2J3-3

掲示板のまとめブログにおける編集行為の分析

The Analysis of Editing Activities in Summary Blogs for BBSs

武田 英明[*1*2]
The program committee of the XXth annual conference of JSAI

沼 晃介[*3]
Second Author's Name

[*1] 国立情報学研究所 National Institute of Informatics
[*2] 総合研究大学院大学 Graduate University for Advanced Studies
[*3] 株式会社サードインパクト The Third Impact, Inc.

Social media has the indispensable role in our society. Editing social media is the important to make social media to distribute easily. In this paper, we collect and analyze the summary blogs ("matome blogs") in order to explicate editing ability of social media. We found that most of blogs are within 100 responses whereas the original BBSs vary so much in length. It suggests the editing policy to keep the length reasonable. Overlap of the selected responses among different blogs are significantly high. It also suggests that the editing policy is not so different among them.

1. はじめに

現在は掲示板やブログといったソーシャルメディアは社会で大きな役割を占めるようになった。ソーシャルメディアは誰でも情報発信することができ、既存のマスメディアではなかった様々な新しい可能性を提供している。一方、ソーシャルメディアは基本的にコミュニケーションや小規模なコミュニティでの情報共有のために利用されているものであり、そのままマスメディアの代替手段になるわけではない。

その違いの 1 つは編集行為のあり方である。マスメディアでは情報を一定の方針で収集・統合して、受け手が便利なように情報を編集している。ソーシャルメディアでは多数の情報提供元がある上、情報の収集・統合はより重要である。そこで情報をまとめるという行為(編集行為)がソーシャルメディア上で重要な役割を占めるようになった。

本研究ではソーシャルメディアでの編集行為がどのようものであるかを調べることを目的とする。具体的には2ちゃんねるのまとめブログを収集して、オリジナルの情報を比較することで、どのような編集行為が行われているかを分析する。

2. データ収集

2.1 まとめブログの収集

まとめブログは現在多数存在するが、その中で比較的人気の高いブログのデータを得るため以下の方法でデータを収集した。

2013/1/31 時点の Livedoor ブログ「ニュース総合→まとめ」のランキング上位 100 ブログを対象に、各 100 エントリーを取得した。ブログの実際の総エントリー数により、実際に取得されたのは 9597 エントリーであった。上記 100 ブログのうち、2ちゃんねるスレッドのまとめを 1 エントリーでも扱っているの(スレッドの明記のあるもの)は、85 ブログであった。全体の 9597 エントリーのうち、分析の対象となる 1 エントリーに対し 1 スレッドが対応してまとめられている記事は、6314 エントリーであった。

この中でスレッドの重複情報を表1に示す。収集範囲では意外に重複度は高くなく、77%はそのブログにしかまとめが存在しないスレッドであった。以下ではこの中の 6 回以上の重複に関するブログエントリーを対象とする。

連絡先:武田英明、国立情報学研究所、東京都千代田区一ツ橋 2-1-2, takeda@nii.ac.jp

2.2 元スレッドとの対応の調査

次にこれらのブログエントリーと2ちゃんねるの元スレッドとの対応を求めた。

対象となるスレッドは 89 スレッドだが、dat ファイルの取得できたものはこのうちの 40 スレッドであった。この 40 スレッドについて、60 のブログにより計 320 エントリー(まとめブログのページ)があった。

2.3 レス抽出

取得したエントリーに対しレスの抽出を行う。レスとは掲示板における個別の発言のことを指し、編集行為の基本単位となる。

ブログごとにタグ付けスタイルが大きく異なるため、単純なパターンでレスを抽出することはできない。また、要素の使い方が HTML の文法に対して正しくないものが多い。さらには、元のスレッドから一部のレスのみを抽出するとともに並び替えを行う(これがまとめの編集作業である)ため、単純な dat との照合も行えない。掲示板上での画像 URL を展開し埋め込んだり、注釈やアフィリエイトリンクなどを差し込むこともある。

今回はレス選択と文字修飾を中心に抽出することを優先して、閉じタグを補足したり記事内を div で分割したりするなどいくつかのヒューリスティクスは組み込んだものの、概ね元の DOM 構造に沿って解析を行った。

320 のエントリのうち、レスを修飾するタグを完全に取得できなかったものは 53 エントリーあった。これらについては HTML が

表 1 収集したブログでの重複度の分布

重複	ブログ数	割合	重複	ブログ数	割合
1	3315	77.2%	11	1	0.0%
2	550	12.8%	12	3	0.1%
3	190	4.4%	13	4	0.1%
4	101	2.4%	14	1	0.0%
5	48	1.1%	15	0	0.0%
6	36	0.8%	16	0	0.0%
7	21	0.5%	17	0	0.0%
8	13	0.3%	18	0	0.0%
9	5	0.1%	19	0	0.0%
10	4	0.1%	20	1	0.0%

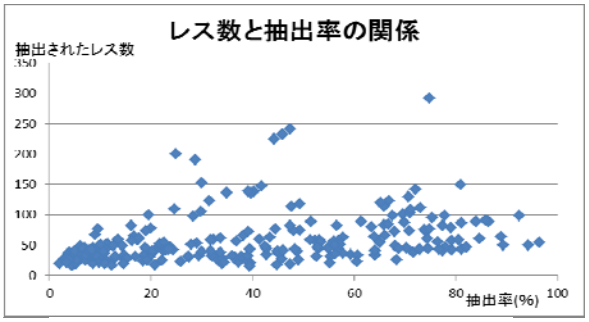

図１　レス数と抽出率の関係

手書きによりランダムに誤っているなどの事情により文字修飾部分を取り出すことはできなかった。加えて、エントリー内に記述された元スレッドの URL に誤りがあり dat との対応が取れずに解析できなかったものも含む。

3. レス抽出の分析

以上のようにして収集したデータから、元スレッドがどのように編集されているかを分析していく。

3.1 抽出の割合に関する分析

39 スレッドについて、エントリー中にいくつのレスが選択されているかの傾向を見る。

まずエントリーごとに、スレッドの全レス数を基準としてそのエントリーに選ばれ表示されているレスの分量の割合（抽出率）を計測する。全エントリーの平均抽出率は 36. 9%であった。

20%以下のレスを抽出しているエントリーが多く、80%以上のレスを含んでいるエントリーは少ない傾向にあるものの、分布はばらついている。267 エントリーにおける抽出率の標準偏差は 25.93 である。

同一スレッドを参照するエントリーを母集団としてレスの抽出率の標準偏差と平均値を求めた。40 スレッド中で最も抽出率の標準偏差が小さかったもので 0.40、最も標準偏差の大きかったもので 22.88 であり、すべてのスレッドにおいてばらつきが全エントリーを母集団とするより小さくなっている。

次に同一ブログに含まれるエントリーを母集団としてレスの抽出率の標準偏差と平均値を求めた。対象となる 267 エントリー中に複数の記事を持つブログは 47 ブログであった。ブログごとの抽出率の標準偏差の最小値は 2.33、最大値は 38.28 であった。スレッドごとに集計するよりもばらつきが大きくなっているが、概ね全体エントリー内で見るよりはばらつきが小さい傾向にあった。

以上から、エントリーへのレスの抽出割合には、ブログごとの編集方針の特徴もある程度現れるものの、スレッド自体の影響が大きいと考えられる。

抽出率と抽出されたレスとの関係の分布図を図1に示す。全体としては比例傾向にあることがわかるが、一方100レス以下のものが多いこともわかる。すなわち、元スレッドの長さよりも結果が一定の長さになるように編集していることがうかがえる。

3.2 抽出されたレスとリプライ数との関係

抽出されたレスの重複度とそのレスにつけられたリプライ数の関係を図2に示す。x軸はそのレスに対するリプライの数を示しており、そのリプライ数のレスがエントリ群の中でどれだけ採用されているかの割合を示している（左軸）。参考までにその個々のリプライ数に該当するレスの数を右軸に示す。対象となったレス数が少ないためばらつきが多いが、全般的に多くのリプライを受けているレスはより多くのエントリで採用やすくなる傾向がうかがえる。

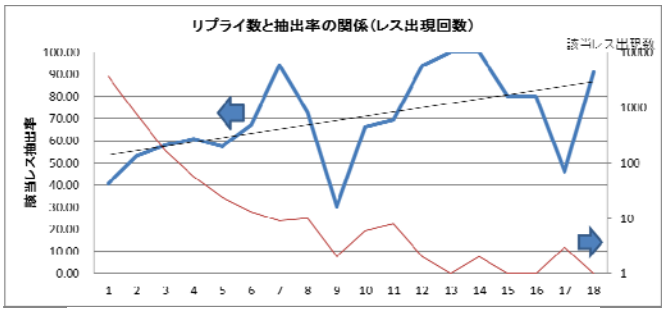

図２　リプライ数と抽出率の関係

3.3 レスの重複度の分布の分析

次にレスの重複度が 1 つのスレッドの中でどのように分布するかを調べた。図3は 39 のスレッドをそのまとめを作っているエントリーの抽出率（どれだけのレスを元スレッドからとっているか）の平均の順に並べたときに、どれだけのレスが多重に抽出されたかを示す図である。例えば 80－100%のカテゴリーはそのスレッドを利用しているエントリーの 80-100%で利用されたレスを示している。1エントリーでも使われたレスを総数として 5 分割したカテゴリーの分布が棒グラフになっている。参考のために、エントリー数（右軸）も折れ線グラフで示している。

重複の度合いは全体的にかなり高い。例えば、19 番目のスレッドは 6 エントリーがあり、平均抽出率は約 40%である。そのなかで使われているレスの 20%以下だけが1つのエントリーにしか使われていないレスで、その他は複数エントリーで使われている。

平均抽出率が高ければ重複の度合いが増えるのは当然であるが、一方エントリー数にはあまり影響されていない。すなわち、エントリーが増えたからといっても抽出の度合いはあまり変化していない。このことはブログ間の編集のふれがあまりにないことを示唆していると思われる。

4. まとめ

ソーシャルメディアでの編集行為を知るために、2ちゃんねるまとめブログを収集して、データの重複度などを分析した。今回は共通性について考察したが、今後は差異についても分析を行う予定である。

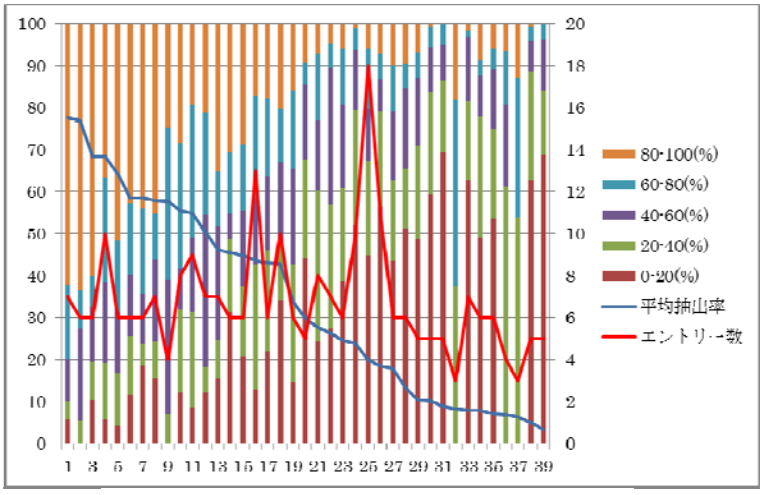

図３　抽出されたレスの分布