Vol.2013-NL-211 No.14
Vol.2013-SLP-96 No.14
2013/5/24

言語モデルを使ったクエリログからの性別推定

坪坂 正志[1,a)]

概要：インターネットサイトにおいて, ユーザの性別などのデモグラフィック情報を推定することはパーソナライズ検索, レコメンド, ターゲティング広告への応用などがあり重要なタスクである. 本論文ではユーザのデモグラフィック情報を推定するタスクのうち, クエリログから性別を推定するというタスクを扱う. このタスクについては既存研究はいずれも英語クエリを扱っており [8], [10], [13], 我々の調査では日本語クエリに対して適応した研究は見つかっていない. 一般に日本語の Web クエリを扱う際には必ずしもスペースで区切られておらず, 未知語やスペルミスが多く含まれ形態素解析にかけるのにも適さないという問題があり, 通常の Bag of words を使った手法を適用するのが難しいという問題が存在する. このため, 本研究では性別ごとに作成した文字 n-gram 言語モデルを使って性別を推定する手法を提案する. また, クエリ単体ではなく, ユーザが行なっている一連のクエリ情報を利用して性別を推定する手法も合わせて提案する.

キーワード：言語モデル, Web 検索

1. はじめに

インターネットサイトにおいて, 訪問ユーザの性別, 年齢情報などを知ることは検索結果のパーソナライズ, レコメンド結果の改良に用いることができる有益な情報である [1], [11], [13].

インターネットサイトの主要な収益源であるオンライン広告の分野においても一般的に性別や年齢を指定して配信するターゲティング広告の方が指定せずに配信するノンターゲッティング広告よりもインプレッション辺りの単価が高く, 訪問ユーザの性別情報などを知ることは重要である.

しかしながら性別, 年齢情報は通常一部のユーザからしか取れず, それだけを用いた場合はカバレッジが低いという問題がある. そのため, Web 上において性別を推定する研究は重要なタスクとなっている [4], [7], [8], [10], [13].

本研究では, Web 上の性別推定タスクのうちクエリから性別を推定するというタスクを扱う. この問題に対してクエリ中のスペースで区切った単語をベースに性別を判別する二値分類問題として解くことが考えられる. しかし, 日本語クエリの場合 “色彩を持たない多崎つくると、彼の巡礼の年” などのスペースが入っていないクエリなどが多く存在するため, 推定できないクエリが多いという問題がある. 通常このようなときには形態素解析器を用いて形態素単位に区切った特徴量を用いるという方法があるが Web クエリの場合は日々新たに出現する未知語が多かったり, スペルミスなどの影響もあるため, この方法を用いるのは難しい.

このため, 本研究では文字 n-gram 言語モデルを性別ごとに作成して, クエリの女性らしさ, 男性らしさを計算することによって性別を推定する方法を提案する. また, クエリ単体ではなく, ユーザが行なっている一連のクエリ情報を利用して性別を推定する手法も合わせて提案する.

2. 関連研究

Web 上のユーザに対して性別を推定するタスクについてはいくつかの研究がなされている [4], [7], [8], [10], [13].

Hu ら [8] の研究では, ウェブ検索ログにおいて飛び先 URL のコンテンツの内容およびサイトのカテゴリ情報を特徴量として性別推定を行なっている. Jones ら [10] の研究では検索クエリを Bag of words に分解して, SVM で性別推定を行なっている. Weber ら [13] の研究では性別, 年齢, 年収などの違いによって, ユーザグループのクエリが異なるかどうか, 検索画面においてクリックする URL に違いがあるかどうかについて調査している. [8], [10] では特徴量としてはスペースで区切った単語レベルの特徴量を利用しており, 本研究とは異なる. [13] ではクエリ全体を使っており, クエリ中の部分文字列とかの影響を考慮できていない.

Burger ら [4] の研究においては Twitter のツイートから

[1] ヤフー株式会社
Yahoo Japan Corporation
[a)] mtsubosa@yahoo-corp.jp

性別を当てるタスクを行なっている．ツイートには絵文字などが含まれており, また中国語などスペースで区切られていない言語のツイートが含まれているため, 文字 3-gram を特徴量として識別モデルを作成している．単語ではなく, 文字レベルの特徴量を性別推定に利用しているという点では本研究と類似しているが, 言語モデルを利用していないことまたタスクが異なるという点で本研究とは異なる．

Goel ら [7] の研究においては, ブラウザのドメインレベルでの訪問履歴を特徴量として Linear SVM でユーザの性別, 年齢などの属性情報を推定するタスクを行なっている．

Peng ら [12] は文章分類のタスクにおいて言語モデルを使った方法を提案している．彼らの手法ではクラスごとの文章集合から文字 n-gram モデルを作成して, クラス未知の文章に対しては各クラス言語モデルにおいて最も確率が高いクラスを割り当てる．本研究では彼らの手法をベースにクエリからの性別推定を行なっている．

大規模な言語モデルをクエリ処理に利用した研究としては [9] がある．ここでは言語モデルをスペル訂正, クエリセグメンテーションのタスクに利用している．奥野, 颯々野 [15] は大規模な日本語ブログコーパスにおいて, スムージング手法, カットオフを変えた時の言語モデルの性能について評価している．

3. 推定手法について

本節では本論文で使う言語モデルを使った推定手法について述べる．

3.1 言語モデル

言語モデルは与えられた文, 単語列, 文字列などに対してそれらが起こる確率を与えるモデルである [6]．機械翻訳, 音声認識などでも応用されている [3], [5].

ここでは与えられた文字列 $w_1^N = (w_1, \ldots, w_N)$ の生起する確率 $P(w_1^N)$ のモデル化に n-gram モデルを採用する．n-gram モデルは文字列に $n-1$ 次のマルコフ性を仮定するモデルであり, $P(w_1^N)$ を

$$P(w_1^N) = \prod_{i=1}^{N} P(w_i|w_{i-n+1}^{i-1}) \tag{1}$$

で表す.

$P(w_i|w_{i-n+1}^{i-1})$ の推定は, データが十分にあれば最尤法により

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i})}{C(w_{i-n+1}^{i-1})} \tag{2}$$

で推定できる．ここで $C(w_i^j)$ は学習コーパス中に文字列 w_i^j が出現する頻度を表す.

しかし, n が大きい時, 取りうる w_{i-n+1}^i の数が非常に大きくなることから, 単純に最尤推定値を用いるとテストデータ中の $C(w_{i-n+1}^i)$ がほとんど 0 となり, 系列の出現確率が 0 になる問題が発生する．

そのため, 高次の n-gram の確率をより低次の n-gram 確率を使って補完する方法が提案されている．代表的な手法として Dirichlet スムージング, Kneser-Ney スムージングなどがある [6].

3.1.1 Dirichlet スムージング

Dirichlet スムージングはゼロ頻度問題に対応するためのモデルの一つであり, $P(w_i|w_{i-n+1}^{i-1})$ の事前分布として, ハイパーパラメータが $(n-1)$-gram 確率に比例する Dirichlet 分布を仮定するモデルである．$P(w_i|w_{i-n+1}^{i-1})$ は

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i}) + \alpha P(w_i|w_{i-n+2}^{i-1})}{C(w_{i-n+1}^{i-1}) + \alpha} \tag{3}$$

と表される．このとき 1-gram 確率は最尤推定 $P(w) = \frac{C(w)}{V}$ で求める．ここで V は全単語数となる．

Dirichlet スムージングの利点として, Kneser-Ney モデルなどと比べて実装が容易かつ計算コストが低いことが上げられる．本研究では言語モデルに Dirichlet スムージングを採用した.

3.2 言語モデルを使った分類について

与えられた文字列 q をカテゴリ $c \in C = \{c_1, \ldots, c_{|C|}\}$ に分類することを考える．このときカテゴリ c を最も事後確率が高いものを選ぶことにすると,

$$c^* = \operatorname*{argmax}_{c \in C} P(c|q) \tag{4}$$

$$= \operatorname*{argmax}_{c \in C} P(q|c)P(c) \tag{5}$$

$$= \operatorname*{argmax}_{c \in C} P(q|c) \tag{6}$$

となる．上の式変形ではベイズの定理と $P(c)$ はクラスによらず一定という仮定を使った.

$P(q|c)$ としては, クラスに属するコーパス中から作った言語モデルを用いる．ここで言語モデルとして文字 n-gram モデルを採用することによってタスクや言語に依存しない分類が行える [12].

今, 男性が行ったクエリ集合 Q_M と女性が行ったクエリ集合 Q_F から言語モデルを作ったとすると, “スイーツ” のような女性が行いそうなクエリに対しては $P(q|Q_F) > P(q|Q_M)$ となり, このクエリは女性のものとして分類される．また “競馬予想” のような男性が行いそうなクエリに対しては $P(q|Q_F) > P(q|Q_M)$ となり, このクエリは男性のものとして分類される.

また, 与えられたクエリの女性らしさのスコアは

$$P(Q_F|q) = \frac{P(q|Q_F)}{P(q|Q_F) + P(q|Q_M)} \tag{7}$$

で計算する.

表 1 利用データについて

ユーザ数	14,842,461
クエリ数	894,174,541
単語数	3,295,271,252
異なり単語数	123,209,025
文字数	9,938,255,685
異なり文字数	63,840

表 2 圧縮データ構造を用いた場合の空間使用量

言語モデル (頻度情報を含む)	空間使用量
男性 7-gram モデル	3.8G
男性 7-gram モデル (圧縮後)	1.1G
女性 7-gram モデル	2.4G
女性 7-gram モデル (圧縮後)	731M

3.3 ユーザ情報を利用した推定

前項ではクエリ単一が与えられた時の性別を予測する方法について述べたが, 実際のアプリケーションにおいてはcookie 情報を利用することによって, そのクエリを行ったユーザが他にどのような検索を行なっているかの情報を利用できる. 同じクエリでも, そのユーザが普段どういうクエリを行なっているかによって, クエリの女性らしさというのは変わってくると考えられる.

ここではユーザの行った他のクエリ情報 $u = \{u_1, \ldots, u_n\}$ が与えられた場合のクエリが女性である確率を

$$P(Q_F|q,u) = \alpha P(Q_F|q) + (1-\alpha)\frac{1}{|u|}\sum_i P(Q_F|u_i) \quad (8)$$

で定義する.

4. 利用するデータ

本論文では性別推定のためのデータセットとして, 3/1 から 3/31 までの Yahoo!検索におけるクエリのうち，性別情報がとれるログインユーザのクエリを用いた. また, ページングの影響などを取り除くため, 一人のユーザが同じクエリを複数回行なっている場合, 一回としてカウントしている. 利用したデータの数については表 1 にまとめた. また, ユーザのうち 90% をモデルの学習用に利用して, 残りの 10% を評価用に用いた.

n-gram の出現頻度の保持は Trie をコンパクトに保持するデータ構造である MARISA [14] を用いた[*1] [*2].

表 2 に示すように, 効率的なデータ構造を利用することによって, 利用するメモリ量を節約できていることが分かる.

[*1] https://code.google.com/p/marisa-trie/
[*2] なお, marisa-trie のライブラリではキーに対して値を格納できないので, ID と値の情報は別ファイルでベクトルとして保存してある

表 3 n-gram モデルにおけるクエリ単体の精度

n	上位 1%	上位 5%	上位 10%
1	1.08	1.01	1
2	1.59	1.54	1.47
3	1.43	1.57	1.52
4	1.43	1.51	1.49
5	1.56	1.35	1.30
6	1.56	1.33	1.24
7	1.60	1.33	1.24

表 4 n-gram モデルにおけるユーザ情報を使った場合の精度

n	上位 1%	上位 5%	上位 10%
1	1.08	1.01	1.00
2	1.72	1.59	1.51
3	1.95	1.64	1.58
4	2.08	1.58	1.54
5	2.11	1.50	1.44
6	2.12	1.50	1.34
7	2.13	1.41	1.27

5. 実験

5.1 クエリ単体の場合の精度

表 3 にクエリ単体の場合の精度についてまとめた. 式 (7) のスコアの上位 1% クエリについては n が大きいほど精度が高くなる傾向が見られたが, 上位 10%でみると必ずしも n が大きくても精度は高くならなかった.

なお, クエリにおける男女情報の割合およびモデルの精度についてはすべて非公開情報であるため, 精度はすべて基準からの比率で表している.

5.2 ユーザ情報を利用した場合の精度

表 4 にユーザ情報を利用した場合の精度をまとめた. 精度の基準値は単体の場合と同一のものを用いている. スコアについては (8) の値を使った. 精度の傾向としては表 3 と同様の傾向が見られるが, 上位 1%クエリにおいては全体的にクエリ単体を利用する場合に比べて 30%程度の精度向上が見られる.

6. まとめと今後の課題

今回の研究ではクエリデータから作成した言語モデルを使うことにより, 性別の推定が行えるということを検証した. また, 単一のクエリではなくユーザの一連のクエリを利用すれば, さらなる精度の向上が見られることが確認できた.

今後の課題としては, 精度の改善がある. 一つは他のスムージング手法などを適応するというのがある. もう一つは利用するデータを増やすということがあげられる. 言語資源の量を増やすことにより一層の精度の改善が期待でき

る．データを増やす方法としては二つあり，一つは利用するログの期間を伸ばすこと，もう一つはログインしていないユーザのデータを利用することがある．しかし，この場合一台のサーバで処理しきれないため，モデルの分散化を考える必要がある．また，予測時とデータの性質が違ってくるため，ドメイン適用の手法を使うなどの必要性があると考えられる．

また，今回は性別の予測を行ったが年齢の推定も応用として考えられる．しかし，この場合性別と違い年齢は二値ではなく，年齢が近いユーザの性質は近いという順序構造を持っているため，それに合わせたモデリングを行う必要がある．方法としては例えば回帰の要素を取り入れたモデルを用いるというのが挙げられる [2]．

参考文献

[1] Agarwal, D. and Chen, B.-C.: fLDA: matrix factorization through latent dirichlet allocation, *Proceedings of the third ACM international conference on Web search and data mining* (2010).
[2] Blei, D. M. and McAuliffe, J. D.: Supervised topic models, *Proceedings of the Neural Information Processing Systems* (2007).
[3] Brants, T., Popat, A. C., Xu, P., Och, F. J. and Dean, J.: Large language models in machine translation, *Proceedings of the 2007 Conference on Empirical methods in natural language processing* (2007).
[4] Burger, J. D., Henderson, J., Kim, G. and Zarrella, G.: Discriminating gender on twitter, *Proceedings of the 2011 Conference on Empirical methods in natural language processing* (2011).
[5] Chelba, C., Bikel, D. M., Shugrina, M., Nguyen, P. and Kumar, S.: Large Scale Language Modeling in Automatic Speech Recognition, Technical report, Google (2012).
[6] Chen, S. F. and Goodman, J.: An empirical study of smoothing techniques for language modeling, *Computer speech and language*, Vol. 13, pp. 359–394 (1999).
[7] Goel, S., Hofman, J. M. and Sirer, M. I.: Who does what on the web: A large-scale study of browsing behavior, *Proceedings of the Sixth International Conference on Weblogs and Social Media* (2012).
[8] Hu, J., Zeng, H.-J., Li, H., Niu, C. and Chen, Z.: Demographic prediction based on user's browsing behavior, *Proceedings of the 16th international conference on World Wide Web* (2007).
[9] Huang, J., Gao, J., Miao, J., Li, X., Wang, K. and Behr, F.: Exploring web scale language models for search query processing, *Proceedings of the 19th international conference on World wide web* (2010).
[10] Jones, R., Kumar, R., Pang, B. and Tomkins, A.: "I know what you did last summer": query logs and user privacy, *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (2007).
[11] Li, L., Chu, W., Langford, J. and Schapire, R. E.: A Contextual-Bandit Approach to Personalized News Article Recommendation, *Proceedings of the Nineteenth International Conference on World Wide Web* (2010).
[12] Peng, F., Schuurmans, D. and Wang, S.: Language and task independent text categorization with simple language models, *Proccedings of HLT-NAACL* (2003).
[13] Weber, I. and Castillo, C.: The demographics of web search, *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010).
[14] 矢田　晋：Prefix/Patricia Trie の入れ子による辞書圧縮，言語処理学会第１７回年次大会 (2011).
[15] 奥野　陽，颯々野学：大規模日本語ブログコーパスにおける言語モデルの構築と評価，言語処理学会第１７回年次大会 (2011).