1A3-03

SVMと新聞記事を用いたWeblogからの意見文抽出

Opinion Extraction from Weblog using SVM and Newspaper Article

川口 敏広[*1] 松井 藤五郎[*2] 大和田 勇人[*2]
Toshihiro Kawaguchi Tohgoroh Matsui Hayato Ohwada

[*1]東京理科大学大学院 理工学研究科 経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

[*2]東京理科大学 理工学部 経営工学科
Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science

In this paper, we propose two methods to classify Weblog article into review article and non-review article using Support Vector Machines (SVMs) and to extract opinions from review article using newspaper article.

1. はじめに

現在，Web サイトでは，個人が自由に意見を発信でき，そこにはさまざまな人の多様な意見が存在する．それらの意見を読むことにより，企業にとってはマーケティングやクレーム処理，個人にとっては製品購入の際の意志決定支援として利用することができる．しかし，こうした意見を手作業で集めるには時間やコストがかかってしまうため，意見を自動的に抽出する研究がさかんに行われている．

立石ら [立石 01] は意見を < 対象，属性，評価値 > の 3 つ組で定義し，3 つ組表現の特徴語辞書を作成し，この表現と意見らしさのパターンマッチによって，意見文かどうかを判定している．しかし，この手法ではアフィリエイトやカタログなどのような意見文を含まない文章からも抽出されてしまう．また，この手法では 3 つ組の辞書をドメイン毎に人手で構築する必要があり，ドメイン毎にこれを行うのは時間的，量的にも現実的でない．

そこで本研究では，まず SVM を用いて記事を主観的な意見を含むレビュー記事と非レビュー記事に分類し，次に新聞記事から抽出した辞書に基づいてレビュー記事から意見文を抽出することを提案する．これにより，主観的な内容の意見文だけをドメイン毎の辞書を人手で構築せずに抽出することができる．

本論文では，Support Vector Machine (SVM) を用いて Weblog の記事をレビュー記事と非レビュー記事に分類する方法と，新聞記事から抽出した辞書に基づいてレビュー記事から意見文を抽出する方法を述べる．

2. 関連研究

立石ら [立石 01] は，Web ページから意見文を抽出することを目的とし，意見を < 対象，属性，評価値 > の 3 つ組で定義し，3 つ組表現の特徴語辞書を作成し，この表現と意見らしさのパターンマッチによって，意見文かどうかを判定している．

Yu ら [Yu 03] は Wall Street Journal 記事 (WSJ 記事) を用いて，文単位で事実と意見の分類を行っている．Yu らは訓練データに人手でラベル付けせずに，ニュースとビジネスを事実文書，社説と編集者への手紙を意見文書と仮定し，意見文書内の文をすべて意見文，事実文書内の文をすべて事実文と仮定することによって分類器の訓練を行った．テストデータには WSJ 記事を用いており，文単位の分類では F-measure で 80%程度の精度を得ている．

藤村ら [藤村 05] は，語のスコアを計算することで，電子掲示板の評判情報を文単位で肯定と否定に分類している．肯定的 (否定的) な評判には，肯定的 (否定的) な概念を持った語が多く含まれているはずであるという仮定を基に，肯定的な評判と否定的な評判の差をとることで，式 (1) によりスコアリングをしている．

$$score(w_i) = \frac{P_P(w_i) - P_N(w_i)}{P_P(w_i) + P_N(w_i) + k} \quad (-1 \leq score(w_i) \leq 1) \tag{1}$$

$$Score(sentence) = \sum_{ALLw_i} score(w_i) \tag{2}$$

$$\begin{cases} if \quad Score(sentence) > 0 \rightarrow positive \\ \quad\quad Score(sentence) < 0 \rightarrow negative \end{cases} \tag{3}$$

一般的な語はどちらの文にも同様に出現するはずであるから，その影響は打ち消される．$P_P(w_i)$ は肯定的な評判に語 w_i が出現する確率，$P_N(w_i)$ は否定的な評判に語 w_i が出現する確率を表しており，これらの確率を利用し語 w_i のスコア $score(w_i)$ を求めている．また，$Score(sentence)$ は文のスコアを表している．最終的な分類器は，文中の語のスコアの総和を求め，その値が正ならば肯定的な文，負の値ならば否定的な文としている．

3. 提案手法

本章では，SVM を用いた Weblog 記事のレビュー記事と非レビュー記事への分類方法と，新聞記事から抽出した辞書に基づくレビュー記事からの意見文抽出の方法について述べる．

3.1 SVM を用いたレビュー・非レビュー記事の分類

本研究では，goo ブログ Search を用いて記事を収集する．検索したい製品の製品名を検索語としてヒットした記事から，テキスト形式の本文と HTML ファイルを取得する．取得した記事を読み，レビュー記事か非レビュー記事かを判断し，人手でラベル付けを行う．訓練データを形態素解析器 MeCab にかけ，必要な品詞を取り出し特徴語辞書を作成する．特徴語辞書，アフィリエイトリンクの有無，品詞の割合を SVM の属性として記事の分類を行う．

連絡先: 川口敏広，東京理科大学大学院，理工学研究科，経営工学専攻，千葉県野田市山崎 2641，04(7124)1501，j7405613@ed.noda.tus.ac.jp

SVM の分類指標となる特徴語には，評価表現になりうる可能性が高い品詞の形容詞，形容動詞，動詞を用いる．訓練データの記事に含まれる形容詞，形容動詞，動詞を抽出し，特徴語辞書を作成する．

アフィリエイトとは自分のサイトに企業サイトへのリンクを張り，ユーザがそこを経由して製品を購入したりすると，サイトの管理者に報酬が支払われるシステムである．アフィリエイトを行っている主な企業は，楽天，amazon，A8.net，bidders，ACCESS TRADE，VALUE COMMERCE，イオン，Link Share などが挙げられ，通信販売が可能な製品にはほぼ全てに対して適用することができる．

現在，アフィリエイトを専門に行うブログが爆発的に増加しつつある．実際にブログを対象にした検索エンジンを用いてある製品名で検索したとき，検索される記事がアフィリエイトを行っている記事である確率は約 30%にも及ぶ．

ここで注目したいのは，アフィリエイトを行っている記事の内容を見ると，アフィリエイト専門のブログである可能性が非常に高いということである．つまり，製品を購入し，レビュー記事を書く人はアフィリエイトをほぼ行わないという傾向が見られる．一方，amazon のアフィリエイトに関しては，ブログの機能として元々備わっている場合が多く，アフィリエイトを専門に行う人以外でも一般的に用いられていることが多い．

アフィリエイトを行っているかどうかは，HTML ソース中のアフィリエイト用リンクの存在有無で判断できる．例えば，「http://pt.afl.rakuten.co.jp」というリンクがあれば，楽天のアフィリエイトを行っていると判断できる．前述のように，アフィリエイトを行っている記事はアフィリエイト専門の記事である可能性が高いため，アフィリエイトリンクが存在すれば，アフィリエイト記事，すなわち非レビュー記事である可能性が高いと考えられる．以上より，アフィリエイトリンクの存在有無は，レビュー記事と非レビュー記事の分類の際，大きな指標として利用できると考えられる．そこで，前述の amazon を除く 7 つの企業のアフィリエイトリンクを記述したアフィリエイトリストを作成し，これらのリンクの存在有無を考慮して分類を行う．

レビュー記事は書き手によって評価が行われるため，主な評価表現である形容詞や形容動詞，動詞が多く存在する．反対に非レビュー記事であるアフィリエイト記事では，名詞や形容詞が多く存在し，また製品の仕様のみが記述された記事では，動詞がほとんど使用されないという傾向が見られる．このように記事の種類によって使用される品詞が大きく異なり，記事の種類ごとに使用される品詞は似た傾向があると考えられる．

分類において，品詞の割合を考慮する式を以下のように定義する．

$$P(x) = \frac{C(x)}{C(noun) + C(verb) + C(adjective)} \quad (0 \leq P(x) \leq 1) \tag{4}$$

noun は名詞，*verb* は動詞，*adjective* は形容詞を表し，x は名詞，動詞，形容詞のいずれかになる．$C(x)$ は記事中における x の出現回数であり，$P(x)$ は x の記事中における割合である．名詞，動詞，形容詞の割合を求め，この 3 つの値を SVM の分類指標として属性に追加することにより，品詞の割合を考慮する．

3.2 新聞記事を用いたレビュー記事からの意見文抽出

本研究には訓練フェーズとテストフェーズがあり，それぞれのデータを形態素解析器 MeCab にかけ形態素解析を行う．訓練フェーズでは，得られた形態素から必要な品詞を取り出し特徴語リストを作成する．その特徴語リストを基に特徴語にスコアリングし，特徴語辞書を作成する．特徴語辞書を用いて，形態素解析を行ったテストデータに文単位でスコアリングをする．得られた文のスコアに基づき，文を分類規則に沿って意見文と非意見文に分類する．

本研究では，意見文とは「個人の主観が述べられている文」と定義をする．また比較として，評判文の定義を述べると「ある特定のモノに対する個人の主観が述べられている文」と考える．つまり，評判文というのは，意見文の部分集合であり，意見文を網羅的に収集することで，評判情報をもれなく収集することができる．立石らの手法でのパターンマッチでは収集できない，広範囲の意見文を網羅的に収集することを狙いとする．

新聞記事は身近で手軽に利用できるデータだが，現在，意見文抽出にあまり使われていない．一般に，新聞記事は文語表現で事実を伝えるメディアだと思いがちであるが，様々なジャンルの記事を掲載しており，ジャンル毎にそれぞれ傾向がある．例えば，経済や国際では文語表現や事実を多く含んでいるが，社説や家庭では，他のソースからの引用も多く口語表現や意見文などを多く含んでいる．そこで本研究では，新聞記事のジャンル毎の傾向を利用し，社説を意見記事，国際を非意見記事と仮定する．そして，意見記事に含まれる文を全て意見文，非意見記事に含まれる文を全て非意見記事と仮定することで，新聞記事を訓練データに用いクラスラベル付けの問題を解決する．また，その信頼性から訓練データ作成時に記事の選別が不必要であることや，膨大なデータ量からの特徴語の網羅性も大きな利点である．

特徴語のスコアリングは式 (5) を用いる．

$$score(w_i) = \frac{P_O(w_i) - P_F(w_i)}{P_O(w_i) + P_F(w_i) + k} \quad (-1 \leq score(w_i) \leq 1) \tag{5}$$

(6)

$P_O(w_i)$ は意見文で特徴語 w_i が出現する確率である．同様に $P_F(w_i)$ は非意見文で特徴語 w_i が出現する確率である．k は経験的に 0.001 とする．

文の分類には以下の式 (7)，(8) を用いる．

$$Score(s) = \sum_{ALL w_i} score(w_i) \tag{7}$$

$$\begin{cases} if \quad Score(s) > 0 \rightarrow \text{意見文} \\ \quad\ Score(s) \leq 0 \rightarrow \text{非意見文} \end{cases} \tag{8}$$

各文に含まれる特徴語のスコアの総和が 0 より大きければ意見文，0 以下ならば非意見文と分類する．文単位では文に特徴語が 1 語も現れない場合が少なからずあり，その場合の文のスコアは 0 となり非意見文と分類される．例えば，「うーん.」や「IXY DIGITAL600」などのように，感嘆表現や名詞句だけで 1 文を構成している場合がそれにあたる．

4. 評価実験

本節では，4.1 で SVM を用いたレビュー記事と非レビュー記事の分類の評価実験を，4.2 で新聞記事から抽出した辞書に基づくレビュー記事からの意見文抽出の評価実験について述べる．

表 1: 評価実験の種類

	実験 1	実験 2	実験 3	実験 4	実験 5
分類対象数 (件)	305	218	218	218	218
レビュー記事 (件)	109	109	109	109	109
非レビュー記事 (件)	196	109	109	109	109
次元数	1227	1227	1234	1230	1237
アフィリエイト考慮	無	無	有	無	有
品詞の割合考慮	無	無	無	有	有

表 2: 評価実験結果

	実験 1	実験 2	実験 3	実験 4	実験 5
Accuracy	78.03	78.44	85.32	83.49	83.48
Recall	79.17	87.81	88.89	90.11	89.52
Precision	52.29	66.06	80.73	75.23	86.24

表 3: 応用実験結果

	IXY DIGITAL 700	iPod nano
Accuracy	83.06	80.99
Recall	70.97	74.78
Precision	94.29	85.57

表 4: テストデータ 1 の分類結果

	Accuracy	Precision	Recall	F-measure
訓練データ 1	72.39	72.01	83.40	77.29
訓練データ 2	72.50	72.59	82.23	77.11

表 5: テストデータ 2 の分類結果

	Accuracy	Precision	Recall	F-measure
訓練データ 1	61.98	68.39	67.52	67.95
訓練データ 2	73.00	73.63	85.35	79.06

表 6: テストデータ 1 における特徴語カバー率

	テストデータ 1 に対する特徴語カバー率
訓練データ 1	54.07%
訓練データ 2	93.56%

表 7: テストデータ 2 における特徴語カバー率

	テストデータ 2 に対する特徴語カバー率
訓練データ 1	54.03%
訓練データ 2	92.89%

4.1 評価実験 1

訓練データとして Canon のデジタルカメラ「IXY DIGITAL 600」に関する記事 305 件 (レビュー記事 109 件，非レビュー記事 196 件) を収集した．特徴語辞書に使用した特徴語は「IXY DIGITAL 600」に関する記事 305 件の本文中に存在した形容詞，形容動詞，動詞の原型 1227 語とした．

実験は 5 種類のパターンで行った．表 1 にその詳細を示す．実験 1 では記事 305 件を特徴語辞書に登録した形容詞，形容動詞，動詞の合計 1227 語を分類指標として実験を行った．実験 2 では実験 1 の分類対象のレビュー記事と非レビュー記事の数を 109 件に統一して実験を行った．実験 3 では実験 2 に 7 種類のアフィリエイトの有無を考慮して実験を行った．実験 4 では実験 2 に品詞の割合を考慮して実験を行った．そして，実験 5 では実験 2 にアフィリエイトの有無と品詞の割合を考慮して実験を行った．いずれの実験も leave-one-out による交差検定を用いて評価を行った．5 種類の実験結果を表 2 に示す．

この実験より，分類対象を統一すること，アフィリエイトの有無を考慮すること，品詞の割合を考慮することのいずれも精度を向上させる結果が得られた．

応用実験として,「IXY DIGITAL 600」に関する記事 218 件を訓練データとし,「IXY DIGITAL 700」に関する記事 183 件 (レビュー記事 70 件，非レビュー記事 113 件) と「iPod nano」に関する記事 221 件 (レビュー記事 97 件，非レビュー記事 124 件) をそれぞれテストデータとして分類を行った．SVM に与える属性は，評価実験で最も精度の高かった形容詞，形容動詞，動詞の 1227 語の特徴語の存在有無と，7 種類のアフィリエイトリンクの存在有無，3 つの品詞の割合を用いた．表 3 に実験結果を示す.「IXY DIGITAL 700」に関しては，Accuracy，Precision，Recall の値すべてが 70%を越え，高い精度で分類できた．また,「iPod nano」に関しても同様に Accuracy，Precision，Recall が 70%を越えた．

4.2 評価実験 2

実験には，CD-毎日新聞 '94 と goo ブログ Search 用いてレビュー記事を収集した．レビュー記事は「アフィリエイトサイトへのリンクを含んでいない，評判文を含むブログ記事」と定義した．レビュー記事はブログの記事の投稿の単位であるエントリ単位で収集し，本文のみを使用した．ブログ記事は Canon のデジタルカメラ IXY DIGITAL とテレビ番組で日本一に輝いたケーキ店キルフェボンに関する記事である．

訓練データ 1 は，文単位に人手でラベル付けされた IXY DIGITAL700 の意見文 636 文，非意見文 451 文を使用した．訓練データ 2 は，社説記事の全 3136 記事中の全 53185 文を意見文，国際記事の全 6770 記事の全 54108 文を非意見文とクラスラベルを仮定し，使用した．

テストデータ 1 には，文単位に人手でラベル付けされた IXY DIGITAL600，700 の意見文 512 文，非意見文 397 文を使用した．また，テストデータ 2 には人手でラベル付けされたキルフェボンの意見文 157 文，非意見文 106 文を使用した．

特徴語の品詞として，形容詞，形容動詞，動詞を使用した．形容詞，形容動詞については，主に日本語でモノの評価を表す表現であり，文が意見性を持つかどうかを判断するために採用した．動詞については，そのモノに対しての個人の行動を表す表現であり，文の意見か，または非意見かを判断するために採用した．

実験は 2 回行った．1 つ目は訓練データ 1，2 でテストデータ 1 を分類する実験，2 つ目は訓練データ 1，2 でテストデータ 2 を分類する実験である．表 4 にテストデータ 1 の分類結果を，表 5 にテストデータ 2 の分類結果を示す．

表 4 では，テストデータと訓練データ 1 は同じドメインのデータだが，新聞記事を用いた訓練データ 2 でも同程度の精度を得ることができた．一方，表 5 では，テストデータは両方の訓練データと異なるドメインであり，新聞記事を用いた訓練データ 2 の方が，全体的に精度が 10%程度上昇した．特に

Recall は 18%程度精度が良く，意見文を網羅的に抽出できたことがわかる．また，クラスラベル付けの観点から見ると，訓練データ 1 は人手，訓練データ 2 は仮定に基づいた自動付与であったが，精度に影響は見られない．

特徴語が分類対象のドメインに特有の特徴語をどれだけ含んでいるかを測定するために,「(訓練データの特徴語∧テストデータの特徴語)/テストデータの総特徴語数」により特徴語のカバー率を算出した．表 6，7 に示す．表 6，7 より，特徴語のカバー率はドメインに関わらず，訓練データ 1 は 54%程度，訓練データ 2 では 93%程度でった．

5. 考察

5.1 考察 1

評価実験 1 から以下の 4 点を考慮することでより効率的な分類を行えることが明らかになった．

- 分類対象である正事例と負事例の数を統一
- アフィリエイトの有無
- 記事中に用いられる品詞の割合
- アフィリエイトの有無と記事中に用いられる品詞の割合

分類対象である記事のレビュー記事と非レビュー記事の数を統一することで，Accuracy，Precision，Recall の全ての評価値が向上した．これは元々の正事例と負事例の比がおよそ 1:2 であることから，SVM が負事例を重視した分類を行ってしまったと考えられる．

アフィリエイトの有無を考慮することで，全ての評価値が向上した．非レビュー記事の内，アフィリエイト記事が占める割合が高いため，とても有効な指標になったと考えられる．特に Recall がおよそ 15 %向上するという結果が目立っている．Recall は正事例であるレビュー記事をどれだけ見分けられたかという評価値であるため，アフィリエイトリンクが存在しなければレビュー記事である可能性が高いというように SVM が分類を行ったと考えられる．

記事中に用いられる品詞の割合を考慮したことで Recall が 10 %近く上昇したことが特に目立っているが，ここでは特徴語辞書に登録された語だけではレビュー記事であると判断できなかった文章量の少ない記事をレビュー記事と判断するのに役立ったと考えられる．

アフィリエイトの有無と品詞の割合を両方考慮することで，それぞれの属性が異なる記事をレビューと判断するのに役立ち，両方の属性を追加することでさらに精度が向上したと考えられる．

応用実験に関しては，訓練データとドメインが違うにも関わらずレビュー記事と非レビュー記事を分類できた．製品レビューというものは，書き手が製品に対して形容詞や形容動詞などを用いて評価を行っており，製品レビュー全般が同じような形式で書かれているためであると考えられる．また，今回の分類実験では特徴語辞書に名詞を用いなかった．名詞はドメインに依存した単語を多く含むので，名詞を用いずに SVM を学習させたことにより，異なったドメインの製品に対しても同様な分類が行えたと考えられる．

5.2 考察 2

評価実験 2 から以下の 2 点が明らかになった．

- 分類に必要な特徴語を十分に網羅しており，ドメインに依存しない意見文抽出が可能である
- クラスラベル付けの問題を解決することができる

抽出精度は，分類対象のドメインから作成した訓練データを用いた場合と同程度であり，また，Yu らの WSJ 記事の分類とも F-measure で同程度であり，十分な精度を達成することができたと考えられる．また，カバー率の測定において，特徴語を詳しく分析すると，そのドメインに特有な特徴語というのは，テストデータの全特徴語の 2 割程度であると推測され，この 2 割の単語を網羅的に含んでいるかどうかが，分類精度に影響を与えたと思われる．新聞記事はこの 2 割の単語を含んでおり，ドメインに依存せずに分類する能力があると言える．

また，クラスラベルを新聞記事の特徴を利用して付与したが，手動で付与の場合と比べても精度は問題なく，うまく機能したと言える．実際には，クラスラベルの仮定に反する文も少なからず含まれているが，式 (5) のスコアリングにより，その影響を取り除き，特徴語辞書の作成ができた．また，人手でクラスラベルを付与した場合でも，実際にはクラスの分類定義に当てはまらない文が多くあるので，このようなノイズと新聞記事を用いた場合のノイズは同程度だと考えられる．

6. まとめ

本研究では，SVM を用いて Weblog の記事をレビュー記事と非レビュー記事に分類する方法と，新聞記事から抽出した辞書に基づいてレビュー記事から意見文を抽出する方法を述べた．

１つ目に，レビュー記事を得るためにブログ記事の特徴に着目し，記事をレビュー記事と非レビュー記事に分類する手法を提案した．実験結果より，アフィリエイトの有無と品詞の割合を考慮することで，いずれも分類精度を向上させ，特に Recall の上昇が目立つ結果が得られた．また，本研究の分類手法を訓練データに用いた製品とは異なる製品の記事に適用したところ，高い精度で分類できた．従って，提案手法は，精度および汎用性の点で有効である．

2 つ目に，新聞記事の特徴を利用した意見文抽出手法を提案した．新聞記事を用いることで，特徴語辞書のドメイン依存の問題，訓練にかかる時間・コストの問題を解決することができた．また，Recall が平均 84%程度と，意見文を網羅的に抽出できた．文語体で主に事実が書かれる新聞記事でも，ジャンルごとの傾向を利用することで，意見文抽出への利用価値が高いことが示され，今後のさらなる利用が期待される．

参考文献

[立石 01] 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，NL-144-11，pp.75-82，2001.

[藤村 05] 藤村滋，豊田正史，喜連川優．文の構造を考慮した評判抽出手法，電子情報通信学会 第 16 回データ工学ワークショップ (DEWS2005)，2005.3.

[Yu 03] Hong Yu，Vasleios Hatzivassiloglou．Toward Answering Opinion Questions: Separating Fact from Opinions and Identifying the Polarity of Opinion Sentences，Empirical Methods in Natural Language Processing(EMNLP2003)，2003.