ブースティングに基づくBlog著者年齢推定

泉　雅貴† 　三浦　孝夫†

† 法政大学 工学研究科 電気工学専攻　〒184–8584 東京都小金井市梶野町 3-7-2
E-mail: †masataka.izumi.xz@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

あらまし　本稿では、Blog 記事著者の年齢推定を行うため、ブースティングに基づく推定手法を提案する。Blog 記事は特異な表現を含むことが多く、通常の文書のために提案された手法の適用が難しい。本研究では、Blog 記事から得る各素性から局所的な重み付き分類器を生成し、それらを総合的に考慮することにより年齢推定を行う手法を提案する。最後に、実験によりその有効性を検証する。

キーワード　データマイニング, テキストマイニング, 情報抽出, プロファイリング

Boosting based Age Estimation of Blog Authors

Masataka IZUMI† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †masataka.izumi.xz@gs-eng.hosei.ac.jp, ††miurat@k.hosei.ac.jp

Abstract　In this investigation, we propose a probabilistic approach for estimating the Age of Blog authors by means of Boosting. The key idea is that we generate local classification machines from each feature in Blog article, and we estimate the Age of the BLOG authors by unifying them. We show the effectiveness of our approach by experimental results.

Key words　Data Mining, Text Mining, Information Extraction, Profiling

1. はじめに

近年、WebLog(Blog) の利用者は爆発的に増大し、現在日本では 2000 万人以上の利用者がいると言われている(注1)。Blog には、映画やファッション、ゲームなど世の中のさまざまな商品や物に対する著者自身の主観的な意見や考えなどが多く記述されている。さらに、Blog 記事は日々更新され、その情報は常に流行性をもつ特徴が挙げられる。これらの特徴より現在 Blog 記事をマーケティングなどの情報源と捉え、現在流行しているものや、商品に対する意見や評判情報などを獲得する研究が盛んに行われている [1] [2] [3]。しかし、このような情報は、著者の年齢や年代、性別、職業などの著者属性に大きく依存し、これらの属性を考慮した情報抽出を行わなければならない [4]。そのため、予め Blog 記事を著者の属性別に分類する必要性がある。

これらの著者の属性素性には年齢、年代、性別、職業の他にも、出身地、家族構成などがある。本論文では、「著者の年齢」に注目し、Blog 記事本文より著者の年齢を推定する手法を提案する。現在、Blog ではそのページ上に著者のプロフィールを設け、これを参照することによってその著者の性別や趣味などを知ることができる。しかし、年齢や年代といった情報は大半が匿名であるため、本研究の有効性が期待できる。

ここで、本研究に関連した研究をいくつか述べる。池田ら [4] は、Blog 記事から著者の”性別”を推定している。彼らの主な手法としてバイナリ分類器である Support Vector Machine (SVM) を用いており、判定精度は 90%以上を得ているが、本研究ではマルチ分類を行うため手法が異なる。本研究では Blog 著者の”年齢”推定を行うが、Schler ら [5] および先行研究 [6] では Blog 著者”年代”推定法の研究を行っている。Schler ら [5] は、英語の Blog 記事に対し、著者の年代を 10 代 (13 歳-17 歳),20 代 (23 歳-27 歳),30 代 (33 歳-42 歳), その他の 4 クラスを予め決め、各 Blog 記事をこの 4 クラスのいずれかに分類する。分類器として、Multi-Class Real Winnow (MCRW) を使用し、素性として、出現頻度に基づいた書き方による素性 (品詞、機能語、特殊語等), 内容による素性 (内容語) を用い、正解率 76.2%を得ている。先行研究 [6] では、日本語の Blog 記事に対し、Schler ら [5] 同様に予め著者の年代を 10 代 (13 歳-17 歳),20 代 (23 歳-27 歳),30 代 (33 歳-42 歳), その他の 4 クラスと決め、各 Blog 記事をこの 4 クラスのいずれかに分類する。ここでは、英語と日本語との違いを考慮し、主な分類手法としてナイーブベイズ分類器を用い、エントロピにより特徴語を抽出する手法を提案している。分類結果として、F 値 71%の精度

(注1)：Wikipedia, http://ja.wikipedia.org/wiki/Weblog

を得ているが、両者とも著者の年代を推定していることから本研究とは異なる。

本タスクでは Blog 記事からその著者の年齢を正確に推定することは極めて困難である。そのため、推定した年齢に対し許容誤差を設け、実験によりその範囲内における信頼性を求める。本研究では、年齢間における Blog 記事の特徴を考慮し、ブースティングに基づく年齢推定手法提案する。主な手法として、推定対象の Blog 記事から得る素性ごとに局所的な重み付き分類器を作成する。そしてこれらの局所的な分類器を統合し、最終的な分類器を作成する。このとき、素性として Blog 記事中の共起語を用い、分類器の重み付けにはエントロピを用いる。また、Blog の特性を考慮し、1 つの局所的分類器の中で近接する年齢間で得られる推定値に対しスムージングを行う。

本論文では、まず第 2 章で年齢間の Blog 記事の特徴について述べ、第 3 章でこれに基づいた提案手法を述べる。次に第 4 章、第 5 章において、本提案手法の有効性を示す実験および考察を行い、最後に第 6 章で結論を述べる。

2. 年齢間における Blog 記事の特徴

本研究では年齢推定タスクを行うため、まず始めに年齢間における Blog 記事の特徴について述べる。Blog は、著者自身を主張する場として提供されており、記事の書き方や内容は著者により大きく異なる。しかし本研究では年齢別に Blog 記事を見た場合、そこにはいくつかの特徴が存在すると仮定し、以下では例題を基にそれぞれの特徴を述べる。

2.1 各年齢間で使用される語

ここでは、Blog 著者らによる記事を各年齢間で比較した場合の特徴について述べる。本研究ではこれらの Blog 記事に関して、近い年齢同士の著者が書く記事には、その記事の内容や単語の使い方などは類似し、年齢差が遠くなるにつれそれらは大きく異なる特徴を持つという特徴を仮定する。たとえば中学生や高校生の間で書かれる内容は、主に学校や部活、恋愛などの記事が多い。これに対し 30 代や 40 代では仕事のことや、自分の家庭についての記事を多く含む。先行研究 [6] では、年代ごとの単語の利用率の分布の偏りを、エントロピを用いて明確化した。これによると、10 代では顔文字などの特殊文字が多く利用され、20 代では内用語,30 代では「です。」や「ます。」などの書き方が多い。

表 1 は、Blog 記事 5000 件から Blog 全体で使用頻度が高い上位 5000 語を用いて、年齢ごとにその利用率(注2) を求め、余弦法により年齢間の類似度を求めたものである。この表を見ると、年齢が遠くなるにつれ類似度の値が小さくなり、近い年齢ほど似たような語を使っていることが分かる。この結果より、各年齢を総合的に見れば、近い年齢間には似たような Blog 記事の書き方が存在する。

(注2)：利用率

$$P_j(w_i) = \frac{\text{年齢 } j \text{ 中の記事のうち単語 } w_i \text{が含まれる記事数}}{\text{年齢 } j \text{ 中の記事数}}$$

	15歳	20歳	25歳	30歳	35歳	40歳
15歳	1	0.56	0.59	0.51	0.54	0.45
20歳		1	0.69	0.65	0.49	0.41
25歳			1	0.72	0.66	0.50
30歳				1	0.82	0.68
35歳					1	0.6
40歳						1

表 1　上位 5000 語における利用率の類似度

しかし、この特徴は各年齢の Blog 記事を総合的に考慮した場合であり、実際には同じ年齢内の著者においてもさまざまな記事の書き方が存在する。そのため、Blog 解析でよく利用される文章ベクトルモデルを用いて、個々の Blog 記事を 1 つの特徴ベクトルで表現したとき、各年齢記事集合をクラスタとした場合、そのクラスタ半径は非常に大きく、各クラスタ間には重複が多数存在する。上述 5000 件記事に対して、図 1 で示す各数値はそのクラスタの年齢を現している。実際には図 1 よりもクラスタ半径が大きく、複数の重複部分が存在する。そのため、推定対象の Blog 記事を 1 つの特徴ベクトルで表現した場合、その記事がどのクラスタに属するかを見極めることは困難である。

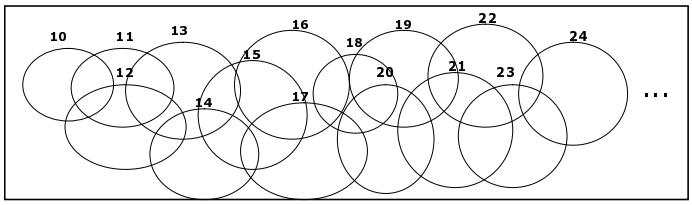

図 1　各年齢のクラスタイメージ

そのため、本手法では各記事を 1 つの特徴ベクトルで表現する手法は用いず、Blog 記事から得られる素性ごとに特徴を考え、これらの結果を統合することで推定を行う手法を提案する。また、近い年齢同士の Blog 記事間では同じような書き方をもつという仮定から、各素性から得られる推定値に対し、近い年齢間においてスムージングを行う。これに関しては次章により詳しく述べる。

2.2 年齢を特定する特徴語

次に、Blog 記事中に含まれる年齢を特定するための特徴語に関して述べる。本研究では Blog 記事に含まれる特徴語は、その著者の年齢を一意に決定できる語は少ない一方、年齢幅を特定する語が多数存在するという特徴を仮定する。表 2 は、Yahoo!Blog より取得した 17 歳の Blog 記事である。

明日から中間試験･･･
ぜんぜん勉強してないや　ﾜﾗ
最近は部活のことしか頭にないんですw
明日は、世界史と数学のテスト
来週は統一模擬もあるよ(泣)

表 2　17 歳の Blog 記事

この表では、年齢を推定に用いる有効な単語として、”中間試験”,”部活”,”世界史”,”数学”,”テスト”,”統一模試”などが挙げられる。しかし、いずれも 17 歳と明確に特徴付けられる語はなく、各単語は年齢幅を特徴付ける語として表現される。たとえば、”中間試験”はおおよそ中学生から高校生、”世界史”は高校生を表す語であることが分かり、それぞれ素性により特徴を現す年齢幅が異なる。そのため、Blog 記事から年齢推定を行う場合、各素性が持つ特徴量を柔軟に考えなければならない。さらに、単語単体よりも単語を組み合わせたほうが、より明確な年齢推定に有効であることがわかる。すなわち、”数学”という単体の語より”世界史+数学”と組み合わせて考えるほうがより年齢を特定しやすい素性として捉えることができる。そのため、特徴語としては、N グラムや共起語などを用いたほうが有効である。

3. Blog 著者年齢推定手法

本章では、日本語の Blog 記事に対する著者の年齢推定手法の提案を行う。本研究では、Blog 記事中から得られる各素性をそれぞれ独立に考え、素性ごとに大まかな年齢推定を行う。そして最終的に得られた全ての結果を総合的に判断することで、記事の年齢を推定する手法を提案する。このとき、この手法に類似した未知データ分類手法として、ブースティングが挙げられる [14]。ブースティングは、複数の局所的な重み付き分類器を作成し、最終的にこれらの分類器から出力される推定結果を総合して推定結果を得る手法である。弱仮説となる複数の分類器を総合することにより弱データに対し有効な結果を得ることができる。すなわち、学習データより作成した局所的な各分類器を $g_1(x), g_2(x), ..., g_m(x)$ とすると、最終的な分類器は、以下のように定義される。

$$V(x) = \sum_{i=0}^{m} w_i g_i(x)$$

このとき、w_i は、各分類器に対する重みを表しており、分類器の信頼性に基づき値が決まる。最終的な推定量は、重み付けされた各分類器の推定値を合計し、その値の最大値をとる引数を出力結果とする。

$$\hat{x} = \arg\max_{x} V(x)$$

本研究では、このブースティングに基づき、Blog 記事に含まれる各素性に対しそれぞれ独立に年齢推定を行い、これを局所的な分類器として見なすことで、Blog 記事中に複数存在する特徴量を考慮した推定結果が得られる。以下ではまず推定に用いる素性を定義し、次に各素性による分類器について述べる。そして最後に、これらの分類器を統合した推定手法について述べる。

3.1 年齢推定に用いる素性

本研究では、Blog 記事本文のみを用いてその著者の年齢推定を行う。このとき、年齢を特定する素性として、単語、N グラム、共起語などが挙げられる。前章で述べた Blog 記事の特徴を考慮すると、単語単体よりも単語の組み合わせを用いるほうが有効である。このため、素性として N グラムや共起語を用いることにより推定に対し効果的に働く。しかし N グラムを用いた場合、語の連続性を仮定しており、特に Blog 記事では著者によりさまざまな書き方が存在するため、これらを分類に用いた場合、再現性が失われる可能性がある [6]。このため、本研究では共起語を素性として用いる。共起語は、Blog 記事中に固定長のウインドウ幅を設定し、そのウインドウ幅内に出現する語のペアを共起語として用いる。さらに、ウインドウ幅を Blog 記事中の文頭から文末までスライドさせることにより、Blog 記事中の全共起語を抽出する。

3.2 素性を用いた局所的分類器

本節では、Blog 記事から得る各素性に対する分類器の作成方法を述べる。このとき、予め年齢が分かっている Blog 記事を教師データとして用い、これに基づいて各年齢に対する推定量を導き出す。ここで用いる推定量とは、推定対象の Blog 記事がどの年齢に属するかを表す指標で、推定量が大きいほどその年齢に属する確率が大きくなる。本節では、分類器の作成法を、素性利用率に基づく推定量の決定、スムージング、分類器に対する重み付けの手順で説明する。

3.2.1 素性利用率に基づく推定量

本節では、各素性に基づく局所的な分類器の推定量について述べる。本研究では、推定量の計算において教師データ中における素性の利用頻度に着目する。すなわち、素性が教師データ中のある年齢に突出して利用頻度が高い場合、その素性はその年齢に分類される確率が高いとみる。これに基づき、素性ごとの各年齢に対する推定量を以下で定義する。ある素性を f_i とし、年齢クラス $C = \{c_1, c_2, ..., c_n\}$ としたとき、素性 f_i の年齢 c_j に対する推定量は以下の式により計算される。

$$g_i(c_j) = \frac{1}{Z_i} * \frac{N(c_j, f_i)}{N(c_j)}$$

$$Z_i = \sum_{j} \frac{N(c_j, f_i)}{N(c_j)}$$

$N(c_j)$ は年齢 c_j 中の全ての Blog 記事数を表し、$N(c_j, f_i)$ は年齢 c_j 中の Blog 記事のうち、素性 f_i を含む記事数を表す。Z_i は正規化項である。これにより、素性 の利用頻度が高い年齢の推定量は大きくなり、逆にあまり利用されない年齢の推定量は小さくなる。この式により得られる $g_i(c_j)$ を素性 f_i からの局所的分類器として見なし、最終的な分類器の一部として利用する。

3.2.2 スムージング

本節では、前節により得た推定値に対しスムージングを行う。実際に教師データより得られた推定値の分布は、図 2 の実線で示すような滑らかでない分布を得る。図 2 は、素性から得た推定値の分布のイメージを示しており、横軸が年齢、縦軸が推定量となっている。しかし、第二章で述べた近接する年齢間の単語利用率は類似するという特徴を考慮すると、得られる推定値の分布は滑らかであることが好ましい。すなわち、本研究では、近接する年齢間の推定値において大きな差が生じた場合、これをサンプリング誤差によるものと見なし、スムージングを行う。この処理を行うことで、教師データ量が乏しい場合でも有効な

推定値が得られる。年齢 c_j の前後 k の年齢をスムージング対象範囲として、以下の式により年齢 c_j の推定値を補正する。

$$\acute{g_i}(c_j) = \frac{g_i(c_{j-k})+\cdots+g_i(c_j)+\cdots+g_i(c_{j+k})}{2k+1}$$

この式より、k を大きくすればより滑らかな分布が得られ、k を小さくすればより教師データに近い推定値を得る。この処理を、全ての年齢に対し行うことで、図 2 で示す点線のような分布が得られる。以下では、$g_i(c_j) = \acute{g_i}(c_j)$ とし、このスムージング処理を行った推定値を用いる。

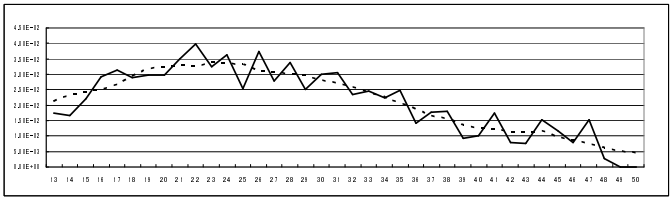

図 2　局所的分類器から出力される推定値サンプル

3.2.3　分類器に対する重み付け

本節では、素性ごとに作成した分類器に対し重み付けを行う。一般的にブースティング手法では、各分類器に対しその分類器の信頼性に基づき重み付けするが、これに対し本研究においては、分類器の確信度に基づき重み付けを行うことを提案する。すなわち、ある分類器において、年齢間における推定値の分布に大きな偏りがある場合、年齢を決定する分類器として大きく作用するように重み付けを行う。たとえば、”中間+テスト”という共起語から推定される年齢はおよそ、中学生から高校生までと推定できる。これに対し、”数 A+テスト”という共起語からは高校生に限定することができる。このため、”数学 A+テスト”を素性として用いた分類器に対して高い重みを付加する。本研究では、推定値の分布の偏りを測るための指標としてエントロピを用いる。エントロピは熱力学や情報理論などさまざまな分野で利用されており、特に自然言語処理においては、単語の出現頻度などの偏りを測る指標として使用される [7]。このエントロピを用いて、分類器における推定値の分布の偏りを以下の式により計算する。

$$E_i = \sum_j \left(-\frac{g_i(c_j)}{T_i}\log_2\frac{g_i(c_j)}{T_i}\right)$$

$$T_i = \sum_j g_i(c_j)$$

上式により分類器に対するエントロピを求めることで、推定値の分布の偏りが大きい分類器ほどその値が小さくなる。このエントロピに基づき、以下の式により分類器に対する重み付けを行う。

$$w_i = P * \log\left((ME - E_i) * Q + 1\right)$$

$$ME = \log_2 n$$

ここで n は推定候補となる年齢数であり $|C| = n$ である。また、ME は、最大エントロピ値を表し、P および Q は重み付けを行う際のパラメータ値を表す。P は重み全体の値に付与し、Q は最大エントロピ値付近の値をとる分類器に対しどれだけ差を付けるかを決定する。この重み付けにより、推定値に大きく偏りがある場合、その分類器に対し、大きな重み付けが付加され、各素性の確信度に基づき柔軟な推定が行える。

3.3　局所的名分類器の統合

本節では、素性ごとに作成した分類器を統合し、最終的な分類器を得る手法について述べる。一般的なブースティングでは、重み付けされた各分類器から得る推定値を合計することにより、その値を最終的な推定値として用いる。本研究において同様な手法を用い、最終的な分類器を以下の式により定義する。

$$V(c_j) = \sum_j w_i g_i(c_j)$$

g_i は各素性に基づく分類器であり、w_i はその分類器の重みを表す。この推定値を基に、最大の推定値を取る年齢を著者の年齢として出力する。

$$\hat{old} = \arg\max_{c_j} V(c_j)$$

このように、各素性に対し独立に分類器を生成し、最終的に統合することにより、素性により特徴を現す年齢幅が異なる場合や、その度合いが異なる場合においても柔軟に推定を行うことができる。しかし一般的に、Blog 記事からその著者の年齢を正確に推定することは難しい。たとえば、16 歳の著者が 18 歳的な記事の書き方をする場合や 14 歳的な書き方をするような場合があり、著者が書いた記事から推定した年齢と、著者自身の年齢は必ずしも一致しない。そのため、本研究では、推定した年齢に許容誤差範囲を設け、次章で述べる実験によりその範囲内における信頼性を求める。

4.　実　　験

本章では、本提案手法の有効性を検証するため、素性比較、重み付けおよびスムージングの有用性、文章ベクトルモデルを用いた分類手法比較の実験をそれぞれ行う。このとき、年齢推定タスクにおいては前述のように、推定結果に誤差が生じるため、本研究では推定結果に対し許容誤差範囲を与える。システムの性能評価として、本実験ではその範囲内における推定精度を指標とし、以下の式を用いる。

$$\text{正解率} = \frac{\text{推定した年齢が実際の年齢} \pm T \text{ の範囲に入る記事数}}{\text{推定した全 } Blog \text{ 記事数}}$$

上式において T は許容誤差範囲を表し、各実験ではこの T を 0〜9 まで変化させた場合の正解率をそれぞれ示す。本実験で用いる Blog 記事は、Yahoo!Blog より予め年齢が分かっている 13 歳から 50 歳までの Blog 記事 50274 件を取得し、このうち 50000 件を教師データ、274 件をテストデータとして用いる。このとき、形態素数が極端に少ない記事は推定に不適切であるため、50 形態素以上を含む記事のみを取得する。本実験で取得した Blog 記事の例を図 3 に示す。また、今回取得する Blog 記事はランダムに抽出を行ったため、図 4 に示すような本

10代Blog記事

【16歳】
さて。
さきほど暇だったのでポップンキャラ書いてみました。
ロキじゃぁー。
かわいぃよ。
っほ。(笑
・・・
【17歳】
今日、５級☆首飾り+3と、５級☆鎧頭+1を手に入れましたw
首飾りは猫さんから借りましたが、頭は買いました＾＾
最近1Mもなくしたのに、900Kの出費・・・・ｲﾀｲ・・・・
でも、このとーりw武器と指輪×?、耳飾以外すべて印になりましたw
・・・

20代Blog記事

【25歳】
本日は中学時代の同窓会でした。
前回は４年くらい前だったかな？
それにしても変わる子は変わるんですね?！！
昔はそうでもなかった子がいきなり可愛くなってたりするし＾＾びっくりですw
・・・
【27歳】
今日は私だけ布の裁断の仕事に行った。
はじめは2人で行く予定だったけど、まだスペースがないとのことで、1人で向かった。4ヶ月半ぶりの出勤！！ﾄﾞｷﾄﾞｷした。
場所はシャドウェルというところで、アケミさんという方のフラットだった。
・・・

30代Blog記事

【34歳】
今日は下の妹の妹の子の姪が泊まりにきています。
わかりにくいので相関図にしてみました。
３美ちゃんです。３年生の時くらいから、
春休み、夏休み、冬休みなどに泊まりにきます。
ただこの子！好き嫌いが多いんですよ！
・・・
【38歳】
毎日暑いですねぇ。(^_^);
異常気象ですね、やっぱり地球温暖化のせいでしょうか？
全国民がエアコン止めて、
自動車に乗らなければ外気温５度は下がるよなと真剣に考えて（も誰もやらないって俺を含めて）る今日この頃皆さんいかがお過ごしでしょうか。
・・・

図 3　取得した記事例

人の記述でない記事 (アフリエイトや他の記事をコピーしただけの記事等) が含まれる。このため、明らかに本人による記述がなされていない記事に対し人手により除去を行う。本実験では、テストデータ 274 件に対し、32 件のデータ除去を行った。また、本実験にパラメータ値として、スムージングを行うパラメータ値を $k = 2$、および重み付けを行う際のパラメータ値を $P = Q = 10$ として実験を行う。

入庫予定情報です！！
年式　車名　グレード　色　走行距離
H10　プレオ　RM　シルバ-　114,992
H10　アリスト　Sベルテックスエディション　パール　75,931
H13　オデッセイ　M　パール　79,918
アップルで買うと20万円安い！
比べてみれば分かります。。
買取り専門だからできるこの安さ！
＊＊＊＊＊＊＊＊＊＊＊＊＊
クルマ売るならアップルで！
　車買取り専門店
http://www.apple-japan.jp
＊＊＊＊＊＊＊＊＊＊＊＊＊

図 4　アフリエイト記事例

4.1　推定に用いる素性比較

本節では、提案手法で用いる素性の有効性を検証する実験を行う。提案手法では、Blog 記事から得る共起語を素性として用いる。そのため、素性として単語のみを用いた場合の推定結果と比較を行うことで、その有効性を確かめる。また、共起語抽出にはウインドウ幅を用いるため、ウインドウ幅を変化させた場合の分類比較も同時に行う。表 3 においてこれらの実験結果を示す。

許容誤差年齢	±0	±1	±2	±3	±4	±5	±6	±7	±8	±9
単語のみ	0.07	0.20	0.33	0.49	0.57	**0.60**	0.68	0.77	0.80	0.83
共起語(3)	0.10	0.24	0.38	0.50	0.60	**0.70**	0.77	0.81	0.86	0.88
共起語(5)	0.08	0.23	0.35	0.48	0.59	**0.67**	0.76	0.79	0.85	0.88
共起語(10)	0.05	0.19	0.29	0.45	0.56	**0.66**	0.74	0.77	0.82	0.86

(カッコ内は共起語抽出時のウインドウ幅)

表 3　素性別による推定結果

表 3 の各項目の数値は、それぞれ推定許容誤差を認めた場合の正解率を示す。表 3 より、単語のみを使用した場合と共起語 (ウインドウ幅 3) を利用した推定精度を比較すると、共起語を用いた場合のほうが推定精度が高いことが分かる。特に、推定許容誤差± 5 で比較すると、共起語 (ウインドウ幅 3) を利用した場合推定精度 70%であるのに対し、単語のみを使用した場合 60%となる。しかし、ウインドウ幅を大きくすると推定精度が下がり、共起語 (ウインドウ幅 10) を用いた場合、推定許容誤差± 0 から± 4 まで一番推定精度が低い。

4.2　スムージングおよび重み付けの有効性

本節では本手法で用いる重み付けおよびスムージングの有効性の検証実験を行う。本実験では、比較対象として、スムージングおよび重み付けを行わない手法、重み付けのみを行う手法、そしてスムージング＋重み付けを用いる手法において推定結果を比較する。このとき推定に用いる素性として共起語 (ウインドウ幅 3) を用いるものとする。この実験結果を表 4 に示す。

許容誤差年齢	±0	±1	±2	±3	±4	±5	±6	±7	±8	±9
なし	0.08	0.21	0.29	0.35	0.40	0.50	0.57	0.66	0.72	0.77
重み付け	0.04	0.18	0.30	0.40	0.50	0.62	0.71	0.76	0.81	0.84
スムージング＋重み付け	0.10	0.24	0.38	0.50	0.60	0.70	0.77	0.81	0.86	0.88

表 4　各手法による推定結果

表 4 より、スムージングおよび重み付けを行わない場合、推定許容誤差± 5 のときで 50%と一番精度が低い。また、重み付けのみの手法とスムージング＋重み付け両方を用いた手法を比較すると、両方を用いた場合は平均して 7%の精度向上を得る。

4.3　文章ベクトルモデルを用いた手法との比較

本節ではベースラインとして文章ベクトルモデルを用いた分類手法との比較実験を行う。第 2 章で述べた Blog の特徴を考慮すると、本研究において Blog 記事を一つの特徴ベクトルで表現する手法は有効ではない。そのため、本節では文章ベクトルモデルを用いた手法と本手法との推定精度比較を行う。本実験で用いる文章ベクトルモデルを利用手法は、自然言語処理において文章分類等で一般的に用いられる余弦類似度を用いた推定法を用いる [13]。本実験においては、各 Blog 記事をベクトルで表現するため、出現頻度が高い上位 5000 単語を記事のベクトル要素とし、各単語の出現頻度をその要素値とする。すなわち、Blog 記事 D に対し、その記事ベクトルを以下で定義する。

$$D = \{p_D(t_1), p_D(t_2), ..., p_D(t_n)\}$$

ここ $p_D(t_i)$ では、Blog 記事 D 内において、単語 t_i の出現頻度を表す。このベクトルを用いて、教師データ中の年齢ごとに重心ベクトルを求め、これを年齢の特徴ベクトルを求める。このとき、推定対象となる記事のベクトルと各年齢の特徴ベクトルとの類似度を計算し類似度が最大のものを推定年齢とする手法をとる。すなわち、年齢 i の重心ベクトルを F_i、推定対象の記事ベクトルを D_{test} とすると、類似度は以下の式により計算される。

$$Sim(F_i, D_{test}) = \frac{F \bullet D_{test}}{|F||D_{test}|}$$

この類似度計算を各年齢ごとに行い、最小値を取るものをその記事の推定年齢とする。

$$\hat{old} = \arg\min_i Sim(F_i, D_{test})$$

また、本実験で用いる本提案手法は、共起語 (ウインドウ幅 3) を素性で用いたものを使用する。この比較実験結果を表 5 に示す。

許容誤差年齢	±0	±1	±2	±3	±4	±5	±6	±7	±8	±9
ベクトルモデル	0.01	0.10	0.18	0.22	0.28	0230	0.32	0.34	0.39	0.43
本手法	0.10	0.24	0.38	0.50	0.60	0.70	0.77	0.81	0.86	0.88

表 5　ベクトルモデルを用いた手法との比較

表 5 より、ベクトルモデルを用いた手法ではほとんど精度が得られず、推定許容誤差± 5 のときを比較すると、ベクトルモデルでは 30%の正解率であるのに対し、本手法では 70%を得る。これより本タスクにおいて、単純なベクトルモデルを用い、その類似度により年齢を推定する手法は適切でないと見られ、本研究における手法の方が有効であることが分かる。

5. 考　　察

本章では各実験結果に対する考察を行う。まず、素性別に行った比較実験を見ると、単語のみを素性に使った場合に比べて、共起語を用いたほうがより高い精度を得た。これより、単語単体より単語の組み合わせを用いたほうがより年齢の特徴を掴みやすいということが分かる。しかし、共起語を用いた場合、そのウインドウ幅を大きくしすぎると推定精度が下がる。これは、ウインドウ幅を大きくすることで、単語間の関連性が弱くなるためであり、本実験結果より共起語を素性として用いた場合、そのウインドウ幅は１～３程度が適切である。

次に、スムージングおよび重み付けを用いた比較実験を見ると、実験結果よりスムージング、分類器に対する重み付け両方が有効な手法であることを得た。スムージングは、近接する著者間において単語の使い方が似ているという仮定の下で提案を行った。そのため、この手法が有効であることが実験で証明されたことから、本研究における仮説が正しいものであることを示す。

文章ベクトルモデルを用いた手法との比較実験では、ベクトルモデルを用いた手法ではほとんど分類精度を得ることができなかった。これは、結果より１つの記事の特徴や同じ年齢の著者が持つ特徴を１つのベクトルとして表現できないため、本手法のように素性ごとにその特徴を捉える方法のほうが有効であることが言える。

本手法の精度をみると、一番精度が高かった結果として、共起語 (ウインドウ幅 3) を用いた素性および、推定値に対するスムージング、重み付けを用いる手法が有効であることが分かった。その推定精度は、推定誤差± 5 歳を認めた場合 70%の信頼性を得た。ただし推定誤差± 3 歳以下の場合、推定精度が 50%以下となる。この原因の一つとして、著者が各記事の年齢と著者自身の年齢が必ずしも一致しないことが挙げられる。また他の理由として、今回実験で用意したテストデータは、予め本人の記述がなされていない記事 (アフリエイトや他の記事のコピーのみの記事) の除去を行った。しかし、図 5 に示すような会話文を含む記事や、ほとんどがリスト形式の記事、および人手による推定においても困難な記事もテストデータには多数含まれる。そのためこれらの記事に対して大きな推定誤差が生じ、精度低下の原因に繋がった。そのため、本手法の推定精度の向上を図るためには、予め図 5 に示すような推定に困難な記事を除去するフィルタリングを行う必要がある。

14歳のBlog記事(26歳に誤判定)

難易度が高いのでは？‥
フルフェイスメイクです
☆用意するもの☆
*マットタイプの下地（ベースカラー）
*ファンデか肌色パウダー（どちらにせよ赤味を隠してクール肌にできるもの）
*コンシーラー（リキッドタイプがいいかも）
*アイブロウ（髪色に揃えて）
*ロングタイプとボリュームタイプの真っ黒なマスカラ
*シャドー（黒色と赤色）
‥‥

31歳のBlog記事(20歳に誤判定)

少しずつ落ち着いてきた
そして今週も
先輩には会った。。。
俺「先輩、俺‥普通が良かった。。こんな眼いらなかった
　そのせいで、喧嘩すんなら潰した方がいいんじゃないかって思う」
先輩「嫌、この前は俺がわりぃ‥酒と鬱憤がごちゃごちゃになってよ
　　なぁんも出来ない自分が悔しかったんだよ」
先輩「なぁ‥央介。。
　　普通ってのはさ、、"みんな"って事じゃないだろ？
　　誰も彼もが同じじゃねぇんだから。。
　　お前がいると、俺の彼女は笑うんだぜ？
　　そのお前が笑えない状態でどうすんだよ！」
‥‥

49歳のBlog記事(36歳に誤判定)

楽しみにしていた大会も終わってしまい寂しい気分
結果は100フリー　1分16秒72？　50フリー33秒90
100フリーはベストから1秒ちょい遅れ
50も1秒半ほど遅いけど去年のひどい落ち込みに比べると嬉しい方　Tちゃんも25フリー　50フリーとも1秒から2秒遅い　k本婦人がこの日1番の伸び
　100フリーで5秒アップ
久々マスターズ復活のS木氏は25フリーでこの日1番の華麗なフライング
本人いわく空中でピストルの音が聞こえたそうな
真後ろで観ていて　わ?って叫んでしまったくらい　スタンドにいたチームのみんなも叫んでいたそう
ところがどっこい失格のアナウンスがないでわないか
速報を見にいくと4着のところに名前が
‥‥

図 5　システムが誤判定した記事

6. 結　　論

本研究では Blog 記事本文を用いて、Blog 記事の著者年齢推

定を行った。Blog 記事は一般記事とは異なりさまざまな書き方をしているため、本タスクは極めて困難である。しかし本研究では Blog 記事の特徴を考慮しブースティングに基づく手法を提案した。この結果、推定許容誤差± 5 歳を認めた場合の正解率は 70%を得ることができた。Blog 記事には多数のノイズとなるデータが含まれ、実際に人手による年齢推定が困難な記事も多数含まれる。この中において許容範囲± 5 歳における推定精度 70%を得たことにより、本提案手法は有効な手法であることが言える。さらに、本手法を評判情報抽出システム [11] [12] の一部として利用した場合、このタスクでは予め評判情報が書かれた記事に対するフィルタリングを行うため、本実験で得た精度以上の精度向上が期待できる。今後は、この評判情報抽出システムに本手法を組み込んだ場合の推定精度を検証すると共に、著者の年齢と性別両方を考慮した推定手法を考案する予定である。

文　　献

[1] 鈴木泰裕， 高村大也， 奥村学 - ”Weblog を対象とした評価表現抽出,” 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-SW&ONT-A401-02,2004

[2] 古林紀哉, 平野耕一, 高橋淳一-ブログ記事の自動分類により消費者意識の側面を捉える試み, NRI 技術創発, 2006

[3] 森本和伸， 林貴宏， 尾内理紀夫-興味発見を支援する Blog 記事推薦システム情報処理学会論文誌, 2006

[4] 池田 大介, 南野 朋之, 奥村 学, blog の著者の性別推定, 言語処理学会第 12 回年次大会, 2006

[5] Schler, J. et al. : Effects of Age and Gender on Blogging , AAAI05, 2005

[6] Masataka Izumi , Takao Miura : ”Entropy-based Age Estimation of Blog Authors” 32nd IEEE Intn’l Computer Software and Applications Conference(COMPSAC) Workshop on Barriers towards Internet-Driven Information Services (BINDIS), Turku, Finland

[7] 堀部 安一 : ”情報エントロピ論”, 森北出版株式会社, 1989

[8] 北 研二 :”確率的言語モデル”, 東京大学出版会, 2004

[9] Tanabe,K.: Speech Patterns of Japanese Girls or Gals -Symbol of Identity and Opposition to Power, OPAL 3,Queen Mary, Univ. of London ,2005

[10] Berger, A.L., Della Pietra, S.A. Della Pietra, V.J. : A maximum entropy approach to natural language processing, Computational Linguistics, 19969

[11] 池田 佳代, 田邊 勝義, 奥田 英範-体験表現を手がかりにした Blog の体験情報の抽出 電子情報通信学会, WI2-2005-36 pp, 2005

[12] 鈴木 泰裕, 高村 大也, 奥村 学-Semi-Supervised な学習手法による評価表現分類言語処理学会第 11 回年次大会, 2005

[13] 戸田智子，福田直樹，石川博- Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出データ工学ワークショップ 2007 年

[14] Robert E.Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. Machine Learning, 39(2/3):135-168