Chapter 3

メッセージ分析の手法

3.1 対立の機械的検出

知性に訴えるメッセージと感性に訴えるメッセージのように，性格の異なるメッセージは，用いられている用語が異なるものと予想される．また，相互理解の上になされる議論は，同種の用語を用いてなされると考えられる．対立が発生している場合，性格の異なる記事が議論の過程で互いに投稿されると考えられる．記事の性格は，そこに現れる用語によって特徴付けられるため，メッセージ中に出現する用語の特徴を抽出し，議論の過程でこれがどのように変化するかを調べれば，その議論がいかなる性格であるかを知ることができよう．メッセージに用いられる用語とその出現度数によって規定されるメッセージの性格は，比較的容易に抽出・解析することができるものと考えられる．

電子的コミュニケーションにおいては，メッセージの機械処理が容易であることから，さまざまな解析手法が提案されている．参加者相互の関係に注目した研究として，個々のメッセージに対するフォローアップの数を解析してコミュニケーションの場の特性を把握する試み [川上 1993, 瀬尾 1996] がなされている．しかし，これらはメッセージの数のみを扱い，メッセージの内容には触れられていない．

議論の推移の把握を目指した研究として，電子メイルを介して行われる協同作業のメッセージから，問い合わせと返答のペアを機械的に抽出し，会話の完結度を解析したもの [村越 1998a, 村越 1998b]，話題の推移に関連した重要メッセージの抽出 [井佐原 1997, 内元 1997, 内元 1998]，話題の消長の自動検出 [斎藤 1998] などが研究されており，記事の内容から議論のダイナミズムを解析する試みとして注目される．これらの手法は特

定のパターンにマッチするメッセージを検索しており，特定の技術やシステムの検索などを想定した手法といえる．

特定の目的を持ったメッセージの検索には少数のキーワードが充分な手がかりと考えられる．しかし，通常ネットニュースに投稿される記事には多種多様の一般的な用語が用いられており，その解析に際してはより抽象性の高い概念にキーワードを統合する必要があると考えられる．

多数のキーワードを少数の指標に統合するメッセージ内容の分析手法として，多変量解析に基づく手法が知られている [Krippendorff 1980]．これらの手法においては，多数のメッセージを含む母集団を対象に，個々のメッセージにおけるキーワードの出現頻度を因子分析，もしくは主成分分析し，各々のメッセージのスコアをメッセージの特性を代表する指標とみなす．

多変量解析を利用したメッセージ分析手法の具体例として，メッセージを特徴に応じて配置して検索を容易にする手法 [Deerwester 1990] や，電子的メッセージの分類への利用 [小川 1999] が提案されている．これらはいずれも静的な指標の抽出に止まっている．しかし，この手法を出抽出された指標を動的に分析すれば，議論の過程におけるメッセージ特性指標の変動から，議論そのものの特性も把握できる可能性がある．

対立を抽出するという目的から，一連の議論を構成するメッセージを母集団として分析すると，その内部における特性指標は，以下の，議論の過程に含まれる要因によって変動するものと考えられる．

- 時間の経過に伴う話題の変化
- 異なる話題を扱う枝分れした議論
- 論点，立場を異にする参加者間での論戦

上記最後の要因は対立に対応する．すなわち，二つの立場の間で対立が生じている場合，特性指標が異なるメッセージが短い周期で交互に現れる確率が高いことが期待される．この仮説が正しければ，議論の過程における特性指標の短周期の変動を測定することにより，対立を検出することができるはずである．そこで，これを機械的に検出する手法を開発し，実際のネットニュース上の議論に対してこの手法を適用して対立を検出することを試みた．

3.2 分析対象テキストの抽出

Table 3.1: ネットニュース記事の構成

Path: 記事の転送経路
From: 投稿者のユーザ ID
Newsgroups: ニュースグループ
Subject: サブジェクト (表題)
Message-ID: <メッセージ ID>
Date: 投稿年月日
References: <応答先メッセージ ID>
記事本文
--
署名

表 3.1にネットニュースの記事の構成を示す．記事はヘッダー部と本文で構成され，ヘッダー部の最後は空行で識別される．ヘッダー部の記述は RFC-1036[Horton 1987] で定められた様式に従い，投稿者のユーザ ID，メッセージ ID，投稿年月日，および，他の記事に対する応答記事である場合は応答先のメッセージ ID その他が記述されている．

キーワードの出現頻度は，ヘッダー部を除外した記事本文を対象として計測した．また，本文中で他の記事を引用している部分も，行頭文字により識別し，計測対象から除外した．ハイフン 2 文字もしくは 3 文字のみからなる行以降はシグネチャとみなして除外した．更に，定型的挨拶部分を削除するため，同一投稿者の記事の 20%以上かつ 3 件以上が先頭または末尾に同一の文字列を含む場合，一致部分を削除した．この操作により全文が削除される記事も存在した．これらは定期投稿文書および誤操作によって多数回投稿されたと思われる記事であり，解析対象から除外して妥当と考えた．

3.3 スレッドの形成

この解析はスレッド (一連の議論の流れ) を対象に行った．ネットニュースに投稿される記事は，新規の話題を提供する記事と，既に投稿された他の記事にフォローアップする記事とに大別される．フォローアップ記事はヘッダ部に Reference 行を含み，この部分の最後にフォローアップの対象となる記事のメッセージ ID が記載されている．この情報を用いることで，記事を参照関係にしたがう木構造として関連付けることができ

る．この木構造において，新規の話題を提供する記事はルートノードに相当し，フォローアップのなされなかった記事は末端ノードに相当する．

ネットニュースメッセージの規格を定めた RFC-1036[Horton 1987] には，Reference 行は以下のように規定されている．

> (References 行の) このフィールドは，この記事を投稿するきっかけとなった記事のメッセージ ID のリストである．これは，すべてのフォローアップ記事に必須であり，新しいサブジェクトの記事にはあってはならない．ニュースリーダはユーザがフォローアップ記事を投稿するためのコマンドを備えるべきである．このコマンドは，元のサブジェクトが “Re: ” または “re: ” で始まる場合は元の記事と同じサブジェクト行を，そうでない場合はサブジェクトの先頭に “Re: ” の 4 文字を挿入したサブジェクト行を生成しなければならない．元記事のヘッダーに “References” 行がない場合は，“References” 行は元記事のメッセージ ID(かぎ括弧を含む) を含まなくてはならない．元記事が “References” 行を持つ場合は，フォローアップ記事は元記事の “References” 行の内容，一つのブランク，元記事のメッセージ ID からなる “References” 行がなくてはならない．
>
> “References” 行の目的はユーザ側のニュースリーダが記事を会話毎にまとめられるようにすることである．これはニュースグループに様々な会話が混在することを可能とし，ニュースグループの購読を続けたまますべての会話を停止することもできるようにする．ニュースリーダはこのヘッダーを使わなくてもよいが，すべての自動的に生成されるフォローアップは，これを使うシステムのために “References” 行を生成しなくてはならず，手で入力する際にも同じようにこれを含めることを奨励しなくてはならない．
>
> 元の “References” 行が長すぎる場合は，これを全て含めなくてもよいが，過去に遡る参照として妥当な数を含めなくてはならない．

この規約に従えば，フォローアップ記事には Reference 行が必須である．しかし，記事のヘッダ部分は投稿者が編集することも可能であり，この規定は必ずしも守られていない．また，フォローアップの対象とされた記事が転送の過程で失われ，フォローアップ記事のみが到着する場合

もある．更に，解析対象外の記事 (他のニュースグループに投稿された記事や，解析期間以前に投稿された記事) に対する複数のフォローアップ記事が解析対象に含まれる場合もある．解析対象以外の記事は本来無視すべきである．しかし，解析対象以外の記事が一連の議論に含まれる場合には，一つにまとめられるべきスレッドが，二つ以上のスレッドに分割されるという，失われた記事と同様の問題を生じる．

このような問題に対処するため，解析対象に含まれない記事を示すメッセージ ID が Reference 行に現れた場合，これをメッセージ ID とするダミー記事を挿入することとした．フォローアップが続けて行われる場合，Reference 行は複数のメッセージ ID を含む場合が多い．Reference 行の末尾に記載されたメッセージ ID に該当する記事が存在しない場合，これをメッセージ ID とするダミー記事を生成し，ダミー記事の Reference 行として，元の Reference 行からダミー記事自身のメッセージ ID を除去したものを含める．これを木構造に追加して同様な操作を繰り返すことにより，失われた記事の結合関係が回復され，失われた記事による木構造の細分化が抑制される．なお，ルートノードであってかつ子が一つのダミー記事は，木構造に影響を与えないため削除した．

こうして形成された一つの木構造を一つのスレッドとみなして，解析の対象とした．今回の解析範囲において，スレッド総数は 5,998 であり，そのうち二つ以上の記事から構成されるスレッドは 3,197 であった．また，最終的に挿入されたダミー記事は 2,340 件であり，ダミー記事を含む解析対象記事総数は 52,634 件であった．

3.4 キーワードの検出

今回の解析では，漢字の 1 以上の連なり，すなわち単漢字および漢字の熟語，をキーワードとした．キーワードは，スレッド毎に，出現度数が高いものを機械的に選び出した．抽出するキーワードの数は，スレッドを構成する記事数の 1/4 とし，これが 150 以上となる場合は 150 とした．

キーワードの抽出は以下の手順にしたがって行った．

1. キーワードリストの初期値として，スレッド内で出現度数の高い単漢字を，充分な数 (抽出キーワード数の 2 倍とした) だけリストに登録する．

2. 解析対象テキストをスキャンして，リストに登録されたキーワードの出現度数とキーワードの連なりの出現度数とを計数する．

3. キーワードとその連なりを一括し，出現度数でソートする．

4. 出現度数の上位の連なりについて，キーワードを接続した新たなキーワードを作成して，キーワードリストに加える．出現度数の下位のキーワードが連なりを接続して作成されたものである場合はこれを分解する．

5. 2から4の操作を，100回を上限として，変化がある間繰り返す．

手順4においてキーワードの競合を避けるため，同一のキーワードが出現度数上位の二つ以上の連なりの前後に同時に含まれる場合，出現度数が最も高い連なりのみを接続操作の対象とした．また，同一の連なりについて接続・分解が繰り返されることを防ぐため，接続操作は出現度数の順位が抽出キーワード数よりも2以上上位の連なりを対象とし，分解操作は出現度数の順位が抽出キーワード数よりも2以上下位のキーワードを対象とした．

次に，キーワード抽出のアルゴリズムを，スレッド3における実際のキーワード抽出の過程参照して，以下に示す．

それぞれの記事に含まれる漢字文字列は，あらかじめヘッダ情報を格納したファイルに記録されており，ファイル上の位置を示す情報が，記事の構造体に保持されている．漢字文字列は，処理対象外となるヘッダ部，引用部，シグネチャ部を除去し，改行記号を除去し，漢字コード以外を区切り記号 (“-”) に置き換えたものである．二つ以上の区切り記号が連続する場合は一つにまとめている．漢字文字列の一例を示すと，「JK: 遅-恐縮-署名-引用-非常識-署名-言-部分-文字-直筆-書-文末-文-書-絵-書-常識的-考-署名部分-所属氏名以外-表記-思-上記-引用-方-私-非常識-引用」となる．

スレッド3は576記事を含むため，その1/4の144のキーワードを抽出する．抽出に先立ち，スレッド中の全ての記事で用いられている単漢字の使用度数を計数し，出現度数が上位288位までの (キーワード抽出数の2倍) の漢字をキーワードリストに登録する．抽出キーワード数の二倍の単漢字をリストに登録する理由は，このキーワード抽出アルゴリズムが，あらかじめ登録されたキーワードのみを対象として処理を行うことによる．キーワードの合成により，一つの合成キーワードがリストに追加されるが，これに伴い，その要素のキーワード出現度数は減少するため，初期状態において下位にリストされていた単漢字が抽出結果に含まれることも起こり得る．

キーワードリストは，キーワード文字列と，キーワードの出現度数と，そのキーワードに始まる連なりのリストへのポインタを含む．連なりの

リストは，その連なりを構成する二つのキーワードへのポインタと，連なりの出現度数を含む．記事のスキャンに先立って，キーワードの出現度数をゼロに，連なりのリスト長をゼロに初期化する．

キーワードの抽出は，それぞれの記事の漢字文字列を先頭からスキャンして行う．それぞれの位置からの文字列をキーワードリストのそれぞれの文字列と比較し，両者が一致する最長のキーワードを探す．一致するキーワードが検出された場合，キーワード出現度数を 1 だけ増加すると共に，記事の漢字文字列の次の位置から同様の操作を行い，キーワードの連なりを求める．キーワードの連なりが発見された場合，それが既にキーワードの連なりのリストに含まれていれば，その出現度数を 1 だけ増加し，連なりのリストに含まれていない場合は新たな連なりのセルをリストに追加して初期化する．

スレッドに含まれる全ての記事の漢字文字列のスキャンが完了したら，キーワードとキーワードの連なりを一括して出現度数でソートする．出現度数の順位が抽出キーワード数以内の連なりについては，連なりを構成するキーワードを接続した新たなキーワードセルを作成し，キーワードリストに追加する．また，出現度数の順位が抽出キーワード数以下のキーワードが複数のキーワードからなる場合は，これを個々のキーワードに分解する．スキャンと接続・分解の操作は，キーワードリストが安定するまで繰り返し行う．

以下のリストは，スレッド 3 のキーワード抽出における第一回の接続・分解の結果である．最初の行は，キーワード「記」と「事」を接続して新たなキーワード「記事」を生成したことを示す．括弧内はそれぞれの出現度数であり，キーワード「記」および「事」の出現度数は，それぞれ，366 回と 494 回であり，これらの連なりである「記事」の出現度数が 323 回であったことを意味する．

```
chk_bind test 言葉 (203) : 発言 (81),  発言 removed
chk_bind test 関西 (167) : 西弁 (121),  西弁 removed
bind 記 (366)-事 (494) => 記事 (323)
bind 自 (435)-分 (528) => 自分 (293)
bind 言 (940)-葉 (205) => 言葉 (203)
bind 問 (273)-題 (215) => 問題 (167)
bind 関 (281)-西 (169) => 関西 (167)
bind 指 (186)-摘 (156) => 指摘 (156)
bind 意 (383)-味 (172) => 意味 (146)
bind 理 (264)-解 (258) => 理解 (129)
```

```
bind 投 (128)-稿 (126) => 投稿 (126)
bind 議 (141)-論 (237) => 議論 (113)
bind 相 (131)-手 (246) => 相手 (105)
bind 場 (192)-合 (187) => 場合 (102)
bind 一 (320)-般 (102) => 一般 (98)
bind 意 (237)-見 (355) => 意見 (98)
bind 必 (117)-要 (136) => 必要 (94)
bind 常 (130)-識 (199) => 常識 (92)
bind 判 (138)-断 (96) => 判断 (92)
bind 馬 (83)-鹿 (83) => 馬鹿 (81)
bind 情 (92)-報 (82) => 情報 (75)
bind-cut loop changed 19 key words
```

以下は第2回の繰り返しであり，初回に生成されたキーワード「関西」に更に「弁」が接続され漢字3文字からなるキーワードが生成されている．2行目の連なり「発言」は，出現度数81回であり，初回の接続対象にもなり得たものである．しかし，キーワード「言」が，初回に接続された「言葉」と，前後に同時に含まれるため，「言葉」のみが初回の接続対象とされた．連なりの出現数の低いものも，第2回の繰り返しで新たに接続対象になりえる．これは，第1回での接続に使用されたキーワードの双方が，単体での出現頻度が低く，キーワード抽出数以下の順位に後退したためである．

```
bind 関西 (167)-弁 (174) => 関西弁 (121)
bind 発 (120)-言 (737) => 発言 (81)
bind 心 (133)-者 (207) => 心者 (74)
bind 質 (86)-問 (106) => 質問 (73)
bind 社 (110)-会 (138) => 社会 (71)
bind 人 (971)-間 (165) => 人間 (71)
bind 方 (453)-言 (656) => 方言 (70)
bind 関 (114)-係 (67) => 関係 (67)
bind-cut loop changed 8 key words
```

以下は第3回の繰り返しである．第2回で生成されたキーワード「心者」が，更にキーワード「初」と接続され，「初心者」を生成している．また，キーワード「関西弁」の追加により，キーワード「関西」の出現度数が低下し，キーワード「関」および「西」に分解されている．

```
bind 初 (157)-心者 (74) => 初心者 (74)
cut 関西 (46) => 関 (93)-西 (48)
bind-cut loop changed 2 key words
```

以下の第4回の繰り返しも同様で，「初心者」の追加によりキーワード「心者」が分解される．

```
cut 心者 (0) => 心 (59)-者 (133)
bind-cut loop changed 1 key words
```

第5回の繰り返しにおいては，接続・分解の対象は存在せず，繰り返しを終了する．

```
bind-cut loop changed 0 key words
```

このケースでは，5回の繰り返しにより操作が終了した．まれに，同一のキーワードについて，接続・分解操作が繰り返し行われ，キーワードリストが安定しない現象が見出されたため，キーワードの接続・分解操作の判断基準に幅2の不感帯を設けると共に，繰り返し回数の上限を100とした．

3.5 主成分分析

それぞれのスレッドを母集団として，スレッドを構成する記事を個体，キーワードの出現度数を変量とする主成分分析を行い，各記事の主成分スコアを求めた．解析は固有値が1より大きい主成分全てについて行ったが，図表には固有値の上位4主成分の結果のみを表示した．

各主成分の意味は，それぞれの主成分毎に，キーワードを主成分負荷でソートして高い負荷を示すキーワード群と低い負荷を示すキーワード群を抽出すると共に，記事を主成分負荷によりソートして主成分負荷の高い記事群と主成分負荷の低い記事群を抽出し，これらの内容を吟味することにより判定した．

解析には，それぞれの記事に含まれるキーワードの出現度数を用いた．この場合，全ての解析において第1主成分の寄与率が際だって高い値を示した．表3.2に示すように，第1主成分スコアの低い記事は本体行数が少なく，この成分は大きさの因子と考えられる．主成分分析において大きさの因子が第1主成分に現れることは珍しくない．ネットニュースで

Table 3.2: 記事の本体行数と第1主成分スコア順位

スレッド	スコア上位記事の本体行数			スコア下位記事の本体行数		
	1位	2位	3位	1位	2位	3位
1	143	120	146	6	12	14
3	229	195	133	11	27	12
6	158	96	131	9	61[1]	9

は，特に，長い記事と短い記事が混在しているため，大きさの因子が際だって高い寄与率を示したものと考えられる．

記事の長さの影響を除去するため，記事中の漢字数に対するキーワードの出現度数の割合を用いた主成分分析も試みた．この場合，大きさの因子が高い寄与率を示す現象は回避される．しかし，この手法は，キーワードをほとんど含まない短い記事の主成分スコアが異常値を示す例が多く，ノイズが増大する結果となったため，比率を用いた分析は不適当と判断した．なお，度数を用いた分析にも，大きさの因子が変動した場合，全ての主成分スコアが変動するという問題が残っている．

3.6 自己相関係数およびパワースペクトル

この論文で提案する手法は，議論の過程における主成分スコアの変動から，議論に含まれる対立を検出するという着想に基づいている．変化する現象を分析する手法として，スペクトル解析が一般的であるが，分枝が多数存在するネットニュースのスレッドでは，十分な長さの変量系列が得られないという問題がある．

自己相関関数は，時間に関する不規則変量のτ時間隔たった二つの変動の積のアンサンブル平均として定義されるが [日野 1977]，ラグτにおける関数値はτ時間隔たった二つの変量間の共分散に他ならない．共分散は，統計的性質が一定とみなされる範囲内であれば，τ時間隔たった二つの変量の全ての組から計算可能であり，分枝の存在する系でも大きなサンプル数を用いて解析できる．この論文では，スレッド毎に1〜8世代隔てた記事間のスコアの共分散を計算して自己相関関数を求め，これから自己相関係数とパワースペクトルを計算し，世代による変動の有無を調べた．

フーリエ変換は，通常の高速フーリエ変換アルゴリズムを用いて行った．自己相関関数が偶関数であることから，負のラグにおける自己相関関数関数値とラグ0における自己相関関数値 (分散に相当) も含め，17点

の自己相関関数値に対してフーリエ変換を行った．また，最も注目すべき 2 世代周期の振動の検出を可能とするため，ラグ間の中心点の自己相関関数値を前後の平均によって補い，データ点数を 2 倍に拡大して計算した．この手法については，自由度に矛盾がないこと，およびメイクアップデータを用いて 2 世代周期の振動が検出できることを確認している．

実際のスレッドにこの手法を適用した結果，一部の自己相関係数にラグ 2 を周期とする明らかな振動が認められた．また，これに対応するパワースペクトルにも，2 世代を周期とする点で上昇が認められ，2 世代を周期とする主成分スコアの振動が存在すると判断された．相関係数 0.2 に対する 95%信頼区間は，スレッドを構成する記事数が 400 の場合，0.10 〜0.29 と見積もられ [国沢 1966]，振動の検出は統計的にも有意と考えられる．

2 世代を周期とするパワースペクトルの値の評価はに際しては，ベースライン補正を行った．ベースラインは，波数 1/16 世代から 7/16 世代のパワースペクトル値を線形回帰することによって求めた．回帰にあたっては，波数 8/16 世代からの距離の自乗に反比例する重み付けを行い，目的とするピークに近いベースラインを重視するようにした．