1N-7

携帯電話検索ログの検索語以外の要素を用いた年代推定法の検討

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† 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 90.3%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携帯電話検索ログを分析し，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 代以降を判別することは，可処分所得が大きく異なることから，携帯電話を用いた Web サービスを展開する上で極めて大きな意味を持つ．そのため，携帯電話検索ログから 20 代以降を判別する推定を検討した．

2．年代の推定法

我々はこれまで，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た年代推定法を検討した[2]．そこでは，検索ジャンルを考慮した際に有用であるという見通しが得られている．

しかし，10 代と 20 代以降という少ないクラスの推定にも関わらず精度が未だに低い．

その要因のひとつに，根拠としているデータが少ないことが考えられる．機械学習手法を用いた推定では，年代が既知である検索者のデータをもとに推定するが，年代情報の取得は困難なため，多くの検索者の情報を根拠に推定をすることは現実的ではない．そのため，多くの検索語でデータ量が不足し，それを推定に使用しているため，精度が低いと考えられる．

携帯電話検索では，固有名詞による検索が多く，検索語の種類が非常に多くなっている．それが精度の低いもうひとつの要因となっている．検索語以外の検索ログの要素には，種類が少ないものが存在する．そのため，多くのデータを集めることが容易になり，精度の向上が見込める．

そこで，本稿では携帯電話検索における検索語以外の検索ログの要素の年代推定への適用可能性について検討する．

3．携帯電話検索ログの要素

一般に検索ログには，検索者側の携帯電話から送信される要素とサーバ側で主に生成される要素がある．送信される要素には，検索語を含む検索クエリ，Referer，UserAgent(以下，UA)などがあり，生成される要素には，検索日時，検索結果，Proxy サーバなどの情報がある．

Referer は，参照元 URL のことであり，Web 検索においては，検索ボタンを押したページの URL となる．また，UA は使用しているブラウザの種類を表し，携帯電話検索では用いた携帯電話端末の種類を特定することができる．

今回は，検索者の行動が反映されやすいと予想される検索日時，Referer，UA を用いた機械学習による推定について検討する．また，種類をさらに限定するために，検索日時を 1 時間ごとに分割した検索時間帯として使用する．

4．ナイーブベイズ分類器

ナイーブベイズ分類器は，クラス $c_i(1 \le i \le n1)$ の事前確率 $P(c_i)$ と素性 $x = (x_1, x_2, \ldots, x_j, \ldots, x_{n2})$ $(0 \le j \le n2)$ 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，分類に効いている素性のみを推定に利

Study of Age Estimation based on Mobile-WWW Search Log Element except Search Word
† Masahiro SANO(m109039@shibaura-it.ac.jp)
† Yukio TOKUNAGA(tokunaga@shibaura-it.ac.jp)
‡ Kiyoshi SUGIYAMA
* Jitsuei KAITANI * Yoshihiko KIMURA
† Shibaura institute of Technology
‡ Tokyo Polytechnic University
* Froute Corporation

用することによって分類精度を高めるため，タームギャップ指標を導入する．$x_j \in x$ となる素性 x_j に対し，その条件付き確率 $P(x_j | c_i)$ 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する．この値がしきい値 T_{TG} を越えていれば，その素性を分類に用いる．これにより，分類に効いているものだけを推定に利用する．種々の値を用いた予備検討の結果を踏まえ，しきい値 T_{TG} の値は 0.4 に設定した．

さらに，根拠が少ない場合でも正確に推定するために，検索回数が少ない場合，平均の確率に近づき，検索回数が多い場合には，実際の確率に近づくような重み付き確率を用いる．平均の確率 P_a，通常の確率 P_r，検索回数 N_s によって重み付けを行った重み付き確率 P_w を式(2)のように定義する．

$$P_w = \frac{P_a + P_r \times N_s}{1 + N_s} \qquad (2)$$

これを用いることにより，根拠の少ない要素の影響を少なくさせることができる．

５．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 ヶ月間の携帯電話検索ログを使用した．機械学習を用いた年代推定法の検討で，検索ジャンルを考慮して推定することが有用であるという見通しが得られている[2]．そのため，本稿では froute.jp において最も検索回数の多い画像検索を対象とする．

結果の評価には，20 代以降の検索者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(3)のように定義する．

$$(\text{適合率}) = \frac{U_{success}}{U_{estimated}} \quad (\text{再現率}) = \frac{U_{successs}}{U_{all}} \qquad (3)$$

$U_{success}$: 推定に成功した検索者数

$U_{estimated}$: 20代以降と推定した検索者数

U_{all} : 検証に用いた20代以降の検索者数

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 として検索時間帯，Referer，UA をそれぞれ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．

年代が既知の検索者を半分に分割し，一方の検索者を学習データ，もう一方を検証データとして使用した．

６．年代推定の結果と考察

様々な検索ログの要素により，検索者の年代を推定した結果を表 1 に示す．

表１　各検索要素を用いた推定結果

検索要素	タームギャップ	適合率	再現率
検索時間帯	なし	59.36%	77.88%
	あり	66.27%	13.72%
Referer	なし	60.68%	77.24%
	あり	61.33%	14.83%
UA	なし	61.35%	49.76%
	あり	64.12%	23.80%
(検索語)	なし	59.78%	30.88%
	あり	64.13%	24.65%

タームギャップ指標を用いない場合，今回試みた検索要素すべてで，再現率を中心に推定精度が向上している．ナイーブベイズ分類器から，ちょうど放課後にあたる 16 時台から 21 時台に 10 代の検索が多いという特徴や，20 代以降の検索者は比較的古い携帯電話で検索しているという特徴など，様々な特徴をもとに推定が行われていることがわかった．このことから，検索語以外にも検索行動の違いが存在し，推定に有効な要素が存在することが明らかになった．

また，タームギャップ指標を用いた場合，UA による推定の場合は，検索語と同程度の再現率であるが，それ以外の検索要素では大きく下回る結果となっている．つまり，UA は年代による検索行動の違いがはっきりと表れる値が多く，他の検索要素は少ないことがわかる．このことから，検索要素ごとに検索の年代による特徴の表われ方が異なることがわかる．

７．まとめ

以上の結果から，検索語以外の要素による年代推定が携帯電話検索ログにおいて有効であるという見通しが得られた．

今後は，各検索要素による推定を統合する方法を検討する必要がある．また，年代推定に有効になるように，検索ログの各要素に対して検討を行う必要がある．

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.2010 http://www.soumu.go.jp/johotsusintokei/field/tsuushin02.html

[2]佐野勝浩, 徳永幸生, 杉山精, 貝谷實榮, 木村義彦．“モバイル検索ログを用いた機械学習による年代推定法の検討”．第 72 回情報処理学会全国大会 1R-4. (2010)