ARG WI2 No.13, 2018
WI2-2018-17

評価値毎の単語頻度分布に基づく レビューからの多様な意見文の抽出

立命館大学情報理工学部
小平 綾香　山西 良典　西原 陽子
{*is0306ih@ed, ryama@fc, nisihara@fc*}*.ritsumei.ac.jp*

概要　本稿では，レビューの評価値毎に出現頻度が異なる単語を評価視点とし，評価視点についての多様な意見文を抽出するための手法を提案する．レビューを閲覧するとき，評価値が極端に高いまたは低いレビューに含まれる評価視点は，レビュー対象の評価・判断に役立つことが多い．しかしながらレビュアーには多様性があり，多くの人が高く評価している評価視点こそが，特定のコンテキスト下のレビュアーの低評価の要因となっている場合も存在する．このような相反する意見文を比較することで，より自分自身のニーズに合った商品やサービスの情報を得られる可能性がある．そのためには，単純に出現頻度が高い評価視点に着目するのみではなく，異なるコンテキストで評価された多様な意見を見る必要がある．提案手法では，出現頻度の高い評価値のレビューのみを分析するだけでは埋もれてしまう評価視点を，評価値ごとの単語の出現頻度の分布に基づいて抽出し，多様な意見文を抽出する．

キーワード　情報抽出，評判分析，単語出現頻度

1　はじめに

宿泊先の決定時に実際にその場所に赴いて様子を伺う事は出来ないため，利用した人々の意見を知ることが出来るレビューサイトは重要な情報収集の手段である．レビューサイトの多くはレビュアーによる意見文の記入と5段階評価システムなどを用いて評価を行なっている．5段階評価システムなどは，数字で判断がしやすい反面，評価値が5や1など高評価・低評価が明確な意見文や頻繁に言及される評価視点が注目される傾向にあり，中程度の評価や評価項目にない評価視点への意見は見落とされがちである．

しかしながら，ユーザのニーズ次第ではそれらの見落とされる情報こそが重要になる可能性がある．多数の類似した意見の中に埋もれていて，相反する評価がされている意見に着目することで，多数の意見の中では言及されていることが見えにくい特定のコンテキストや評価視点が抽出可能になると考えられる．それらの視点を含む文を抽出することで，単語が持つ極性にとらわれない多様な意見の抽出を行うことができる．また，それぞれのレビュアーが重要だと感じた評価を得ることができるため，ユーザが自身が重要だと思う視点について比較しながらレビューの閲読が可能になる．

本研究では，レビューの評価値に限らず，ユーザが自身のニーズに合ったアイテムを決定するために必要な情報の抽出を研究の最終的な目的とする．本稿では，評価値毎の単語の出現頻度に着目し，多様な意見文の抽出と分析を行う方法を提案する．

2　関連研究

レビューの分析に関連した研究として，那須川らは文章に含まれる評価表現と好評と不評のどちらを示すかという極性の自動抽出を行っている [1]．対象とした文書の分野に応じて異なる評価を抽出できるが，ある単語が好評と不評の両極の意味で使われる場合は対応が困難と考えられる．立石らは，評価軸ごとの満足度・信頼度を含む意見の分類を行っている [2]．多数の情報の中から入力されたキーワードについての意見のみを知ることができるが，少数意見や明確な意見を表す文でない場合に排除されてしまう点で本研究とは目的が異なっている．また，小林らは，意見を対象，属性，評価に分けることを目的としている [3]．主観での意見に加えて状態や事実なども評価の対象となるが，文が設定された3つの要素で構成されているものに限定している．

単語の出現頻度に着目した分析として，和多らは特定のテーマの文書群において特徴的な単語の抽出をしている [4]．異なる分野の文書間において単語の乖離度の分布から，分析を行うことができる．単語頻度を用いて他の文書との比較を行う点において提案手法と類似しているが，評価情報を利用していない事や同一分野内での比較を行なわない点は本研究とは異なっている．ユーザのコンテキストに着目した研究としては，中山らはレビュー中の評価が成り立つための条件の分類・抽出をしている [5]．評価の理由や状況を抽出することが出来るが，それぞれがどのように評価されているか評価表現などの比較は行われていない．本稿では，レビューに付与された段階評価の値を用いて多様な意見の比較を行うことを目的とする．レビュー中の評価の比較については，

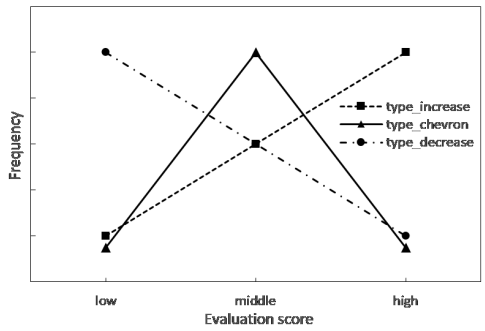

図 1 単語の出現頻度分布パターンのイメージ．type_increase は高評価になるほど高頻度となる単語，type_decrease は高評価になるほど低頻度となる単語，type_chevron は中評価で最も高頻度で高・低評価では低頻度となる単語をそれぞれ示す．横軸は評価値，縦軸は単語の頻度を示す．

Hoque et al. が Web 上でのコミュニティでの会話について，極性を可視化 [6] することで，意見の比較を容易にするしくみを提案している．提案手法では，各意見についての極性を提示するのではなく,「多様な意見が存在する評価視点」を発見することを目的としている．

3 分析手法

図 1 に，提案する分析手法で抽出を目指す評価視点の評価値毎の出現頻度分布パターンを示す．出現頻度がこのような分布に従った単語を抽出することによって，type_increase は「高評価になるほど高頻度」，type_decrease は「高評価になるほど低頻度」，type_chevron は「中評価が最も頻度が高く，高・低評価では低頻度」といった 3 種類の出現頻度分布パターンに該当する評価視点を抽出することができる．したがって,「多数の人が高評価なレビューで言及している評価視点が，低評価のレビューではどのように言及されているのか」といった，評価視点についての多様な意見文が抽出可能であると考えられる．

3.1 レビューデータの分割と各コーパス中での単語出現頻度の算出

あるレビュー対象 h についてコーパス c があるとする．本稿では評価値には 1 ～ 5 の離散値が用いられているものとする．レビュー対象 h についてのレビュー文を評価値ごとに分類し，評価値 1 の文を集めたコーパス（コーパス low），評価値 2〜4 のコーパス（コーパス middle），評価値 5 のコーパス（コーパス high）を用意する．レビュー対象 h についてのコーパス c のレビュー中で用いられている名詞 w の出現頻度 $Freq_c^h(w)$ を計算する．ここで，用意したコーパスごとにレビュー数や使われる名詞の数には差が存在するため，コーパス中の全名詞数によって $Freq_c^h(w)$ を正規化する．正規化したコーパス毎の名詞 w の出現頻度 $StFreq_c^h(w)$ は，式 (1) に従って算出される．

$$StFreq_c^h(w) = \frac{Freq_c^h(w)}{N_c}, \tag{1}$$

ここで，$c \in \{low, middle, high\}$ であり，N_c は，各コーパス中での全名詞の出現数を示す．

3 種類のコーパスに共通して出現する名詞 w を分析対象とする．提案手法では，$StFreq_c^h(w)$ を用いた関係式によって意見文に多様性が存在する可能性が高い評価視点 w の検出を行う．

3.2 出現頻度分布パターンに応じた単語群の抽出

図 1 中の type_increase を満たす名詞を抽出するために，以下の条件式 (2) と (3) を用いる．

$$\frac{StFreq_{high}^h(w)}{StFreq_{middle}^h(w)} > 1.0, \tag{2}$$

$$\frac{StFreq_{middle}^h(w)}{StFreq_{low}^h(w)} > 1.0. \tag{3}$$

式 (2) と式 (3) の条件をどちらも満たす名詞 w を高評価になるほど高頻度となる名詞として抽出し，得られた単語群を set_{inc}^h とする．

図 1 中の type_decrease を満たす名詞を抽出するために，以下の条件式と (4) を (5) 用いる．

$$\frac{StFreq_{low}^h(w)}{StFreq_{middle}^h(w)} > 1.0, \tag{4}$$

$$\frac{StFreq_{middle}^h(w)}{StFreq_{high}^h(w)} > 1.0. \tag{5}$$

式 (4) と式 (5) の条件をどちらも満たすことで，高評価になるほど低頻度となる単語群 set_{dec}^h を獲得できる．

最後に，図 1 中の type_chevron を満たす名詞を抽出するために，以下の条件式 (6) と (7) を用いる．

$$\frac{StFreq_{middle}^h(w)}{StFreq_{high}^h(w)} > 1.0, \tag{6}$$

$$\frac{StFreq_{middle}^h(w)}{StFreq_{low}^h(w)} > 1.0. \tag{7}$$

式 (6) と式 (7) の条件をどちらも満たすことで，中評価で最も高頻度で高・低評価では低頻度な単語群 set_{chev}^h を獲得できる．

得られた単語群 set_{inc}^h，set_{dec}^h，set_{chev}^h に含まれる名詞を含む意見文を各コーパスから抽出することで，単語の出現頻度分布パターンに基づいた多様な意見文の抽出を実現する．

4 実験と考察

実験では，分析対象のレビューデータには楽天株式会社が国立情報学研究所の協力により研究目的で提供して

表 1　ある宿泊施設のレビューから抽出された単語群の例. 各 set において，$StFreq_c^h(w)$ が高いものから上位 10 件を示す.

set_{inc}^h	set_{dec}^h	set_{chev}^h
部屋	フロント	の
利用	時	便利
宿泊	人	立地
こと	サービス	駅
今回	よう	品川駅
大変	さ	前
東京	ため	階
泊	スタッフ	非常
出張	もの	1
チェックイン	翌朝	一

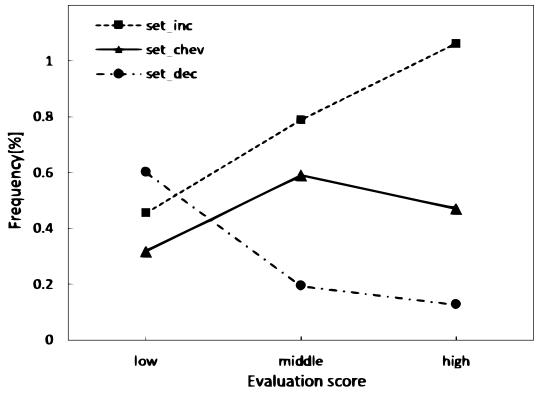

図 2 評価実験で得られた単語の出現割合の平均.

いる「楽天公開データ」に含まれる，楽天トラベルのレビューデータの一部を用いた．用意したレビューデータの中から無作為に選択した 20 施設のホテルレビューを 1 文ずつ分割した全 69,803 文を対象とした．これらのレビュー文についての情報として,「投稿本文」「総合評価値」「施設番号」の 3 項目を用いた.

4.1　単語群の抽出

表 1 に，提案手法によって得られたある宿泊施設のレビューから抽出された各単語群から $StFreq_c^h(w)$ が高いものから上位 10 単語を例として示す．set_{inc}^h では「部屋」や「宿泊」などどのような施設であっても共通する単語が見られる一方で，set_{dec}^h では,「フロント」「サービス」といったコミュニケーションに関連した単語が見られる．また，set_{chev}^h では,「駅」や「立地」などの宿泊施設外の事柄についての単語が特徴的に見られる.

図 2 に，得られた単語群 set_{inc}^h，set_{dec}^h，set_{chev}^h に含まれるそれぞれの単語の $StFreq_c^h(w)$ について，各宿泊施設および各コーパスでの平均値の割合のグラフを示す. なお，$StFreq_c^h(w)$ の最大値は 6.153，最小値は 0.003 であった．値の算出方法は式 (1) と同様に行った．set_{inc} では，評価値が下がるにつれて約 0.3 %ずつ $StFreq_c(w)$ が減少している．set_{dec} では middle での $StFreq_c(w)$ の平均値が突出しており，コーパス middle，high での $StFreq_c(w)$ に対して 0.5 %以上と大きな差異が見られる．set_{dec} に含まれる単語は，コーパス low 以外ではあまり出現せず，評価が低いレビューに特徴的に現れる単語であるということがわかる．概ね，図 1 で示したイメージに類似した分布傾向となっており，分析対象とする単語をデザイン通り抽出可能な数理モデル化に成功したと考えられる.

4.2　意見文の抽出

抽出された単語群それぞれについて，各コーパスから宿泊施設毎に単語群中の単語が含まれる意見文を抽出した．得られた文集合から無作為に 30 文ずつを取り出して考察した．表 2 に，それぞれの単語群中の単語を含む意見文の一部を例示する．意見文は考察に利用した文章の中で施設に関わらず無作為に取り出した.

set_{inc} 中の単語を用いて取得した意見文では，コーパス high からは「部屋がきれい」「満足」などの全体的な評価をしており，1 文中で複数の項目の評価をしているものも多い．しかし，旅行の詳細や対応に関する内容は，施設によって左右される傾向にある．コーパス low から得られた意見文，評価視点に対して評価の詳細な理由が述べられている傾向があった．このような高評価なレビュー中で評価された視点について低評価のレビューを参照することで，評価視点についての詳細な意見が取得された.

set_{dec} では，全体的に不評な意見が多いが，コーパス low からの取得とそれ以外では抽出される意見文の性質に差異が見られた．コーパス low では「残念」「悪かった」という評価部分のみの文章や，やり取りをした際の会話内容の一部などが多かった．一方で，コーパス middle やコーパス high から抽出された意見文では，“何がどう悪かったのか” が把握可能な文が抽出されている．同一の評価視点に対して好評な意見も多く抽出されており，不評な意見や要望のみを見て判断するのではなく，多様な観点に基づいてレビュー対象を吟味することが可能になると期待される.

set_{chev} を用いた場合では，施設ごとに異なった内容であり set_{inc} で抽出された意見文にはない特徴的な評価視点が出力された.「アメニティ」や「エレベーター」など部屋や施設内の細かなポイントで評価が記述されていることが多く，意見文の記述にホテルの特徴や個人のこだわりが反映されている傾向が示唆された．好評，不評に限らず多様な内容で意見が述べられているため，個人のニーズに合った情報が含まれている可能性が期待されるが，これらの意見文の中からユーザのコンテキストに応じて必要な情報を更に適切に絞り込むことが必要と考

表 2　単語群をもとに取得された意見文．表中の文は，すべてレビューデータから抽出された本文のままを示す．下線は，意見文の抽出の際にキーワードとなった各単語群中の単語を示す．

単語群	コーパス	取得した意見文
set_{inc}	high	駅からも近く、ホテル自体もとても綺麗でした お部屋はカジュアルツインの予約だったのにデラックスツインを用意いただいてラッキーでした 結果は大満足です
	middle	お部屋は狭かったのですが、ベッドはよかったので疲れもとれました 部屋の広さ・立地など価格に見合った満足感だと思います GW の 5 月 3 日直前予約でお部屋が空いておられたのは、とってもラッキーでした
	low	朝食は、値段の割に品数が少ないと思いました 素泊まりプランで利用をさせていただきました 朝食のウリであるワッフルも品切れしたまま補充なし
set_{dec}	high	ホテルの問題ではないですが、意外に駅から遠かったです 朝食はブッフェ式で品数はそんなに多いわけではないが、味・内容ともに断満足でした 多少トイレが古いタイプの感じですが，十分に使えます
	middle	ちょっとさすがに部屋のお風呂に入るのが気持ち悪いなと思い、大浴場へ 上の階の方(子供？) が飛び跳ねているのか寝れませんでした 今回初めての和室でしたが、風呂場の換気扇の音の大きさに驚きました
	low	今まで相当数のシティホテルに宿泊しましたが、残念ながら不満でした その間もホテルからは何のアナウンスも無く警報も止まりません 他の客は案内有りなのになぜですかね？ 楽しい気分を害された感じで最悪でした
set_{chav}	high	車は少し離れのパーキングに駐車となりますが、ホテルに荷物を預けてパーキングに行くことが出来ます 娘と何度か利用させて頂いていますが、とてもお気に入りのホテルです お風呂もエレベーターも宿泊者しか使えないようにしてありました
	middle	チェックアウト前に 荷物を送ろうと荷物を詰め そのブースに行きました ビジネスでしたが、繁華街も近く（ごはんマップなども参考にしました）良かったです シャワーとトイレが一体型のため、子供は使いづらかったようです
	low	良かったのは立地だけ、自社の駐車場に車を止めるのに 900 円も徴収するし 部屋はビジネスホテルのそれ 全体的にスタッフの印象、すごく悪かったです

えられる．

5　おわりに

本稿では，レビューの評価値毎に出現頻度が異なる単語に着目し，多様な意見文を抽出するための手法を提案した．また，そこから得られた意見文の内容について考察した．分析の結果，目的とする分布の単語を抽出に成功し，ユーザが重要と評価する視点に沿った意見の抽出も可能であることが示唆された．

一方で，名詞の詳細分類を更に絞り込む必要があるほか，評価値毎で単語の出力数の偏りや文の分割方法によっては文章の意味が把握できないなどの問題があった．今後は，評価や文章数の偏りを考慮した比較方法や，文章の分割・単語の選択方法の変更，ユーザの状況を示す条件文のような特徴を取り入れたレビュー分析手法へと発展させていく．

謝辞

本研究では，楽天株式会社が国立情報学研究所の協力により研究目的で提供している「楽天公開データ」を利用した．本研究は，一部，すかいらーくフードサイエンス研究所の助成のもと行われた．記して謝意を表す．

参考文献

[1] 那須川哲哉, 金山博. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会研究報告自然言語処理（NL）, Vol. 2004, No. 73, pp. 109–116, Jul 2004.

[2] 立石健二, 石黒義英, 福島俊一. インターネットからの評判情報検索. 情報処理学会研究報告自然言語処理（NL）, Vol. 2001, No. 69, pp. 75–82, Jul 2001.

[3] 小林のぞみ, 乾健太郎ほか. 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, Jul 2005.

[4] 和多太樹, 関隆宏, 田中省作, 廣川佐千男. 単語の出現頻度に着目した病院評判情報の分析. 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 2005, No. 50, pp. 15–20, May 2005.

[5] 中山祐輝, 藤井敦. レビューテキストを用いた条件付き意見文の抽出. 言語処理学会第 20 回年次大会発表論文集, pp. 888–891, Mar 2014.

[6] Enamul Hoque, Shafiq Joty, Luis Marquez, and Giuseppe Carenini. Cqavis: Visual text analytics for community question answering. In *Proc. of the 22nd International Conference on Intelligent User Interfaces*, pp. 161–172, 2017.