単語の分散表現を用いた意見文クラスタリングに対する一考察

平野　真理子† 榊　剛史† 小早川　健‡

†株式会社　ホットリンク ‡日本放送協会　放送技術研究所

†m.hirano@hottolink.co.jp †t.sakaki@hottolink.co.jp ‡kobayakawa.t-ko@nhk.or.jp

概要　意見文クラスタリングでは、意見の観点または評価の種類によって分類・まとめ上げをする方法が模索されてきたが、現実はその両方を取り扱って初めて意見全体の俯瞰を行うことが可能になる。また、文書クラスタリングとして最も王道な手法は、文書-単語行列を作成し、ベクトル空間上の類似度によりクラスタリングを行うものであるが、特徴量に単語を用いると、次元数が肥大するとともに、似た意味を持つ表層の異なる単語が異なる次元を形成する問題があった。本稿では意見文クラスタリングの課題と目標を明確にする一方、単語の分散表現を用いて単語に近似的な意味づけを行い、文書-近似意味行列を作成して文書クラスタリングを行った。その上で、人手で書き出した意味要素をどの程度再現できるかの考察を行った。

キーワード　意見文クラスタリング，単語分散表現

1 はじめに

ブログ等で個人発信をする手段が一般化して久しい今日、個人の感想や考えを記した意見文テキストは豊富に存在するようになり、フリーフォーマットであるがゆえに解析が困難ながらも、誘導のない自発的な意見を収集することが容易になった。これらを適切に解析することで、消費者のニーズや評価を理解すれば、効率的な生産が行えるなど、消費者、生産者ともにWin-Winの関係がもたらされることが期待される。

上記の目標のためには、複数種類の意見を、相互のボリューム感を含めて全体の傾向をとらえるという、指標を定めて件数の推移を観察するのとは異なるアプローチが必要である。一方、全体を把握するにしても、全てのテキストを読むことは量的に困難であるため、文書クラスタリングが精度よく行われることが求められている。

書き言葉でも話し言葉でも、あることについて意見を述べる意見文は、基本的には「○○（意見対象）が××（評価表現）」という構造をとる。キーワードを設定した際、意見対象が1つであるならば、評価表現でのみクラスタリングを行うことが可能であるが、少し広めの概念をキーワードに設定した場合は、キーワードの下位レベルで複数の観点・論点が生じ、それぞれに対して評価・評論がなされる。評価に関しては、評判分析などではポジティブ／ネガティブの二値分類が行われるが、実際には、ポジティブの中でも“かわいい”“便利”など、意味合いが異なる要素が複数出現することがあり、それぞれに興味があるため、二値以上に分類したい場合がある。

このように、意見文の集合を俯瞰するための意見文クラスタリングにおいては、分類すべきクラスタの数を事前に決定できないことに加え、分析軸として評価対象と評価表現の2軸が混在するのが大きな特徴である。

本稿では、意見文を評価対象、評価表現双方を考慮しながらクラスタリングする試みに於いて、単語分散表現を用いることの有効性ついて議論する。

2 関連研究

評価表現での意見文クラスタリングに関し、二値以上の分類を行っているものとしては、例えば橋本（2011）ら[1]の研究の中で感情表現による分類として行われているが、評価対象については考慮されていない。また評価対象（観点）に基づいた意見文クラスタリングについては、鷹栖（2013）ら[2]が行っているが、こちらは逆に評価対象のみを考慮したクラスタリングとなっている。

3 提案手法

3.1 実験データ

意見文のなかでも、トピック（意見対象、観点）が多く出現し、意見内容（評価表現）がニュアンスも含めて多様である、討論番組に対する意見・感想について実験を行うこととした。2010年12月25日放送の、「日本の、これから　就職難をぶっとばせ」に対する意見文 1,057文を実験対象のテキストデータとした。このテキストデータは、Web上で男女各年代層600人に、アンケートの一部として自由記述文で回答してもらう形で収集を行い、解析のために次の条件を課した。

1. 1つの回答欄には句点を用いて複数の文を書いてもよいが、1つの主旨しか述べない（複数の主旨がある場合はそれぞれ別の回答欄に記入する）
2. 1つの回答欄に書ける文字数は150字程度とする

本稿では、1つの回答欄に書かれた1つの主旨の意見を、便宜上“1文”ということとする。

3.2 人手による意味要素書き出し

検証を目的とし、1,057 文に対して、人手により書かれている内容の意味要素書き出しを行った。既に述べたように、意見文は評価対象と評価表現の 2 軸が存在し、その組み合わせで意見が述べられるため、また、評価表現も 1 文に複数の要素が含まれる場合もあるため、評価対象、評価表現にかかわらず、出現した要素全てを書き出し、文書全体で出現した全要素を各次元に持つ意味要素ベクトルを作成した[1]。アンケート調査における「MA（複数回答可）」の集計のイメージに近い。結果、意見対象（言及要素）として 12、評価表現（評価内容詳細）として 53 の合計 65 次元のベクトルとなった。この要素は必ずしも単語によるものではなく、あくまで意味合いで書き出し・まとめ上げを行ったもので、これらの要素の組み合わせでクラスが決まると想定される。

3.3 文書単-語行列の作成

文書クラスタリングを機械的に行うには、文書を何らかの処理を経てベクトルで表現することが必要である。文書-単語行列は、文書全体に出現した語彙を次元に持つ、文を、単語の出現位置を考慮しない単語の集合として扱う bag-of-words モデルとして扱ったものである。本稿では生成された文書-単語行列に対して、更に tf-idf により、各単語の文中での重みづけを行っている。

3.4 次元圧縮と単語の分散表現の導入

文書-単語行列はスパースであるため、次元圧縮が必要である。また意味は近いが、表記揺れを含め、表層形は異なる単語が別の次元として独立してしまう[2]という問題を緩和することも視野に入れ、「特異値分解」と「単語分散表現での畳み込み」で効果を比較した。

畳み込みにより生成される i 番目の文の意味要素ベクトル $\boldsymbol{x}_i$ は次式で表すこととする。

$$\boldsymbol{x}_i = \sum_j w_{i,j} \boldsymbol{W}_j$$

$w_{i,j}$ 文書-単語行列の i 行 j 列番目の要素の値（出現頻度に tf-idf の処理を行ったもの）、$\boldsymbol{W}_j$ は単語 j の分散表現である。

3.5 ベクトル要素としての妥当性の検証

文書クラスタリングにおいては、1. ベクトルの作り方 2. ベクトルの距離（類似度）の定義の仕方 3. クラスの組み上げ方、の 3 段階で検討が必要になる。本稿では、まずベクトルに単語分散表現を導入した妥当性を検証するため、クラスサイズを 20 以下に制限した Nearest Neighbor 法様のクラスタリングを行い、各クラスに含まれる文の意味要素ベクトルの平均と、二乗誤差の平均を算出。全てのクラスでその値を加重平均し、クラスタリング結果の意味要素の再現性（クラス内の意味要素ベクトルの分散が小さいほどよい）の指標とした[3]。特異値分解を行ったもの、単語分散表現で畳み込みを行ったもので指標を比較すると、僅かに単語分散表現で畳み込みを行ったものが指標の値を改善した（表 1 参照）。

表 1 行列の生成法別クラスタリング指標

行列の生成方法	指標 I
単語分散表現による畳み込み（200 次元）	9.633
特異値分解による次元圧縮（200 次元）	9.983
<参考>tf-idf のみ（2,878 次元）	11.262

3.6 クラスタリング手法の比較

事前にクラス数を決めないクラスタリング方式として、BIC 情報量を基準に再帰的に k=2 の k-means 法を繰り返す x-means 法（トップダウンクラスタリング）を、人手作成の意味要素ベクトルに対して実施し、人手による意味要素ベクトルの次元に近い 50 クラスを再現する文書-単語行列の列数を決定。階層型クラスタリングの枝刈り（足切）によって生じるクラス数および k-means のクラス数を同様に 50 とし、x-means、階層型クラスタリング（ボトムアップクラスタリング）、k-means（凝集型クラスタリング）の 3 種の方法でクラスタリングを行った後、前節と同様に指標を計算した（図 2 参照）。

表 2 クラスタリングの組み上げ法別クラスタリング指標

クラスタリング法	クラス数	行列の利用次元数	指標 I	
			特異値分解	分散表現
x-means	28（自動）	50	37.259	48.297
階層型	50（固定）	50	20.873	21.714
k-means	50（固定）	50	18.462	20.172

4 おわりに

意見文クラスタリングの困難な点と目指すべき目標、評価方法の例を示した。文書単語行列に単語分散表現を導入することは、文書単語行列の次元圧縮に加え、単語表層の差異の吸収に貢献する可能性が示唆された。組み上げられた各クラスの意味内容の再現性を比較してもその傾向が見られた。

参考文献

[1] 橋本和幸, 中川博之, 田原康之, 大須賀昭彦：センチメント分析とトピック抽出によるマイクロブログからの評判傾向抽出, 電子情報通信学会論文誌-D, J94-D(11), pp. 1762-1772, 2011.

[2] 鷹栖弘明, 小林聡, 内海彰：Twitter における観点に基づいた意見文クラスタリング, 言語処理学会第 19 回年次大会発表論文集, A4-3, 2013.3.

[1] 例えば、例えば、「いろいろな立場の人が出演しそれぞれの立場で発言していたため、問題点をよく理解することが出来た」という感想なら、“出演者構成よい”“わかりやすい”の 2 要素が出現 1 となる

[2] 例えば、テレビ番組に対する評価としては、“勉強になった”と“参考になった”はほぼ同じ意味合いである

[3] クラスタリングは単語由来、評価は人手由来の行列を使用