4N－2　文末態度表現を用いた文書分類のためのWeb Pageにおける文末態度表現の使われ方の調査*

土井 晃一†
学術情報センター
doy@rd.nacsis.ac.jp

1　はじめに

近年、インターネットによる検索が盛んに行われるようになった。しかし、検索エンジンによる検索は、しばしば適切な結果をもたらさない。一般の検索エンジンは、いわゆるキーワードを主体とした検索やディレクトリ・サービスなどを主体としている。普通、付属語 (助詞・助動詞など) は不要語として無視しているようである。本論文では、普通検索には使われない付属語、その中でも特に文末態度表現 (表 2、表 3、表 4 にあるような、文末にある書き手の態度を表す語) に話を絞る。

文末態度表現は文体と関連する。文体は文章のジャンルと関連する。つまり、文末態度表現を研究することで文章のジャンルがわかる可能性がある。

我々は、文章のジャンルとして日記などの個人の主観的な情報に着目した。これらは個人の主観的な情報が数多く盛り込まれている。例えば、何かの商品のことを知りたいと思ったときに、その商品の販売元の home page を見れば、スペックなどの公式の情報は得られるが、その商品の便利さ、有用性など使った人にしたわからない情報は得がたい。日記などの web page には後者の情報があふれているので役立つ。日記などの　web page を検索できるようになることは、このような点からも重要と考えられる。

また、逆に、百科事典的な情報を得たいのに、日記のようなものばかり出てきて困るというときフィルターとして利用することが考えられる。

本論文では、文末態度表現による web page の調査を行い、さらに日記などの主観的データの比率を求めた。

2　測定環境

以下の測定は Livelink search 検索エンジンを用いて行なった。検索対象は ne ドメインの web page と image data からなる web page を除く、日本の web page である。対象となる web page の数は 608,983 ページである。

これらの web page を　nkf v1.7 を用いて、jis、sjis、MIME code　の file を euc へ変換した。

文末記号としては、「。」/「.」/「?」/「!」各々の全角と半角 (ただし半角の「。」は nkf によって全角の「。」に変換されているため存在しない。現実には、page の decode に失敗したものがあるので存在はする。しかし、以下の測定には影響は無かった。) を採用した。

3　統計解析

日本語の文末態度表現の洗い出しは、あ。、い。... のように「あいうえお」順で page を表示し、その page を人間が見ることによって、文末態度表現を洗い出した。さらに、が。ぱ。などの濁音・半濁音・促音についても行った。さらに、日本語の文法書 [1] のムードの章から文末態度表現を洗い出した [2]。その結果、全部で 310 の文末態度表現を取り出した。

その各々の文末態度表現に関して、

1. その文末態度表現が現れるページ数 (以下「ページ数」と呼ぶ)
2. その文末態度表現の総数 (以下「ワード数」と呼ぶ)
3. その文末態度表現が現れるページでの文末記号の数 (以下「まるの数」と呼ぶ)

を計数した。次に各々の比を取り、基礎統計量を出した。結果を表 1 に示す。ここで、x_1 は、一ページ当たりの文末表現の数、つまり、(ワード数/ページ数) である。x_2 は一ページ当たりの文の数、つまり、(まるの数/ページ数) である。x_3 は文の中で文末表現が占める割合、つまり、(ワード数/まるの数) である。この基礎統計量からわかる事をまとめておく。歪度は 0 に近いと正規分布に近くなる。いずれも正の値なので、すそが右に延びていることを意味している。これは、一方では、頻度が非常に高いもの (例えば、ます。、た。) などが少数存在し、一方では、頻度が低いもの (例えば、なきゃあ。、とってきてやあ。) が多数存在し、頻度が低いものが特異な分布をするためと考えられる。尖度は 3 に近いと正規分布に近くなる。いずれも 3 よりもはるかい大きいのですそが長いことを意味している。この理由も前述と同じ理由によるものと考えられる。

また、これらの文末態度表現でヒットしたページ数は、594,515 ページであった。これは、全ページ数の 97.62%におよぶ。これらの文末態度表現で大半の web page がカバーできることがわかった。

次に、表 1 の三変量を基に因子分析を行った。その結果、因子分析は役立つ (.543)、変数の間に有意な関係がある ($p = .000 < 1\%$) ことがわかった。また、主成分は一つだけ抽出され、その累積パーセントは 58.61% となった。また、相関係数はいずれも、$p = .000 < 1\%$ で有意であった。式は $z_1 = .443x_1 - .353x_2 + .498x_3$ となった。このことから、x_1, x_2, x_3 は相関が高く、新たな文末態度表現も同様な式が当てはまると考えられる。

次に、できあがったデータを x_1、x_2、x_3 をキーにしてソートしてみた。ページ数・ワード数ともに小さいため、データとして意味を持たないものを除くと x_1 の大きい側、x_2 の 11 番目から、x_3 の大きい側が有意であった。その結果を各々表 2、表 3、表 4 に示す。

表 2 を見ると、一度使われると何度も使われる文末態度表現であることは容易に想像が付く。表 3 を見ると、ページ数、ワード数共に大きく、長い文章ではよく使われる文末態度表現であることがわかる。戻る。は web page 特有の文末態度表現である。リンク元に戻るリンクとして記述されているものである。表 4 を見ると、文末態度表現として比率が高いものが示されており、いずれもよく使われる文末態度表現である。

4　日記などの比率

また、web page の中で日記などの占める割合を測定してみた。測定には収集した全 web page からランダムに web page を取りだし、日記などかどうかを人手で判定するという方法を取った。比率検定であるので、信頼係数 95%の区間推定の誤差を 5%に抑えるためには、サンプル数は 385 ページである。この 385 ページの web page について、二段階の基準を設け、日記などであるかどうかを判定した。基準として採用したのは、

狭い基準　人間が見て、主観的な情報であることがはっきりわかるもの

広い基準　人間が見て、主観的な情報であると思われるもの、あるいは、一回リンクをたどることにより主観的な情報に到達できるもの

とした。公式な主観情報 (公式ページで主観的であるかのように書かれたページ) は、日記などとは判定しなかった。また、この基準による判定の確信度は 90%以上である。

測定の結果、

狭い基準　$7.79\% \pm 5\% (2.79\% < p < 12.79\%,$ 棄却水準 5%)

広い基準　$24.42\% \pm 5\% (19.42\% < p < 29.42\%,$ 棄却水準 5%)

であった。これをページ数に換算すると、

狭い基準　$604568 * .0779 = 47096 (16867 < N < 77324)$

広い基準　$604568 * .2442 = 147636 (117407 < N < 177864)$

であった (単位はページ)。

ロボットで収集すると、更新が頻繁なためロボットが別の page と思ったり、リンクが循環していたりして、ページが重複するものだが、本収集セットでも約 10%の重複があった。日記などのページに対する影響は、サンプルセットを見る限り見られなかったので、0.9 で割ることにより補正すると、

狭い基準　$8.66\% \pm 5\% (3.66\% < p < 13.66\%,$ 棄却水準 5%)

広い基準　$27.13\% \pm 5\% (22.13\% < p < 32.13\%,$ 棄却水準 5%)

であった。これをページ数に換算すると、

狭い基準　$604568 * .08656 = 52331 (22103 < N < 82560)$

広い基準　$604568 * .27133 = 164037 (133809 < N < 194266)$

であった (単位はページ)。

5　おわりに

本稿では、web page での文末態度表現の使われ方の調査と、日記などの主観的データの web page での比率を調査した。主成分分析により、文末態度表現は一定の傾向を示すことがわかった。また、日記などの比率は、狭く取って 8.656%、広く取って 27.133%であることがわかった。今後は、いよいよ文書分類に取り組みたい。

参考文献

[1] 益岡 隆志・田窪 行則著.「基礎日本語文法 -改定版-」. くろしお出版, 1995.

[2] 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. 自然言語処理研究会, 1999.

*A Research about Attitude Expressions at the ends Sentences on Web Pages for Document Classification

†(株) 富士通研究所からの客員助教授

	一ページ当たりの文末表現の数 (ワード数/ページ数)(x_1)	一ページ当たりの文の数 (まるの数/ページ数)(x_2)	文の中で文末表現が占める割合 (ワード数/まるの数)(x_3)
平均値	1.57	440.89	6.00E-03
中央値	1.33	362.37	4.85E-03
歪度	3.08	2.92	6.50
尖度	12.52	11.69	58.07
範囲	5.07	2359.00	.10
最小値	1.00	10.00	.00
最大値	6.07	2369.00	.10

表 1: 基礎統計量

文末表現	ページ数 (B)	ワード数 (C)	まるの数 (D)	$x_1 = C/B$	$x_2 = D/B$	$x_3 = C/D$
です。	183179	943485	21464684	5.15	117.18	0.0440
である。	33617	183048	4287467	5.45	127.54	0.0427
ます。	283499	1637946	29567248	5.78	104.29	0.0554
にゃあ。	28	164	19051	5.86	680.39	0.0086
た。	164190	996301	18950402	6.07	115.42	0.0526

表 2: x_1 の大きい順

文末表現	ページ数 (B)	ワード数 (C)	まるの数 (D)	$x_1 = C/B$	$x_2 = D/B$	$x_3 = C/D$
ます。	283499	1637946	29567248	5.78	104.29	0.0554
ください。	124216	229257	13434984	1.85	108.16	0.0171
た。	164190	996301	18950402	6.07	115.42	0.0526
戻る。	4296	5920	497343	1.38	115.77	0.0119
です。	183179	943485	21464684	5.15	117.18	0.0440

表 3: x_2 の 11 番目から

文末表現	ページ数 (B)	ワード数 (C)	まるの数 (D)	$x_1 = C/B$	$x_2 = D/B$	$x_3 = C/D$
する。	44324	151114	6607120	3.41	149.06	0.0229
である。	33617	183048	4287467	5.45	127.54	0.0427
です。	183179	943485	21464684	5.15	117.18	0.0440
た。	164190	996301	18950402	6.07	115.42	0.0526
ます。	283499	1637946	29567248	5.78	104.29	0.0554

表 4: x_3 の大きい順