ブログ情報とブログユーザ間のリンク構造を用いた著者の年齢推定

酒井 啓道　　小町 守　　松本 裕治
奈良先端科学技術大学院大学
{hiromichi-s, komachi, matsu}@is.naist.jp

1　はじめに

ブログユーザは急激に増加しており，総務省の調査では2008年の時点で日本国内のブログの総数は約1,690万件に及ぶと推計されている[*1]．ブログは商品やサービスに関する意見・感想が多く書かれているため，マーケティングの情報源として期待されており，意見情報抽出などの関連した研究やサービスが多く行われている．

年齢などの属性は，アンケート等によるマーケティングの際に重要な情報とされてきたため，ブログを用いたマーケティングの際にも有用な情報になりうる．しかし，多くのブログユーザは匿名性を守るためや，ブログのサービスとして属性情報を表示するための項目が存在していないなどの理由により，自身の年齢について明らかにしていない．そのため，ブログ著者の年齢を推定することは，マーケティングで扱える情報の増加につながり非常に有益だといえる．

既存のブログ著者の年齢推定タスクでは，記事などのブログ情報とリンク先ユーザの年齢などのリンク情報を単独で用いており，ブログ情報とリンク情報の両方を考慮した研究は行われていなかった．そこで本研究では，ブログ情報とリンク情報の両方を用いることによってブログ著者の年齢を推定する．また，自身の属性情報を公開しているブログユーザが少ないため，利用できるラベル付きデータが少ないという問題がある．本論文ではラベルなしデータを用いることにより精度向上を目指す．

本論文の主な貢献は以下の2点である．(1) ブログ情報とリンク情報の両方を年齢推定タスクに適応し，有効性を示した．(2) ブログ情報とリンク情報の両方を用いる提案手法においても，ラベルなしデータを利用する効果があることを示した．

2　ブログ著者の年齢推定に関する研究

Dongら [1] は，ブログ，電話での会話，オンラインフォーラムの投稿という3つの異なるジャンルの投稿者の年齢を線形回帰によって推定した．Dongらはブログ著者の年齢推定においては，ブログ記事の単語1-gram，品詞1-gramと2-gram，Linguistic Inquiry and Word Count (LIWC)，ブログ著者の性別というブログ情報に関する素性のみを用いている．本手法も同様に線形回帰を用いてブログ著者の年齢を推定するが，リンク情報も用いるという点が異なる．

Ikedaら [3] は，半教師あり学習の手法である Alternating Structure Optimization (ASO) を用い，ブログ著者の年齢と性別の推定を行った．ASOは Andoら [4] によって提案された手法であり，Ikedaらの手法はこれの応用とみなすことができる．Ikedaらは，副分類器を用いてラベルなしデータから素性を抽出し，これを利用した結果，性別，年齢ともに精度の向上が見られた．本手法もラベルなしデータを用いるが，Ikedaらが素性抽出のためにラベルなしデータを使ったのに対し，本手法では推定した年齢を使うという点が異なる．

Ianら [2] は，ブログサービスの一つである LiveJournal[*1]の著者の年齢と居住地をリンク情報のみを用いて推定した．LiveJournalではリンク機能として“friends”がある．Ianらは“friends”登録されたユーザ同士は似た属性を持っているという仮定のもと，ターゲットの年齢とターゲットの“friends”の平均年齢との対応を線形回帰を用い推定した．本手法もリンク情報としてリンク先のユーザの平均年齢を用いるが，ブログ情報も用いるという点が異なる．

[*1] http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-2.pdf

[*1] http://www.livejournal.com/

アルゴリズム ブログ推定器とリンク推定器を用いた年齢推定

Require: ラベル付きのユーザのセット L，ターゲットユーザのセット T，ブログ推定器の重み w_{blog}

1: ブログ推定器の学習 (L)
2: リンク推定器の学習 (L)
3: $w_{link} \Leftarrow 1 - w_{blog}$
4: **for** $t \in T$ **do**
5: t の推定年齢$_{blog} \Leftarrow$ ブログ推定器 $(t_{\text{ブログ情報}})$
6: **for** $t_l \in \{t$ のリンク先 $\}$ **do**
7: **if** t_lの年齢が不明 **then**
8: t_lの推定年齢$_{link} \Leftarrow$ リンク推定器 $(t_{l\ \text{リンク情報}})$
9: $t_{\text{リンク情報}} \Leftarrow t_{\text{リンク情報}} \cup \{(t_l$の推定年齢$_{link})\}$
10: **end if**
11: **end for**
12: **if** t のリンク情報が空 **then**
13: t の推定年齢 $\Leftarrow t$ の推定年齢$_{blog}$
14: **else**
15: t の推定年齢$_{link} \Leftarrow$ リンク推定器 $(t_{\text{リンク情報}})$
16: t の推定年齢 $\Leftarrow t$ の推定年齢$_{blog} \times w_{blog} + t$ の推定年齢$_{link} \times w_{link}$
17: **end if**
18: **end for**

3 ブログ情報とリンク構造を使った年齢推定

ブログの著者の年齢を推定する際には，ブログ情報だけでなくリンク情報も有効であるが，これまでブログ・リンクの両方の情報を利用し年齢を推定した研究は行われていなかった．本手法ではブログ情報に基づく推定器（ブログ推定器）とリンク情報に基づく推定器（リンク推定器）を利用することにより，ブログ著者の年齢を推定する．ブログ推定器では，ブログ情報としてブログ記事，プロフィールを入力とし，リンク推定器では，リンク情報としてリンク先ユーザの年齢の多重集合を入力とする．

提案手法では，ラベル付きのユーザのセット L，ターゲットユーザ（年齢を推定したいユーザ）のセット T，ブログ推定器の重み w_{blog} を入力とする．ただし，w_{blog} は 0 以上 1 以下とする．ブログ推定器・リンク推定器を L を用い学習する（行番号 1-2）．リンク推定器の重み w_{link} を $1 - w_{blog}$ とする（行番号 3）．全てのターゲットユーザ t に対し，ブログ推定器による年齢推定（行番号 5）を行う．ターゲットユーザ t のリンク先のユーザ t_l に対し，年齢がわからない場合は推定を行い（行番号 7-8）推定した年齢を $t_{\text{リンク情報}}$ に追加する（行番号 9）．ただし $t_{\text{リンク情報}}$ は多重集合なので，同じ年齢の重複は許される．このようにリンク先のラベルなしユーザの年齢を推定することにより，ラベルなしユーザの情報を利用する．最終的にできた $t_{\text{リンク情報}}$ を使い，ターゲットユーザ t のリンク推定器による年齢推定を行う（行番号 15）．ブログ推定器による推定結果とリンク推定器による推定結果の重み付き平均を最終的な年齢として出力する（行番号 16）．ラベルなしユーザを用いない場合は行番号 6 から 11 までがスキップされる．ラベルなしユーザを使う手法，使わない手法どちらにおいても，ターゲットユーザのリンク先の中に年齢がわかるユーザが存在しない場合，リンク推定器は年齢を推定することができない．その場合は，ブログ推定器の出力が最終的な出力となる（行番号 12-13）．

4 ブログ著者の年齢推定の評価実験

4.1 実験データ

実験データとして，2012 年 4 月から 12 月にかけて収集した Yahoo!ブログ[*2]と Yahoo!プロフィール[*3]を用いた．Yahoo!ブログと Yahoo!プロフィールは相互に連携して使用することができるため，Yahoo!ブログの著者の属性情報，ユーザ間のリンク情報は全て Yahoo!プロフィールを利用して取得した．以下で実験データの収集方法，統計情報を説明する．

まず，次の 2 条件を満たすブログを 8,000 件取得した．

1. Yahoo!プロフィールに性別について記入があり，年齢が 14 歳以上 45 歳以下
2. 2012 年に投稿したブログ記事の合計文字数が 1,000 文字以上

図 1 に収集したデータの年齢の分布を示す．Yahoo!ブログのユーザは 16 歳が最も多く，20 代以降はほぼ同程度となり，LiveJournal 等の他のブログデータと同様の傾向が見られた [2]．図 2 にユーザの年齢とユーザのリンク先の平均年齢，ユーザのリンク先の年齢の標準偏差を示す．ユーザの年齢とリンク先の平均年齢では，16 歳から 20 歳には比較的近い値となるが，それ以降差が大きくなり，30 代の後半以降はほとんど対応していない．

次に取得したユーザ 8,000 件のリンク先ユーザ（1 次リンク）のプロフィールを 170,561 件，さらにそのリンク先ユーザ（2 次リンク）のプロフィールを 1,454,990

[*2] http://blogs.yahoo.co.jp/
[*3] http://profiles.yahoo.co.jp/

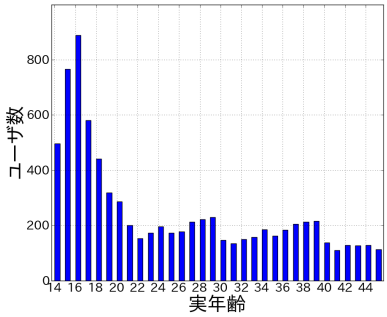

図 1　実験データの実年齢の分布

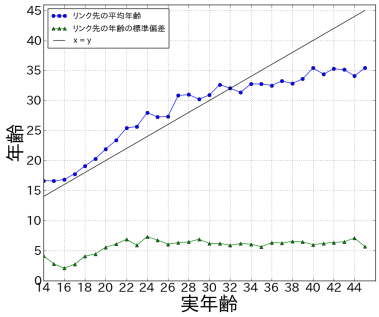

図 2　実年齢とリンク先の平均年齢，標準偏差

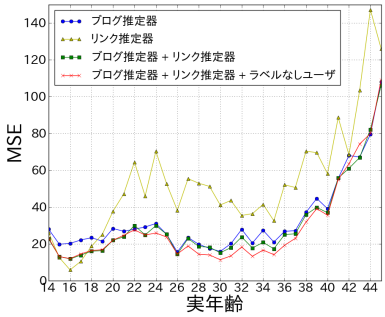

図 3　各手法における実年齢と精度

件を収集した．ただし，1 次リンク，2 次リンクの中には，属性情報を記入していないユーザも含まれる．

4.2　ブログ推定器とリンク推定器

本論文ではターゲットユーザのブログ情報を用いるブログ推定器とターゲットユーザのリンク情報を用いるリンク推定器の 2 つを利用する．

ブログ推定器では support vector regression (SVR) を用いた．素性として，ブログでは “ブログ記事” の単語 1-gram，品詞の 1-2gram を用いた．プロフィールでは，記入項目である “表示名” の文字 1-5gram，“名前” の文字 1-5gram，“自己紹介” の単語 1gram，文字 1-5gram，“好きなもの” の文字 1-5gram を用いた．ただし，プロフィールでは “表示名” 以外の項目に記入していないユーザもおり，そのような場合は用いることができない．SVR の学習には，LIBLINEAR1.92[*4]を用いた．また，学習・推定の際には，2012 年に書かれたすべての記事を使用した．

リンク推定器では Ian ら [2] と同様にユーザのリンク先の平均年齢を用いる．Yahoo!プロフィールではリンク機能として “友だち”，“お気に入り”，“ファン” が存在すが，本実験ではリンク情報として全てを区別せずに利用した．

4.3　実験設定

ベースラインとして，ブログ情報のみを使った手法とリンク情報のみ使った手法を設定した．どちらの実験も 4.1 節で説明した 8,000 件のブログを用いた．なお，リンク情報のみを使った手法では，リンク先の中に年齢がわかるユーザが存在しない場合推定することができない．そのため，リンク情報のみを使って推定したユーザは 8,000 件中 6,029 件であった．

[*4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

提案手法として 3 節で説明したブログ情報とリンク情報を用いる手法，ブログ情報とリンク情報に加えラベルなしユーザを用いる手法を設定した．全ての実験において 5 分割交差検定法を用い，学習に 3 セット，パラメータ調整に 1 セット，テストに 1 セットとした．ブログ推定器の重みはパラメータ調整用のセットを用い決定した．

4.4　評価方法

推定した年齢の評価方法として，平均二乗誤差 (Mean squared error:MSE) と相関係数 (correlation:r) を用いた．ここで，n は事例数，x_i は事例 i の正解値，y_i は事例 i の予測値，$\bar{x}$ は全事例の正解値の平均値，$\bar{y}$ は全事例の予測値の平均値を表す．

平均二乗誤差の計算方法は次の通りである．

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \qquad (1)$$

相関係数の計算方法は次の通りである．

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}} \qquad (2)$$

4.5　実験結果

各手法の精度を表 1 に示す．リンク推定器の MSE=35.3，r=0.78 と比べ，ブログ推定器では MSE=28.2，r=0.83 となりリンク推定器よりも精度が高かった．ブログ情報，リンク情報を単独で用いる手法と比較し，ブログ情報とリンク情報を使う提案手法では MSE=23.8，r=0.85 となり，提案手法の有効性が確認できた．また，提案手法に加えラベルなしユーザを利用した場合，MSE=22.7，r=0.86 と最も精度が高くなり，ラベルなしデータを利用する有効性を示すことができた．

各手法において，ユーザの実際の年齢と MSE の比較を図 3 に示す．ブログ情報，リンク情報を単独で用いる手法と比べ，ブログ情報とリンク情報を使う提案手法は，ほぼ全ての年齢において精度が上回っている．提案手法に加えラベルなしユーザを利用した場合も同様に単独で用いるものより精度が高く，特に 26 歳から 40 歳においてはラベルなしユーザを利用しない手法より高い精度が確認できる．

5 考察

提案手法ではブログ情報，リンク情報を単独で用いる手法よりも高い精度が確認できた．これは，一方の推定器で誤った推定をした場合に，他方の推定器をあわせることにより誤りが軽減されるためだと思われる．また，実験データの傾向として，年齢が高くなるに従い，ユーザについているリンクの数は減少する．提案手法においては，年齢が高くなると利用できるリンクの数が減少するためリンク推定器の効果が小さくなり，ブログ推定器とより近い精度となる．

ラベルなしユーザを利用する手法の場合，利用しない場合に比べ，特に 26 歳から 40 歳において精度の向上が確認できる．これは，リンク先の中にラベル付きユーザが少ない比較的高い年齢において，ラベルなしユーザを推定することにより利用できるリンクの数が増えるためだと考えられる．そのためもともと大量のラベル付きユーザのいる低年齢層ではあまり効果が見られなかった．

全ての手法において，年齢が高くなるにつれ MSE が急激に大きくなることがわかる．同様の現象は先行研究でも報告されている [3]．原因として，ブログ推定器では，今回用いた単語 n-gram などの素性では年齢の高いユーザの特徴を捉えきれなかったことが考えられる．また，リンク推定器においても，36 歳付近以降 MSE が大きく上昇する．図 2 のようにリンク先の平均年齢においても 36 歳付近から対応がとれておらず，単純にリンク先の平均年齢を用いる手法では高年齢のユーザの年齢推定は難しいと考えられる．しかし，高年齢ユーザにおいても，自身と近い年齢のユーザがリンク先に存在する場合はあるため，単純な平均年齢ではなく，よりターゲットユーザに似ているだろうユーザに重みをつける等の対応が考えられる．

表 1　手法ごとの精度の結果

手法	MSE	r
ブログ推定器	28.2	0.83
リンク推定器	35.3	0.78
ブログ推定器 ＋ リンク推定器	23.8	0.85
ブログ推定器 ＋ リンク推定器 ＋ ラベルなしユーザ	**22.7**	**0.86**

6 おわりに

本研究では，ブログ著者の年齢推定において，ブログ情報とリンク情報の両方を用いる手法を提案した．実験ではブログ情報のみを使う手法，リンク情報のみ使うを手法，ブログ・リンクの両方の情報を使う手法，ブログ・リンクの両方の情報に加えラベルなしユーザを使う手法を比較した．その結果，ブログ・リンクの両方の情報を使う手法では，ブログ情報のみ・リンク情報のみを用いる手法より精度の向上がみられ，提案手法の有効性が示された．また，ブログ・リンクの両方の情報に加えラベルなしユーザを用いる場合が最も精度が高く MSE が 22.7 となった．

今後の課題としてリンク推定器の構築方法が挙げられる．本手法では 3 つのリンク機能を，同じリンク情報として全て対等に扱った．しかし，例えばリンク機能の “友だち” では他のリンク機能よりも近い属性のユーザとつながりやすいなどの特徴があると思われる．このようなリンク機能の違いを考慮することにより，リンク推定器の精度が改善されると考えている．

参考文献

[1] Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123, 2011.

[2] Ian MacKinnon and Robert H. Warren. Age and geographic inferences of the livejournal social network. In Edoardo Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, and Alice X. Zheng, editors, *Statistical Network Analysis: Models, Issues, and New Directions*, volume 4503 of *Lecture Notes in Computer Science*, pages 176–178. Springer Berlin Heidelberg, 2007.

[3] Daisuke Ikeda, Hiroya Takamura, and Manabu Okumura. Semi-supervised learning for blog classification. In *Proc. of AAAI*, pages 1156–1161, 2008.

[4] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:1817–1853, 2005.