E-054

ブログ著者の年齢推定に有効な素性の抽出

Feature Extraction for Estimating Blogger's Age

篠山 学†
Manabu Sasayama

1．はじめに

ブログは情報発信のひとつの方法として非常によく用いられるようになった．ブログの特徴として，ブログの記事を携帯電話などから手軽に投稿できることが挙げられる．そのため，ブログには即時性があり，ブログ著者の率直な感想や意見が記事に反映される．例えば，映画を観た直後に携帯電話からその映画の感想を投稿する，といった場合である．この感想や意見が書かれた記事から，映画やレストランの評判などの情報を得ることができる．このとき著者の年齢が得られれば少ないコストでマーケティングが可能になる．しかしブログ著者は年齢に関する情報を公開していることが出身地などの情報に比べて少ない．そこで著者の年齢を推定する必要がある．

ブログの著者属性推定に関する先行研究には，著者の性別の推定に関する研究[1]や性別，年齢，居住地の推定に関する研究[2]，年齢の推定に関する研究[3]などがある．池田ら[1]はブログ著者の性別を推定するために，素性に機能語や一人称を用いている．大倉ら[2]は単語から χ 二乗値を計算し，χ 二乗値の高い単語を素性として Complement Naive Bayes を用いて属性値を推定する手法を提案している．性別と居住地は高い精度で推定できている．しかし年齢は性別や居住地に比べ低い精度だった．そのため，テキストからの年齢推定は困難であると報告している． 泉ら[3]は共起語を素性としてブースティングを用いた年齢推定手法を提案している．

本研究では，年齢推定に有効な素性を抽出することを目的とする．具体的には単語 bi-gram や単語 tri-gram を素性の対象とする．また単語だけでなく品詞 bi-gram や品詞 tri-gram，単語と品詞を組み合わせた bi-gram や tri-gram を素性に用いる．これは特定の年齢を表す単語はほとんどないが年齢を表す言い回しや品詞の並びは存在すると考えるためである．例えば「攻撃」や「力」という単語はすべての年齢層が使用すると考えられるが，「攻撃-力」という単語 bi-gram は主にゲームに関連した語であるため 10 代が使用すると考えられる．なお，本研究では共起語は素性として用いない．共起語は n-gram と異なり，語順に無関係である．そのため著者ごとに異なる書き方をしても再現性を保つことができる．しかし語順も年齢に関係していると考える．例えば文章を書く機会が限られている 10 代に比べ 20 代や 30 代はレポートや報告書などを書く機会が増え，添削される機会も多くなるためである．

以下，本論文では第 2 章で年齢推定について説明する．第 3 章では提案する素性について説明し，第 4 章では提案した素性が年齢推定に有効であることを示し，結果を考察する．第 5 章ではまとめと今後の課題を述べる．

†香川高等専門学校詫間キャンパス情報工学科

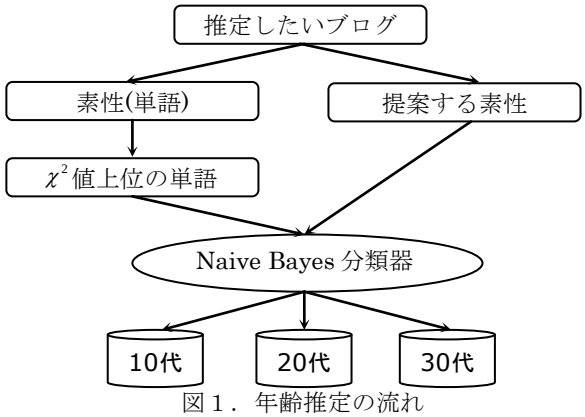

図 1．年齢推定の流れ

2．年齢推定

2.1 年齢推定の流れ

本研究では Naive Bayes を用いてブログ著者を 10 代，20 代，30 代に分類する．年齢推定の流れを図 1 に示す．まず推定したいブログから素性を取り出す．素性には 2 種類あり，ひとつは単語 uni-gram の素性，もうひとつは提案する素性である．単語 uni-gram の素性は，χ 二乗値を計算し，上位にある単語 uni-gram のみを用いる．提案する素性は単語 bi-gram や品詞 bi-gram などである．詳しくは 3 章で述べる．次に取り出した素性を元に Naive Bayes 分類器で 10 代，20 代，30 代に分類する．分類器の学習にはラベル付きデータを用いる．

3. 年齢推定に使用する素性

3.1 χ 二乗値を用いた単語 uni-gram の素性

年齢の推定に有効な語を選択するために大倉ら[2]は χ 二乗値を用いている．本研究でも χ 二乗値を用いる． χ 二乗値について簡単に説明する． χ 二乗値はある集合を複数に分割したとき，分割された集合同士がどの程度一致しているかを示す値である．本研究ではブログを集合とする．ブログの中にある単語 t が含まれているか否かによる分割と，ブログの著者の年齢が c であるか否かによる分割がどの程度一致するかを計算する．この値が大きいほど単語 t が年齢推定に有効な語といえる． χ 二乗値は以下の式で表される[4].

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A は単語 t を含み年齢が c であるブログの数，B は単語 t を含み年齢が c でないブログの数，C は単語 t を含まず年齢が c であるブログの数，D は単語 t を含まず年齢が c でないブログの数，N は全ブログの数である． χ 二乗値が大

表 1. χ 二乗値の計算結果

10 代		20 代		30 代	
χ 二乗値	単語	χ 二乗値	単語	χ 二乗値	単語
456.13	ω	88.27	数学	309.81	ω
430.16	学校	83.27	テスト	239.64	仕事
422.33	テスト	78.79	宿題	226.58	笑
299.45	仕事	77.53	体育	216.84	先日
291.29	´	72.86	史	202.47	子供

きい順に用いた．評価実験に使用したデータを用いて計算した χ 二乗値を表 1 に示す．

3.2 提案する素性

本研究で提案する素性は，単語 bi-gram と単語 tri-gram である．これは年齢を表す言い回しが存在すると考えるためである．また単語 bi-gram の一部を品詞に変更した素性や単語 tri-gram の一部を品詞に変更した素性も用いる．これは年齢によって品詞の並びにも変化が現れると考えるためである．例えば「一ヶ月ぶりに原宿へ行った」ことを 10 代や 20 代は「一ヶ月ぶりの原宿！」などのようにくだけた表現をすることがある．このとき「接尾辞(ぶり)+の+名詞(原宿)」の並びが出現する回数は 30 代よりも 10 代や 20 のほうが多くなると考えられる．単語 bi-gram の一部を品詞に変更した素性は「品詞+形態素」，単語 tri-gram の一部を品詞に変更した素性は「品詞+の+品詞」，「品詞+で+品詞」，「品詞+に+品詞」である．

4. 評価実験

4.1 実験目的

本研究では，提案した素性の有効性を確認するため素性を組み合わせたブログの年齢推定実験を行った．

4.2 実験条件

実験に使用したデータについて説明する．ブログデータは Ameba ブログから収集した．収集期間は 2010 年 12 月から 2011 年 3 月までの 4 ヶ月間である．収集したブログは約 2 万ブログであった．その中からプロフィール欄に年齢の記述があるブログを簡単なパターンを用いて抽出した．抽出したブログを人手により 10 代、20 代、30 代へ分類した．その結果 10 代は 1377 ブログ，20 代は 1647 ブログ，30 代は 875 ブログの合計 3902 ブログだった．推定には各 50 ブログを用いる．学習に用いたデータは推定に用いる各 50 ブログを除いた全てのブログで，10 代が 1327 ブログ，20 代が 1597 ブログ，30 代は 825 ブログである．

単語 uni-gram(形態素)の χ 二乗値の上位 60 語と提案した素性を組み合わせて年齢推定を行った．単語 bi-gram と単語 tri-gram は χ 二乗値の上位 120 個と 240 個を用いた．「品詞+の+品詞」は 188 個，「品詞+で+品詞」は 163 個，「品詞+に+品詞」は 196 個を用いた．なお，形態素解析には JUMAN を利用した．

4.3 実験結果

実験結果を表 2 に示す．形態素 60 語に「品詞+の+品詞」の素性を組み合わせた場合が最も精度が良くなった．また，形態素 60 語に「品詞+形態素」の素性を組み合わせた場合も精度が上昇した．それ以外の場合は，20 代の精度が悪くなった．

表 2. 実験結果

素性の組み合わせ	10 代	20 代	30 代
形態素 60	35	19	35
形態素 60 + bi-gram 120	38	17	36
形態素 60 + bi-gram 240	38	16	37
形態素 60 + bi-gram(品詞+形態素) 60	35	20	36
形態素 60 + tri-gram 120	34	16	36
形態素 60 + tri-gram 240	34	18	37
形態素 60 + tri-gram(品詞+の+品詞)188	**35**	**21**	**38**
形態素 60 + tri-gram(品詞+で+品詞)163	33	19	35
形態素 60 + tri-gram(品詞+に+品詞)196	33	17	38

4.4 考察

形態素 60 語に「品詞+の+品詞」の素性 188 個を組み合わせた場合が 20 代で合計 1 ブログ，30 代で合計 3 ブログ精度が良くなった．形態素 60 語に単語 tri-gram の素性 240 個を組み合わせた場合と比較すると，「品詞+の+品詞」を用いた場合のほうが明らかに精度が良くなっている．これは「品詞+の+品詞」の使用方法が年齢によって異なっていることを示している．このことから，品詞の並びが年齢に関係している可能性が示せたといえる．その他の組み合わせでは精度が悪くなった．形態素 60 語に単語 bi-gram の素性を組み合わせた場合や形態素 60 語に単語 tri-gram の素性を組み合わせた場合などである．これは単語 bi-gram や単語 tri-gram では必ずしも年齢を表す語にならないことを示している．

5. おわりに

本研究ではブログ著者の年齢を推定するために，年齢を表す言い回しや品詞の並びがあると考え，素性として単語 bi-gram と単語 tri-gram を用いることを提案した．推定実験を行い，形態素 60 語と単語 tri-gram の一部を品詞に変更した素性「品詞+の+品詞」を組み合わせた場合に最も精度が高くなった． しかし 20 代の推定については改善したものの精度は低いままであった．今後はさらに精度が高くなる素性の組み合わせや 20 代を判別できる素性の組み合わせを調査したい．

参考文献

[1] 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 会年次大会(2006).

[2] 大倉　務，清水　伸幸，中川　裕志，“スケーラブルで汎用的なブログ著者属性推定手法”，情報処理学会 NL 研究会 NL-180(2007).

[3] 泉　雅貴，三浦　孝夫，“ブースティングに基づく Blog 著者年齢推定”， Forum on Data Engineering and information Management(2009).

[4] YumingYang and JanO.Pedersen，“A Comparative Study on Feature Selection in Text Categorization”，Proceedings of the Twentieth International Conference on Machine Learning(1997).