レビュー文章の自動分類におけるテキストの前処理手法の検証

西川 崇哉[1]　岡田 真[2]　橋本 喜代太[3]
[1,2]大阪府立大学大学院理学系研究科情報数理科学専攻
[3]大阪府立大学人間社会学部
{[1]ss301013@edu, [2]okada@mi.s, [3]hash@lc}.osakafu-u.ac.jp

1. はじめに

近年，インターネットの発達によってweb上に個人が情報を発信する機会が増え，“Amazon.co.jp”，“価格.com”，*“TripAdvisor”*などのカスタマーレビューを投稿できるサイトが多くなっている．これらのサイトに寄せられるカスタマーレビューの文書は，個人の意見や評判を含んでいるという点で有益な情報源である．レビュー文書は一般的な人にとっては商品やサービスを選択したり，購入決定の指標になる．また，企業にとっては顧客のニーズを知ることのできる情報源となる．

しかし，それらは膨大な量のデータとなるため全てを閲覧するには時間や労力が大きい．そこでレビューデータを有効活用するための研究が近年盛んになされている．これら既存の研究では，必要な情報を抽出したり，入手したデータの傾向によって意見などを分類したりする．分類というアプローチを取る場合，レビューを肯定的か否定的かの2種類で分類する研究などがある．

しかしながら，文書自体の内容のみではなく文章の書き手の状態を推測して分類することも必要である．例えば，客層ごとに意見を整理し，それに基づきマーケティングをおこなうことは企業においてもよくなされている手法である．本研究ではこのような分類をおこなうための前処理に関連する実験をおこなった．

本研究では*“TripAdvisor”*という旅行情報ポータルサイトのレビューを用いて，筆者の利用状況・状態に応じての文書分類をおこなうことを考える．具体的には，*“TripAdvisor”*のレビューの記述項目のオプションに，旅行へ“一人で行った”や“家族で行った”などを記述する項目があるが，これらの項目ごとに文書が自動分類できないかと考え，機械学習を用いた自動分類のアプローチを取ることにした．近年，自然言語処理の分野で機械学習を使う技術は盛んにおこなわれており，この分類においてもうまく機能することを期待して導入した．機械学習をおこなうには，文書データを数値化してベクトルにする必要があり，そのための準備として前処理が必須である．そこで，ベクトルの作成の前処理としてさまざまな手法をおこない，どのようなデータが有用であるかを実験により検証した．

以下，2章で関連研究について，3章で*“TripAdvisor”*について，4章では提案手法について説明し，5章で実験と考察について述べる．最後にまとめと今後の課題について述べる．

2．関連研究

大量の文書データを対象とした研究のアプローチは大きく2つの方法がある．1つは意見の抽出，もう1つは意見の分類である．

前者は，意見がどのような要素によって成り立つのか考察し，意見の構成要素を定義する．その構成要素に基づいて，要素間の関係をテキストから抽出する課題に取り組むものである．飯田ら[1]は意見を<対象，属性，評価値>の3つ組で定義し，文書から適切に抽出できるか試した．

後者は文書集合を内容ごとに分類するものである．近年盛んにおこなわれている研究として，意見の内容を肯定意見と否定意見の2種類に分類するものがある．これは極性判定と言われる．また，橋本[2]らは新聞記事の記事内容ごとに，池田ら[3]はブログの記事内容にタグが付いているデータからタグなしのデータを，ともに機械学習によって分類している．これらの研究は文書内容から推測して，文書の内容を基に文書を分類している．

本研究では，文書の内容を基に筆者の状況などを推定し，分類することを試みた．具体的な例として，ある宿泊施設をビジネスで利用した場合と子供を連れて家族連れで利用した場合では，同じ施設でもその評価に差が出ることが予想される．このような場合には，評価の内容に加え，どのような状況で利用したかという情報を取得し，それに基づいて評価を分類する必要がある．そこで今回は利用者の状況を考慮してレビュー文書の分類ができるか検証した．

3. *“TripAdvisor”*

分析の対象とするレビューデータは *“TripAdvisor”* [4]という旅行ポータルサイトから入手した．*“TripAdvisor”*は世界中のホテル・観光名所・レストランに関する５千万件の口コミ情報を扱い，ユーザーが投稿した写真などが掲載されているサイトである．図１に *“TripAdvisor”*のレビューの例を示す．このサイトのレビューには文書以外に利用者の利用目的・状況の記載や，指標ごとの五段階評価などが付加されているものもある．利用者の利用目的・状況は“ビジネス”，“カップル・恋人”，“家族旅行”，“友達”，“一人旅”の５種類の利用者タグで表すことができる．利用状況に依存して，同じホテルを利用した場合でも評価の内容は変わると考えられる．利用者タグのない文書を読んで利用状況の判別をおこなうのは手間のかかる作業であり，状況を考慮して文書を自動的に分類するシステムが必要とされる．

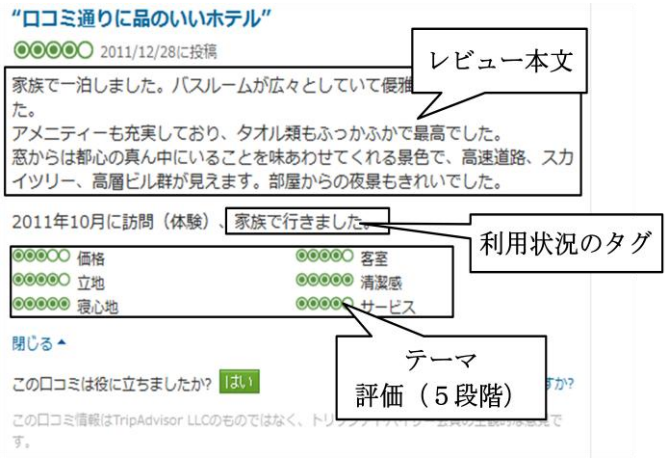

図１．*“TripAdvisor”*のレビューの例

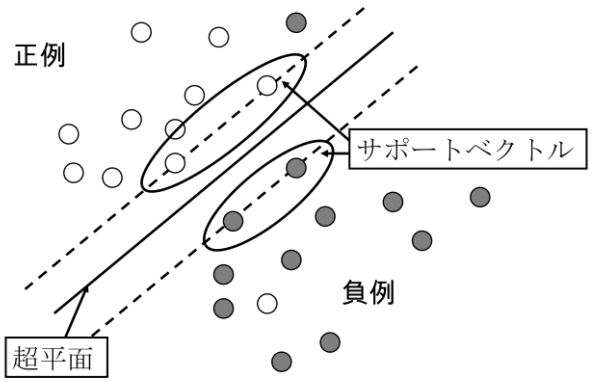

図２．サポートベクターマシンの概念図

4. 提案手法

今回は機械学習を用いて文書の分類をおこなう．機械学習には決定木やニューラルネットワークがあるが，本研究ではサポートベクターマシン(SVM)を使用する．SVM は近年の自然言語処理関連の研究でも盛んに使われている．

4.1 SVM

SVM は Vapnik[5]によって提案されたデータを２つのクラスに分類する教師あり学習アルゴリズムである．図２にサポートベクターマシンの概念図を示す．正例と負例の含まれたベクトルの学習データが与えられ，それらの正例と負例を区切る超平面を計算する．その超平面によって未知データのクラスを推測する．SVM は高次元のベクトル空間であっても超平面で分類することでき，高い汎化能力をもつことが知られている．自然言語処

理のベクトル化は高次元になることが多いため，自然言語処理の研究において SVM が用いられることは多い．

4.2　ベクトル化

機会学習はルールベースの手法と違い，ルールを考察する必要はない．しかし，機械学習は数学的な枠組みの上に成り立っているので，データを数値化してベクトルを作る必要がある．文書からベクトルを作成するために，文書をある規則に従った単位で区切り，bag-of-words の考えによって頻度ベクトルまたは二値ベクトルに変換する．今回は文書を区切る単位を次のようにし，二値ベクトルを作成した．

文字 N-gram

連接する N 文字を一つの単位として区切る方法である．どのような言語や文書に対しても適用でき，作成の速度も速いという利点を持っている．欠点はベクトルの次元が大きくなるということである．

単語　N-gram

連接する N 単語を一つの単位として区切る方法である．英文の場合には単語と単語の間にスペースが入るため，スペースごとに区切られた文字列を抽出すれば単語ごとのベクトルが容易に作れる．しかし日本語はスペースで区切る「わかち書き」の習慣がないため，形態素解析器 mecab[6]を用いることによって単語ごとに区切る必要がある．単語 N-gram の利点は内容語のみを抽出し用いることができることである．欠点は，形態素解析は常に正しく解析できるわけではなく，特に辞書に登録されていない未知語ひらがな単語の抽出に難があることである．

今回は文字 2-gram，3-gram と単語 1-gram，2-gram の 4 種のデータを作成してそれぞれの結果を検証した．

4.3　ベクトル要素の剪定

学習データのベクトル要素中で，出現頻度があまりにも高い要素と低い要素は学習に悪影響を及ぼす場合があると我々は予想した．例えば高頻度語はさまざまな文書に入っており独自性が低く，分類に役立たないデータの場合がある．また低頻度語はあまりにも独特な情報や雑音の情報となる場合が多いと考えた．ここで分類精度を維持しつつベクトルの次元を小さくすることができれば，分析時間の短縮とさらに多くのデータによる学習が可能になると考える．

そこで今回は，学習用データのうち頻度の上位 10%，20%，30%を切り捨てたデータと下位の 10%，20%，30%を切り捨てたデータと切り捨てなしの計 7 種類をそれぞれ作成して実験をおこなった．

5. 実験と考察

5.1　実験

*“TripAdvisor”*のレビュー文書を一人または複数人（家族・友達・カップル）の 2 つのクラスにどの程度の精度で分類できるか調べた．

*“TripAdvisor”*のレビューのうち利用目的・状況が書かれているため，一人の利用か複数人による利用かが明らかなものを選び出し，あらかじめクラス分けをおこなって，各レビューに一人の利用か複数人の利用かのラベルを付加しておく．

それらデータのからランダムに抜き出した 10000 件のデータのうち，5000 件を学習データ，5000 件を実験データにした．

学習データを 4 章で説明した文字 2-gram，3-gram と単語 1-gram，2-gram の 4 種と，データごとに頻度を変えた 7 種類の計 28 通りのデータに対して分類の実験をおこなった．

5.2　考察

表 1 に実験結果を示す．まず，その分類精度は全体的に 60%前後に収まった．各手法（文字 2-gram，3-gram と単語 1-gram，2-gram）の最良の結果同士を比較しても大きな差はなかった．切り捨てたデータとその正解率に関して観察をおこなうと，低頻度語を切り捨てたデータは全体的

に正解率が大きく低下しており，高頻度語の低下は小さかった．よって今回の実験においては，分析に有用なデータは低頻度語に多いことが分かった．

ベクトルの次元のサイズにおいては表 2 に示す．連接数が増えるとサイズ大きくなり，文字 2-gram と単語 1-gram，文字 3-gram と単語 2-gram がほぼ同じサイズであることが分かった．

今回の実験ではベクトルのサイズが同程度である場合には，精度の面で文字 N-gram の方が有利となったが，各条件やレビュー文書の内容によりこの結果は変わると思われる．

表 1．実験結果（精度）

			文字 N-gram		単語 N-gram	
N-gram 連接数(N)			2	3	1	2
頻度	全要素使用		64.5	65.3	63.8	63.6
	上位未使用	10%	64.7	64.5	64.0	63.5
		20%	64.1	63.9	64.0	62.3
		30%	63.9	64.1	64.2	62.7
	下位未使用	10%	61.6	61.2	56.7	61.2
		20%	60.3	60.6	51.9	59.1
		30%	59.7	59.7	52.3	57.9

表 2．実験結果（ベクトルの次元の大きさ）

			文字 N-gram		単語 N-gram	
N-gram 連接数(N)			2	3	1	2
頻度	全て使用		7828	13991	8718	13975
	上位未使用	10%	7045	12591	7846	12576
		20%	6262	11191	6974	11179
		30%	5479	9793	6102	9782
	下位未使用	10%	7046	12952	7847	12578
		20%	6263	11193	6975	11180
		30%	6103	9794	6103	9783

6.まとめと今後の課題

本稿では，機械学習によって利用者の状況ごとにレビューを自動分類できるか実験をおこなった．さらに前処理の条件を複数決定し，それらのうちどの処理が有効かを実験により検証し，考察をおこなった．

今後の課題としては，精度に関しては本研究で作成した連接数以外のパターンも作成し，解析精度がどの程度変わるか検証すること，また単語 N-gram 使用時に同類語・同義語を日本語 WordNet などの概念辞書を用いて統一することで，どのような影響があるかを検証することなどが挙げられる．またデータのサイズに関して，計算速度を実際に計測することも課題の一つである．

参考論文

[1]飯田龍，小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：“ 意見抽出を目的とした機械学習による属性-評価値対同定”，情報処理学会自然言語処理研究会，NL-165-4，pp.21-28，2005.

[2]橋本泰一，村上浩司，乾孝司， 内海和夫， 石川正道：“文書クラスタリングによるトピック抽出および課題発見”，社会技術研究論文集，Vol.5，pp.216-226，2008.

[3]池田 大介，高村 大也，奥村 学：“blog 分類のための半教師有り学習”，情報処理学会自然言語処理研究会，NL-183，pp.59-66， 2008.

[4]http://www.tripadvisor.jp/

[5]V. N. Vapnik，The Nature of Statistical Learning Theory，2nd ed.，Springer-Verlag，New York，2000.

[6]http://mecab.sourceforge.net/