他者のコメントの引用を考慮したオピニオンマイニング

岡山　有希　　白井　清昭

北陸先端科学技術大学院大学 情報科学研究科

{okayama, kshirai}@jaist.ac.jp

1 はじめに

近年，ウェブを対象としたオピニオンマイニングの重要性が増している．特に，あるトピックに対する賛成意見や反対意見を集約し，それらを整理した上でユーザに提示するシステムは，トピックに関する多くの意見を俯瞰的に知ることができるために有用である．例えば，Opinion Reader は，与えられたトピックに対する主観情報を集約し，賛成・反対に分類した上で，それらを可視化するシステムである [1]．Opinion Reader は，トピックの「論点」を表わすキーワードを抽出し，それらを固有度と重要度に応じて 2 次元空間上に配置することで主観情報の可視化を行う．Shibuki et al. は，与えられたトピックに対する調停要約 (Mediatory Summary) を生成する手法を提案している [6]．調停要約とは，トピックに対する肯定的意見と否定的意見を対比させた要約であり，トピックを表わす文とその否定文をクエリとしたパッセージ検索に基づいて生成される．水野らは言論マップを生成する手法を提案している [5]．彼らは，文間の意味的関係を分類する文間関係認識技術に基づき，検索された文がトピックに対する賛成・反対意見であるのかを分類するだけでなく，賛成・反対の根拠を含むかを認識し，それらを俯瞰的に示している．

本研究では，あるトピックに対して賛成もしくは反対を表明しているウェブページを検索し，それらのページ数とともに，賛成意見や反対意見を提示することを目指す．この際，他者の記事やコメントを引用したウェブページの扱いに特に焦点を当てる．

ウェブ上で社会問題に対して自分の意見を述べる際，他者の意見が引用される場合があるが，他者のコメントの引用は，ウェブページの著者が賛成・反対の立場を取るのかを判断する際に問題となりうる．例えば，「A さんは『○○に賛成している』と述べているが，私は反対だ.」といった記述があるとき，引用箇所 (『○○に賛成している』) の表現からそのウェブページが賛成の立場を取ると誤って判定してしまう可能性がある．また，単に他者のコメントを引用するだけで，著者が自身の立場を明確に記述しないウェブページも存在する．この場合，著者は他者のコメントを引用することで賛成・反対を間接的に表明している場合もあるが，単にニュースを紹介しているブログ記事など，著者自身は賛成・反対の立場を明示していない場合もある．したがって，著者が明確に自分の見解を述べているページと，他者のコメントを引用して間接的に賛成・反対を表明している (かもしれない) ページを同等に取り扱ってよいかには疑問が残る．ウェブ上のテキストは容易にコピーすることができるため，他者のコメントの引用の取り扱いはオピニオンマイニングにとって重要である．

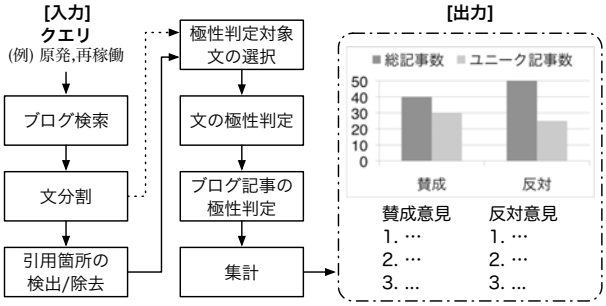

図 1: システム概要

本論文は，ブログ記事を対象とし，あるトピックに対して賛成または反対の意見を表明している記事数をカウントし，ユーザに提示するシステムの構築を目的とする．その際，他者のコメントの引用箇所を推定し，引用箇所を賛成・反対の判定に利用しないことで，著者が明確に自分の見解を述べているブログ記事の数を示す点に特徴がある．

2 提案手法

2.1 システム概要

提案システムの概要を図 1 に示す．本システムの入力は，調査対象となるトピックを表わすクエリ (「原発 再稼働」など) とする．

「ブログ検索」では，検索エンジンを利用してクエリのキーワードを含むブログ記事を検索する．

「文分割」では，いくつかのヒューリスティクを用いてブログ記事内から広告などを除く本文の範囲を検

RT，転載，転載開始，転載終わり，引用，掲載，NAVER まとめ，社説，毎日新聞，日本経済新聞，読売新聞，朝日新聞，New York Times

図 2: 引用を示唆するキーワード (抜粋)

出し，さらにその範囲内のテキストを HTML タグや句読点などで文単位に分割する．

「引用箇所の検出/除去」では，ブログ記事に出現する個々の文について，それが他のウェブページから引用されたものかを判定する．また，引用と判断された文をブログ記事から除去する．この処理の詳細については 2.2 項で述べる．

「極性判定対象文の選択」では，クエリのトピックに対して賛成もしくは反対を表明していると思われる文を選別する．詳細については 2.3 項で述べる．

「文の極性判定」では，前のモジュールで選択された個々の文が賛成もしくは反対を表明しているのかを判定する．その結果を基に,「ブログ記事の極性判定」ではブログ記事全体がトピックに対して賛成もしくは反対の立場を取るのかを判定する．それぞれの処理の詳細については 2.4，2.5 項で述べる．

最後に,「集計」では賛成・反対と判定されたブログ記事数をカウントし，グラフ形式で表示するとともに，賛成意見文と反対意見文をユーザに提示する．この際，ブログ記事の全文を極性判定に用いた場合 (図 1 の点線のように引用箇所の除去を行わない場合) を【総記事数】，引用箇所以外の文のみを対象に極性判定を行った場合を【ユニーク記事数】とした 2 通りの集計結果を示す．後者はブログの著者が明示的に賛成・反対を表明している記事数を指す．

以下，各モジュールの詳細について述べる．以降の説明では，クエリを $Q = \{\cdots, q_i, \cdots\}$,「文分割」によって得られたブログ記事における文の集合を $S = \{\cdots, s_i, \cdots\}$ と記す．

2.2 引用箇所の検出

このモジュールでは，文 s_i が他者のコメントの引用に当たるかを判定し，引用に該当する文の集合 $C(\subset S)$ を求める．C の決定は,「引用ブロックの検出」と「文単位の引用判定」の 2 つの手続きに基づく．

引用ブロックの検出

1. HTML ページの DOM(Document Object Model) において，引用を示唆するキーワード k_{cite} を含むノードを検出する．本研究では，308 個のキーワードを k_{cite} としてあらかじめ人手で用意した．その一部を図 2 に示す．新聞記事からの引用を検出するため，主要な新聞の名称は全て k_{cite} としている．次に，k_{cite} を含むノードおよびそれと隣接する兄弟ノードの支配下にある文は引用であるとみなし，C に加える．ただし，k_{cite} を含むノード下のテキスト長が 50 文字以下のとき，ならびに k_{cite} を含むノードの親ノードがブログ記事全体をカバーする場合は，引用箇所として検出しない．
2. 罫線で囲まれたブロックを検出する．ここでは罫線を ⟨hr⟩ タグ及び同じ文字の 4 回以上の連続 (「＝＝＝＝」「＊＊＊＊」など) と定義する．罫線で囲まれたブロックの中に k_{cite} が含まれる場合，そのブロック内の文を C に加える．
3. ⟨blockquote⟩ タグは無条件に引用箇所と判定し，そのタグ内の文を C に加える．

文単位の引用判定

上記の方法では引用を示唆するキーワードを引用箇所検出の手がかりとしているが,「転載」や新聞名のようなキーワードなしに他者のコメントが引用される場合もある．このようなとき，引用元のウェブページにリンクが張られていることが多い．そこで，ブログ記事内の文とほぼ同じ文がリンク先のウェブページに存在するとき，その文は他者のコメントの引用であるとみなす．具体的な手続きを以下に示す．

1. ブログ記事からのリンク先ウェブページを取得する．ただし，ページ内リンクや同一ブログへのリンクは除外し，外部サイトのリンク先ページのみ取得する．
2. リンク先ページのテキストを文単位に分割する．
3. クエリ q_i を含む 7 文字以上の文について，それと類似した文がリンク先ページに存在するかを調べる．形態素を単位とした文間の編集距離を求め，編集距離が元の文長の 30%未満のとき，2 つの文は類似しているとみなす．リンク先ページに類似文が見つかった場合，その文を C に加える．

2.3 極性判定対象文の選択

ブログ記事から引用箇所を除いた残りの文集合から，極性判定対象文の集合 $S_t(\subset S \setminus C)$ を求める．ここでは，クエリの近くにある文はトピックに対して賛成・反対を表明している可能性が高いという考えに基づき，クエリを含む文およびその近傍の文を極性判定対象文として選択する．ただし，複数のクエリが互いに離れ

た位置に存在するときは，入力トピックと関係のない文である可能性もあるため，Q における複数のキーワードが互いに近い位置に存在するという条件も加える．具体的な手続きは以下の通りである．

1. Q 中の全ての q_i が出現する文の範囲 $[i, j]$ を求める．ただし，$[i, j]$ の範囲において，q_i が出現する文は互いに距離 2 以内に位置するものとする．
2. その前後 2 文，すなわち $[i-2, j+2]$ の範囲にある文を S_t の要素とする．

例えば，$Q =$ { 原発, 再稼働 } としたとき，図 3 の例では，文 s_{23} に「原発」，s_{25} に「再稼働」が含まれるため，クエリにおける全てのキーワードを含む文の範囲は $[23, 25]$ となり，その前後 2 文までの範囲の文 $s_{21} \sim s_{27}$ が S_t の要素となる．

s_{20} **s_{21}** **s_{22}** **s_{23}** **s_{24}** **s_{25}** **s_{26}** **s_{27}** s_{28}

∪ ∪

原発 (q_1) 再稼働 (q_2)

図 3: 極性判定対象文の選択例

2.4 文の極性判定

S_t 中の文 s_i に対して極性スコア $Score(s_i)$ を計算する．$Score(s_i)$ は文 s_i の極性を表わし，肯定的なときに $+1$，否定的なときに -1 の値をとる．その定義を式 (1) に示す．

$$Score(s_i) = \begin{cases} \prod_{w \in V} polar(w) \times neg(w) & \text{if } V \neq \phi \\ \prod_{w \in N} polar(w) & \text{if } N \neq \phi \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

式 (1) における V は，文 s_i 中の単語のうち，日本語評価極性辞書 (用言編) [4] に含まれる用言の集合であり，$polar(w)$ は同辞書における極性 ($+1$ or -1) を表わす．また，$neg(w)$ は否定語のスコアであり，w と同一文節中に否定語が出現すれば -1，それ以外は $+1$ とする．一方，N は日本語評価極性辞書 (名詞編) [2] に含まれる名詞の集合である．すなわち，ここではまず用言の評価語を，次に名詞の評価語を手がかりに文の極性を判定している．

2.5 ブログ記事の極性判定

ブログ記事を A とし，その極性スコア $Score(A)$ を S_t 中の文の極性スコアの重み付き和と定義する (式 (2))．

$$Score(A) = \sum_{s_i \in S_t} Weight(s_i) \times Score(s_i) \quad (2)$$

重み $Weight(s_i)$ は以下のように決定する．

1. 文 s_i において，検索クエリ q_i と評価語の間に係り受け関係があるとき，重みを 5 とする[1]．
2. 文 s_i に含まれる検索クエリの数が 3, 2, 1 個のとき，重みを 3, 2, 1.5 とする．

$Score(A)$ が 1 以上のとき，ブログ記事 A はトピックに対して「賛成」を，-1 以下のときには「反対」を表明していると判定し，それ以外は「中立」と判定する．

3 評価実験

本節では提案手法の評価実験について述べる．表 1 は，評価に用いたクエリの一覧である．これらのうち，Q_1 と Q_2 は提案手法の設計・開発の際に使用したクエリであり，評価データとしては使用しない．残りの 6 つのクエリに対し，Yahoo! ブログ検索 [2] によってそれぞれ上位 50 件のブログ記事を取得し，2 節で述べた手法で各ブログ記事の極性を判定した．ただし，Q_6, Q_7, Q_8 における「(賛成 or 反対)」はトピックに対する意見文を効率良く検索するために追加したキーワードである．「賛成」や「反対」という語を含まない意見文があることも考慮し，極性判定対象文を選択 (2.3 項) する際には，「賛成」「反対」を除くキーワード群の近傍にある文を選択した．

表 1: クエリ一覧

Q_1: 原発 再稼働	Q_2: 消費税 増税
Q_3: 赤ちゃんポスト 批判	Q_4: オリンピック 東京 開催
Q_5: TPP メリット	
Q_6: 英語教育 小学校 (賛成 or 反対)	
Q_7: 胃瘻 (賛成 or 反対)	Q_8: 女性専用車 (賛成 or 反対)

3.1 引用箇所検出の評価

2.2 項で述べた引用箇所検出手法の評価結果を表 2 に示す．同表は，ブログ記事内の各文に対する引用か否かの判定の精度 (P)，再現率 (R)，F 値 (F) である．引用箇所検出の精度は高いが，再現率は平均で 70%とやや低い．これは，引用を示唆するキーワードが存在しなかったり，引用とそれ以外の文の境界が曖昧なとき，引用箇所を正しく検出できなかったケースが多かったためである．なお，全 6 クエリに対して検索されたブロク記事の文の総数は 44,106 であったが，そのうち約半分の 22,387 文は引用文であり，ブログ記事においては他者のコメントの引用が多いことがわかった．

[1] 文節の係り受け解析は CaboCha を用いた．http://code.google.com/p/cabocha/

[2] http://search.yahoo.co.jp/blog/

表 2: 引用箇所検出の評価

	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	ALL
P	.950	1.00	.957	.899	.885	.991	.942
R	.663	.623	.705	.822	.677	.726	.711
F	.781	.767	.812	.859	.767	.838	.810

3.2 極性判定の評価

2.4 項で述べた文の極性判定手法を評価した．結果を表 3 に示す．ここでは正解率，すなわち文に対する「賛成」「反対」「中立」の判定結果が正解ラベルと一致している割合を評価基準とした．表 3 における A_1 は，文がクエリのトピックに対して賛成または反対を表明し，それがシステムによる判定結果と一致したときを正解とみなしている．一方，A_2 は，文がトピックとは異なる対象に対して肯定的または否定的見解を述べているときでも，システムの判定結果と一致していれば正解としている．

表 3: 文の極性判定の評価

	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	ALL
A_1	0.16	0.11	0.16	0.17	0.20	0.18	0.180
A_2	0.22	0.37	0.24	0.40	0.48	0.46	0.426

本論文では，文中における用言および名詞の評価表現のみを手がかりとした比較的単純な手法で極性判定を行っているため，その正解率は十分に高いとは言えない．特に，名詞の評価表現が文中に出現したために，本来は「中立」と判定するべき文を「賛成」または「反対」と誤って判定した場合が多かった．評価表現分析に関する最近の研究成果 [3] を応用するなどして，極性判定の正解率を向上させる必要がある．

次に，2.5 項で述べたブログ記事の極性判定手法を評価した．ここでは正解率，すなわちブログ記事に対する「賛成」「反対」「中立」の判定結果が正解ラベルと一致している割合を評価基準とした．結果を表 4 に示す．同表における S_{all} はブログ記事の全文を用いて極性判定をする手法，S_c は引用箇所を検出し，引用と判定された文を極性判定に使わない手法，S_{c_g} は S_c と同じく引用文を極性判定に使わないが，引用箇所の検出は人手で行った手法を表わす．

文単位の極性判定の正解率が悪いため，ブログ記事の極性判定の正解率もそれほど高くはない．また，提案手法 (S_c) における「賛成」「反対」「中立」のそれぞれの F 値は 0.18，0.43，0.74 であり，「中立」の記事を正しく判定したケースが多いことがわかった．

表 4: ブログ記事の極性判定の結果

	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	ALL
S_{all}	0.68	0.54	0.68	0.60	0.40	0.42	0.553
S_c	0.70	0.52	0.72	0.60	0.50	0.48	0.587
S_{c_g}	0.70	0.54	0.74	0.60	0.54	0.52	0.607

S_{all} と比べて，S_c と S_{c_g} の正解率は向上している．この結果から，引用箇所を検出し，それをブログ記事の極性判定に用いない提案手法のアプローチは適切であるといえる．正解率が改善した主な要因は，S_{all} では「中立」の記事を誤って「賛成」または「反対」と判定していたのに対し，S_c および S_{c_g} では正しく「中立」と判定されたブログ記事が多かったためである．すなわち，提案手法は，他者のコメントの引用を自動的に検出し除外することにより，他者のコメントに含まれる評価表現による誤判定をある程度抑制している．

4 おわりに

与えられたトピックに対し，賛成や反対の立場を表明しているブログ記事の数を集計するオピニオンマイニング・システムにおいて，他者のコメントの引用を自動検出し，それらを極性判定に用いない手法を提案した．評価実験の結果，他者のコメントの引用文を極性判定に使用しないことでブログ記事の極性判定の正解率が向上することを確認した．今後の課題として，文およびブログ記事の極性判定の正解率の向上，提案システムの実用的な評価などが挙げられる．

参考文献

[1] 藤井敦. Opinionreader: 意思決定支援を目的とした主観情報の集約・可視化システム. 電子情報通信学会論文誌 D, Vol. J91-D, No. 2, pp. 459–470, 2008.

[2] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会発表論文集, pp. 584–587, 2008.

[3] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[4] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[5] 水野淳太, 渡邉陽太郎, エリックニコルズ, 村上浩司, 乾健太郎, 松本裕治. 文間関係認識に基づく賛成・反対意見の俯瞰. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3408–3422, 2011.

[6] Hideyuki Shibuki, Takahiro Nagai, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. A method for automatically generating a mediatory summary to verify credibility of information on the Web. In *Proceedings of the COLING*, pp. 1140–1148, 2010.