モバイル検索ログを用いた機械学習手法による年代推定法の検討

佐野 勝浩† 徳永 幸生† 杉山 精‡ 貝谷 實榮* 木村 義彦*
芝浦工業大学大学院 工学研究科† 東京工芸大学‡ エフルート株式会社*

1．はじめに

ユーザの性別，年代，職業などの人口統計学的な属性データであるデモグラフィック情報は広告などの様々な分野で活用されている．

しかし，近年プライバシ問題に対する意識が高まっており，ユーザの個人情報を入力することに対する抵抗感が強くなっている．そのため，デモグラフィック情報の取得は一般に容易ではない．

一方，現在の携帯電話の普及率は 85.8%と高く[1]，携帯電話は1人1台の時代を迎えつつある．そこで，携帯電話を用いて Web 検索された際のログであるモバイル検索ログを分析し，デモグラフィック情報を推定することを試みる．

デモグラフィック情報の中でも年代情報は，年代により購入する商品の価格帯が異なるため，広告分野において重要な情報である．特に 10 代と 20 代以降では価格帯の差が大きく見られる．そのため，まずはモバイル検索ログから 20 代以降を判別する推定を行う．

2．年代の推定法

我々はこれまで，年代ごとに知っている可能性の高い固有名詞を集めた「年代別固有名詞データベース」を作成し，利用することで検索者の年代を推定してきた[2]．この推定法に簡易的な機械学習を推定に組み合わせることで，推定の精度を向上できる見通しが得られている．

また，関連研究としてブログ記事の筆者の年代を推定する研究が広く行われている．その研究のひとつに，機械学習手法のひとつであるナイーブベイズ分類器を用いてブログ記事の筆者の年代等を推定することを通して，男女別・地域別・年代別傾向を分析した研究がある[3]．

An examination of age estimation based on machine learning database gleaned from a mobile- WWW search log
† Masahiro SANO(m109039@shibaura-it.ac.jp)
† Yukio TOKUNAGA(tokunaga@shibaura-it.ac.jp)
‡ Kiyoshi SUGIYAMA
‡ Jitsuei KAITANI ‡ Yoshihiko KIMURA
† Graduate School of Engineering Shibaura institute of Technology
‡ Tokyo Polytechnic University
* Froute Corporation

このことから，機械学習手法のひとつであるナイーブベイズ分類器がモバイル検索の検索者の年代推定法として有効である可能性は高い．しかし，ブログ記事は文章として書かれているのに対し，モバイル検索ログでは単語単位で書かれることが多いなど異なる点が多くある．

そこで，本稿ではモバイル検索におけるナイーブベイズ分類器の年代推定への適用可能性について検討する．

3．ナイーブベイズ分類器

ナイーブベイズ分類器は，クラス $c_i(1 \le i \le n_1)$ の事前確率 $P(c_i)$ と素性 $x = (x_1, x_2, \ldots, x_j, \ldots, x_{n2})$ $(0 \le j \le n_2)$ の条件付き確率 $P(x \mid c_i)$ が与えられたときに，クラスの条件付き確率 $P(c_i \mid x)$ を最大化するクラス $\hat{c}$ を求める問題として定式化され，式(1)のように表される．

$$\hat{c} = \arg\max_{ci} P(c_i) \prod_j P(x_j \mid c_i) \qquad (1)$$

また，ナイーブベイズ分類器の精度を高めるために，タームギャップ指標を導入する．

タームギャップ指標は分類に効いている素性だけを利用することによって分類精度を高めるための指標であり，$x_j \in x$ なる素性 x_j に対し，その条件付き確率 $P(x_j \mid c_i)$ を降順に並べたときの隣接する条件付き確率の常用対数の差の最大値をタームギャップ指標と定義している[3]．この値が閾値 T_{TG} を越えていれば，その素性を分類に用いることにより，分類に効いているものだけを推定に利用する．

4．推定に使用するデータと評価方法

検索者の年代推定には，エフルート株式会社が運営する検索サイト froute.jp (http://froute.jp) の 2009 年 1 月～9 月の 9 ヶ月間のモバイル検索ログを使用した．一般の検索では，ジャンルに関わらず検索する総合検索が主流だが，froute.jp では，総合検索に加えて 14 のジャンルに関して，ジャンルを限定して検索できる．そのため，ジャンル別検索も考慮して推定を行った．

結果の評価には，20 代以降のユーザを抽出する場合の適合率と再現率を用いる．適合率と再現率は式(2)のように定義する．

$$(\text{適合率}) = \frac{U_{success}}{U_{estimated}} \quad (\text{再現率}) = \frac{U_{successss}}{U_{all}} \qquad (2)$$

$U_{success}$：推定に成功したユーザ数

$U_{estimated}$：20代以降と推定したユーザ数

U_{all}：検証に用いた20代以降の検索ユーザ数

5．検索者の年代推定

クラス c_i として検索者の年代が 10 代のクラスと 20 代以降のクラス，素性 x として検索語を用いたナイーブベイズ分類器により，検索者の年代を推定する．

学習データには，ジャンル別検索を含めたすべてのモバイル検索ログを用いた．ユーザの年代情報は半分に分割し，一方を学習データ，もう一方を検証データとして使用した．

タームギャップ指標のしきい値 T_{TG} であるが，しきい値を上げると，適合率は上がり，再現率が下がる傾向がある．このように，適合率と再現率はトレードオフの関係にある．このしきい値 T_{TG} を決定するため，すべてのジャンル別検索を含めたモバイル検索ログを用いて，しきい値を変化させたときの適合率，再現率の変化を表 1 に示す．

表 1　しきい値 T_{TG} を変化させた結果

しきい値 T_{TG}	適合率	再現率
0.0	52.83%	52.95%
0.5	52.57%	23.48%
1.0	56.92%	23.92%
1.5	58.02%	24.60%
2.0	58.02%	24.60%
2.5	58.02%	24.60%

10 代と 20 代以降の 2 クラスの推定なので，適合率が 50%に近いとランダム抽出と変わらない結果となる．また，$T_{TG} \geqq 1.5$ では他の結果と比べ適合率，再現率が共に高くなっているため，本稿ではしきい値として $T_{TG}=2.0$ を採用した．

6．年代推定の結果と考察

ジャンル別検索の検索者が多い順に 5 つのジャンル別検索ログと総合検索ログに対し，検索者の年代を推定した結果を表 2 に示す．

表 2　年代の推定結果

ジャンル	適合率	再現率	20代以降の割合
総合検索	58.87%	23.27%	54.95%
画像	67.59%	24.45%	58.67%
YouTube	51.95%	29.32%	51.04%
辞書	62.12%	27.36%	56.13%
動画	67.24%	23.42%	52.65%
ニコニコ動画	59.78%	17.21%	50.20%
(全検索ログ)	58.02%	24.60%	55.19%

比較のために，すべてのユーザを 20 代以降と決定した場合を考える．その場合，適合率は 20 代以降の割合と同じになる．よって 20 代以降の割合と推定結果を比較する．

総合検索や全検索ログでは，適合率は平均 3.38%しか向上していない．しかし，画像，辞書などジャンル別検索ごとに推定すると，適合率は平均 8.00%と大きく向上していることがわかる．この結果から，総合検索など様々なジャンルを混在させると年代による検索語の傾向が出にくい傾向にあると考えられる．

ジャンル別検索の推定結果を詳細に見ると，動画検索は適合率が 14.59%向上した一方，YouTube 検索では適合率が 0.91%しか向上していない．この結果から，年代による検索語の傾向が出やすいジャンルが存在すると考えられる．

7．まとめ

ナイーブベイズ分類器による年代推定を通して，検索語による機械学習手法がモバイル検索ログにおいて，ジャンルを考慮した際に年代推定に有効であるという見通しが得られた．

今後は，より質の高い年代推定を行うために，検索ジャンルを考慮した年代推定法を検討するとともに，検索語以外のパラメータの使用や他の機械学習手法を用いることを試みる．

参考文献

[1]携帯・PHS の加入契約数の推移，総務省情報通信統計データベース，Sep.2009
http://www.johotsusintokei.soumu.go.jp/field/tsuushin02.html

[2]佐野勝浩，徳永幸生，杉山精，尾下順治，星川剛彦．“モバイル検索ログを用いた年代別固有名詞データベースによる年代推定”．第 71 回情報処理学会全国大会 6P-4. (2009)

[3]松本真宏，三浦麻子．“ブログ記事における男女別・年代別・地域別傾向の分析”，第 21 回人工知能学会全国大会，2F4-9 (2007)