3E3-3

Markov Logicを用いたテキストからのユーザ属性推定

Inferring User Profile from Text using Markov Logic

平野 徹　牧野 俊朗　松尾 義博
Toru Hirano　Toshiro Makino　Yoshihiro Matsuo

日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation

User profile inference is the task that we select a label from candidates for each attributes. For example, we select Man or Woman for an attribute Gender. Most previous work built individual classifiers for each attributes, such as Gender, Age, Location, and etc. However, attributes of a single user are dependent each other. The relatedness with the other attributes can also provide classifiers useful information. Therefore, we propose a method which collectively infers user profile using Markov Logic. Experimental results show the proposed method outperformed prior methods, and increased accuracy by 4.1 points for Age and 4.0 points for Occupation.

1. はじめに

近年，Twitter や Facebook 等のソーシャルメディアが急激に増加している．これらのソーシャルメディアには，商品やサービスに対する意見や感想が投稿されていることから，ソーシャルメディアをマーケティングに活用することに注目が集まっている [池田 12].

従来主流であったアンケートによるモニタ調査は，モニタ数や質問項目数に応じて費用がかかるため，多くの情報を得ようとするとコストが高くなりがちであった．また，調査開始から集計までに時間がかかるため，リアルタイムに意見や感想を調査することができなかった．ソーシャルメディアを用いたクチコミマーケティングによって，大量の意見や感想をリアルタイムに低コストで調査することが可能になる．しかしながら，クチコミしているユーザがどんな人物であるかわからないデメリットがある．商品やサービスに対する意見や感想はユーザの性別，年代，職業などの属性に応じて異なる．そのため，属性の分布傾向を調べたり，属性ごとに意見や感想を集計して分析したりするセグメント分析がマーケティングで行われている．従来のモニタ調査であれば質問項目を設けて属性を調査することが可能であったが，ソーシャルメディアにおいては属性が明記されていないことが多く，セグメント分析ができなかった．

この問題を解決するため，ユーザの投稿内容から性別，年代等の属性を推定する研究が行われている [Cheng 10, Burger 11, 池田 12]．これらの研究では，特定の属性値を持つユーザに特徴的に現れる単語を用いて分類学習器で推定する．また，投稿内容からではなく，ソーシャルグラフから属性を推定する研究も行われている [Mislove 11, Wen 11, 蔵内 12]．これらは，ソーシャルグラフにおける近隣ユーザは互いに似た属性を持つと仮定してユーザの属性を推定する．

ユーザ属性推定タスクは，属性ごとに定められたラベルの中から 1 つを選択する問題 (多値分類問題) として定式化される．上記の先行研究では，SVM 等の分類学習器を利用して属性ごとに独立して推定を行う手法が提案されている．しかしながら，属性を個別に扱う分類器で推定する手法では，例えば，ユーザの職業が “専業主婦/主夫” のとき既婚/未婚が “未婚” になることは考えにくいが，このような矛盾した結果を出力してしまうことがある．また，属性の中でも性別は高精度に推定できるが，年代や職業などの属性に関しては改善の余地がある．

本稿では，上記の問題を解決するため，属性間に存在する依存関係に着目する．依存関係が考慮できれば，矛盾のない結果の出力が期待できる．さらには，ある属性を推定する際に，別の属性との関係が推定の手がかりになり，推定精度の向上が期待できる．例えば，ユーザの職業が “高校生” と推定されたときは，そのユーザの年代が “10 代” である可能性が高くなり，ユーザの年代が “10 代” と推定されたときは，職業が “会社員” でない可能性が高くなると考えられる．

このような相互の依存関係を捉えて投稿内容からユーザ属性を推定する手法として，本稿では，統計的関係学習の枠組みの一つである Markov Logic を用いて，属性の全てを同時に考慮しながら集合的に推定する手法を提案し，その有効性について議論する．

2. 関連研究

ソーシャルグラフから属性を推定する研究では，ソーシャルグラフにおける近隣ユーザは互いに似た属性を持つと仮定してユーザの属性を推定する手法がとられている．Wen らは，会話の回数や共通の友人数からソーシャルグラフにおけるエッジの重みを算出し，近隣ユーザの属性値を重みづけ平均する手法を提案している [Wen 11]．また Mislove らは，グラフをクラスタリングし，各クラスタに代表的な属性を付与する手法を提案している [Mislove 11]．これらは，一部のユーザの正解属性が与えられたとき，残りのユーザの属性を推定する手法である．一方，蔵内らは，投稿内容から自動推定された誤りを含む属性が与えられたとき，ソーシャルグラフを用いて誤った属性推定結果を補正することで精度向上させる手法を提案している [蔵内 12]．ただし，大半の近隣ユーザで誤った推定結果が入力されると，ソーシャルグラフを用いても推定結果の補正ができないため，入力となる推定結果もより高い精度が求められる．

投稿内容から属性を推定する研究では，各属性値を持つユーザ集合において特徴的な単語を利用する手法がとられている．これは例えば，“部活”，“試験” という単語を利用したユーザは “学生” である確率が高いなどとするものである．

連絡先: 平野 徹, 日本電信電話株式会社 NTT メディアインテリジェンス研究所, 神奈川県横須賀市光の丘 1-1, 239-0847, hirano.tohru@lab.ntt.co.jp

池田らは属性中のクラスごとに特徴語を抽出して素性とし，SVM で学習・推定する手法を提案し，年代・性別・居住域に関する推定を行っている [池田 12]．この手法は，年代・性別・居住域の属性に限定した手法ではなく，他属性の推定にも適用できる汎用的な手法である．一方，居住地推定に特化した研究 [Cheng 10] や，性別推定に特化した研究 [Burger 11] がある．Cheng らは居住地推定のための格子ベースの近隣平滑モデルを，Burger らは性別推定の手がかりとして，投稿内容だけでなくスクリーンネーム等を用いた手法を提案している．

本研究は推定する属性を限定しない汎用的な手法の確立を目指しているため，属性ごとに推定手法を変えない池田らの手法をベースラインとする．池田らの手法は属性ごとに独立して推定しているが，本研究は全ての属性を同時に考慮しながら集合的に推定する点が異なる．

3. Markov Logic

局所的な分類学習に対して，統計的な変量の間にある大域的な相互関係を協調しながら学習する統計的関係学習として，Markov Logic が近年急速に広まりつつある [Richardson 06]．これは一階述語論理と Markov Networks を組み合わせたもので，一階述語論理式に，ある程度の罰則をもって矛盾を許容する枠組みと考えることができる．また，Markov Networks を一階述語論理式によって表現するテンプレート言語とも解釈できる．自然言語処理の分野においても，実体解析 [Singla 06]，情報抽出 [Poon 07]，述語項構造解析 [Yoshikawa 11] など，大域的な制約の利用が重要な分野において利用されてきている．

Markov Logic では，推定すべき問題に必要な述語 (predicate) を定義する．例えば，ユーザの書いたテキストを入力して，そのユーザの職業と年代を推定することを考える．前者を表現する述語として occupation(i) を，後者を表現する述語には age(i) を定義する．この 2 つの述語は，推定を行う際にはその情報が与えられない引数を含むことから，潜在述語 (hidden predicate) として定義される．それに対するものとして，学習と推定の両方においてその情報が与えられる引数のみを含む，観測述語 (observed predicate) を定義する．ここでは，観測述語 word(w) を定義する．これはユーザが書いたテキスト中に単語 w が 1 回以上出現することを表す．

これらの述語を利用して，重み付きの一階述語論理式を構築する．例えば，次のような式を作ることができる．

$$\text{word}(\text{“出張”}) \Rightarrow \text{occupation}(\text{“会社員”}) \quad (1)$$

この式は “出張” いう単語を書いたユーザの職業は “会社員” であるという素性を表現していることになり，対応する重みによって，この式の確かさが表現される．一般に，大きな重みを持つ論理式ほど，そのモデルにおいてより高い確率で成立すると言える．Markov Logic では，この重みをコーパスからの学習によって自動的に獲得する．式 (1) で表現されているのと同等の意味合いは局所的な分類学習器を利用しても学習することが可能である．しかし，Markov Logic は次のような表現もできる．

$$\text{age}(\text{“10 代”}) \Rightarrow \text{occupation}(\text{“高校生”}) \quad (2)$$

この式は年代が “10 代” である時，そのユーザの職業は “高校生” であることを表しており，局所的な分類学習器では捉えることができない種類の大域的な意味合いを持っている．

表 1: 潜在述語 (hidden predicate)

述語	定義
gender(i)	ユーザの性別は i である
age(i)	ユーザの年代は i である
occupation(i)	ユーザの職業は i である
location(i)	ユーザの居住域は i である
married(i)	ユーザの既婚/未婚は i である
alcohol(i)	ユーザの飲酒の有無は i である
smoking(i)	ユーザの喫煙の有無は i である

表 2: 観測述語 (observed predicate)

述語	定義
word(w)	テキスト中に 1 回以上出現する単語 w

本研究では Markov Logic の実装として Markov thebeast[*1] を利用し，学習・推論には以下を選択した．重みの識別学習には，直接の計算コストが高い条件付確率尤度の計算が必要なため，オンライン学習手法であるマージン最大化学習アルゴリズムである MIRA[Crammer 03] を利用する．推論は最大事後確率 (MAP) 推定問題となるが，この MAP 推定を正確かつ効率的に行うために，base solver に整数線形計画法 (ILP) を用いた Cutting Plane Inference(CPI)[Riedel 08] を利用する．

4. 提案手法

本節では集合的なユーザ属性推定モデルについて，提案する Markov Logic Networks を中心に説明する．本研究では，ユーザの書いたテキストを入力とし，ユーザの性別・年代・職業・居住域・既婚/未婚・飲酒の有無・喫煙の有無の 7 つの属性を推定する．これらの推定のために，表 1 の潜在述語を定義した．次に，この 7 つの潜在述語を推定するための素性である局所論理式と大域論理式について説明する．

4.1 局所論理式 (Local Formula)

Markov Logic の定義では，ただ一つの潜在述語を持つ論理式を局所論理式 (local formula) と呼び，局所的素性を表現するのに用いられる．局所論理式で表現する素性は，先行研究 [池田 12] と同様に，単語素性を利用する．ただし，先行研究では単語の頻度情報を用いていたが，予備実験にて，テキスト中に 1 回以上出現した単語を同一に扱うとより高精度にユーザ属性を推定できたため，頻度情報は利用しない．上記の素性を与えるための観測述語を表 2 に示す．

ここで定義された観測述語はただ一つの潜在述語とともに次の式 (3)-(9) で利用される．

$$\text{word}(w) \Rightarrow \text{gender}(i) \quad (3)$$
$$\text{word}(w) \Rightarrow \text{age}(i) \quad (4)$$
$$\text{word}(w) \Rightarrow \text{occupation}(i) \quad (5)$$
$$\text{word}(w) \Rightarrow \text{location}(i) \quad (6)$$
$$\text{word}(w) \Rightarrow \text{married}(i) \quad (7)$$
$$\text{word}(w) \Rightarrow \text{alcohol}(i) \quad (8)$$

*1 http://code.google.com/p/thebeast/

表 3: アンケート項目

属性	選択肢
性別	男性 (808)，女性 (682)
年代	10 代 (97)，20 代 (357)，30 代 (470)，40 代以上 (566)
職業	会社員 (634)，自営業/個人事業 (154)，公務員 (46)，団体職員 (23)，アルバイト/パート (159)，専業主婦/主夫 (145)，大学生 (157)，高校生 (42)，無職 (97)，その他 (33)
居住域	北海道/東北 (137)，関東 (675)，北信越 (52)，東海 (163)，近畿 (244)，中国/四国 (110)，九州/沖縄 (109)
既婚/未婚	既婚 (670)，未婚 (820)
飲酒の有無	飲む (1,192)，飲まない (298)
喫煙の有無	喫う (288)，喫わない (1,202)

$$\text{word}(w) \quad \Rightarrow \quad \text{smoking}(i) \tag{9}$$

一階述語論理で記述されたこれらの式は素性のテンプレートであり，変数には具体的な値が割り当てられ，個々の素性として展開される．例えば前述の式 (1) は式 (5) から展開された式である．

式 (1) は “出張” いう単語を書いたユーザの職業は “会社員” であるという素性を表現しており，この展開された式に対して学習により重みが付与される．

4.2 大域論理式 (Global Formula)

局所論理式に対し，2 つ以上の潜在述語を含めることで大域的な推定を可能にするのが大域論理式 (global formula) である．大域論理式を以下に示す．

$$\text{gender}(i_1) \wedge i_1 \neq i_2 \quad \Rightarrow \quad \neg\text{gender}(i_2) \tag{10}$$
$$\text{age}(i_1) \wedge i_1 \neq i_2 \quad \Rightarrow \quad \neg\text{age}(i_2) \tag{11}$$
$$\text{occupation}(i_1) \wedge i_1 \neq i_2 \quad \Rightarrow \quad \neg\text{occupation}(i_2) \tag{12}$$
$$\cdots$$
$$\text{smoking}(i_1) \wedge i_1 \neq i_2 \quad \Rightarrow \quad \neg\text{smoking}(i_2) \tag{16}$$

$$\text{gender}(i) \quad \Rightarrow \quad \text{age}(j) \tag{17}$$
$$\text{gender}(i) \quad \Rightarrow \quad \text{occupation}(j) \tag{18}$$
$$\text{gender}(i) \quad \Rightarrow \quad \text{location}(j) \tag{19}$$
$$\cdots$$
$$\text{age}(i) \quad \Rightarrow \quad \text{occupation}(j) \tag{24}$$
$$\cdots$$
$$\text{occupation}(i) \quad \Rightarrow \quad \text{age}(j) \tag{30}$$
$$\cdots$$
$$\text{smoking}(i) \quad \Rightarrow \quad \text{location}(j) \tag{56}$$
$$\text{smoking}(i) \quad \Rightarrow \quad \text{married}(j) \tag{57}$$
$$\text{smoking}(i) \quad \Rightarrow \quad \text{alcohol}(j) \tag{58}$$

上に示した論理式のうち，式 (10)-(16) の 7 個は，推定する性別等の属性毎に出力するラベルを 1 つに制約するための，重みが無限大に設定される論理式である (Hard Constraint)．これらの Hard Constraint は，ユーザ属性推定を多値の分類タスクとしたことによるものである．

一方，式 (17)-(58) の 42 個は，出力されるユーザ属性間の一貫性を保ちつつ，ユーザ属性推定の性能を向上させるための論理式である (Soft Constraint)．例えば，年代と職業との間の相関関係を学習するため，年代から職業に対する論理式 (24) と，職業から年代に対する論理式 (30) とを定義している．同様に，推定する 7 つの属性間全てで論理式を定義した．これらの Soft Constraint の重みの割り当てはコーパスからの学習によって決定される．

5. 評価実験

ユーザの書いたテキストから，そのユーザの属性を推定するタスクにおいて，局所論理式だけを利用する局所モデルと大域論理式も利用する大域モデルを比較し，提案手法の有効性を検証する．

なお，前節で述べた大域論理式の Hard Constraint は，Markov Logic でユーザ属性推定を行う際のタスク設定からくる制約であるため，局所モデルでも利用する．つまり，局所モデルと大域モデルとの差分は，Soft Constraint のみである．学習と推論に関しては Markov Logic エンジンである Markov thebeast を利用した．

局所モデルとの比較に加えて，先行研究である池田ら [池田 12] の手法とも比較する．池田らの手法は，AIC(赤池情報量基準) を用いて各属性に特徴的な 2,000 単語を選択し SVM で推定する手法である．

5.1 実験データ

本実験では，ツイートデータを用いて，あるユーザが投稿した直近の 150 ツイート (1 日 5 ツイート ×30 日を想定) および自己紹介文を入力として，そのユーザの性別・年代・職業・居住域・既婚/未婚・飲酒の有無・喫煙の有無の 7 つの属性を推定する．実験に際して，ツイートをしている 1,490 ユーザに対して上記 7 項目のアンケートを実施した．アンケートは各属性に用意された選択肢から 1 つを選択するものである．各属性の選択肢を表 3 に示す．選択肢の作成に当たっては先行研究 [池田 12] や一般的な市場調査を参考にした．各属性の選択肢は，性別が 2 個，年代が 4 個，職業が 10 個，居住域が 7 個，既婚/未婚が 2 個，飲酒の有無が 2 個，喫煙の有無が 2 個であり，アンケート回答数は括弧内の数値の通りである．

アンケートに回答した 1,490 ユーザのツイートデータを入力とし，アンケート回答結果を正しく推定できるか評価した．なお実験は 5 分交差検定で実施した．

5.2 実験結果

性別・年代・職業・居住域・既婚/未婚・飲酒の有無・喫煙の有無の 7 つの属性推定において，全ての属性を同時に考慮しながら集合的に推定することによりどの程度推定性能が向上するか調査した．実験結果を表 4 に示す．なお正解率とは，

表 4: ユーザ属性推定の正解率 [%]

	大域モデル	局所モデル	先行研究
性別	**85.5**	81.6	73.9
年代	**56.6**	52.5	48.1
職業	**49.1**	45.1	38.4
居住域	**47.9**	**47.9**	41.3
既婚/未婚	**72.0**	68.4	62.6
飲酒の有無	**80.9**	79.0	76.4
喫煙の有無	**80.4**	79.7	78.5

システムの推定結果がどの程度正しいかを示す尺度であり，次式の通りである．

$$\text{正解率} = \frac{\text{正しく属性値を推定できたユーザ数}}{\text{全ユーザ数}}$$

表 4 では，各属性で最も高い正解率の数値を太字で示している．提案する大域モデルは，局所モデルや先行研究と比べ，全属性で最も高い正解率であり，その有効性が確認できた．

大域モデルと局所モデルを属性ごとに比較すると，大域モデルは，年代で 4.1 ポイント，職業で 4.0 ポイント，性別で 3.8 ポイント，既婚/未婚で 3.6 ポイント，飲酒の 有無で 1.9 ポイント，喫煙の有無で 0.7 ポイント向上していることがわかる．一方で，居住域では局所モデルと差がない．この原因を探るべく学習されたモデルパラメータの分析を行った．学習によって大きな重みが割り当てられた属性間の論理式の一部を以下に示す．

$$\begin{aligned}
\text{occupation}(\text{“高校生”}) &\Rightarrow \text{age}(\text{“10 代”})\\
\text{occupation}(\text{“会社員”}) &\Rightarrow \neg\text{age}(\text{“10 代”})\\
\text{alcohol}(\text{“飲む”}) &\Rightarrow \neg\text{age}(\text{“10 代”})\\
\text{married}(\text{“未婚”}) &\Rightarrow \neg\text{age}(\text{“40 代以上”})\\
\text{age}(\text{“10 代”}) &\Rightarrow \text{occupation}(\text{“高校生”})\\
\text{age}(\text{“40 代以上”}) &\Rightarrow \neg\text{occupation}(\text{“大学生”})\\
\text{married}(\text{“未婚”}) &\Rightarrow \neg\text{occupation}(\text{“専業主婦”})\\
\text{alcohol}(\text{“飲む”}) &\Rightarrow \text{gender}(\text{“男性”})\\
\text{age}(\text{“40 代以上”}) &\Rightarrow \text{married}(\text{“既婚”})\\
\text{age}(\text{“10 代”}) &\Rightarrow \text{alcohol}(\text{“飲まない”})\\
\text{smoking}(\text{“喫う”}) &\Rightarrow \text{alcohol}(\text{“飲む”})
\end{aligned}$$

例えば，職業が “高校生” ならば年代が “10 代” であるや，職業が “会社員” ならば年代が “10 代” でないなどに大きな重みが割り当てられている．これらを分析すると，1. 居住域と他属性間の論理式に小さな重みが割り当てられていること，2. 大きな重みが割り当てられた論理式に年代・職業が多いこと，がわかった．前者は，居住域と他属性に依存関係がないことを意味している．そのため大域モデルと局所モデルで性能に差が出なかったと考えられる．後者は，年代・職業が他属性と依存関係になりやすいことを意味している．上記の実験結果に鑑みると，他属性と依存関係になりやすい属性ほど大域モデルの効果が大きいと考えられる．

大きな重みが割り当てられた論理式には，局所モデルでは生じていた矛盾する推定結果を防ぐための式が含まれている．これらの式によって，例えば，飲酒が “飲む” なのに年代が “10 代” や職業が “高校生” なのに年代が “30 代” といった矛盾する推定結果を大域モデルでは改善できた．

6. おわりに

本稿では，ユーザの書いたテキストから，そのユーザの属性を推定するタスクに取り組み，従来手法のように属性ごとに独立に推定するのではなく，推定する属性の全てを同時に考慮しながら集合的に推定する手法を提案した．評価実験では，提案手法は局所モデルや先行研究と比べ，最も高い正解率でユーザ属性を推定できることを確認できた．

今後は，更なる推定精度向上を目指し，先行研究で報告されている属性に特化した素性の導入を検討したい．また，今回の 7 属性以外の属性 (趣味や性格など) の推定にも取り組みたい．

参考文献

[Burger 11] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G.: Discriminating Gender on Twitter, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309 (2011)

[Cheng 10] Cheng, Z., Caverlee, J., and Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768 (2010)

[Crammer 03] Crammer, K. and Singer, Y.: Ultraconservative Online Algorithms for Multiclass Problems, *Journal of Machine Learning Research*, Vol. 3, pp. 951–991 (2003)

[池田 12] 池田 和史, 服部 元, 松本 一則, 小野 智弘, 東野 輝夫 : マーケット分析のための Twitter 投稿者プロフィール推定手法, 情報処理学会論文誌 コンシューマ・デバイス&システム, Vol. 2, No. 1, pp. 82–93 (2012)

[藏内 12] 藏内 雄貴, 内山 俊郎, 内山 匡 : マルコフ確率場を用いたソーシャルネットワークからのユーザ属性推定, 第 5 回 Web とデータベースに関するフォーラム WebDB Forum 2012 (2012)

[Mislove 11] Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N.: Understanding the Demographics of Twitter Users, in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (2011)

[Poon 07] Poon, H. and Domingos, P.: Joint Inference in Information Extraction, in *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*, pp. 913–918 (2007)

[Richardson 06] Richardson, M. and Domingos, P.: Markov Logic Networks, *Machine Learning*, Vol. 62, No. 1-2, pp. 107–136 (2006)

[Riedel 08] Riedel, S.: Improving the Accuracy and Efficiency of MAP Inference for Markov Logic, in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (2008)

[Singla 06] Singla, P. and Domingos, P.: Entity Resolution with Markov Logic, in *Proceedings of the Sixth International Conference on Data Mining*, pp. 572–582 (2006)

[Wen 11] Wen, Z. and Lin, C.-Y.: Improving User Interest Inference from Social Neighbors, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1001–1006 (2011)

[Yoshikawa 11] Yoshikawa, K., Asahara, M., and Matsumoto, Y.: Jointly Extracting Japanese Predicate-Argument Relation with Markov Logic, in *The 5th International Joint Conference on Natural Language Processing*, pp. 1125–1133 (2011)