ユーザレビューを用いた内部環境ラベル付き未知情報特定法

1X07C068-6　鈴木和磨
指導教員　後藤正幸

1　研究背景と目的

質的なマーケティング分析ツールである SWOT 分析 [1] は，経営に影響を与える要素を内部環境 (強み，弱み) と外部環境 (機会，脅威) の 4 つの枠組みで分析する方法であり，企業の置かれている現状を把握する方法として有用である．この分析では，企業側には見えていない消費者視点での強みや弱みを含めて網羅することが望ましいが，そのためには何らかの形でユーザから情報を取得しなければならない．

一方，近年ユーザレビューサイトが増加し，大量に消費者の意見を取得することが可能になった．この中には，消費者が考える評価対象の強みや弱みが含まれている．これらのユーザレビューから，企業側が網羅できていない“ 強み ”，“ 弱み ”情報が自動抽出できれば，SWOT 分析における強み，弱みの情報が補完可能となり，分析者の支援ツールとして有用となると考えられる．

よって本研究では，“ 強みと弱みのユーザレビュー文の抽出 ”及び“ 未知情報の特定 ”の 2 段階からなるフィルタリング手法を提案し，企業側が未知である強み，弱み情報を抽出する方法を示す．第 1 フェーズでは，分類性能を向上させるために“ 強み ”や“ 弱み ”の表現で出現し易い特徴語が存在することに着目し，この特徴を活かした重み付きベクトルでユーザレビュー文を表現する．それらに対し RVM[2] を用いた分類を行い，強みと弱みを表す文を抽出する．第 2 フェーズでは，第 1 フェーズで抽出した文を対象とし，文中の名詞のマッチングによって未知情報の抽出を行う．提案手法を，宿泊施設を対象としたユーザレビューに適用することで，その有用性を実証する．

2　提案手法

2.1　問題設定

分析者は，学習データとして与えるユーザレビュー文の SWOT 分析を行い，あらかじめ人手によって各ユーザレビュー文を強み，弱み，その他に分類しているものとする．そして，“ 強みと弱みに分類された学習用ユーザレビュー文が指摘している対象を表す名詞 ”を評価対象語と定義し，これらの評価対象語に対する指摘は企業にとって既知の意見であると考える．

今回の分析対象である宿泊施設のユーザレビューでは，“ 部屋 ”，“ 風呂 ”，“ レストラン ”などが評価対象語となる．

ここで，本研究で抽出する未知の意見とは，“ 強みと弱みを表す ”かつ“ 学習データ作成時に出現していない未知の評価対象語を含む ”ユーザーレビュー文であると定義する．なお，1 つのユーザレビューには複数のラベルや多数の評価対象語が含まれていることがあるため，ユーザレビュー単位の分析では，その後の作業が煩雑となる可能性がある．そこで，本研究では分析を行う単位を文とする．

2.2　2 段階フィルタリング手法

本研究では，“ 強みと弱みの抽出 ”と“ 未知情報の特定 ”の 2 フェーズから成り立つ 2 段階フィルタリング手法を図 1 のように提案する．

第 1 フェーズの“ 強みと弱みの抽出 ”ではまず，学習データ中の全文の単語頻度分布に基づき相互情報量によって単語重要度を算出する．その後，RVM[2] を用いて強み，弱み，その他に分類を行うことで企業の強みと弱みを表す文を抽出する．ここでは，前節で述べた SWOT に分類された全文を学習データに用いる．

第 2 フェーズの“ 未知情報の特定 ”では，“ 強みと弱みの抽出 ”によって収集した文から，評価対象語のマッチングによって未知の評価対象語を含む文の抽出を行う．

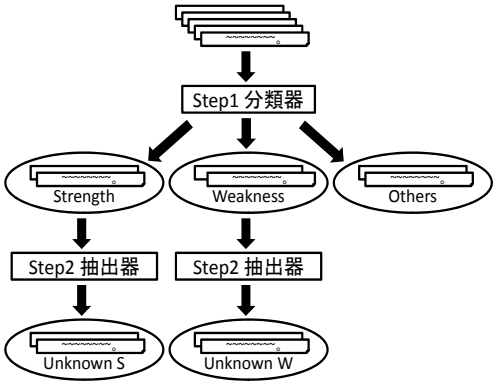

図 1 提案手法の手順

この手法を用いることで，ユーザレビュー文から専門的知識を必要とせずに，企業が未知の評価対象語を含む，強みと弱みを表す文を自動的に抽出することができる．

(1) 強みと弱みの抽出 (フェーズ 1)

ユーザレビューは，“ 素晴らしい ”，“ 綺麗 ”，“ 良い ”のような強みだけに出現する単語，“ 悪い ”，“ 古い ”のような弱みだけに出現するような分類に寄与する単語があるという特徴をもつ．分類を行う際には，単語の出現頻度のみのベクトルを用いるのではなく，各単語の重要度による重み付けを行った重み付き単語頻度ベクトルを入力とすることで分類性能の向上が期待できる．

本研究では重み付けの尺度として知られている相互情報量を用いる．ここで，J を全文中に出現する単語の種類数，単語を $w_j(1 \leq j \leq J)$，強み，弱み，その他のカテゴリを $c_k(1 \leq k \leq 3)$ と表すと，単語 w_j の相互情報量 $I(w_j)$ は以下の式で算出することができる．

$$I(w_j) = \sum_{k=1}^{3} P(w_j, c_k) \log \frac{P(w_j, c_k)}{P(w_j)P(c_k)}. \tag{1}$$

$P(w_j, c_k)$：全文中で単語 w_j を含み，かつカテゴリ c_k に属する文の割合
$P(w_j)$：全文中で単語 w_j を含む文の割合
$P(c_k)$：全文中でカテゴリ c_k に属する文の割合

ここで，ユーザレビュー文の数を M，それぞれのユーザレビュー文を $\boldsymbol{d_i}(1 \leq i \leq M)$ とし，単語 w_j の出現頻度 t_{ij} を要素にもつ単語頻度ベクトル $\boldsymbol{d_i} = (t_{i1}, t_{i2}, ..., t_{ij}, ..., t_{iJ})$ で表す．このとき，t_{ij} を (1) 式で算出した相互情報量によって重み付けした要素をもつ重み付き単語頻度ベクトル $\boldsymbol{x_i} = (v_{i1}, v_{i2}, ..., v_{ij}, ..., v_{iJ})$ と定義する．ここで，$\boldsymbol{x_i}$ における単語 w_j の特徴量 v_{ij} は，

$$v_{ij} = t_{ij} \cdot I(w_j), \tag{2}$$

で与えられ，$\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_i}, ..., \boldsymbol{x_M}$ を学習することによって，RVM の分類器を構築する．ここで，判別対象である新たなデータを$\boldsymbol{x}$とする．

従来，RVM は 2 値分類の手法であるため，多値分類に対応させるために一対多分類方式を用いた．各カテゴリに対して，カテゴリ c_k とそれ以外のカテゴリに分ける判別器を作成し，入力 $\boldsymbol{x}$ に対する各々の判別器の出力を $\boldsymbol{R} = (R_{c_1}, R_{c_2}, R_{c_3})$ とすると，

$$\hat{c} = \arg\max_{c_k} R_{c_k}, \tag{3}$$

となるカテゴリ $\hat{c}$ に分類する．ここで，R_{c_k} は入力 $\boldsymbol{x}$ がカテゴリ c_k に属する確率である．

(2) 未知情報の特定 (フェーズ 2)

フェーズ 2 では，フェーズ 1 で強みと弱みに分類された文から，未知の評価対象語を含む文を特定する．

学習データに含まれる名詞の集合を $\boldsymbol{S_L}$，フェーズ 1 によって強みと弱みに分類された全ての文中の名詞の集合を $\boldsymbol{N}$ とする．ここで，企業側が学習用ユーザレビュー文にない強み，弱みを既知情報として持っている場合は，$\boldsymbol{S_L}$ に既知評価対象語の追加処理を行う．本フェーズの目的は企業が未知の情報として，$\boldsymbol{N}$ から $\boldsymbol{S_L}$ を除いた名詞集合を出力することである．$\boldsymbol{N}$ から $\boldsymbol{S_L}$ を除くことで，$\boldsymbol{N}$ 中に学習データに出現しない未知の評価対象語が含まれていれば，これを必ず検出することができる．これらを含む文を，未知情報を含む文であると判定する．

3 評価実験

3.1 分析対象

本研究では，宿泊予約サイト“ じゃらん.net ”[3] 内のユーザレビューを分析対象とする．分析対象施設は京都のホテルとし，2010 年 1 月 13 日から 6 月 10 日までに投稿された 989 文のユーザレビュー文を分析に用いた．

3.2 実験方法

SWOT 分析により 1642 文を人手で強み，弱み，その他に分類し，評価対象語を抽出した．これらの評価対象語を含む文は，企業が既知の意見であると想定する．

未知意見の抽出可否を検証するため，抽出された評価対象語から 10 個，20 個，30 個の語を取り除き，それぞれの集合を $\boldsymbol{S_U}$ とし，これらを未知の評価対象語集合と想定することで実験を行う．

実験では，それ以外の文を学習データとし，$\boldsymbol{S_U}$ の要素を含む文を企業の未知の意見とみなし，これらをテストデータとして，その抽出可否について検証した．

3.3 評価方法

フェーズ 1 の評価は，重み付けを行わない場合と行う場合の分類再現率，分類精度で行う．また，フェーズ 2 の評価は，出力結果に $\boldsymbol{S_U}$ 以外の名詞がどの程度出力されているのか，また，出力されるべき未知の評価対象語がどの程度出力されているのかを表す評価対象語再現率，評価対象語精度で行う．

$$\text{分類再現率} = \frac{\text{正しく分類された文数}}{\text{テストデータの文数}}, \tag{4}$$

$$\text{分類精度} = \frac{\text{正しく分類された文数}}{\text{強み弱みに分類された文数}}, \tag{5}$$

$$\text{評価対象語再現率} = \frac{\boldsymbol{S_U}\text{中の出力された語数}}{\text{未知評価対象語数}}, \tag{6}$$

$$\text{評価対象語精度} = \frac{\boldsymbol{S_U}\text{中の出力された語数}}{\text{出力された語数}}. \tag{7}$$

3.4 実験結果と考察

(1) 強みと弱みの抽出

表 1 に重み付けを行わない場合と行う場合の分類再現率，分類精度を示す．その結果，重み付けを行う提案手法は，全てのパターンで分類精度，再現率が共に高くなっており，その有効性が示せた．また，強みと弱みの分類再現率を評価対象語毎にみてみると，強みの平均は 0.876，弱みは 0.567 となり強みの方が高い結果となった．これは，学習データ内に強みを表す文が多く，弱みを表す文はその半数以下というカテゴリ間のデータ数の偏りによる RVM の分類性能劣化のためと考えられる．

表 1　実験結果 (強みと弱みの抽出)

重み付け	未知評価対象語数	10	20	30
なし	分類再現率	.685	.763	.764
	分類精度	.725	.798	.784
あり	分類再現率	.704	.789	.828
	分類精度	.745	.826	.850

(2) 未知情報の特定

表 2 に評価対象語再現率と評価対象語精度を示す．フェーズ 2 では，取り除く名詞に $\boldsymbol{S_U}$ の要素は含まれていない．そのため，フェーズ 1 で強みと弱みを表す文が全て正しく抽出できた場合，2.2.2 節で述べた通り未知の評価対象語集合 $\boldsymbol{S_U}$ の要素である未知評価対象語は全て出力される．しかし，表 2 の評価対象語再現率より，全て出力されている訳ではないことが分かる．この原因は，フェーズ 1 でユーザレビュー文が正しいカテゴリに分類されていないことであり，これを改善するためには，フェーズ 1 の分類再現率の向上が必要となる．

表 2 の評価対象語精度より，$\boldsymbol{S_U}$ 以外の名詞が多数出力結果に含まれていることが分かる．そのため，出力結果からどれが未知評価対象語なのかを判断することは難しい．ただし，今回は“ 浴室 ”や“ 浴衣 ”などに対する指摘を未知評価対象語としており，$\boldsymbol{S_U}$ 以外の出力された名詞は，“ 換気 ”や“ 襟元 ”のような，未知評価対象語に関係すると考えられる名詞であった．このような名詞を下位の未知評価対象語として階層的に出力することができれば，指摘されている細かい点まで知ることができ，評価対象語精度を向上させることができると考えられる．

表 2　実験結果 (未知情報の特定)

未知評価対象語数	10	20	30
評価対象語再現率	.900	.900	.867
評価対象語精度	.164	.171	.176

4 まとめと今後の課題

本研究では宿泊施設のユーザレビューに対して，2 段階フィルタリング手法を用いることによって，ユーザレビュー文から強みと弱みを表す文を抽出し，そこから未知の評価対象語を抽出することができた．

今後の課題としては，SWOT 分析における強み，弱みの網羅性を更に高めるために，既知評価対象語に対する未知の指摘内容も抽出できるようにすることである．

参考文献

[1] Philip Kotler, “A Framework for Marketing Management，First Edition,” *Prentice Hall*, 2001.

[2] M.E.Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning Research*, pp. 211–244, 2001.

[3] じゃらん.net，“http://www.jalan.net/”