5R-4

Weblog における書き手の興味に関する推定分布生成手法の提案

松山学 ‡　大囿忠親 ‡　伊藤孝行 ‡　新谷虎松 ‡
‡ 名古屋工業大学　大学院工学研究科　情報工学専攻
e-mail: {manabu, ozono, itota, tora}@ics.nitech.ac.jp

1　はじめに

近年注目され始めている情報源として Weblog がある．Weblog は通常の Web ページとは異なり，速報性，リアルタイム性のある新鮮な情報が発信されることから，掲示板同様に有用な情報源と考えられるようになってきている．また，Weblog を心理学の立場から見ると Weblog には書き手の興味が存在すると考えられている [1]．BlogWatcher[2] では，複数の Weblog を自動収集，監視することで，有用な情報を得る事ができ，記事やキーワードによる検索を行うことが可能である．

本稿では，一般的に行われている記事やキーワードによる検索，分類とは違い，書き手の興味による検索，分類を行うことを目的としている．そのため，Weblog に存在する書き手の興味を興味クラスタと考え，興味クラスタごとに推定分布を生成する手法を提案する．書き手の興味クラスタを生成するため，本稿では，書き手は Weblog を持つ事を前提としている．本推定分布を利用する事により，書き手の興味に即した Weblog 検索や分類を行う事が可能である．

本稿では，２章で書き手の興味に関するクラスタリング手法，３章で書き手の興味推定分布生成手法について述べ，４章で考察を示し，５章で本稿をまとめる．

2　興味に関するクラスタリング

2.1　前処理

本稿では，解析するテキストデータとして MovableType[1]利用者の書き出し用データ用いた．また，形態素解析には chasen[2]を用いた．まず，書き出しデータに対して形態素解析を行い単語の頻度を取得する．本稿では，単語を「名詞」と「未知語」と定義する．Weblog は日記であるため口語表現や未知語が多数存在するため，形態素解析に対して以下のルールを定義する．

・連続するアルファベットは連結し単語とする．
・連続する名詞，未知語は連結し一つの単語とする．

このルールにより「マツケンサンバ」や「blog」などの単語を取得することが可能となる．

2.2　クラスタリング

具体的なアルゴリズムを示す．閾値 α, β は予備実験により定めている．

(1) 重み付け：前処理により得られた単語に対して TF-IDF[3]により重み付けを行う．

(2) 重要語の選出：TF-IDF 値が高い語を重要語とし全単語数の α %に達するまで重要語の上位語を取り出す．取り出された集合を C とする．

(3) クラスタリング１：重要語 $c \in C$ をそれぞれクラスとする．初期クラス数 M は重要語数 C_{max} と等しい．

(4) クラスタリング２：クラスの組み合わせ $M(M-1)/2$ 個について，各組を併合したときに最大の共起の相互情報量 $I(\pi)$ を与えるものを選択する．これにより，クラス数は $M-1$ 個となる．相互情報量 $I(\pi)$ を式 (1) に示す．

$$I(\pi) = \sum_{c_1 \in C_1, c_2 \in C_2} P(c_1, c_2) \log \frac{P(c_1, c_2)}{P(c_1)P(c_2)} \qquad (1)$$

ここで，π はクラス集合 $\pi = \{C_i\}$，c_i はクラス C_i に含まれる重要語，$P(c_i)$, $P(c_i, c_j)$ は確率を示す．

(5) クラスタリング３：適当なクラス数になるまで (4) を繰り返す．

以上の操作により重要語におけるクラスタを生成する．これを興味クラスタ P としクラスタ P_i 中の重要語 c_j を $p_{ij} \in P_i$ とし，単語全てに対して p_{ij} との相互情報量 $I(\pi)$ が閾値 β よりも高いものをクラスタに加える．

3　推定分布生成

興味に関してのクラスタリングで生成された各クラスタ P_i　$(i = 1, .., P_{max}$: P_{max}は生成されたクラスタ数) 中に含まれる単語集合を $X(P_i) = \{x_{ij}\}$　$(j = 1, .., P_{i-max}$: P_{i-max}はクラスタ P_iの単語総数) とする．書き出しデータから得られた単語の全集合 W に

Technique for estimated distribution about user's interest in Weblog

Manabu MATSUYAMA, Tadachika OZONO, Takayuki ITO,and Toramatsu SHINTANI

Graduate School of Engineering,Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555, JAPAN.

[1]http://www.movabletype.org/

[2]http://chasen.aist-nara.ac.jp/hiki/ChaSen/

[3]本稿では，語 w_i に関する TF-IDF を以下のように定義する．

$$tfidf(w_i) = tf(w_i) \cdot \log \frac{N}{df(w_i)}$$

ただし，$tf(w_i)$ は語 w_i の出現頻度，$df(w_i)$ は語 w_i の文書頻度，N は全文書数を示す．なお本稿では 1 文書を 1 記事としている．

表 1: 各クラスタの重要語

A	B	C	D
すみれ	検索	PE'Z	日本
主人公	キーワード	ライブ	試合
世界	ブログ	アルバム	ニュース
レビュー	blogWatcher	ロック	友達
漫画	サイト	音楽	忘年会

表 2: 各クラスタの代表例

A	B	C	D
ミュー	blog	チケット	大小保
村上春樹	評判検索	音楽	サッカー
先生	奥村研究室	ROVO	ビリヤード
本田吾郎	日記	奥田民生	ミラン
サンデー	楽天	jazz	サンズ

対してクラスタごとに式 (2) により重み付けを行う．

$$key(w_i)_{P_j} = \begin{cases} tf(w_i) & (w_i \notin X(P_j)) \\ \gamma \cdot tf(w_i) & (w_i \in X(P_j)) \end{cases} \quad (2)$$

ここで，γ は予備実験により定められる．この key 値を利用して興味クラスタごとに式 (3) により推定分布を求める．N は単語の全集合 W の単語総数を表す．

$$\hat{\theta}_i = \frac{key(w_i) + 1}{\sum_{i=1}^{N} key(w_i) + N} \quad i = 1, ..., N \quad (3)$$

4 考察

4.1 評価

被験者が持つ記事総数 50 記事，単語総数 9549 語 (内，単語として扱うのは 1428 語) の Weblog に関して本提案手法によってクラスタリングした結果を表 1，表 2 に示す．なおクラス数は 4 つ生成された．表 1 では各クラスタの重要語を上位 5 つを提示し，表 2 では重要語と結びついた代表的な例を提示する．

実行結果を見て被験者にクラス名を付けもらった．クラス A は"漫画，小説"，クラス B は"研究，blog"，クラス C は"音楽，ライブ"，クラス D は"サッカー，ビリヤード"となった．ここで，クラス A からクラス C まではあるカテゴリに関してまとまったものが得られたと言えるが，クラス D に関してはサッカーの記事に関してや趣味のビリヤードに関してなど複数のトピックが混じり合う結果となった．理由としては閾値の設定，クラスになり得るだけの記事が存在しない，記事中に複数のトピックが書かれていることが原因として

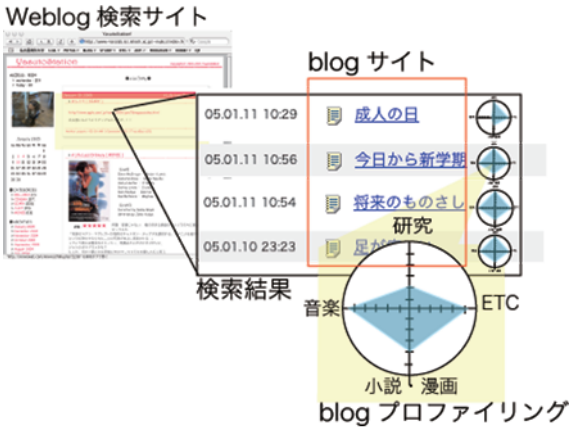

図 1: Weblog プロファイリング実行例

考えられる．しかし，クラス D を"etc"というクラス名と考えると，書き手自身は他の 3 つの内容を良く理解しているため，本推定分布を利用した検索や分類などには十分適用できると考える．

4.2 本推定分布の応用例

本推定分布を用いた応用例の一つとして Weblog プロファイルリングが考えられる．Weblog プロファイリングとは，他人の Weblog 内容を自分の評価項目と照らし合わせることで効率的に把握することである．Weblog プロファイルを検索サイトで利用した際の実行例を図 1 に示す．各カテゴリと評価したい Weblog との分類問題と考えることで簡単に応用可能である．つまり，未知 Weblog テキスト $d*$ に対して $P(p_i|d*)$[4]を最大化するクラス p_i が最適なクラス分類となる．

5 おわりに

本稿では，Weblog において書き手の興味をクラスタと捉えて推定分布を作成する手法の提案を行った．本推定分布を利用することにより，本稿で提案した応用例や Weblog 検索，分類に適応可能である．

参考文献

[1] 山下清美, "ウェブログの心理学", 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-03, 2004.

[2] 奥村学，南野朋之，藤木稔明，鈴木泰裕， "blog ページの自動収集と監視に基づくテキストマイニング", 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004.

[4] ベイズの定理より

$$P(\theta|D) = \frac{P(D|\theta)p(\theta)}{P(D)}$$

となる．ここで，θ に本推定分布を利用することにより計算可能となる．