熟知度とセンチメント分析に基づくブロガー中立度判定手法の検討

Method of estimate blogger's neautrality based on sentiments and knowledgeable analysis about specific topic

志甫谷 匠 † **中島 伸介** †† **角谷 和俊** ††† **稲垣 陽一** ††††
Takumi Shihoya Shinsuke Nakajima Kazutoshi Sumiya Youichi Inagaki

1 はじめに

昨今，CGM の普及により，一般の人々が情報を発信することが容易になり，欲しい情報を簡単にに手に入れやすくなった．一般の人々が特定のトピックについての情報を得ようとする場合，そのトピックについて熟知しているユーザが発信した情報を参照する方が，そのトピックについて詳しくないユーザが発信した情報を参照するよりも，有益な情報を得られる可能性が高い．これは，そのトピックに対して，広い知識を有しているユーザが，他のユーザと比べて，有益な情報を発信している可能性が高いからである．中島ら [1] の研究では，ブロガーの熟知度を対象熟知領域に関連する話題を含んだエントリの投稿数に基づいて算出し，熟知ブロガーの判定を行っている．

しかし，ブロガーが特定のトピックに対して数多くの記事を投稿するということは，熟知度が高くなる一方で，トピックに対するブロガーの思い入れが強くなっている可能性が高く，発言の客観性・冷静さが損なわれている可能性がある．例えば，“ プロ野球 ” というトピックにおいて，巨人，阪神，中日などの各球団に関するサブトピックが存在するが，熟知ユーザが投稿する内容が，特定球団のトピックに依存している場合，そのユーザ発信する情報には，自ずと特定球団を応援する感情が込められる可能性が高くなる．そして，このユーザの立場は特定の球団の立場に偏ったものである可能性も高くなるため，その熟知ユーザが発信した情報が，野球というトピックにおいて必ずしも有益であるとは言いがたい．なぜなら，野球というトピックに精通していないユーザが野球というトピックの知識を立場の偏った熟知ユーザから得ようとする場合に，情報を受け取ったユーザに誤解が生じる可能性があるからである．

本研究では，このような立場の偏りが発生する原因を以下の 2 点にあると仮説を立てた．

- 述べられている意見の冷静さによって信頼性が変化する
- 特定トピックに対する知識の偏りによって信頼性は変化する

そこで本研究では，中立度という尺度を定義し，熟知度が高く，かつ中立度の高いブロガーを発見し，より客観的で有益な情報の取得を目的とする．このことにより，有用性があると想定されるのは以下の 2 点である．

- トピックに熟知していないユーザが，トピック内で中立な立場のブロガーから意見を聞ける
- ブロガーの中立度を測ることで，中立度の高いブロガーを多く含む熟知グループを推定出来る

以下に，本稿の構成を記す．2 節では，専門性や熟知度を扱った研究や，センチメント分析を行った研究についてまとめる．3 節では，本研究で用いる概念の定義を行う．具体的には，中立度と熟知度の定義を行い，既存熟知グループの細分化についても触れる．4 節では，ブロガーの中立度判定手法について説明する．加えて，本研究で想定する中立度判定を用いたアプリケーション例の説明を行う．5 節では，まとめと今後の課題について述べる．

2 関連研究

本節では，専門性や熟知度に関する研究やセンチメント分析を行った研究，立場の違いに着目した従来研究を挙げていく．

2.1 専門性・熟知度に関する研究

中谷らは [2]，特定分野における専門用語を，非専門家ユーザにとって理解しやすいように wikipedia のリンク構造とカテゴリ構造を用いて，ユーザが入力した検索語からその語に関連する専門用語を抽出する手法を提案している．高橋らは [3]，Web テキストと修辞表現の間の適合度を判定する手法を提案している．このことにより，一見専門性が高いように見える修辞表現を用いた Web テキストでも，内容と修辞表現が伴っていないかどうかを判断することを可能にしている．竹原らは [4] 特定のトピックに関連するキーワードの発言頻度の多さによって，ブロガーの熟知度を判定し，どれだけ熟知度の高いブロガーからのリンクを持っているのかによって web ページのトラスト値を算出する手法を提案している．

2.2 センチメント分析の研究

濱砂らは [5] 任意のトピックに対するニュース記事がどのような観点から見られているのかを抽出し，各ニュース記事に対するセンチメント値を 4 軸 8 方向で表現している．さらに，これらのセンチメント分析の結果を，日本地図上にマッピングすることで，地域ごとの観点の違いを視覚化している．俵本らは [6] 感情を分布で表現するという分布感情モデルを提案している．このモデルは，感情分析を行うための汎用的なモデルとして提案されている．この研究では，ある特定事象に対して人々が抱きや

† 兵庫県立大学大学大学院 環境人間学研究科，Graduate School of Human Science and Environment, University of Hyogo
†† 京都産業大学コンピュータ理工学部，Faculty of Computer Science and Engineering, Kyoto Sangyo University
††† 兵庫県立大学 環境人間学部，School of Human Science and Environment，University of Hyogo
†††† 株式会社きざしカンパニー，kizasi Company,Inc

すい感情の分析が行われている．

2.3　立場の違いに着目した研究

井上らは [7] 時事問題における論点を抽出し，その論点にを二次元で可視化している．さらに，時事問題の論点に対して賛否両方の立場の意見を QA サイトから抽出することで，賛否両論の立場を網羅的に洗い出すことで，利用者の意思決定を支援するシステムを開発している．本研究では，特定のトピックに対して語っている複数の立場の中から，中立度が高いブロガー，つまり信頼性の高い情報発信者を発見することを目的としている点で井上らとは，立場が異なる．

3　中立度の定義と熟知グループの細分化

本節では，本研究における“中立度”という言葉の定義を行うとともに，中立度の算出方法について説明する．さらに，優良でかつ中立な立場のブロガーを発見する条件について触れるとともに，既存の熟知グループ分類を細分化する手法についても触れる．

3.1　中立度の定義と算出方法

本研究で扱う中立度とは，特定トピックに関して，特定の立場に偏らずに意見を述べている度合いを示す尺度である．中立度の測定方法は，特定トピックを語っている特定の立場とセンチメントの傾向を，各熟知グループのセンチメント履歴から分析し，特定のブロガーのセンチメント履歴と比較することによって決定する．特定のグループのセンチメント履歴の傾向と相関が見られる場合に，中立度が低くなる．中立度は，以下の式によって求められる．

$$Neautral_n = \sum_{k=1}^{m}(1 - Re_k) \tag{1}$$

$$Re = \frac{\sum_{i=1}^{n}(Bi - aveB)(Gi - aveG)}{\sqrt{\sum_{i=1}^{n}(Bi - aveB)}\sqrt{\sum_{i=1}^{n}(Gi - aveG)}} \tag{2}$$

m は，対象トピックを語っている熟知グループの総数を表現している．n は，多数存在するトピックの内の 1 つを表現している．Re_k は，対象トピック k を語っている熟知グループのセンチメント履歴と中立度を測定する対象ブロガーが書いた対象トピック k に関する記事におけるセンチメント履歴の相関値を示している．相関値は，最小値が-1，最大値が 1 とするため，$\sum_{k=1}^{m}(1 - Re_k)$ の値がきくなるのは，熟知グループのセンチメント履歴と対象ブロガーのセンチメント履歴との相関が低くなる場合である．

3.2　優良で中立なブロガー発見のための条件

本研究の目的は，優良な知識を有するだけでなく，特定トピックに対する中立度の高いブロガーの発見である．このようなブロガーを発見するための条件は，熟知度と中立度という二つの尺度が高いことである．優良でかつ中立なブロガーは，以下の式によって表現される．

$$KN_n = Knowledge_n \cdot Neautral_n \tag{3}$$

KN_n は，ブロガーが n というトピックに対して，どれだけ優良な意見を持ち，かつ中立な立場のブロガーなのかを数値で表現するための式である．$Knowledgeable_n$ と $Neautral_n$ は，それぞれあるトピック n に対する熟知度と中立度を表現しており，熟知度と中立度が高ければ，その分だけブロガーが優良でかつ中立な立場である度合いが高くなる．

3.2.1　熟知度の定義

本研究で扱う熟知度とは，特定のトピックに対して精通している度合いを示す尺度である．中島らによる研究 [1] を基に，特定トピックに関するブロガーの熟知度を測定する．この既存の熟知度判定手法において，特定トピックに関連する共起語をブロガーがブログ上で用いられる頻度が高い場合に，熟知度が高くなる．

この尺度が中立度を決定する要素だとする理由は，特定のトピックに関して広い知識を持っているブロガーは，そうでないブロガーと比べて中立度が高いと予想されるからである．例えば，野球というトピックに関して考えてみると，1 つの球団について深く知っているブロガーが，その年のリーグ優勝チームを予想するのと，全球団について深く知っているブロガーが，その年のリーグ優勝チームを予想するのとでは，後者の方が，中立に判断しているのだと言える．なぜなら，後者は，野球というトピックに対して，前者と比べて広い知識を有しているので，そのトピック全体を知った上で，意見を述べているからである．

3.3　熟知グループ細分化

本研究では，既存の熟知グループの細分化について検討を行う．これは，熟知グループの細分化によって，既存の熟知グループでは見られないグループの傾向を詳細に掴むためである．例えば，従来の野球グループの傾向では，野球グループ内に存在する各球団のファンの傾向は掴めないが，野球 (阪神)，野球 (巨人) のようにグループを細分化することで，野球グループの中に存在する各球団のファンを分類することが出来るため，それぞれの立場の傾向の違いを掴むことが可能になる．

図 1 は，1 つの熟知グループ内に複数の立場が存在することを示した図である．このような，熟知グループの細分化を行うために，本研究では，トピック依存度という尺度を用いる．これらの熟知グループの細分化の型，およびトピック依存度の求め方に関しては，次小節にて詳しく触れる．

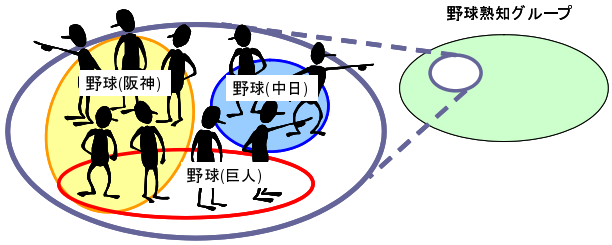

図 1　熟知グループ内に存在する複数立場の例

3.3.1 グループ細分化の型

本研究では，既存の熟知グループをさらに細分化することを提案する．その表現方法は以下のようになっている．

$$Topic(opinion) \quad (4)$$

Topic には，既存の熟知グループで用いられているラベル名が入る．そして，*opinion* 部分には，ブロガーのトピック依存度に基づいて既存の熟知グループのラベル名が入る．例えば，*Topic* に”野球”というラベルが入る場合，*opinion* には，“巨人”“阪神”“中日”などのラベルが入ることになる．現状では，*Topic* 部分に熟知グループのラベル名のみが入るが，今後，“セリーグ＋優勝”などのように複数のトピック名を含めることを検討する．

3.3.2 トピック依存度

本研究におけるトピック依存度とは，特定トピック内のサブトピックに対してブロガーがどれだけ依存しているのかということを示した尺度である．例えば，あるブロガーが野球というトピックにおいて，特に阪神というサブトピックに関するブログ記事を書いている場合，そのブロガーは，野球というトピックの中の阪神というサブトピックに対して依存度が高くなる．本研究では，このようなトピック依存度を以下の式によって求める．

$$depend(\beta) = \frac{NumWords(\beta)}{AllNumWords(\alpha \cdot \beta)} \quad (5)$$

α と β は，熟知グループのラベルを表現しており，同時にブログでのトピックを表現している．例えば，α のトピックを持った記事というのは，α の熟知度を測る際に用いる共起語辞書に含まれる共起語を多く持つ記事を指している．*NumWords*(β) は，トピック依存度を求めたいブロガーが書いたブログ記事の中で α のトピックを持つブログ記事の中に含まれる β の共起語の総数を示している．*allNumWords*(α・β) は，α という熟知グループの共起語辞書内に含まれる β の共起語の総数を示している．*depend*(β) の値は，トピック α を持つ記事の中で，トピック β の共起語を多く使用している場合に，高くなる．

例えば，野球というトピックを α とし，阪神というトピックを β とした場合，解析対象となるブロガーが，野球のトピックだと判定されたブログ記事内において，阪神に関する内容を多く書いていた場合，このブロガーは，野球の中でも，特に阪神のトピックに依存してブログを書いているということがわかる．この場合，ブロガーが属する熟知グループは，野球 (阪神) になる．

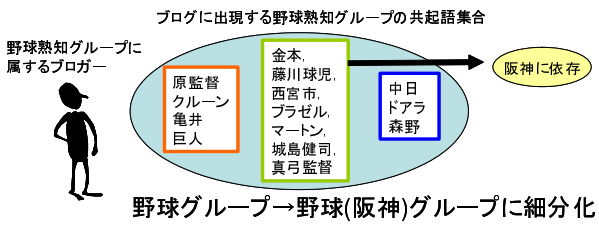

図 2　トピック依存度に基づく熟知グループ細分化の例

4　ブロガーの中立度判定手法

本節では，ブロガーの中立度判定手法について述べる．さらに，ブロガーの中立度判定に基づいた，熟知グループの中立度の高さの推定についても触れるとともに，ブロガーの立度判定を利用したアプリケーション例を提示する．

4.1　手法の概要

本節では，ブロガーの中立度判定手法の概要について説明する．ブロガーの中立度判定手法の手順は，図 3 のようになっている．

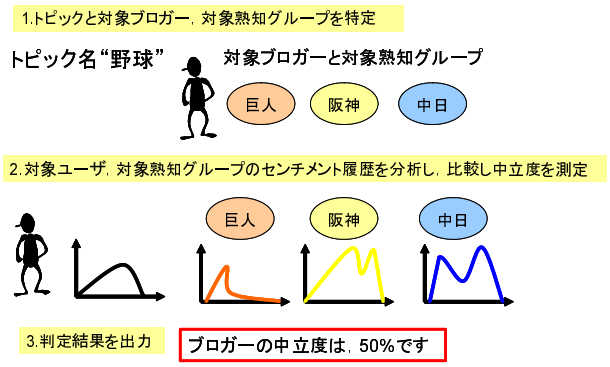

図 3　ブロガーの中立度判定の流れ

手順は大きく分けて 3 つある．1 つめは，対象となるトピック，ブロガーおよび熟知グループを特定するという段階．2 つめは，対象となるブロガーと熟知グループのセンチメント分析を行い，ブロガーの中立度を判定する段階．3 つめは，ブロガーの中立度を実際に出力するという段階である．これらの詳細については次節で述べるが，以上のようにして，ブロガーの中立度判定を行う．

4.2　ブロガーの中立度判定手法

ブロガーの中立度判定手法の詳細について述べる．ブロガーの中立度判定手法は 2 つの段階に分かれている．1 つめは，対象となるトピック，ブロガー，熟知グループの決定である．2 つめは，センチメント分析に基づくブロガーの中立度判定である．

4.2.1　対象トピック，対象ブロガー・熟知グループの決定

まず，中立度判定を行う対象ユーザと対象トピックを指定する．この時対象ユーザは，対象トピックに対して熟知度が高いことが条件となる．なぜなら，中立度が高いユーザの条件は，熟知度が高く，かつ冷静度が高いことであるからだ．

次に，対象トピックに関連する熟知グループを抽出する．この時の関連する熟知グループとは，対象トピックに対して多く語っている熟知グループのことである．以下，これらの熟知グループの集合を対象グループ群と呼ぶ．このように，対象トピック，対象ブロガー，対象グループ群を抽出した後，対象ブロガー，対象グループ群のセンチメント分析を行い，ブロガーの中立度を求める．

4.2.2　センチメント分析に基づくブロガーの中立度判定

センチメント分析に基づくブロガーの中立度判定は，対象グループ群に含まれる各グループの中で，対象トピックについて語っているブログデータを過去 1 か月分収集

し，それらのデータを用いて行う．この時，中島らの研究 [1] にて提案されている，20 個のセンチメント軸を用いる．センチメント分析は，3.1 にて提案した中立度を測定する式 (1) と式 (2) に基づいて行われる．この時の分析対象は，対象グループ群に含まれる各熟知グループと，対象ブロガーである．対象ブロガーに関しても，同様に過去 1 ヶ月分のブログデータを用いて，センチメント分析を行う．

対象グループ群に含まれる各熟知グループ及び対象ブロガーのセンチメント分析を行った後に，対象ブロガーと対象グループ群に含まれる各熟知グループのセンチメントの変化傾向の相関値を求める．相関が見られる場合，対象ブロガーは，特定の熟知グループの立場に類似しているため，特定の熟知グループに立場が偏っていると想定されるので，中立度の値は低くなる．

4.3 ブロガーの中立度判定に基づく熟知グループの中立度推定

ブロガーの中立度判定に基づき，中立度の高いブロガーが多く属する熟知グループの中立度の高さを推定することが出来る．ブロガーの中立度判定に基づき，熟知グループの中立度の判定を行う．

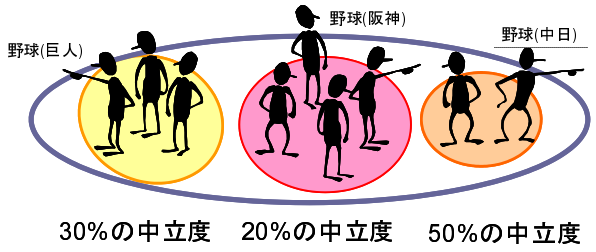

図 4　同一トピックにおける熟知グループの中立度推定の例

図 4 は，同一トピックにおける熟知グループの中立度判定の例を示したものである．この場合，野球というトピックの中に，巨人，阪神，中日という 3 つの立場が存在しており，それぞれのグループ内に，中立度の高いブロガーがどれだけの割合存在するのかによって各グループの中立度を推定する．この例だと，中日グループの中立度が高いので，野球というトピックに関して中立な意見を聞きたい場合に，中日ファンの意見を参考にする方が，他のグループに意見を求めるより中立な意見が得られる可能性は高くなる．

4.4 想定するアプリケーション例

本研究で想定するアプリケーション例は，2 つある．1 つめは，ブロガーの中立度をユーザに示すアプリケーションである．例えば，ダイエットに関するブログ記事をユーザが閲覧している状況で，そのブログ記事を書いたブロガーがどのくらいの中立度でそのブログ記事を書いているのかを提示することで，ユーザは，ブロガーの意見がどれくらい偏ったものなのか，それとも中立な意見なのかを判断することが可能になる．

2 つめは，同一トピックについて語る複数グループの中立度を提示するアプリケーションである．例えば，野球の優勝を争う複数グループ同士の中立度を提示することで，優勝に関係するグループほど，熱狂的である様子を捉えることが可能となる．これは，4.3 で述べた例をそのままアプリケーションとして活用することを想定している．

5 おわりに

本稿では，熟知度とセンチメント分析に基づいたブロガー中立度判定手法の検討を行った．今後は，熟知グループの細分化手法に関して，*Topic* 部分に，熟知グループのラベル以外の名詞を自動で組み込む手法を検討するとともに，同一トピックについて語っている熟知グループのライバル・兄弟関係の推定を自動で行うことを検討していく．さらに，今後評価実験を行い，中立度の高いブロガーを発見することがどういう場面において有効であるのかを検証していく．

謝辞

この研究は、独立行政法人情報通信研究機構の高度通信・放送研究開発委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発 課題ア Web コンテンツ分析技術」および文部科学省科学研究費補助金若手研究（B）（課題番号：20700089）の一環としてなされたものである．ここに記して謝意を表します．

参考文献

[1] Shinsuke NAKAJIMA, Jianwei Zhang, Yoichi INAGAKI, Tomoaki KUSANO and Reyn Nakamoto. Blog Ranking Based on Blogger's Knowledge Level for Providing Credible Information, Proc. of the 10th International Conference on Web Information Systems Engineering. WISE2009, pp.227-234:

[2] 中谷誠，Adam Jatowt，大島裕明，田中克己. Wikipedia のリンク構造とカテゴリ構造を用いた検索語からの専門語の抽出,

[3] 高橋良平，小山聡，大島裕明，田中克己. web テキストと修飾表現との適合度判定手法，第 2 回データ工学と情報マネジメントに関するフォーラム C3-3 2010 年 3 月;

[4] 竹原幹人，中島伸介，角谷和俊，田中克己. Web 情報検索のための Blog 情報に基づくトラスト値の算出方式

[5] 濱砂佳貴，河合由起子，熊本忠彦，田中克己. センチメントマップによる複数ニュースサイトの差異情報可視化手法の提案，DEWS2008 B6-4;

[6] 俵本一輝，川本淳平，浅野泰仁，吉川正俊. 感情解析のための分布モデルと相互強化型解析手法，第 2 回データ工学と情報マネジメントに関するフォーラム C4-1 2010 年 3 月;

[7] 井上結衣，藤井敦. 意見マイニングを志向した QA サイト投稿テキストの解析, 第 2 回データ工学と情報マネジメントに関するフォーラム A8-4 2010 年 3 月;