E-002

文体識別によるブログ推薦システム
Blog Recommendation System based on Style Recognition

瀬川 修 † 坂内 和幸 ‡
Osamu Segawa Kazuyuki Sakauchi

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アプローチとして文体識別によるブログ推薦システムを提案する。

2. 文体識別に用いる特徴量

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。

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 [1]。コンテンツ種別に依存した文頭・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 [1]。

3. 品詞 N-gram を用いた文体識別

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概要を示す。

ベイズの定理より

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \approx P(x|c)P(c) \quad (1)$$

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 c を求めればよい。

$$\hat{c} = \underset{c}{argmax}\ P(x|c)P(c) \quad (2)$$

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)。

$$\hat{c} = \underset{c}{argmax}\ P(x|c)P(c) \approx P(c)\prod_{i=1}^{n} P(x_i|c) \quad (3)$$

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)。

$$\hat{c} = P(c)\prod_{i=1}^{n} P(x_i, x_{i+1}|c) \quad (4)$$

4. 文体識別の評価実験

ここでは、文体識別の評価実験結果を示す (文献 [1] の実験結果を再掲)。

4.1 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)

なお、学習データに出現しない品詞 N-gram の確率は、バックオフ・スムージング (本実験では Good Turing 法 [2]) によって推定・補完している。

4.2 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。

4.3 実験結果

文体識別の正解率を表 2 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。

†中部電力 (株) エネルギー応用研究所
‡TIS(株)

表 2: 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%

4.4 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。

5. ブログ推薦システム

以下では、提案する文体識別手法を用いたブログ推薦システムについて述べる。本システムでは、文体種別として前述の「コラム系」、「日記系」に加え、新たに「かたい (formal)」、「やわらかい (casual)」という種別を設ける。「かたい」、「やわらかい」の文体識別は、提案手法による識別器によって同様な手順で行なう。

「かたい」文体の品詞 N-gram の学習には、新聞記事 3000 文を用い、「やわらかい」文体の品詞 N-gram の学習には、BBS 記事 3000 文を用いた。

次に、推薦システムの実現にあたり、まず文体識別結果を 2 次元平面にプロットすることを考える。3 節の式 (4) から、X 軸座標は次式で求める。

x = コラム系の識別スコア (対数確率) - 日記系の識別スコア (対数確率)　(5)

また、Y 軸座標は次式により求める。

y = かたい文体の識別スコア (対数確率) - やわらかい文体の識別スコア (対数確率)　(6)

前述の実験で用いたブログ 200 記事をプロットした例を図 1 に示す。

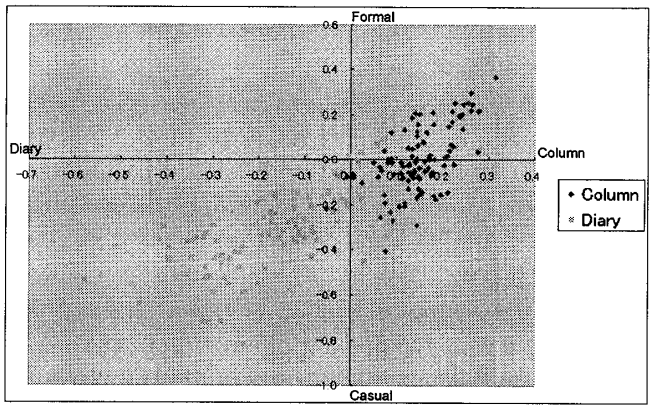

図 1: 文体識別結果のプロット例 (コラム系 100 記事、日記系 100 記事)

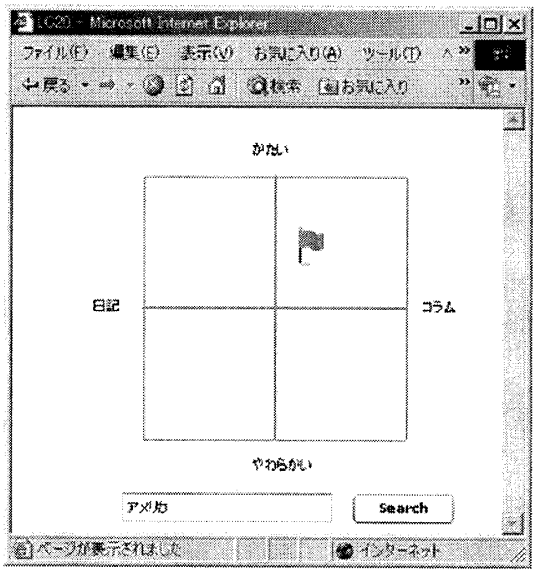

図 2: ブログ推薦システム

ブログ推薦システムの検索インタフェースを図 2 に示す。本システムでは、ユーザが検索キーワードと共に、文体条件の指定を 2 次元平面上のポインタ (青旗) によって行なう (図 2 の例では、「アメリカに関するかたいコラム系記事」という検索条件となる)。システムは、文体識別結果に基づき 2 次元平面にプロットされたブログ記事の中で、ユーザの指定するキーワードを含み、かつユーザが指定した文体条件の座標の近傍にあるコンテンツを適合記事と判定し、推薦結果として提示する。本システムが提供する GUI により、ユーザは嗜好に合った文体の記事をより感覚的に検索することができる。

6. 関連研究

Ni らはニュースなど情報提供が主体のブログ記事と、作者の主観的意見表明が主体の記事の判別を試み、判別結果を用いた推薦システムを提案している [3]。また、Jung らは、ブログ記事のムード (happy, sad, angry, fear) による分類手法の検討を行なっている [4]。

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、ベイズ的枠組みによる文体識別と、その結果を用いたブログ推薦システムについて述べた。

参考文献

[1] 瀬川, 坂内, 高橋, “品詞 N-gram を用いたブログの文体識別”, FIT2008, (2), pp.151-152, 2008.

[2] 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.

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

[4] Y.Jung *et al.*, “A Hybrid mood classification approach for blog text”, PRICAI2006, LNAI4099, Springer, pp.1099-1103, 2006.