5S-7

LSI 文書検索精度に応じた特徴語抽出

桜井　敬介†　　三浦　孝夫‡

法政大学　工学部　情報電気電子工学科

東京都小金井市梶野町 3-7-2

1.前書き

近年、情報源の一つとして Blog(Weblog)がある。Blog にはファッションや化粧品、ゲームなど様々な商品や物に対する主観的な意見や評価が多く書かれており、流行を捉えることができ、商品を選択するときの判断基準になっている。

企業側が自社商品の販売のために、人気のある著者になりすまし、消費者の購買意欲をかきたてる操作をすることも考えられる。そのため、本当に Blog の記事をその著者かどうかを確かめる必要がある。また、Twitter 上で他人になりすまして情報を発信する問題も起きているため、著者を推定することが重要である。このような情報の文章は著者の文章構成に大きく依存するため、これを考慮した処理が必要である。

〈表１　実際の Blog 記事〉

前バオバブ で念願のテレビを落札できたので次はパラディソ でもやってみようかなと思い、出品されてる商品をみてみる・・・・・| 壁 |д･)
そしたら前見た時は出品されてなかったインテリアが新規参入していたっ！！(*_*) しかもテレビ台があるじゃないっ！！！（笑） ペニーオークションユーザのひとりごと

表１は本実験データとして用いる Blog 記事の一例である。著者によって、単語数は少ないが、顔文字や記号などで表現をする文章構成もある。また、人が書く文章には機能語（、　。「」（）！　？　・・・　～　など）の使い方に特徴が現れることがよくある。本稿では機能語の有用性を検証するために、名詞を特徴語とする文書行列を用い、機能語を特徴語とした文書行列と推定結果を比較する。

2.潜在的意味索引付け

本研究では、前章で述べた理由から文書中の機能語に着目した著者推定方法を提案する。　機能語を特徴語とした文書行列を用い、この行列を分類することにより、著者推定を行う。

潜在的意味索引付け（Latent Semantic Indexing, 以下 LSI）を用いて、次元縮小を行うことにより、文書行列を特徴語と概念の関係に変換し、概念と文書間の関係に変換する。これより特徴語と文書や、文書同士を概念で関係付け、概念による文書比較が期待できる。

テキストデータの次元縮小・概念付けをするために、LSI では特異値分解（SVD）により次元縮小のための射影行列を求める[2]。特異値分解を使うと単語文書行列は、２つの直交行列と対角行列の積で表せる。対角成分(特異値)の大きい k 個を選び残りを無視することで、特徴つけに効果的な主要 k 次元に縮小することができる。

3.実験準備

本実験で用いる Blog 記事は Ameba より予め著者が判明している 20 人の Blog 記事 863 件を用いる。記事内に広告などがある場合は純粋に著者が書いた文章のみをデータとして用いるため削除している。これらを正解として再現率 R、適合率 P を用い、評価法として F 尺度を用いる。　F 尺度は 2PxR/(P+R) と定義され、再現率と適合率が共に値が大きいときに、値が大きくなるため F 値は大きいほどよい。

文書行列の特徴語として、形態素解析茶筅を用い記号と未知語のみを抽出し特徴語とし、重み付けに TF*IDF を用いた文書行列とする[2]。次元縮小をしていない文書行列と各次元において縮小した文書行列に対し、k-means 法を用いてクラスに分類する。分類した各クラスを次元別に、予め人為的に著者別に記事を分類した正解クラスと総当りで F 値を用いて比較する。各正解クラスをもとにして最も F 値の高いペアを取り出し、各正解クラスの F 値の平均をその次元の F 値として用い、評価の基準とする。

〈表２　記号・未知語（一部）〉

記号『,』,！, (,), ・,゜,。,、, ,…, ￣
未知語 ≧,д,≦,∀,www, Kra, orz , レフティ,ガチ

表２は実際に茶筅を用いて Blog 記事を品詞分解したときの記号と未知語の例である。表２から分かるように顔文字に使われる記号や、“ガチ”や“orz”などの語も含まれる。これらは Blog の著者の文章構成の特徴であり、新聞記事などとは異なり Blog 記事特有の単語である。これらの単語を特徴語として要素に取り込むことは Blog 記事を解析するてがかりとなる。

4.実験の結果と考察

表３は機能語を特徴語とした文書行列に対する次元縮小をしたときの F 値の平均値を表している。

なお、2263 は機能語見出し語数（次元縮小しない文書行列の次元）である。表３より一番 F 値の平均値が高い次元は 50 である。これは次元縮小をしていない 2263 次元のときより高い F 値を示している。この結果より、LSI を用いて 50 次元に縮小し文書同士を概念で関連付けることにより分類精度が向上することを表す。また、次元を縮小することにより負担軽減、計算効率の向上が図れる。

〈表３　機能語を特徴語としたF値の平均〉

次元	2263	700	600	500	400	300	200	100	90
再現率(%)	68.4	67.9	61.1	55.1	58.5	64.5	64.2	69	60
適合率(%)	38.2	36.2	36.6	49.5	46.6	46.8	38.4	37	51.2
F値(%)	35	35.2	33.2	36.6	36.9	38.3	37.6	39.6	41.4
次元	80	70	60	50	40	30	20	10	
再現率(%)	56.3	60	57.6	62.1	55.5	52	55	54.1	
適合率(%)	46.3	45.7	51.9	54.8	47.1	44.6	42.8	37.6	
F値(%)	40.1	38.9	41.2	42	41.7	39.3	36.1	31.5	

〈表４　名詞を特徴語としたF値の平均〉

次元	9490	700	600	500	400	300	200	100	90
再現率(%)	86.1	84	78.2	73.9	74.4	78.2	69.7	66.5	67.2
適合率(%)	19.4	17.7	20.9	21.2	23.4	26	26.3	31.3	38.2
F値(%)	14	13.9	17.2	17.9	20.6	21.8	22.1	24.8	24.9
次元	80	70	60	50	40	30	20	10	
再現率(%)	68.1	56.7	65.1	67.1	68.5	62.4	67.5	56.3	
適合率(%)	37.8	38.6	41.3	34	39.1	31.9	28.1	25.1	
F値(%)	25	24.2	25.5	25	25.8	24	23.9	24.5	

表４は名詞を特徴語とした文書行列に対する次元縮小をしたときの F 値の平均値を表している。名詞を特徴語とした文書行列の元の文書データは、機能語の文書データと同じものを使用し、茶筅により名詞のみを抽出したものであり、機能語と同様の操作を行う。なお、9490 は名詞見出し語数（文書行列の次元）である。

表３と表４を比較すると各次元において表３の F 値のほうが高い。これより Blog 記事の分類には機能語を特徴語とする方が効果的である。機能語の使い方は著者により特徴があり、著者推定の際に機能語を分類の要素として取り入れることは効果的であることがわかる。

表５は機能語を特徴語とする著者別の再現率適合率の表である。再現率・適合率が共に高い著者 14 の記事を抜粋した例を表６に示す。

〈表５　著者別の再現率、適合率〉

	著者1	著者2	著者3	著者4	著者5	著者6	著者7	著者8	著者9	著者10
再現率(%)	44.1	59.6	43.9	56.2	92.6	62.2	77.5	94.5	19.6	38.4
適合率(%)	36.5	50	32.3	6.2	13.4	78.6	36.2	4.8	78.9	100
F値(%)	39.9	54.3	37.2	11.1	23.4	69.4	49.3	9.1	31.3	55.4
	著者11	著者12	著者13	著者14	著者15	著者16	著者17	著者18	著者19	著者20
再現率(%)	68.9	63.6	20.8	94.6	28.9	86.7	51.5	63.8	37.6	96.1
適合率(%)	17.9	96.5	26.4	98.6	18.8	19.2	39.8	53.5	55.9	10.3
F値(%)	28.4	76.6	23.2	96.5	22.7	31.4	44.9	58.1	44.9	18.6

〈表６　再現率適合率の高い記事〉

☆★☆★☆★☆★☆
余談…
中途半端な文章だけど、長くなったので次回に続きます。
2 回に分けるほどの内容じゃないけど、直感力を高めるトレーニングも書かないとただの雑学…
それにしても俺のブログ、パソコンからだと、恐ろしく読みにくいんですね
なんであんなに改行されちゃうのΣ(￣□￣;)

この多くに“Σ(￣□￣;)”などの顔文字や“★☆”の機能語が多く生じる。これらの機能語が他にはあまり現れず、この著者の記事は分類した際に特定しやすいため再現率・適合率が高い。高い数値の他の著者の場合も、特定の顔文字や、特殊な記号が多く出現する特徴がある。

〈表７　再現率適合率の低い記事〉

また「シャドウロール 」が
中山 7R の三連複が的中しとるやんけー！
ってさっき調べて知った（笑）
最近はかなり忙しくて、
まぁ言うても面接やけどな。

再現率適合率が低い値の著者４の記事例を表７に示す。記事の中には多くの著者に共通して出現する記号が多いため、著者４を特徴的に表すとは言えず、著者推定が難しくなる。

５.結論

本研究では Blog 記事を用いて、著者推定を行った。Blog 記事は著者により様々な文章構成があるため、名詞などの内容語だけを用いた著者推定だけでは不十分である。そのため、本提案手法は有効な手法である。

６.参考文献

[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
[2]北　研二, 津田　和彦, 獅々堀　正幹：“情報検索アルゴリズム”, 共立出版株式会社