E-006

品詞 N-gram を用いたブログの文体識別
A Method of Style Discrimination of Blog using POS N-gram

瀬川 修 † 坂内 和幸 ‡ 高橋 誠 ‡
Osamu Segawa Kazuyuki Sakauchi Makoto Takahashi

1. はじめに

我々は、コンテンツの適合性（ユーザマッチング）という観点から、ブログの推薦技術を検討している。一般にブログには様々なコンテンツが混在しており、ある話題に対して、「格調の高いコラム系」の記事もあれば、「やわらかい日記系」の記事も存在する可能性が高い。このような状況の下で、ユーザの嗜好に合わせたコンテンツの自動判定が実現できれば大変有用性が高い。そこで、本稿では、コンテンツ推薦のための要素技術として確率統計的枠組みに基づく文体識別の手法を提案する。

2. 文書内容によるコンテンツ判定

テキスト本文の解析によるコンテンツ判定は、一般に処理コストが高く、数理的手法による定量的評価が困難である。テキストの表層的な情報から文書の特徴を捉える場合の指標として、例えば、あるドメインに固有の名詞や特定品詞の単語の出現頻度、または、文末表現などが考えられる。

しかしながら、表層の語彙レベルの特徴量は考慮すべきパラメータ数の増大を招き、またドメイン依存性が高いという根源的な問題をはらんでいる。確率統計的な枠組みを適用したとしても、対象ドメインが限定されている場合を除き、汎用性を欠く手法になる可能性が高い。

3. 文書特徴量の検討

前節の検討より、本研究ではコンテンツの文体を表現する最適な特徴量について考察を行った。文体や論旨展開は、表層の語彙レベルに特徴が表出されているのは自明であるが、我々は、語彙の品詞レベルでもある程度の特徴を保持した表現ができるのではないかと考えた。そこで、これらの裏付けとなる予備検討として、国内の Web から収集したブログ記事を用いて、品詞 N-gram の確率値の分布のコンテンツによる差異を観察した。ここでは、コラム系ブログ記事 5630 文、および日記系ブログ記事 5696 文のコーパスより学習した品詞 1-gram、品詞 2-gram の対数確率値を求め、双方の比較を行った。結果を図 1 と図 2 に示す (図の横軸は品詞または品詞 2 つ組みを識別する ID である)。なお、品詞 1-gram で生起しなかった品詞については-6.0 でプロットしている。また、品詞 2-gram については双方のカテゴリで一致した 2 つ組みをプロットしている。

図 1、図 2 から観察されるように、両コンテンツの品詞 N-gram の確率値の分布は、微妙に異なっていることがわかる。以上の予備検討から、コンテンツの種類に依存した文頭・文末の言い回しや、論旨展開に用いる表現などの特徴は品詞レベルに縮退しても、ある程度保持されるのではないかと考えられる。また、特徴量を品詞に縮退させることによって（品詞の種類は活用形を考慮してもたかだか数百オーダ）、少ない学習データでも精度のよい識別モデルを推定できる可能性がある。

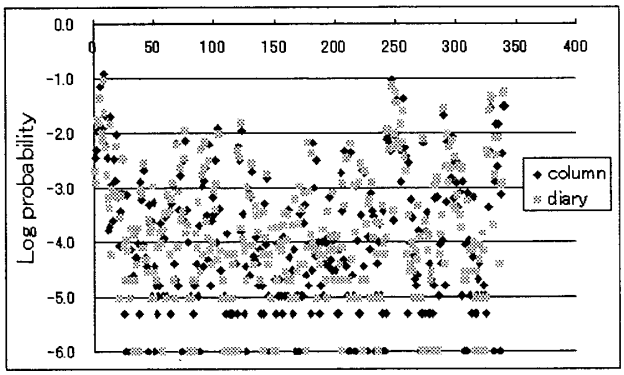

図 1: 品詞 1-gram の確率値の分布

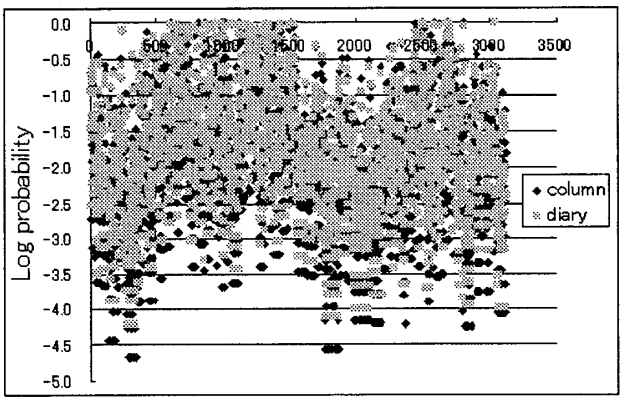

図 2: 品詞 2-gram の確率値の分布

4. 品詞 N-gram を用いた文体識別

前節の予備検討を踏まえ、ここでは、品詞 N-gram とベイズ的パターン識別の枠組み（事後確率最大化）を用いた文体識別手法の提案を行う。

ベイズの定理より

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \approx P(x|c)P(c)$$

ここで、x は観測される品詞系列 $\{x_1, x_2, \cdots, x_n\}$ であり、c は識別カテゴリ $\{c_1, c_2, \cdots, c_m\}$ である。

文体識別のためには、事後確率 $P(c|x)$ を最大にする

†中部電力 (株) 電力技術研究所
‡TIS(株)

c を求めればよい。

$$\hat{c} = \underset{c}{argmax}\ P(x|c)P(c)$$

さらに、文書中での品詞 x_i の独立性を仮定すれば、$\hat{c}$ は次式で与えられる (1-gram によるモデル化)。

$$\hat{c} = \underset{c}{argmax}\ P(x|c)P(c) \approx P(c)\prod_{i=1}^{n} P(x_i|c)$$

また、文書中で連接する品詞の 2 つ組み x_i, x_{i+1} の独立性を仮定した場合、$\hat{c}$ は次式で与えられる (2-gram によるモデル化)。

$$\hat{c} = P(c)\prod_{i=1}^{n} P(x_i, x_{i+1}|c)$$

本手法では SVM などの 2 値分類器と比較して、複数カテゴリが扱えること、またカテゴリ識別のスコアを尤度として得ることができる、などのメリットがある。

5. 評価実験

5.1 識別器の学習

前節で述べた手法に基づき、コラム系記事と日記系記事の 2 つの識別器を構成した。品詞 N-gram の学習データは国内の Web より収集したブログ記事を用いた。その詳細を表 1 に示す。品詞系列を求める際の形態素解析には Chasen を用いている。

表 1: 学習データ詳細

種別	文数 (形態素数)
コラム系ブログ記事	5630 (179798)
日記系ブログ記事	5696 (91619)

参考として、学習により得られた識別器の品詞 N-gram モデルのエントリ数を表 2 に示す。

表 2: 品詞 1-gram と品詞 2-gram のエントリ数

種別	品詞 1-gram	品詞 2-gram
コラム系ブログ記事	304	4495
日記系ブログ記事	313	4821

なお、学習データに出現しない品詞の確率は、バックオフ・スムージング (本実験では Good Turing 法 [1]) によって推定・補完している。

5.2 評価データ

評価データとして、学習に用いていないブログよりコラム系、日記系、それぞれ 100 記事 (10 ブログより 10 記事ずつ) を用いた。評価データの文数と形態素数の平均であるが、コラム系で 21.8 文 (674.9 形態素)、日記系で 22.9 文 (372.8 形態素) であった。

5.3 実験結果

文体識別の正解率を表 3 に示す。なお、カテゴリごとの事前確率 P(c) の信頼できる値の推定には、膨大な量のブログ記事のサンプリングとラベル付けが必要なため、本実験では P(c) は等確率としている。また、記事の長さによる影響を防ぐため、文体識別の尤度スコアは形態素数で正規化している。

表 3: 文体識別の正解率

種別	品詞 1-gram による識別器	品詞 2-gram による識別器
コラム系ブログ記事	98%	98%
日記系ブログ記事	90%	94%

5.4 考察

実験結果から、品詞 2-gram による識別器の方が性能が高く、カテゴリ間で類似した微妙な文体に対しても頑健性が高いことがわかる。

学習データ量については、品詞 N-gram による文書の特徴空間のスパース性から、$N \leq 2$ であれば 5 千文程度の小量の学習データでも性能の高い識別器を構成できることを示唆している。

6. 関連研究

ブログのコンテンツ内容判別については、ニュースなど情報提供が主体の記事と、作者の主観的意見表明が主体の記事の判別を試みた研究 [2] がある。

また、文字単位の N-gram を用いた著者推定 [3] や、助詞の N-gram の分布を用いた著者の識別 [4] などが行なわれている。

7. まとめ

本稿では、文書の特徴表現として品詞 N-gram を用い、ベイズ的枠組み (事後確率最大化) による文体識別を検討した。本手法では、特徴表現として品詞 N-gram を用いることから、文書のトピック内容に依存せず、また識別器の学習に大量の学習データを必要としないなどの有効性が確認された。

参考文献

[1] S.M.Kaz, "Estimation of probabilities from sparse data for language model component of a speech recognizer", IEEE Trans. ASSP, Vol.35, pp.400-401, 1987.

[2] X.Ni *et al.*, "Exploring in the weblog space by detecting informative and affective articles", 16th WWW Conf., pp.281-290, 2007.

[3] 松浦, 金田, "n-gram の分布を利用した近代日本文の著者推定", 計量国語学, 22(6), pp.225-238, 2000.

[4] 金, "助詞の n-gram 分布に基づいた書き手の識別", 計量国語学, 23(5), pp.225-240, 2002.