文体からみた学術的文献の特徴分析

石田栄美(駿河台大学)* 安形輝(亜細亜大学) 野末道子(鉄道総合技術研究所)
久野高志(作新学院大学) 池内淳(大東文化大学) 上田修一(慶應義塾大学)
*e-mail：emi@surugadai.ac.jp

1. はじめに

現在，ウェブ上には，電子ジャーナルに収載される他に多数の学術論文を掲載するウェブページが存在する。2001年に得た約1,100万の日本のウェブページから抽出した 300 件の標本を対象に調査したところ，12 件の学術的文献が含まれていた。単純に全体に当てはめれば約 44 万件の学術的文献がウェブ上に存在することになる。国内では，申告制，登録制の論文アーカイブはあまり機能していない。そこで，これらを組織化する一方，長期保存する方策を考える必要がある。

本研究は，こうした意図のもと，学術的文献を自動判定する方法を考案するための基礎的な調査を行うものである。

ウェブページの中で，テキストを主体とするページの判定は容易であるが，テキスト主体のページ群の中から学術論文を選別するにはいくつかの方法が考えられる。例えば，ウェブページの構造やファイル識別子を判別の手がかりとすることなどであるが，いずれも確実ではない。

そこで，文体に基づく判定を試みる。文体とは，「ことばづかいから見た文章の体裁」[1]であるが，文体的指標という観点から見た場合，大きく定量的指標と定性的指標に分けられる。文体の統計的な研究は，計量文体学，計量文献学と呼ばれ，一般に 1851 年にオーガスタ・ド・モルガンによる使徒パウロの著作の判定に語の長さを用いることができるという示唆に始まるとされる[2]。計量文体学は，著者推定の手段として，主として聖書や文学を対象として行われてきた。

計量文体学のいくつかの教科書 [2][3]や総説 [4][5]，研究 [6][7]に示された文体的指標を表 1 に示した。ここでは，大きく，量，構文，位置，表現，内容に関する指標に分類した。多変量解析を目的とした研究では，数多くの文体的指標が使用されているものの，文体的特徴を表す際に基盤となるのは，文の長さ，単語の長さ，単語の出現率であることは明白である。また，日本語を対象とする場合には，文頭や文末の表現に特徴を求めようとする傾向がある。例えば，日本語のウェブページを対象に，文末表現とページジャンルとの関係を調査した例[8]などがある。

著者推定の場合には，文体の個人差（癖）を強調する指標が重視され，それらの指標の開発と判別のための統計的手法が研究課題とされてきた。

本研究では，一定の特徴を持ったテキスト（学術論文）を他のグループから分離することを目的とする。まず学術論文の文体的特徴を明らかにするために，学術論文と，ウェブ上で比較的多い日記，およびウェブ上にあり文体的特徴が明確と考えられる新聞記事を対比させた。

2.文体的特徴の分析方法

2.1 特徴分析の概要

文体的特徴を明らかにするために,既往研究で用いられている指標のいくつかを，学術論文，日記,新聞記事に対して調査した。既往研究の中には，手作業で調査した指標もある。本研究では，自動判定できる特徴を見つけだすことが目的であるので,自動抽出できる指標だけを用いた。具体的には，文の長さ，単語数，文字種の比率，品詞の比率，文頭・文末表現,各品詞で用いられている表現などである。

表1 既往研究で扱われている文体的指標

		指標	ケニィ	村上	ホームズ	吉岡	安本	陳
			1982	1994	1994	1996	1977	2003
量	長さ	文の長さ	○	○	○	○	○	○
		単語の長さ	○	○	○	○	○	○
		音節			○			
	生起回数	単語の出現率	○	○	○	○		
		同義語		○				
		異なり語		○	○	○		
		漢字					○	○
		名詞					○	○
		接続詞						○
		接続助詞						○
		四字熟語						○
		人格語					○	○
		多出語		○		○		
		句点					○	○
		読点		○				
構文		主語, 述語, 修飾語などの構文に関する情報		○				
位置	文頭	文頭に置かれる単語や品詞の出現率		○				
	文中	読点の位置		○				
	文末	文末に置かれる単語や品詞の出現率		○				
		過去止					○	○
		現在止					○	○
		不定止					○	○
表現		直喩					○	○
		声喩					○	○
		色彩語					○	○
		会話文					○	○
内容		話題						○
		引用						○

分析では,文字種の比率,文頭や文末表現など文字列そのままを扱う場合と形態素解析システムChaSen[9]を用いて切り出された単語を扱う場合がある。ChaSen で切り出された単語には,それぞれ品詞が付与されており，品詞の比率にはこの結果を用いた。

2.2 テキスト集合の作成

学術論文，ウェブ上の日記，新聞記事のテキスト集合を以下の方法で作成し，これらを用いて文体の特徴を分析した。なお，それぞれのテキスト集合の件数が異なるが，これはテキスト 1 件あたりの長さが異なるため，全体でほぼ同じテキスト量になるよう調整したためである。

まず，学術論文，日記，新聞記事のテキスト集合を収集した。学術論文は PDF ファイルで全文が提供されていたものをテキストファイルに変換し，90 件を収集した。その際，分野の偏りをなくするために，人工知能，計算工学，昆虫学分野など 11 分野の論文を収集した。ウェブ上の日記は，日記を公開しているサイトである「はてなダイアリー」[10]から 1 ファイルを 1 件とみなし，2,700 件をランダムに収集した。1 ファイルには複数の日付の日記が含まれている場合が多い。新聞記事は，12 万件の記事が収録されている『CD-毎日新聞'98』からランダムに 1,800 件を選択した。

この中で，学術論文ではタイトル，著者名などを含めた全文，日記では日記本

文とコメント部分，新聞記事は見出し，リード，本文を用いた。

3． 文体的特徴

3.1 基本的なデータ

基本的なデータとして，論文，新聞記事，日記の文の長さなどを表 2，文字種の比率を表 3,品詞の比率を表 4 に示す。

1 件あたりの文の数は論文が圧倒的に多いが，これは全文を対象としているためである。一文の長さは新聞が最も長い。文字種の比率をみると，論文はその他の割合が大きく，カタカナやひらがなの割合が小さい。品詞の比率をみると，論文は助詞，動詞，助動詞の割合が小さく，記号の割合が大きい。記号にはアルファベットも含まれているので，式やアルファベットを含んでいる論文が多いと考えられる。新聞は，名詞の割合が大きく，未知語の割合が小さい。日記は，助動詞，未知語，形容詞，副詞の割合が多い。日記に未知語が多いのは，定型的な表現がないため，形態素解析システムで単語として切り出せないものが多いということが考えられる。

品詞の比率に関しては，論文と新聞は一つの品詞が占める割合が大きく，日記は偏りが小さいという特徴があった。

3.2 文末・文頭表現

用いられている頻度が高い上位 10 の文末表現を表 5 に，文頭表現を表 6 に示した。文末とは，「。」「.」や「?」「!」，改行が 2 行以上続くものとした。

表 5 から，日記は「りました。」「しました。」などの過去形やですます調が多いことがわかる。一方，論文と新聞はである調が多い。また，「を行った。」「がわかる。」などは論文で用いられる特徴的な表現であるといえよう。

表 6 からは，「すなわち，」「本研究では」「その結果，」などは論文特有の表現といえる。

文頭・文末表現に関しては，論文特有の表現がいくつかあることがわかった。

表2 文の数と長さ・単語数

	文数/一件	長さ/一文	単語数/一件
論文	239.0	71.7	3706.3
新聞	14.1	85.6	287.3
日記	93.7	51.6	1085.2

表3 文字種の比率

	カタカナ		ひらがな		漢字		半角カナ／英		その他	
論文	226.1	3.0%	1590.6	21.4%	1511.5	20.3%	378.0	5.1%	3726.2	50.1%
新聞	16.4	4.6%	111.5	31.0%	123.7	34.4%	35.6	9.9%	72.7	20.2%
日記	94.2	5.3%	505.0	28.6%	425.1	24.1%	60.9	3.4%	681.9	38.6%

表4 品詞の比率

	記号		名詞		助詞		動詞		助動詞	
論文	437,430	46.2%	286,276	30.3%	115,225	12.2%	50,308	5.3%	21,952	2.3%
新聞	130,535	15.5%	391,272	46.6%	176,494	21.0%	74,279	8.8%	39,469	4.7%
日記	1,134,843	21.0%	1,511,326	27.9%	1,157,849	21.4%	607,637	11.2%	441,688	8.2%

	未知語		接頭詞		接続詞		連体詞		形容詞	
論文	14,606	1.5%	5,172	0.5%	4,538	0.5%	3,769	0.4%	3,609	0.4%
新聞	3,775	0.4%	8,030	1.0%	1,938	0.2%	2,515	0.3%	5,723	0.7%
日記	238,270	4.4%	26,508	0.5%	33,222	0.6%	32,258	0.6%	87,538	1.6%

	副詞		感動詞		その他		フィラー		合計	
論文	2,920	0.3%	187	0.0%	68	0.0%	31	0.0%	946,091	100.0%
新聞	5,904	0.7%	294	0.0%	0	0.0%	99	0.0%	840,327	100.0%
日記	114,083	2.1%	22,138	0.4%	390	0.0%	5,292	0.1%	5,413,042	100.0%

3.3 接続詞

文頭表現からもわかるように接続詞の表現にも特徴的な傾向があるのではないかと考え，接続詞とされた単語のうち，用いられている頻度が高い表現を表 7 に示した。「また」や「しかし」など論文，新聞記事，日記に共通する表現も多いが，「すなわち」「あるいは」「など」は論文に頻繁に用いられる表現といえよう。

表5 用いられている頻度が高い文末表現(末尾5文字)

論文			新聞			日記		
文末表現	頻度	割合	文末表現	頻度	割合	文末表現	頻度	割合
している。	521	4.13%	している。	633	2.39%	りました。	1737	1.10%
えられる。	397	3.28%	っている。	381	1.47%	きました。	1678	1.07%
れている。	306	2.62%	れている。	265	1.04%	しました。	1399	0.91%
ができる。	283	2.48%	なかった。	246	0.98%	ています。	1294	0.84%
っている。	191	1.72%	発表した。	234	0.94%	なかった。	849	0.56%
であった。	149	1.36%	になった。	232	0.94%	てました。	848	0.56%
のである。	128	1.19%	たという	168	0.69%	そうです。	791	0.53%
なかった。	127	1.19%	となった。	163	0.67%	りません。	752	0.50%
を行った。	113	1.07%	かにした。	144	0.60%	でしょう。	735	0.49%
がわかる。	104	1.00%	していた。	138	0.57%	あります。	726	0.49%

表6 用いられている頻度が高い文頭表現(文頭5文字)

論文			新聞			日記		
文末表現	頻度	割合	文末表現	頻度	割合	文末表現	頻度	割合
したがって	120	12.36%	写真説明	140	0.53%	そういえば	301	0.36%
すなわち，	98	10.09%	葬儀 告別	81	0.31%	というわけ	276	0.33%
このような	86	8.86%	問い合わせ	62	0.23%	というか、	150	0.18%
このように	82	8.44%	……	60	0.23%	っていうか	139	0.17%
しかしなが	70	7.21%	▼シント	44	0.17%	そんなこと	134	0.16%
本研究では	50	5.15%	みんなの	43	0.16%	------	128	0.15%
その結果、	43	4.43%	喪主は長男	40	0.15%	もちろん、	110	0.13%
このため、	41	4.22%	このため、	37	0.14%	·····	98	0.12%
このとき，	41	4.22%	ところが、	33	0.12%	そんなわけ	97	0.12%
本論文では	40	4.12%	これに対し	33	0.12%	といっても	80	0.10%

表7 用いられている頻度が高い接続詞

論文			新聞			日記		
接続詞	頻度	割合	接続詞	頻度	割合	接続詞	頻度	割合
また	925	20.38%	しかし	409	21.10%	でも	4108	#####
および	816	17.98%	また	328	16.92%	そして	2650	7.98%
一方	275	6.06%	ただ	119	6.14%	しかし	2338	7.04%
しかし	249	5.49%	だが	103	5.31%	また	2307	6.94%
例えば	212	4.67%	そして	88	4.54%	で	2020	6.08%
なお	204	4.50%	一方	84	4.33%	しかも	1267	3.81%
すなわち	194	4.28%	でも	74	3.82%	ちなみに	1138	3.43%
あるいは	189	4.16%	例えば	52	2.68%	いや	1105	3.33%
そこで	180	3.97%	ところが	50	2.58%	ただ	1104	3.32%
ただし	136	3.00%	そこで	44	2.27%	だから	890	2.68%

4．おわりに

学術論文，新聞記事，日記の文体的特徴を，各指標を用いて示した。論文の特徴を明確に示す指標はなかったが，文末・文頭表現，接続詞で特有の表現が多く用いられていることがわかった。

分析結果から，一つの指標だけでは学術論文の自動判定は困難であるが，これら複数の指標を組み合わせることで可能だといえる。今後は，実際にウェブ上の学術的文献の自動判定を目指す。

引用文献

1) 日本国語大辞典第二版編集委員会．日本国語大辞典第二版．東京，小学館，2000-2002.
2) アンソニー・ケニィ．文章の計量：文学研究のための計量文体学入門．吉岡健一訳．東京，南雲堂，1996. 244p.（原著は1982 年）
3) 村上征勝．真贋の科学：計量文献学入門．東京，朝倉書店，1994. 154p.
4) Holmes, D. I. Authorship Attribution. Computers and Humanities. Vol.28, p.87-106, 1994.
5) 吉岡健一．計量文体学研究の展望．in: アンソニー・ケニィ．文章の計量．東京，南雲堂，1996. p.196-234
6) 安本美典．語彙の量的構造．数理科学．Vol.15, No.6, p.44-49(1977)
7) 陳志文．新聞の各紙面に見られる文体の類型：主成分分析法による朝日新聞と読売新聞の分析から．国語学研究．No.42, p.54-44(2003).
8) 土井晃一．文末態度表現に注目した Web Page の調査．情報処理学会研究報告自然言語処理．No.130-7, p.49-56(1999).
9) はてなダイアリー，http://d.hatena.ne.jp/ [2004.09.19]
10) 形態素解析システム　ChaSen version 2.3.3 http://chasen.naist.jp/hiki/ChaSen/ [2004.09.19]