなった．また，雰囲気が悪化状態では，未判定発言は不適切発言と受け取られる割合が増加することが明らかとなった．その結果，発言の 8 割を適切に分類する手法を確立し，不適切発言を過失で投稿してしまうユーザに不適切発言であることを提示する教育支援手法の実現が可能であることを示した．例として，このシステムを小中学校での授業に導入することができれば，生徒が書き込む発言一つ一つに対し，人間評価を下さずとも，どういう発言が不適切発言となるのかを学ぶことも可能となる．これによりユーザは，どういう発言が不適切発言となるかを学ぶことができ，結果として不適切発言を過失で行ってしまうことは減少し，電子掲示板で発生する不適切発言の

数が減少し，インターネットを利用したいじめ問題などの解決に役立たせることが期待できる．

References

[1] http://www.2ch.net/ ２ちゃんねる， 2009 年 4 月参照

[2] http://messages.yahoo.co.jp/index.html Yahoo 掲示板, 2010 年 1 月参照

[3] http://mixi.jp/ mixi, 2010 年 1 月参照

[4] 毎日新聞, 2004 年 6 月 2 日朝刊, “ 長崎・佐世保の小 6 女児殺害：ネット掲示板でトラブルか 加害女児を家裁送致へ ”

[5] 読売新聞, 2006 年 10 月 19 日朝刊, “「いじめ」高 2 自殺未遂 ”

[6] http://www.mext.go.jp/b_menu/toukei/001/index48.htm 文部科学省 “ 青少年が利用する学校非公式サイトに関する調査報告書 ”, 2009 年 4 月参照

[7] 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, 平成 13 年 11 月 30 日法律第 137 号

[8] 菊川 暁, ”電子掲示板システム ”, 特許第 3776313 号公報, 2006 年 3 月

[9] http://solution.gaiax.co.jp/ 株式会社ガイアックス 2009 年 3 月参照

[10] http://www.pit-crew.co.jp/ 株式会社ピットクルー 2009 年 3 月参照

[11] 毎日新聞, 2009 年 6 月 27 日朝刊, “ ネットの書き込み自治体が監視続々 ”

[12] http://schecker.jp/ 学校裏サイトチェッカー 2009 年 4 月参照

[13] 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.

[14] 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.

[15] 三品賢一, 土屋誠司, 黒岩眞吾, 任福継, “ N-gram 出現頻度を用いた感情類似度計算 ”, 情報処理学会研究報告, 2007-NL-180(7), 2007.

[16] 平林 幹雄, 江渡 浩一郎, “ N.M-gram：ハッシュ値付き N-gram 索引による全文検索の一手法 ”, 情報処理学会論文誌：データベース, Vol.48, No.SIG 7, 2007.

[17] http://mecab.sourceforge.net/ 形態素解析エンジン“ mecab ”, 2009 年 4 月参照

[18] http://chasen.naist.jp/hiki/ChaSen/ 形態素解析システム茶筌, 2009 年 4 月参照

[19] http://www.rimarts.co.jp/index-j.html 有限会社リムアーツ, 2009 年 4 月参照

[20] P. Graham: A Plan for SPAM, http://paulgraham.com/spam.html 2009 年 4 月参照

[21] 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), pp.464-468, 2006.

[22] 岩永学, 田端利宏, 櫻井幸一, “ ベイジアンフィルタリングを用いた迷惑メール対策における多言語環境でのコーパス分離手法の提案と評価 ”, 情報処理学会論文誌, 46(8), pp.1959-1966, 2005.

[23] 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.

[24] 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 ”, 人工知能学会論文誌, Vol.18, Non.4, pp.165-172, 2003.

[25] 阿部達郎, 柳谷雅之, 野村光紀, 蔓部音士, “ 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.

[26] 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.

謝辞

末筆ながら，本研究を行うにあたり，多くの方々より御助言を頂き，深く感謝いたします.

特に，研究を行う機会を与えて頂き，懇切丁寧な御指導を賜りました東北大学サイバーサイエンスセンター曽根秀昭教授に厚く御礼申し上げます.

また，本文の執筆にあたり有益な御助言，御討論を賜りました同学サイバーサイエンスセンター木下哲男教授並びに同学情報科学研究科加藤寧教授に深く感謝いたします.

さらに，本研究の進め方や方針の提示など，あらゆる方面で多大な御尽力を頂きました，同学サイバーサイエンスセンター水木敬明准教授に深く感謝いたします.

また，本研究を進めるにあたり，御指導・御助言を賜りました千葉工業大学電気電子情報工学科今野将准教授に深く感謝いたします.

また，本研究を行うにあたり，御討論，御協力を賜りました曽根・水木研究室の皆様にこの場を借りて厚く御礼申し上げます.

最後に，精神的に経済的に研究を支えてくださった両親に深く感謝いたします.