

オンラインディスカッションの効率性定量化手法の提案

武吉 朋也[†] 帆足 啓一郎[†] 松本 一則[†] 滝嶋 康弘[†]

KDDI 研究所[†]

1. まえがき

SNS(Social Networking Service)等のオンラインコミュニティサービスの普及により、近年では一般ユーザであってもオンラインでディスカッションを容易に行える。このような状況下で交わされるユーザ同士のディスカッションは、正常に進行されない、荒れたディスカッションとなる場合がある。荒れたディスカッションが多くなると、ディスカッションの場を提供しているサービス運営者は、その場を健全に保つための行動が必要となる。しかしながら、提供サービス内で行われている多数のディスカッションについて、健全に盛り上がっているか、それとも荒れているかを判定するには、現状目視で確認する必要があり、多大なコストを要する。これらから、特に荒れているディスカッションをサービス運営者に自動的に提示することは重要である。

オンラインで行われているユーザ同士のコミュニケーションの解析に関連する既存技術としては、ブログのコメント欄での反応を定量化し検索へ応用する技術[1]が挙げられる。しかしながら、主に発言の数やその時間間隔に着目しているため、ディスカッションが健全に盛り上がっているか、荒れているかを判別できない。そこで本稿では、発言間で新規に出現する情報に着目し、効率的にディスカッションが行われているか、という観点でディスカッションを評価し、荒れているディスカッションを抽出する手法を提案する。

2. 効率性指標

本章では、荒れているディスカッションを検出するために、効率性を測定する指標(以降、効率性指標とも呼ぶ)について述べる。言い争いや平行線の発言が多数発生し、荒れているディスカッションの特徴として、下記2点が挙げられる。

- (1) 発言内に新規の情報が含まれにくい
- (2) 発言間隔が短い

(1)については、言い争いや等が発生すると、自身の主張のみを変更することなく発言し、同一内容の発言を繰り返す傾向にあるため、このような特徴を有する。また、(2)については、他者の発言を理解しない、あるいは初めから理解する意思が無く、自身の意見を発言することのみに注力する傾向にあるため、このような特徴を有する。

本稿では、(1)と(2)の特徴を有するディスカッションについて、正常に進んでいない「効率性の低いディスカッション」と定義(図1に概念図を示す)し、上記特徴の傾向が強いディスカッションを荒れたディスカッションとして抽出する。各ディスカッションの効率性指標として、上記2つの特徴をそれぞれ新規情報含有度、発言間隔度として定量化する。以下、2.1.節で新規情報含有度の算出方法、2.2.節で発言間隔度の算出方法について述べる。なお、以降では効率性を測定するディスカッションデータ数を n 個とし、任意のディスカッションデータを d で表す。また、 d 内の総発言数を m 個とし、 j 番目の発言を c_j で表す($1 \leq j \leq m$)。

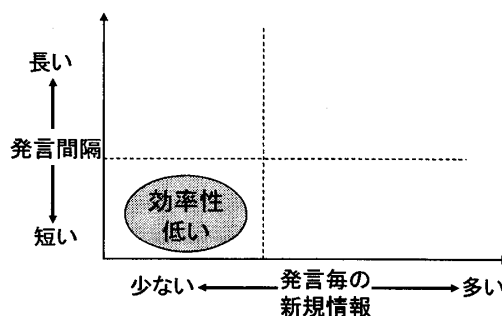


図1 効率性が低いディスカッションの特徴

2.1. 新規情報含有度

ディスカッションデータ d の新規情報含有度 $I(d)$ は、各発言で新規に出現した単語数を、新規の情報と捉え、算出する。この値が大きなディスカッションは平均して各発言で新規の情報が多ことを表す。本稿では、新規情報含有度を下記処理ステップにより算出する。

Step w_1: ディスカッション d 中の1番目の発言、つまり c_1 に形態素解析を適用し、単語を抽出する。さらにこれらの単語からなる集合 $W(1)$ を構築する。

Step w_2: 2以上の j に対し、 j 番目の発言、つまり c_j に形態素解析を適用し、単語を抽出する。そして、 c_j に含まれ、かつ $W(j-1)$ に含まれていない単語の数を集計し、 k_j とする。

Step w_3: k_j 個の単語を $W(j-1)$ に加え、 $W(j)$ とする。

Step w_4: 2以上 m 以下の j に対し、Step w_2 と w_3 を実施し、 k_j の総和 $K(d)$ を求める。

Step w_5: ディスカッション d での平均値、つまり次式により新規情報含有度を算出する。

$$I(d) = K(d) / (m - 1)$$

“Efficiency Quantification Method for Online Discussion”

[†]Tomoya TAKEYOSHI, Keiichiro HOASHI, Kazunori MATSUMOTO and Yasuhiro TAKISHIMA, KDDI R&D Laboratories, Inc.

2.2. 発言間隔度

[1]では、発言間隔の平均値(本稿での発言間隔度)を、ユーザ同士のインタラクション特性のうち「速さ」を構成する要素として用いていたが、本稿では効率性指標として利用する。発言間隔度 $T(d)$ が大きなディスカッションは、平均して各発言間の時間間隔が長いことを表す。本稿では、発言間隔度を下記処理ステップにより算出する。

Step t.1: 2 以上の j について、 c_j と c_{j-1} の時間間隔 t_j を算出し、ディスカッション d 全体での t_j の総和 $S(d)$ を求める。ここで、時間間隔は分単位で求める。

Step t.2: ディスカッション d での平均値、つまり次式により発言間隔度を算出する。

$$T(d) = S(d) / (m - 1)$$

3. 評価実験

本章では、効率性指標が有効に作用するか評価実験を行い、その結果について述べる。まず、実験に用いたデータと正解の定義といった実験条件について述べる。次に、実験結果と考察について述べる。

3.1. 実験条件

本稿では、Wikipedia のノートページ¹上で行われているディスカッションを評価実験に用いる。ここで、ディスカッションの単位は、ノートページ中の「節」で区切られるテキスト範囲とし、発言の単位については発言ユーザ名と発言時刻の組み合わせで区切られるテキスト範囲とした。なお、総発言数が 10 個以上 ($m \geq 10$) であるディスカッション 7769 件を実験用データとした。

さらに、発言間隔度の 5 パーセンタイル値 (224.625²) 以下であるディスカッションを実験用データから抽出した後、発言数が 10 または 11 であるディスカッション 100 件を評価対象データとし、正解ラベルの付与を行った。ラベル付与は評価者 1 名により実施し、下記 4 種類のラベルを付与した。

- (ア) 8 割以上が効率的でない発言
- (イ) 5 割程度が効率的でない発言
- (ウ) 2 割程度が効率的でない発言
- (エ) ほぼ全てが効率的な発言

ただし、上記の「8 割以上」等はおおよその目安であるため、厳密な数値ではない。また、効率的でない発言とは、言い争いや平行線の発言、他のユーザを攻撃する発言等、ディスカッションに不要な発言とした。(ア)、(イ)、(ウ)のいずれかが付与されたディスカッションは効率的でない、荒れている議論として評価を行う。ラベル付与の結果、

63 個が効率的でない議論であった。

3.2. 結果と考察

図 2 に適合率-再現率曲線を示す。適合率と再現率の算出には、効率性指標について任意のパーセンタイル値以下である評価対象データを効率的でないと判定した結果と、ラベル付与結果を比較して算出した。なお、発言間隔度については、1% から 5% まで 1% 刻みのパーセンタイル値を用いて判定した結果である。一方、新規情報含有度については、5% から 100% まで、5% 刻みのパーセンタイル値を用いた結果である。

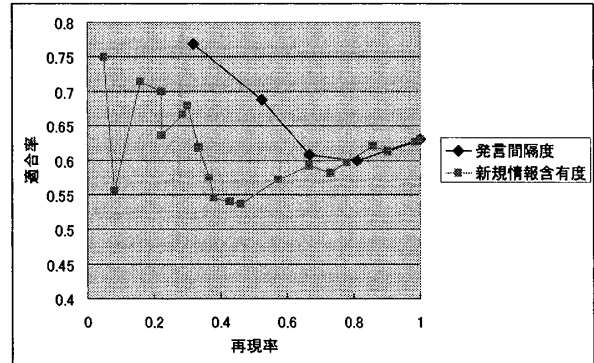


図 2 実験結果

図の結果より、発言間隔度のみを用いる方法が、新規情報含有度を用いる方法よりも適合率が高い結果となった。新規情報含有度が有効に作用しなかった例としては、ユーザ同士が互いの批判を繰り返しているディスカッションが挙げられる。この例では、批判を繰り返した結果として、ディスカッションに関連の無い新規単語が多数出現し、新規情報含有度が高くなったと考えられる。この結果より、単純に単語数を用いるのでは無く、その発言がテーマに関連があるのか等、文脈を考慮して新規情報含有度を定量化する必要があると言える。

4. まとめ

本稿では、オンライン上で行われている多数のディスカッションから、荒れている議論を抽出するために、その効率性を測定する 2 つの指標の利用を検討した。発言毎に追加される新規の情報を、それまでに出現していない単語数で定量化した新規情報含有度について、各発言がディスカッションに関連するか考慮する必要があることを、実験結果から確認した。今後の課題としては、各発言についてディスカッションへの関連度を定量化し、新規情報含有度に適用する手法の検討が挙げられる。

参考文献

- [1] 宮田章裕, 他, “反響特性分析を利用したブログ記事検索手法,” 情処学論, Vol. 48, No. 12, pp. 4041-4050, Dec. 2007.

¹ Wikipedia 日本語版のデータベースダンプのうち、2009 年 12 月 20 日版を下記サイトよりダウンロードし、使用した。

<http://download.wikimedia.org/jawiki/>

² 実験用データ中の 5% が発言間隔度 224.625 以下であることを表す。