書き込みのエントロピーに着目した 速報スレッドにおける盛り上がり要因の分析

前川和輝　　松本和幸　　吉田稔　　北研二
徳島大学工学部知能情報工学科

1. はじめに

ニュース関連の掲示板における書き込み内容は，意見・評判情報や口コミ情報の豊富さにより，マスメディアに匹敵するほど我々の生活に深く関わっている．話題性に富んだニュースであればあるほど掲示板における書き込みは増加し，スレッドの消費速度も速くなる．また，過去の書き込み内容を閲覧することで，新たに掲示板上で意見が述べられたり，メディアに取り上げられることもある．

ブログやTwitter等のSNSで軽犯罪自慢や誹謗中傷など不適切な発言をすることで，その内容が瞬く間にWeb上で拡散される．これにより，2ちゃんねる等の匿名掲示板でスレッドが乱立し，個人では収拾がつかない状態，いわゆるネット炎上という現象が起きる．ネット炎上には匿名電子掲示板が大きく関わっており，そこでの盛り上がりがネットユーザの関心を集める要因になっている．

本研究では掲示板における盛り上がりの要因の分析を目的としている．盛り上がり要因を特定することで，掲示板の一連の書き込みのなかで，どの箇所がもっとも盛り上がっているのかを自動検出できると考える．

盛り上がっているかどうかの指標として，「勢い」というものがある．これは，ひとつの掲示板において書き込み速度（消費速度）を計算することで得られる．しかし，単にスレッドを消費させることが目的と考えられる，内容に乏しい無意味な書き込みまで盛り上がりの程度を上昇させることがあり，本来の意味での盛り上がりを特定できない可能性がある．内容に乏しい無意味な書き込みがなされる原因として，つぎのようなことが考えられる．

- 指定した書き込み番号まで書き込み数を増やす目的
- アフィリエイトのリンクへの誘導目的

このような書き込みによる勢いは，本来の意味での盛り上がりの検出には有効ではない．

ネット炎上した勢いのあるスレッドを実際に観察すると，ある特定の対象に対する誹謗中傷や意見・感想を含んだ書き込みが多くみられる．このように，掲示板におけるネット炎上では，自らの意見や考えを主張する書き込みと，内容のない無意味な書き込みが混在しており，どちらの書き込みも勢いの値に影響を与えてしまう．そのため，書き込み速度以外の指標を組み合わせて用いることにより，本来の意味での盛り上がりを検出する必要がある．

2. 関連研究

電子掲示板を用いた書き込み解析の研究は従来からおこなわれている[1-3]．高間ら[4]は電子掲示板のスレッドにおける話題とその推移を，共起キーワードの時間的変化に着目し，アニメーションにより提示することで話題推移を可視化するシステムを考案している．この研究では掲示板の話題推移を，以下に示す3パターンに分類している．

1. 話題推移パターン
   炎上が発生することも多いが，参加者が自分の考えやその理由を活発に書き込まれ，重要な情報が多く含まれるパターン
2. 話題雑談パターン
   議論でなく雑談であるため炎上が発生しにくいが，スレッドの内容に重要な情報が含まれるパターン

3. 自由雑談パターン
話題とは関係のない荒らし行為がおこなわれているパターン

この研究では，これら 3 パターンの特徴を用いて一定時間内におけるキーワードの共起度からスレッド内の内容に関する要約を生成する．この要約を用いることで，ユーザがスレッドの全体像を把握する負担を減らしている．

本研究では，最終的に「炎上」状態を検出することが目的である．そのため，主に話題推移パターンの抽出を目標として分析をおこなう．

3. 盛り上がりの分析方法

2 ちゃんねるにおける掲示板カテゴリ「ニュース速報」では通常，ひとつのスレッドは短期間で消費されてしまい長期間継続することは少ない．しかし，炎上が起こった場合にはスレッドが継続的に立てられ長期にわたって書き込まれ続ける．本研究では 2 ちゃんねる掲示板の炎上に古くから関わっていると考えられるニュース速報板に着目する．掲示板の盛り上がりを検出する方法として，書き込みの勢いとコメントの情報量との関係を調べる．書き込みの勢いの計算は，式(1)を用いておこなう．

$$\text{書き込みの勢い}(t,n)=\frac{\text{書き込み数}_n}{\text{書き込み}_t\text{から書き込み}_{t+n-1}\text{までの経過時間}} \quad (1)$$

書き込み t から書き込み $t+n-1$ までは，連続する書き込みであり，書き込み数は全部で n である．書き込み集合におけるエントロピー(情報量)は式(2)で計算することができる．図 1 のように，一定数の連続する書き込みをひとまとまり（書き込み集合）として抽出し，抽出範囲を少しずつずらしながら，書き込み内容をもとにエントロピーを計算することで，情報量の多い，または少ない箇所を検出することを考える．今回，文字 n-gram または単語 n-gram，書き込み集合抽出のウインドウ幅 W，ずらし幅(シフト長) S をパラメータとする．

$$\boldsymbol{H(X)=-\sum P_i \log P_i} \qquad \boldsymbol{(2)}$$

(P_iはある n-gram$_i$ が出現する確率)

図 1　書き込み集合抽出例

ここで，ネット炎上したスレッドのなかでも，長期にわたってスレッドが立てられたトピックについて着目する．ある大学生が Twitter 上で不適切な発言をしたために炎上した話題についてのスレッド(2011 年)である．この話題に関するスレッド群について時系列順に勢いとエントロピーをスレッド単位で計算した結果を，図 2 に示す．

もっとも勢いのある箇所ではエントロピーとの関連はあまり見られなかったが勢いが盛り返している図中の②ではエントロピーがもっとも小さくなる箇所が見られた．この箇所では主に図 3 のような同じようなアスキーアートの書き込みが多く見られた．

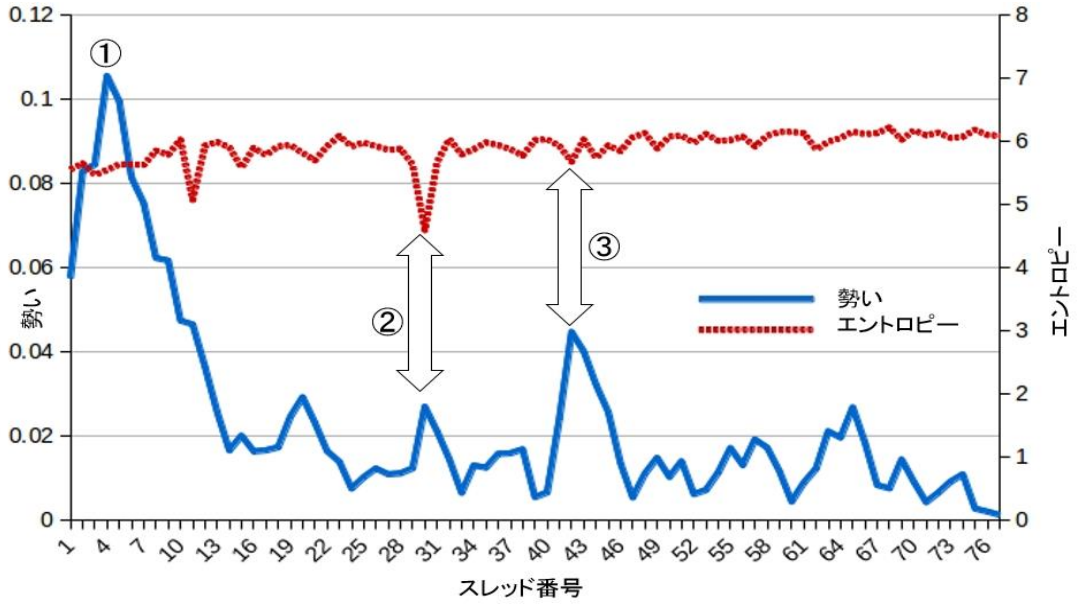

図 2　1 スレッドごとの勢いとエントロピー

図 3　書き込まれたアスキーアートの例

つぎにスレッド内での勢いとエントロピーを検証する．ここでは，パラメータを，W=200, S=200 として勢いとエントロピーの計算結果を図 4 に示す．なお，エントロピーの計算は，単語 1-gram の出現確率をもとにおこなった．

ひとつのスレッド内での勢いとエントロピーの各ピーク（最大値と最小値）の位置が近いため，この 2 つは関係があるように見てとれる．しかし，図 2 において，勢いの値が大きく，エントロピーの値が小さい箇所ではほぼ同一のアスキーアートが連続して書き込まれていることが目立った（図 2 の②の箇所）．この原因として，掲示板訪問者が当該話題に飽きて書き込み数が減少し，落ち着いてきたときに，スレッドが dat 落ち（スレッドが終了して書き込めなくなる状態）してしまうことを防ぐために，意図的にこうした書き込みをしたと考えられる．

また，図 2 の③の箇所では，勢いが盛り返していた箇所のうちでも上昇幅が大きく，エントロピーも小さくない．この箇所の直前で，2 ちゃんねるで話題になったスレッドを掲載するまとめサイトやネットニュースに掲載された可能性がある．常時掲示板を訪問し，書き込みをおこなう人以外が，まとめサイト等を見て掲示板を訪問し，新たに立てられた当該トピックの最新スレッドに，意見等の書き込みを活発におこなったためであると推測される．つまり，勢いの推移をみるだけでなく，エントロピーと照らし合わせてみることで，そのトピックに関する最初の盛り上がりだけでなく，2 度目以降の盛り上がりを正確に検出できる可能性がある．

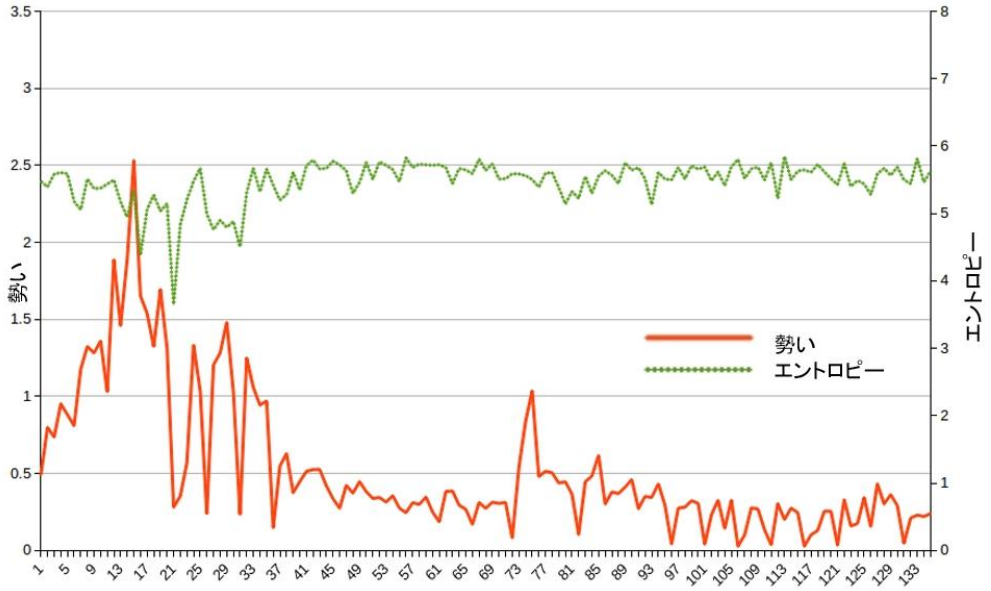

図 4　スレッド内での勢いとエントロピー

4. おわりに

本稿では，掲示板における盛り上がり要因の分析を目的とし，書き込みの情報量（エントロピー）に着目した．分析結果から，勢いのみに着目した場合に見落としてしまう様々な現象を，エントロピーの単純な計算により検出できる可能性があることを明らかにすることができた．また，エントロピーが小さくない状態での勢いの盛り返しは，２度目以降の盛り上がりとして判断できると考えられる．

今後は，勢いとエントロピーの推移をより明確にするために，最適なパラメータの決定方法を提案するとともに，炎上箇所の検出に有効な特徴量の抽出方法を提案したいと考えている．また，従来手法を用いた話題推移パターンの抽出との比較もおこないたい．

参考文献

[1]. 松村真宏，大澤幸生：“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.

[2]. 桜井茂明，折原良平：“掲示板サイト分析における重要議論抽出と特徴表現抽出”，知能と情報(日本知能情報フィジィ学会誌) Vol.19，No.1，pp13-21，2007.

[3]. 松尾豊，大澤幸生，石塚満：“電子掲示板における会話からのトピックの発見と要約”，第16回人工知能学会全国大会，3D1-07，2002.

[4]. 高間康史，小井沼岳：“BBS スレッドにおける話題推移理解を支援する視覚的要約手法”，知能と情報(日本知能情報フィジィ学会誌) Vol.22，No.6，pp.680-690，2010.