時事問題に対する賛否両意見の自動収集手法

井上　結衣　　藤井　敦

筑波大学図書館情報専門学群　　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 上の文書には，意見や評判などの主観情報が多く含まれる．複数の人間が書いた主観情報から人々の考え方に対する傾向や法則を発見できれば，個人や組織の意思決定に役立つ可能性がある．

筆者らは「OpinionReader」[1]というシステムを提案し，意思決定支援を目的とした主観情報の集約と可視化を行った．意思決定とは，ある話題に対する賛否両論を洗い出し，対立させて，より合理的な立場を採用する過程と捉える．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 OpinionReader の出力である．「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付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．

しかし，OpinionReader には改善の余地がある．現在は，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」の意見を入力として利用しているため，対象となる話題や意見の数が制限されている．本研究は，Web から，対象となる話題に関する意見を賛否に分けて自動的に収集する手法を提案する．

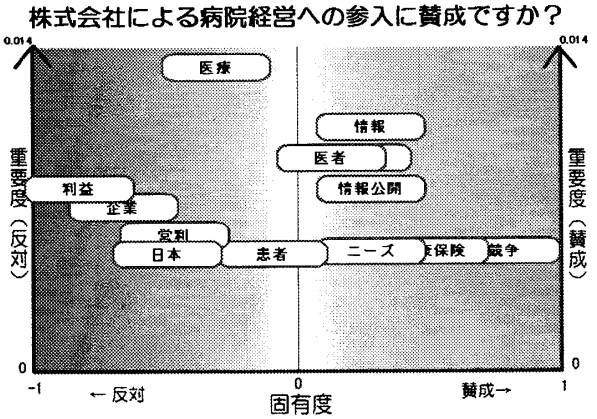

図 1：「株式会社による病院経営への参入」という話題に対する OpinionReader の出力

An Automatic Method of Collecting Pro and Con Opinions for Current Topics. Yui Inoue, Atsushi Fujii. University of Tsukuba.

2. 先行研究

主観情報に関する研究には，文書中の主観的記述を抽出する手法[2]がある．また，主観情報を「肯定」と「否定」のような観点に基づいて分類する手法[3]がある．主観情報の分類に関する既存の手法は，多くが商品や映画に対する批評を対象としており，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話題である．

それに対して，時事問題に対する意見を対象とする場合は，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は難しい．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．

江口[4]は，この問題に取り組んだ．しかし，江口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 から集めるためには，検索エンジンに「話題を表す言葉」と「観点（賛成／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」のような具体的な表現で検索すれば，賛成意見を得ら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．

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．すなわち，情報検索の適合性フィードバックを応用する．以下，賛成意見を収集する場合で説明を進める．しかし，「賛成」を「反対」に置き換えて同様の処理を行い，反対意見も収集する．

3.2 初期検索

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検索する．ただし，以下のような同義表現も用いて検索する．

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 上の意見情報を分析し，決定した．

ただし，検索質問を含んでいても賛意を表明していない表現は抽出から除外する．例えば，「Xに賛成ですか」や「Xに賛成ですなんて」などがある．上記の表現に該当しない場合でも，質問表現がアンカータグでくくられている場合は抽出元に意見はないことが多いため除外する．

3.4 関連語抽出

初期検索で得られた意見集合から，関連語抽出によって特徴的な言葉を抽出する．Opinion Reader 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 Dpro とし，反対意見用の初期検索で得られた段落の集合を Dcon 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．反対意見から関連語を抽出する場合は，Dpro と Dcon を入れ替える．

$$\frac{\text{Dpro における論点 A の出現頻度}}{\text{Dpro と Dcon における論点 A の総出現頻度}} \quad (1)$$

3.5 再検索

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）」に追加して検索する．段落抽出では，検索質問に使用した言葉を 3 語以上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．

4 評価実験

本研究で提案した意見収集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，以下に示す 3 通りの手法を比較した．「X」と「P」はそれぞれ話題と立場（賛成／反対）を表す．

(a) 「X　P」で検索
(b) 初期検索（「XにPです」など）で検索
(c) 初期検索＋再検索（本手法）

「赤ちゃんポスト」で実験した結果，(a)の精度が 32.7%，(b)の精度が 87.0%であった．また，「憲法改正」では(a)の精度が 6.2%，(b)の精度が 93.0%であった．すなわち，初期検索の精度が高いことが分かった．

(c)の精度は「赤ちゃんポスト」では 65.2%，「憲法改正」では 64.4%となり，(b)より下がった．しかし，網羅性を比較するために(b)と(c)のそれぞれで得られた段落集合から抽出した論点数を比較すると，「赤ちゃんポスト」では(b)の 32 件に対し(c)では 56 件に増えた．また「憲法改正」では(b)の 42 件に対し(c)では 51 件に増えた．すなわち，再検索によって論点抽出の網羅性が上がった．

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，J91-D(2)，2008.

[2] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics, pp.1367-1373, 2004.

[3] 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.

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