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

佐野 勝浩† 徳永 幸生† 杉山 精† 尾下 順治‡ 星川 剛彦‡
芝浦工業大学† エフルート株式会社‡

1. はじめに

ユーザの性別，年齢，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を用いれば，適切な購買層に絞って広告をうつことができる．このように，デモグラフィック情報は様々な分野で有用と考えられる．しかし，この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 78.7%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．具体的には，年代別固有名詞データベースを作成し，モバイル検索ログから 10 代と 20 代以降を判別する推定を行った．

2. 年代の推定

モバイル検索ログの分析から，モバイル端末における検索では一般名詞よりも固有名詞による検索が多い傾向が見い出された．また，昔流行した固有名詞で検索している例がみられた．

そこで，20 代以降のユーザは昔のことを思い出して検索すると仮定する．昔の古い固有名詞を 20 代以降のユーザ特有の検索ワードとみなし，その出現数を調べることでユーザが 20 代以降であると推定する．

提案する推定法の構成を図1に示す．

昔の古い固有名詞を判別するために，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(以下，年代別固有名詞 DB)」を作成する．

本稿では，この提案手法の有効性を確認するため，まずは音楽ジャンルでのモバイル検索ログに限定する．この音楽分野の年代別固有名詞 DB を用いて，20 代以降のユーザを推定する．

Age estimation based on chronological proper noun database gleaned from a mobile- WWW search log
Masahiro SANO†
Yukio TOKUNAGA† Kiyoshi SUGIYAMA†
Junji OSHITA‡ Takehiko HOSHIKAWA‡
Shibaura Institute of Technology†
FROUTE Corporation‡

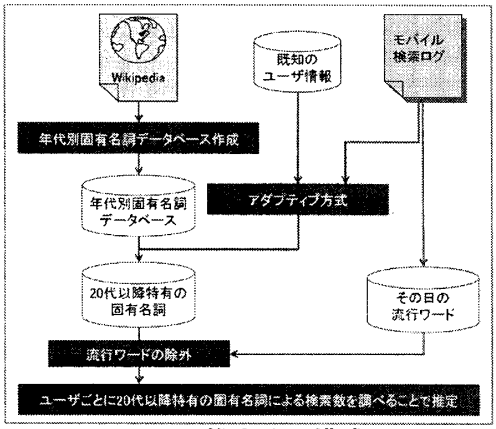

図1 推定法の構成

3. 年代別固有名詞 DB の作成と利用

年代別固有名詞 DB の作成には下記の手順で行った．

① 「フリー百科事典ウィキペディア日本語版(http://ja.wikipedia.org)」から 1970～2007 年の年間売上げ上位 50 位の邦楽のシングル，アルバム名とそのアーティスト名を年ごとに話題になった固有名詞として取得する．

② 対象とする年代の 12 歳の時代から年ごとに話題になった固有名詞を収集し，それを年代別固有名詞 DB とする．

このデータベースを利用し，20 代以降が知っている可能性が高く，10 代が知っている可能性が低い固有名詞を昔の古い固有名詞，即ち 20 代以降のユーザ特有の検索ワードと判定する．本稿では，20 代以降特有の固有名詞を 1 回以上検索したユーザを 20 代以降と推定した．

4. 年代別固有名詞 DB による推定法の補完

年代別固有名詞 DB による推定法のみでは，その作成法から最近出現した固有名詞に対応することができない．そこで，この推定法を補完する以下の2つの方法を導入する．

4.1 流行ワードの除外

最近頻繁に検索されている固有名詞は 10 代も知っている可能性が高い．20 代以降のユーザは昔を思い出して検索することから，こういった固有名詞は推定の際にノイズになると考えられ

る．そこで，最近頻繁に検索されている検索ワードを「流行ワード」とし，それを除外する．

本稿では，検索ワードに対する検索人数を1日あたりで算出し，その上位 500 位までを流行ワードとして推定の際に除外した．ここで，検索人数を使用したのは，1 日に多く検索されている場合を除外するためである．

4.2 アダプティブ方式

最近よく検索されている検索ワードに含まれる 20 代以降特有のものを推定に取り入れることができれば，推定の質を向上できる．そのために，既に年齢が判明しているユーザの検索履歴を用いて，20 代以降のユーザ特有の検索ワードを判別し，推定に加えることで年代別固有名詞 DB の推定を補完する．これをアダプティブ方式と名付ける．

本稿では，推定中の検索ログの日付が変わる際に，過去 45 日間の年齢の判明している全ユーザの検索履歴を参照し，検索ワードごとの出現回数を記録する．そこから以下の計算式を用い，アダプティブ方式の判断値 J_{adp} を算出する．

$J_{adp} = N_{over} - N_{under}$

J_{adp}：アダプティブ方式の判断値

N_{over}：20 代以降の検索回数

N_{under}：10 代の検索回数

流行ワードの除外と同様の理由で，N_{over} と N_{under} は，同じユーザが 1 日に複数回検索していた場合は 1 回しかカウントしないものとする．

本稿では判断値 J_{adp} の上位 20 位の検索ワードを 20 代以降特有の固有名詞として，推定に追加した．

5. 実験内容と評価法

年代別固有名詞 DB を用いて，モバイル検索ログから 10 代と 20 代以降のユーザを判別する推定を行った．また，年代別固有名詞 DB による推定法のみの場合とそれを補完する各手法と組み合わせた場合とで比較した．

モバイル検索ログには 2008 年 6 月～9 月の 122 日間の音楽ジャンルに限った検索ログを用いた．アダプティブ方式における既に判明しているユーザの年齢情報と推定結果の検証には，4947 名のユーザの年代情報を使用した．その内，20 代以降のユーザは全体の 50.92%にあたる 2519 名である．アダプティブ方式ではその半分のユーザの年齢情報を用いて判断値 J_{adp} を算出し，このデータを検証には用いないものとする．

結果の評価には，この推定を 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．適合率と再現率は次のように定義する．

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

$U_{success}$：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}：検証可能な20代以降の検索ユーザ数

6. 実験結果と考察

実験結果の適合率と再現率を表 1 に示す．

表 1　20 代以降の推定結果

組み合わせた手法	適合率	再現率
年代別固有名詞 DB のみ	56.72%	5.84%
流行ワードの除外	63.38%	3.45%
アダプティブ方式	58.25%	13.30%
すべて	60.58%	11.22%

年代別固有名詞 DB のみを使用した推定では 56.72%であるため，20 代以降のユーザ数の割合が 50.92%であることを考慮すると有意な結果が出ていると考えられる．

流行ワードを除外することで，適合率が 63.38%に上がり，再現率が 3.45%に下がっていることから，年代別固有名詞 DB のみを使用した推定の際にノイズとなっている情報を除去し，推定の質が向上していることがわかる．

アダプティブ方式により 20 代以降特有の固有名詞を追加することで，再現率が 13.30%に大きく向上したことから，年代別固有名詞 DB による推定とは違ったユーザが推定できていることがわかる．

すべての方法を組み合わせることで，年代別固有名詞 DB のみを使用した推定から適合率と再現率が共に向上した．

7. まとめ

年代別固有名詞 DB による推定が年代の推定に対して有効であり，流行ワードの除外によるノイズとなる情報の除去やアダプティブ方式による他の根拠からの推定を行うことで適合率と再現率を共に向上できる見通しが得られた．

今後はより質の高い推定を行うために，年代別固有名詞 DB による推定を補助する新たな方法を考案するとともに，その作成法や計算式を検討する．

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，Dec.2007 http://www.johotsusintokei.soumu.go.jp/field/tsuushin02.html