Web 世論からの意見抽出と賛否に基づく分類

井上 結衣
筑波大学図書館情報専門学群

藤井 敦
筑波大学大学院図書館情報メディア研究科

1 はじめに

World Wide Web には，報道記事のように客観性が高い情報だけではなく，意見，評判，感想などの主観情報も存在する．複数の人間が書いた主観情報から人々の考え方に関する傾向や法則を発見することができれば，個人や組織の意思決定に役立つ場合がある．

例えば，種々の商品に対する批評を読んで，購入する商品を決める場合がある．また，ある時事問題に対する賛否両論が含まれる意見群を読んで，その問題に対する自分の態度を決定する場合がある．これらの例における意思決定は，以下に示す手順に分解することができる．

(1) 対象の話題（商品や時事問題）に関する文書を Web から収集する．
(2) 収集した文書から主観的な記述を抽出する．
(3) 抽出した主観的記述を「肯定/否定」や「賛成/反対」などの観点に応じて分類する．
(4) 主観的記述を集約し，さらに可視化する．
(5) 可視化された内容を吟味して,「肯定/否定」から一方を選択する．対象の話題が商品の場合は，肯定を選んだ場合に，その商品を購入する．

上記の手順を全て人手で行うことは高価であるため，「OpinionReader」[7, 8] という意思決定支援を目的としたシステムがある．意思決定とは，ある話題に対する賛否両論を網羅的に洗い出し，対立させて，合理的な立場を採用する過程である．ある話題について賛否両論が対立する場合は「論点」が存在する．OpinionReader は，賛否両論が対立する構図を論点に基づいて可視化する．

図 1 は,「株式会社による病院経営への参入」という話題に対する出力インタフェースの表示内容である.「情報公開」などの論点を 2 次元グラフ上に表示する．グラフの縦軸は論点の重要度を表し，横軸は論点がどれだけ賛成／反対に固有かを表す．論点を選択すると，該当する論点を含む意見が順位つきリストで表示される．以上の機能により，ユーザは大量の意見情報を読まなくてもその話題に関する議論の全容を把握することができる．

OpinionReader では，上記の手順 (4) だけを実装しており，手順 (1)〜(3) は人手か既存の手法によって完了していることを前提としている．Web には，ある話題について賛成か反対かを明示した上で意見を投稿する意見サイトがある．このようなサイトから収集した意見情報

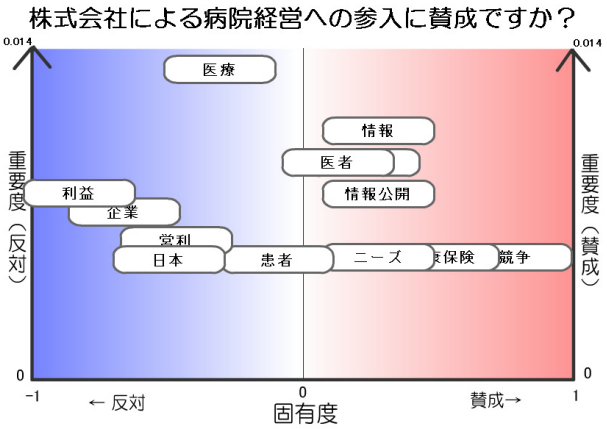

図 1:「株式会社による病院経営への参入」という話題に対する OpinionReader の出力

は，OpinionReader にそのまま入力することが可能である．しかし，特定の意見サイト以外に存在する意見は利用することができないため，対象となる話題や意見の数が制限されてしまうという問題がある．

本研究は，手順 (1)〜(3) の自動化を目的として，ある話題に関する意見を賛成と反対に分けて Web から収集する手法を提案する．

2 本研究の位置付け

1 章で示した手順 (1)〜(3) のそれぞれについて先行研究が存在する．手順 (1) に関して，日記やブログのように主観情報を多く含む文書を選択的に収集する手法 [4] がある．手順 (2) に関して，文書中の主観的な記述を抽出する手法 [2] がある．手順 (3) に関して，主観情報を「肯定」と「否定」のような 2 つのグループに分類する手法 [1, 6] や，多段階に分類する手法 [5] がある．

しかし，手順 (1)〜(3) について総合的に取り組んだ研究事例は少ない．Hu ら [1] は批評の収集から要約までを総合的に行うシステムを提案しているものの，評価実験では特定の Web サイトから選択的に収集した批評を用いている．

主観情報の分類に関する既存の手法は，商品や映画に対する批評を対象としていることが多い．批評の記述には，特定の商品とは無関係に,「満足した」や「不具合」

のような肯定や否定に特有の表現が存在する．他方で，例えば「大きい」という表現が商品によって肯定と否定のどちらでも使用されることがある．しかし，総じて，既存の研究では肯定や否定に関する普遍的な表現を学習することが中心的な課題である．

それに対して，本研究は「赤ちゃんポスト」などの時事問題に対する Web 上の意見，すなわち Web 世論を対象とする．この場合,「賛成」と「反対」という言葉以外には，話題とは無関係にどちらかの立場に特有の表現を見つけることが難しい．このことは，話題の選び方によって賛成と反対が入れ替わることから分かる．例えば,「詰め込み教育」という話題に対する反対意見は,「ゆとり教育」という話題に対する賛成意見になる可能性が高い．

Eguchi ら [3] は，この問題に取り組んだ．しかし，Eguchi らが新聞記事から文単位で意見を検索するのに対して，本研究は雑多な Web から段落単位で意見を検索する点が異なる．また，本研究は，検索モデルに依存しないため，既存の検索エンジンを利用することができる．

3 賛否両意見の収集手法

3.1 概要

ある話題に対する賛成や反対の意見を Web から集めるには，検索エンジンに「話題を表す言葉」と「観点（賛成または反対）」を同時に入力する方法がある．

例えば,「赤ちゃんポスト　賛成」と入力すれば,「赤ちゃんポスト」と「賛成」の両方を含むページが検索される．しかし，この方法では必ずしも対象の話題に対する賛成意見だけが検索されるわけではない．

それに対して,「赤ちゃんポストに賛成です」のように具体的な表現を検索質問とすれば，賛成意見が検索される可能性が高くなる．しかし，賛意を表明する表現は多様であるため，この方法では賛成意見の一部しか検索することができない．また，検索されるページの件数が少ないために多様な意見を収集することができない．

以上を踏まえて，本手法は 2 段階検索に基づく意見収集の手法を提案する．初期検索では，具体的な検索質問を用いて高精度の検索を行う．次に，検索されたページに頻出する言葉を関連語として抽出する．再検索では，関連語を検索質問に追加して網羅性が高い検索を行う．

本研究で提案する意見収集の手法を図 2 に示す．本手法は 2 段階検索を行うため，図 2 は 1 回目の「初期検索」と 2 回目の「再検索」で構成されている．ここまでの処理を賛成と反対で個別に行い，賛成意見と反対意見の候補を収集する．さらに，意見の候補を教師なし手法で賛否に分類した結果が OpinionReader の入力となる．

3.2 初期検索

「赤ちゃんポスト」や「憲法改正」などの話題に関するキーワード X をユーザが与える．次に，Web 上の検索エンジンに「X に賛成です」という検索質問を入力して,「X に賛成です」という表現を含むページを検索する．現在，検索エンジンとして Google[1]を使用している．ただし，以下のような同義表現も用いて検索する．

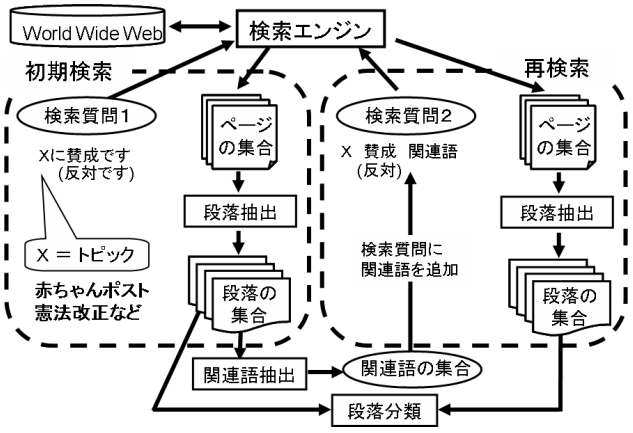

図 2: 意見収集手法の概要: 話題 X に関する賛成意見を収集する場合

X(に | には) 賛成 (です | だ | である | します)

3.3 段落抽出

検索されたページから段落の単位で意見を抽出する．具体的には，検索質問の表現（「X に賛成です」など）を中心とした，100〜300 文字の範囲で，改行で区切られた最も小さな領域を抽出する．抽出する文字数は時事問題に対する Web 上の意見を調査し，決定した．図 3 に，初期検索で検索されたアンケートサイト[2] から実際に段落として抽出した領域を太線の枠で示す．

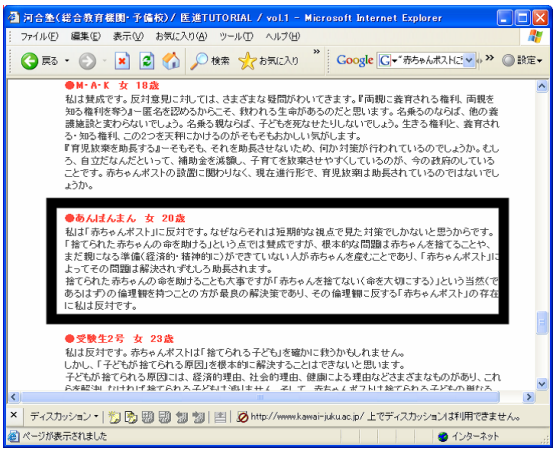

図 3: アンケートサイトからの段落抽出例

しかし，検索されたページには，X に関する賛成意見以外の情報も含まれることがある．まず「X に賛成です」

[1] http://www.google.co.jp/
[2] http://www.kawai-juku.ac.jp/medical/2007/vol_1_6_01.html

という表現を含むにも拘らず，実際には賛意を表していない表現がある．例えば「X に賛成ですか」,「X に賛成ですが」,「X に賛成ですなんて」などがある．このような表現を含む段落は抽出の対象から除外する．

上記の一覧に該当しない場合でも,「X に賛成です」などの表現がアンカーテキストとして使用されて，賛成意見に対するリンクがはられている場合がある．そのような場合は，リンク元のページには賛成意見が書かれていないことが多い．そこで，検索質問と同じ表現が <A> で括られている場合は，その段落を抽出対象から除外する．

3.4 関連語抽出

初期検索で得られた意見の集合から，関連語抽出によって特徴的な言葉を抽出する．OpinionReader は，意見テキストに対する形態素解析と係り受け解析の結果から，規則に基づいて名詞句と動詞句を抽出し，論点として使用する．本研究では，この機能を用いて名詞句と動詞句を関連語として抽出する．

次に，賛成意見用の初期検索で得られた段落の集合を $\mathbf{D}_{pro}$ とし，反対意見用の初期検索で得られた段落の集合を $\mathbf{D}_{con}$ としたとき，適合情報に出現する割合が高い論点を関連語として抽出する．具体的には，式 (1) を用いて論点 A のスコアを計算し，スコアが 0.6 以上の場合に，論点 A を関連語として抽出する．

$$\frac{\mathbf{D}_{pro}\text{における論点 } A \text{ の出現頻度}}{\mathbf{D}_{pro}\text{と } \mathbf{D}_{con}\text{における論点 } A \text{ の総出現頻度}} \quad (1)$$

このスコアは 0 以上 1 以下の値をとる．反対意見から関連語を抽出する場合は，式 (1) の $\mathbf{D}_{pro}$ と $\mathbf{D}_{con}$ を入れ替えて同様の処理を行う．

3.5 再検索

再検索では，関連語抽出によって抽出された関連語の集合を「X　賛成（あるいは反対）」の後ろに追加して検索質問を構成する．関連語を検索質問に追加することで，不要なページをなるべく検索しないようにする．段落抽出では，検索質問に使用した言葉のうち 3 語以上を含む領域を抽出する．領域の判定基準は初期検索と同じである．

3.6 段落分類

段落分類では，初期検索と再検索で収集した段落集合を教師なし手法で賛否に分類する．具体的には，まず，初期検索（3.2 節）で得られた精度の高い段落集合を教師事例とし，サポートベクターマシン（SVM）を用いて分類器を学習する．システムが賛成として収集した段落集合を正例とし，反対として収集した段落集合を負例とする．学習した分類器を用いて，初期検索と再検索で収集した全ての段落を再分類する．また，SVM のスコアに対する閾値を設定することによって，スコアが賛成と反対の中間に近い値をとる段落を分類から除外し，純度を高めることができる．

4 評価実験

4.1 段落抽出精度の評価

段落抽出を機械的に行い，その精度を評価した．トピックは「赤ちゃんポスト」で実験を行った．賛成と反対について初期検索と再検索の結果から上位 5 ページずつ合計 20 ページを対象とした．評価では人手で抽出した段落とどの程度一致するかを調べ，正解と見なす範囲を変えながら精度を評価した．

表 1 に，正解とする範囲ごとに精度を示す.（a）は，人手で抽出した段落と自動抽出した段落が一字一句一致した段落である.（b）は，賛成意見の中に一文程度の関係ない文が含まれているような場合である.（c）は，本来抽出すべき段落の 1/4 以下が不足している場合である．誤って抽出された段落では，半分以上が関係のない文であった．また，立場が逆の意見を含んでいる場合があった．

表 1: 段落抽出精度の評価

正解とする範囲	精度
(a) 完全一致	12.0%（17/142）
(b) 賛成／反対意見に少量の関係ない文が含まれている	36.6%（52/142）
(c) 取るべき段落の不足が 1/4 以下である	47.2%（67/142）
(d) (b) かつ (c) である	50.0%（71/142）

4.2 意見収集精度

本研究で提案した意見収集の手法を実験によって評価した．評価用の話題として「赤ちゃんポスト」と「憲法改正」を用いた．評価では，初期検索と 2 段階検索，さらに 2 段階検索で収集した段落を SVM で分類した結果について，収集された正しい意見の件数（意見件数）と収集精度を比較した．初期検索では，検索された上位のページから段落 50 件を収集した．これは，統計頻度に基づいて関連語を抽出するために必要な段落の件数を経験的に決めた結果である．初期検索によって収集された段落に対して関連語抽出を行い，抽出した関連語を検索質問に追加して再検索を行った．例として，実際に抽出された「赤ちゃんポスト」の関連語を以下に示す．

- 賛成： 愛情，未来，人生，ニュース，虐待，事件，幸せ
- 反対： 反対，無責任，安易，施設，相談，病院，子ども，責任，大人

再検索では，検索された上位 30 ページから段落を収集した．2 段階検索では，初期検索と再検索で得られた段落の集合を結果とした．また，2 段階検索の結果を SVM の閾値を変えながら分類し，それぞれの結果を評価した．

表 2: 意見収集の実験結果

	赤ちゃんポスト		憲法改正	
手法	意見件数	収集精度	意見件数	収集精度
初期検索	87	87.0% (87/100)	93	93.0% (93/100)
2 段階検索	139	65.3% (139/213)	123	64.4% (123/191)
+SVM(0)	140	65.7% (140/213)	124	64.9% (124/191)
+SVM(0.1)	134	72.0% (134/186)	122	69.3% (122/176)
+SVM(0.2)	128	75.7% (128/169)	121	73.8% (121/164)
+SVM(0.3)	116	77.9% (116/149)	117	77.5% (117/151)

表 2 に結果を示す．表 2 において，「+SVM」の丸括弧内にある数値は，SVM のスコアに対する閾値である．

初期検索では，収集精度について高い値が得られた．2 段階検索では，収集精度が初期検索よりも下がり，意見の件数が増えた．このことより，再検索では網羅性が上がることを実証した．2 段階検索で収集した段落の集合を SVM で分類した結果，閾値を上げるにつれて意見の件数が減り，収集の精度は高くなった．このことから，SVM を用いることにより，意見件数と収集精度をまんべんなく上げることができた．

4.3 論点可視化の実行例

また，2 段階検索で収集した意見を OpinionReader に入力し，その出力結果について考察した．図 4 に出力結果を示す．

図 4 の右上に表示されている「愛情」と「事件」は，システムが賛成に固有と判断した論点である．この論点を含む意見には，「子供には愛情が必要だから，愛情のない実親が育てるより赤ちゃんポストのほうが良い」や「コインロッカーに捨てられるなどの事件が減る」など賛成の意見が多かった．左側に表示されている「匿名」，「相談」，「無責任」では，「匿名で預けることができるという点に反対」，「母親が相談できる体制を整えるほうが先」，「無責任な親を増やすだけ」などの反対意見が多かった．

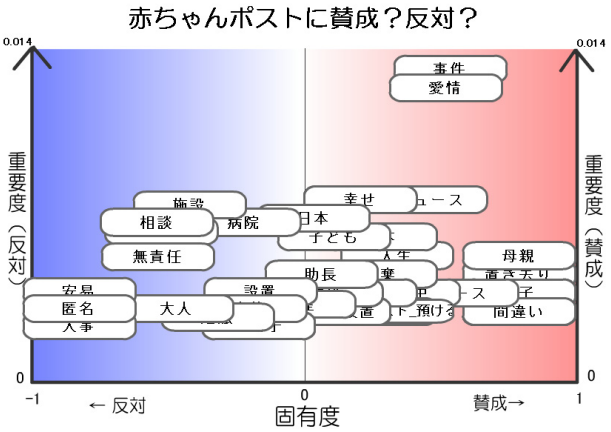

図 4: 2 段階検索で収集した意見を OpinionReader に入力した場合の出力

5 おわりに

筆者らは Web 上の主観情報を可視化することで個人や組織の意思決定を支援するシステムについて研究している．当該システムにおける自動化の度合いを高めるために，時事問題に対する賛成意見と反対意見を Web から選択的に収集する手法を提案した．また，提案手法を実験によって評価した．今後は，対象の時事問題を増やしながら評価を繰り返し，手法のさらなる改善を行う予定である．

参考文献

[1] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004.

[2] Soo-Mim Kim and Eduard Hovy. Determing the sentiment of opinions. *Proceeding of Conference on Computational Linguistics*, pp. 1367–1373, 2004.

[3] Eguchi Koji and Victor Lavrenko. Sentiment retrieval using generative models. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 345–354, 2006.

[4] Tomoyuki Nanno, Tochiaki Fujiki, Ysuhiro Suzuki, and Manabu Okumura. Automatically collecting monitoring and mining japanese weblogs. *In The 13th International World Wide Web Conference*, pp. 320–321, 2004.

[5] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales. *Proceeding of the 43th Annual Meeting of the Association for Computational Linguistic*, pp. 1367–1373, 2005.

[6] Peter. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[7] 佐々木千晴, 藤井敦, 石川鉄也. 意思決定支援のための主観情報マイニング. 言語処理学会第 12 回年次大会発表論文集, pp. 77–80, 2006.

[8] 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, Vol. J91-D, No. 2, pp. 459–470, 2008.