内容と文体による五大全国紙の比較分析

狩野恵里奈
東洋大学社会学部

荒川唯
筑波大学大学院
図書館情報メディア研究科

鈴木崇史
東洋大学社会学部

本研究では五大全国紙の社説を研究対象とし，多変量解析と機械学習を用いて分類実験を行った．従来の社説の比較研究は内容分析によるものが多かったが，近年では計算機を用いた分析による研究も盛んに行われている．本研究では，新聞社別，1月／8月社説別，論調別に，内容，文体上の差異を検討するため，主成分分析とランダムフォレスト機械学習法を用いて研究を行った．結果，どの実験においてもそれぞれ新聞社間に違いがあることが観察された．本研究では，新聞の社説，論調を実証的に明らかにすることで，人々のメディアリテラシーの向上，政治や社会の理解に役立つ知見を提出することを目指す．

1. はじめに

自然言語処理の技術が向上し，文学作品の著者推定や執筆時期の推定など，人文，社会科学分野においても計算機を用いた分析手法によって研究が行われるようになってきた[1]．文学作品のみならず，犯罪捜査や市場調査といった実用的な分野においても計量分析の技術は活用され始めている[1;2]．

一方新聞は，発行部数こそ年々減少傾向にあるが[1]，五大紙の東京都内普及率をそれぞれ足し合わせたものは今なお6割ほどあり[2]，まだまだ人々にとって欠かせない情報源の一つであると言える．

新聞社説の分析は，これまで内容分析によるものが一般的であり，阿部[3]のアメリカ同時多発テロを題材にした朝日・毎日・読売三紙の社説の比較や，諏訪ほか[4]の全国五紙に東京新聞を加えた都合六紙の社説を，愛国心などの五つのテーマごとに比較した研究など数多く存在する．近年では計算機を用いた研究も多くなされており，樋口[5]は計算機を用いても従来の手作業による分析と同じ結果が得られるか，また，新聞記事を分析する際に計算機を用いる長所と短所とは何か，これらを検討することを目的とした研究を行っている．

本研究は比較対象を五大全国紙とし，多変量解析と機械学習を用いて社説の比較を新聞社別，1月／8月社説別，論調別に行う．計算機を用いた新聞記事の比較研究は，従来は全語彙を利用するか内容語を利用するのが一般的であったが，本研究では機能語にも重点を置き，内容語と機能語に分けての比較も行う．内容語と機能語に分けて比較分析を行うことで，社説の分析を行う際に内容語と機能語がそれぞれどのように作用しているかを考察していく．本研究では，新聞の社説，論調を実証的に明らかにすることで，人々のメディアリテラシーの向上，政治や社会の理解に役立つ知見を提出することを目指す．

2. データと分析手法

2. 1 データ

研究対象は読売・朝日・毎日・産経・日経の五大全国紙とする．各紙ウェブ上の新聞記事データベースを用いて，2000年から2010年までの1月1日付け社説および8月15日付け社説を収集した．1月1日付けの社説では，全体的な新聞社の違いを，8月15日付けの社説では特定の話題に対する新聞社の違いを分析する目的で，この2つの日付を分析対象とした．分析対象記事数は日経新聞のみ31記事，他四紙は各22記事ずつ収集し延べ119記事である[3]．

[1]一般社団法人日本新聞協会
http://www.pressnet.or.jp
[2]読売新聞広告ガイド
http://adv.yomiuri.co.jp/index.html

[3]日経新聞は8月15日付け社説のみ同日に2つ

2. 2 分析手法

形態素の出現頻度の計量

五紙それぞれの総記事を，MeCab[4]を用いて形態素に分解し形態素ごとの出現頻度を計量した．さらに1月1日付けおよび8月15日付けに分け，日付別では内容語と機能語に分けての計量を行った. 内容語と機能語の種類分けは MeCab の品詞タグを用いた. 今回の実験では機能語を名詞-非自立，名詞-代名詞，接続詞，連体詞，助詞，助動詞，記号とし，それ以外を内容語とした．

主成分分析

テキストを行，形態素の相対頻度を列とする行列から分散共分散行列を計算し，主成分分析[6]を適用した．テキストの位置関係を可視化し，それぞれの特徴や差異を理解するために主成分分析を用いた．

今回の実験では全語彙を利用したもの，内容語のみを利用したもの，機能語のみを利用したものと三種類の特徴量を用いて主成分分析を行った．

ランダムフォレスト機械学習法

分類実験にはランダムフォレスト機械学習法[7]を用いた．まず，テキストを行，形態素の相対頻度を列とする行列を作成し，1000 回ブートストラップを行った. さらに Breiman[7]に従い，それぞれのブートストラップサンプルから，変数の正の平方根をランダムサンプリングで抽出した．ランダムサンプリングの 2/3 を学習に用い，残りの 1/3 を評価に用いた．分類実験の評価は精度，再現率，F_1値を用いた[8].

今回の実験では利用特徴量を，全語彙 (実験 1-3，10)，内容語のみ (実験 4-6，11)，機能語のみ (実験 7-9，12)の 3 パターン用意し，分類クラスを，新聞社別 (5 クラス，実験 1，4，7)，月別 (2 クラス，実験 2，5，8)，記事別 (10 クラス，実験 3，6，9)，論調別 [5](3 クラス，実験 10-12) の 4 種類用意し, 計 12 の実験を行った．

社説を掲載している年があり，それらを別々の記事としてカウントしたため 31 記事となった．

[4] mecab.sourceforge.net

[5] 論調別では産経新聞論説委員室[9]を参考に，朝日・毎日を同グループに，読売・産経を同グループにし，日経は単独とし三つに分けた．

実験の分類性能を確認することで，それぞれの特徴と差異を明らかにし，また，分類に有効な特徴量を抽出することが可能である．

3. 結果と考察

3.1 形態素の出現頻度の計量

表 1 は新聞社別に記事を 1 月 1 日付けと 8 月 15 日付けに分け，延べ語数と異なり語数の平均，標準偏差 (s.d.)，変動係数 (c.v.) をまとめたものである．延べ語数，異なり語数ともに最も多いのは読売新聞の 1 月 1 日付けの社説となる．日経新聞以外はどの新聞社も，8 月 15 日付け社説より 1 月 1 日付け社説の方が異なり語数の平均値が高くなっている．これは 1 月 1 日付けの記事の方が話題性に富み，年によって違った内容の記事を書いているからと推察できる. また，日経新聞のみ8月15日付け社説の異なり語数の平均値が高いのは, 日経新聞の 8 月 15 日付け社説では，社説が二つに分かれ，一方では終戦記念日にまつわる社説を掲載し，もう一方では経済に関する社説を掲載していることがほとんどであるため，使われる語の種類が増えたと推察できる．

表 1　基礎データ

	延べ語数		異なり語数	
	1/1	8/15	1/1	8/15
朝日	14,722	13,722	5,364	5,015
毎日	13,589	14,313	4,975	4,964
日経	11,879	12,625	4,365	5,296
産経	14,015	11,265	5,249	4,526
読売	20,595	13,120	6,510	4,424
平均	14,960	13,009	5,292.6	4,845
s.d.	3,319.45	1,162.98	782.68	362.44
c.v.	0.22	0.09	0.15	0.07

形態素の出現頻度の計量を行った結果，内容語の各紙上位 3 語は，1 月 1 日付けでは朝日：「し」「日本」「する」，毎日：「し」「いる」「する」，日経：「日本」「し」「経済」，産経：「し」「日本」「的」，読売：「し」「する」「的」となる．日経新聞のみ「経済」が 3 位以内に入っているのが特徴的である．8 月 15 日付けでは，朝日：「し」「日本」「戦争」，毎日；「し」「いる」「日本」，日経：「し」「する」「日本」，産経：「し」

「日本」「いる」，読売：「戦争」「責任」「し」となる．産経新聞以外の四紙は「戦争」が 20 位以内に入る．機能語に関しては，各紙とも「の」「、」「。」が上位 3 語を占めていること以外はあまり違いが見受けられない．形態素の出現頻度の計量に関しては，上位にくる形態素は各紙とも似たようなもので，人の目で見て得られる情報は少ないと推察される．

3.2 主成分分析

図 1 は全語彙を利用した主成分分析の結果である[6]．この他に内容語のみを利用したものと機能語のみを利用した分析を行った．図 1，図 2 から機能語のみを利用した図 2 が全語彙を利用した図 1 とほぼ同様の結果だったため，主成分分析では機能語が大きく影響していることが観察される．分析の結果，読売新聞の 1 月 1 日付け社説が図の左側に散らばっている．また，1 月 1 日付けと 8 月 15 日付けが交わらないのも特徴的である．日経新聞も 1 月 1 日付けと 8 月 15 日付けがはっきり分かれており，後者は図の右端に固まっている．産経新聞も，前出の二紙ほどではないが，1 月 1 日付けと 8 月 15 日付けの間に差がみられる．この三紙とは反対に，毎日新聞と朝日新聞の間にはさほど差がない．

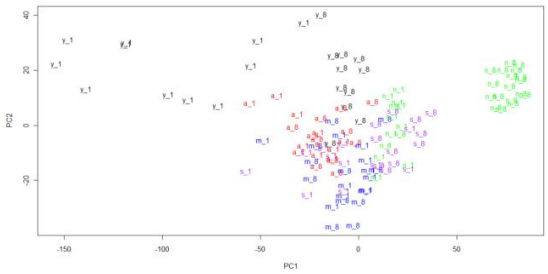

図 1　全特徴量を利用した主成分分析

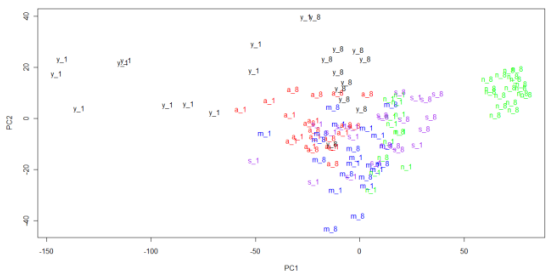

図 2　機能語のみを利用した主成分分析

[6]テキストの位置は新聞社のイニシャル+月であらわされている．

3.3　ランダムフォレストによる分類実験

表 2 は分類実験の評価結果を表す．表 2 から，新聞社ごとの分類においては，内容語を用いた実験 4 よりも機能語を用いた実験 7 の方がより性能が良いため，新聞社間の違いは文体に表れると推察される．しかし記事の内容ごとの分類においては，機能語を用いた実験 9 よりも内容語を用いた実験 6 の方がより性能が良いため，話題性の違いの判別には内容語が重要な役割を果たしていると推察される．

ここからは分類実験の結果について推察する．表 3 は全語彙を用いて新聞社別の分類を行った実験 1 の結果である．産経新聞以外の四紙は 86%以上の割合で本来の新聞社に分類されているが，産経新聞のみ分類性能がやや低い．機能語のみを用いた実験 7 の結果は実験 1 とさほど変化はなく，反対に内容語のみを用いた実験 4 では日経新聞，読売新聞以外の三紙で相互に誤分類される結果となった．このことから新聞社間の違いは文体に表れると推察できる．

次に話題による分類は可能なのかを検証する目的で，月別に分けて実験を行った．表 4 は全語彙を用いて月別に分類を行った実験 3 の結果である．結果，1 月 1 日付けはどの特徴量を用いても結果に大きな変化は見られなかった．一方 8 月 15 日付けの結果は，全語彙を利用した実験と内容語を利用した実験では高い性能で分類されたにも関わらず，機能語を利用した実験では性能が大幅に落ちた．1 月 1 日付けは幅広い話題を提供しているが，8 月 15 日付けでは話題が戦争に集中するため，文体よりも内容に特徴が表れたと推察される．

表 5 は全語彙を用いて論調別の 3 クラスによる分類実験を行った結果である．こちらも 3 パターンの特徴量を用いて実験を行った．結果，朝日・毎日グループの分類が，どの特徴量を用いた場合でも 90%以上の割合で同グループに分類された．読売・産経グループは 75～88%の割合で同グループに分類され，朝日・毎日グループの結果よりは低い性能を示した．このことから，朝日新聞と毎日新聞は両紙間の距離が近いと推察される．五紙での全語彙を利用した実験 1 の分類結果を見てみても，朝日新聞・毎日新聞ともに産経新聞と読売新聞には一つも分類されていないため，この 2 グループの間には違い

があると推察される．

表 2　分類実験結果（%）

	精度	再現率	F_1値
実験 1	88.8651	87.3771	87.7476
実験 2	68.0151	67.4064	66.3599
実験 3	92.2782	92.0245	92.1181
実験 4	74.3694	66.0288	65.8772
実験 5	—	46.7018	—
実験 6	89.7634	88.5080	88.8048
実験 7	87.8092	87.3676	87.4243
実験 8	67.5203	66.9629	—
実験 9	78.9182	79.0275	78.9188
実験 10	91.1990	89.3767	89.8415
実験 11	83.2952	81.7645	81.9750
実験 12	92.9771	91.3517	91.9013

表 3　実験 1 分類結果（新聞社別，全語彙）

	朝日	毎日	日経	産経	読売
朝日	20	2	0	0	0
毎日	2	19	1	0	0
日経	0	0	30	1	0
産経	2	1	3	16	0
読売	0	1	0	1	20

表 4　実験 3 の分類結果（月別，全語彙）

	1 月	8 月
1 月	47	8
8 月	3	61

表 5　実験 10 の分類結果（論調別，全語彙）

	朝日･毎日	日経	産経･読売
朝日･毎日	43	0	1
日経	2	28	1
産経･読売	8	0	36

4. おわりに

本研究は新聞社別，1 月／8 月社説別，論調別に，内容，文体上の差異を検討するため，多変量解析と機械学習を用いて五大紙の比較を探索的に行った研究である．五大全国紙の 1 月 1 日および 8 月 15 日付け社説を対象に，単語の出現頻度の計量，主成分分析，ランダムフォレスト機械学習法による分類実験を行った．今回の実験では，新聞社ごとにそれぞれ違いがあることが示された．また，新聞社間の違いは文体に表れること，話題の違いには内容語が大きく影響を及ぼしていることなどの知見を示すことができた．

今後は，分析対象となる記事数を増やすこと，話題の種類を増やすこと，なぜこのような結果になったのかの原因を追求することを課題とし，引き続き研究を進めていく．

謝辞

本研究は，科学研究費基金助成金若手研究(B)「計算文体論による多種メディアテキスト解析（研究代表者：鈴木崇史，研究課題番号：23700288）」より，一部支援を受けています．ここに記して謝意を表します．

文献

[1] 村上征勝．シェークスピアは誰ですか？-計量文献学の世界，文春新書，東京，2004.

[2] Argamon, S., et al. Stylistic text classification using functional lexical features, *Journal of the American Society for Information Science and Technology*, 58(6), 802–822, 2007.

[3] 阿部康人．9・11 事件以降の『朝日新聞』『毎日新聞』『読売新聞』の一考察–『朝日新聞』『毎日新聞』『読売新聞』の社説を題材に，新聞学，(19)，18-76，2004.

[4] 諏訪哲二・森永卓郎・戸高一成・長山靖生・桜井裕子・ラクレ編集部（編）．社説対決五番勝負，中央公論新社，東京，2007.

[5] 樋口耕一．計算機による新聞記事の計量的分析，理論と方法，(19)，161-176，2004.

[6] 加納学．主成分分析，京都大学大学院工学研究科化学工学専攻プロセスシステム工学研究室，京都，2002.

[7] Breiman L. Random forests, *Machine Learning*, 45, 5-23, 2001.

[8] 徳永健伸．情報検索と言語処理 (言語と計算)，東京大学出版会，東京，1999.

[9] 産経新聞論説委員室．社説の大研究 - 新聞はこんなに違う！，扶桑社，東京，2002.