1T-04

Twitter 投稿文からの嗜好の推定における性格情報の有効性

小原　博明†　　白井　靖人†
静岡大学大学院総合科学技術研究科†

1. はじめに

情報化社会の発達によって誰でも気軽に情報を発信できるようになった。情報発信の主要な手段の 1 つである SNS では多くのユーザが投稿を行うため利用者をターゲットとした情報推薦が多く行われている。

情報推薦にあたって性別、年齢、居住地といった情報は有用であるが、SNS の性質上匿名での利用が多く、プライバシー、不正アクセスなどへの対策といった観点から明示されていない情報であるため取得が難しい。

それ以外に個人が公開している情報の 1 つにエゴグラム[1]という性格の情報がある。エゴグラムには 5 つの自我状態、すなわち、CP(Critical parent)厳しい親、NP(Nurturing parent)優しい親、A(Adult)大人、FC(Free Child)子ども、AC(Adapted Child)従順な子どもがある。各自我状態は abc の 3 段階があり a が高く c が低い。この組み合わせで性格を判断する。

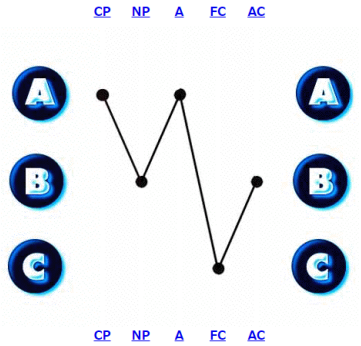

図 1　エゴグラムの例

例えば上記の図 1 のエゴグラムは abacb となる。

本研究では投稿文からの嗜好の推定においてこの性格に着目し、SNS 上でエゴグラムを診断する性格診断サイト[2]を利用した人を対象に嗜好の推定を行い、性格情報を加えたものとそうでないものの比較及び考察を行う。

続く 2 章ではこの研究を行うにあたって参考にした先行研究や関連研究、及び本研究における情報推薦の提案を、3 章ではその具体的な実装方法、4 章では比較結果を、5 章では結果に関する考察について述べる。

The effect of personality information on preference estimation from Twitter
†Hiroaki Obara, Graduate School of Informatics Shizuoka University
†Yasuto Shirai, GraduateSchool of Informatics Shizuoka University

2. 先行研究

テキストから性格の推定を行う研究は南川ら[3]がエゴグラムと個人 blog で利用される単語の対応関係を Naïve Bayes を使用して推定しているものがある。

SNS の投稿文から利用者の情報を推定する手法としては田中ら[4]の Naïve Bayes を利用するものや、池田ら[5]のマーケティングに利用される情報を SVM で推定するものがある。

情報推薦に当たって個人情報は年齢、性別、居住地が特に重要とされている[5]。実際にこれらの情報を集めることは難しいためそれ以外の取得できる情報を考える。Twitter にはプロフィールを記述するスペースがあるので、プロフィールから抽出した単語をユーザの嗜好情報とすることにした。

本研究では twitter の投稿文と性格から利用者の嗜好情報を推定し、性格情報の有無での推定結果を比較する。

3. 実装

以下の手順で実験を行う。

1. データの取得
   twitterAPI を使用し性格診断以降の 100 件のツイートを収集する。
   エゴグラムは 5 つの各自我状態、3 段階の強弱の 243 パターンがあるが、約 100 パターン 200 件のデータを取得した。
2. データの整形・抽出
   取得した投稿文を janome で形態素解析し分かち書きをする。素性は名詞のみを抽出する。抽出した名詞一覧のうち特徴を表す名詞以外のもの(数字や記号、代名詞等)を除去する。
3. 嗜好情報の分類辞書の作成
   プロフィール情報から利用者の嗜好を表すと考えられる名詞を抽出し、gensim でカテゴリー分類辞書を作成する。

4. 特徴語辞書を作る
   2. で作成した結果を読み込み単語ごとの出現率を表す特徴語辞書を作成する。

5. 特徴ベクトルを作る
3. で作成した辞書をもとに各利用者の特徴ベクトルを作成する。
6.1 性格情報なしでの嗜好の推定
性格ごとの取得数が多い 5 人、6 人のものについて学習と推定を行った。
テキストに対し嗜好（1 ユーザが複数の嗜好を持つことがある）のマルチラベル分類を random forest を使用し行う。
6.2 性格情報なしでの嗜好の推定
6.1 性格情報を付加しマルチラベル分類を random forest を使用し行う。

なお、性格なしで学習を行う際は性格情報にニュートラルな情報を付加したいため、エゴグラム bbbbb の性格を付加した。これは全ての自我状態が a と c の中間に位置する為、目立った特徴がない性格と考えられるからである。

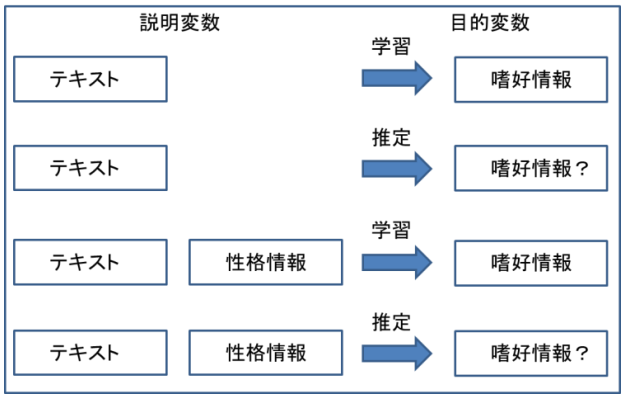

図 2　多ラベル分類問題におけるデータの関係

また、本研究のような 1 つのデータに複数のラベルが付加される多ラベルの分類問題(図 2)に対しては、binary relevance など 2 値問題へ還元する手法もあるがラベル間の相関を考慮していないため、random forest を採用した。

4. 結果

2 つの性格グループにおいて嗜好推定の実験を行った。結果を表 1 に示す。

表 1 嗜好推定結果

	一致率	
辞書	性格なし	性格あり
性格A	49.68	変化なし
性格B	46.35	変化なし
性格A及びB	44.81	49.65

性格 A はエゴグラム bbaba、B は cbaba のグループである。同じ性格の利用者のグループで辞書を作成し、学習に用いたデータを使って推定を行った場合、性格情報を付加しても嗜好の一致率は変化しないが、異なる性格の利用者のグループで特徴語辞書を作成し性格情報を付加した場合に、若干推定結果が向上した。

5. 考察、今後の課題

2 グループ間での嗜好推定を行い、性格情報を付加したグループのほうが嗜好推定の結果が良いことがわかった。推定に影響を与えている要因としては形態素解析では素性として名詞のみを抽出しているが、略語やスラングといった語は正しく分かち書きできないことがある。この問題に対してはこれらの単語を解析対象外にするか、辞書に組み込むといった方法がある。

今後の課題としては、全ての性格について推定をすることができなかったこと、分類辞書の作成に人手がかかること、などが挙げられる。

また、twitter 登録者を対象に実際の嗜好をアンケートを実施して裏づけをとるといったことを検討し、推定に役立てていきたい。

参考文献

[1]新里里春, ジョン・M・デュセイ, エゴグラム―ひと目でわかる性格の自己診断, 1985, 創元社
[2]エゴグラムによる性格診断, 2018-01-12 確認, http://www.egogram-f.jp/seikaku/
[3]南川敦宣, 横山浩之, テキストマイニングによる個人 Blog データからの性格推定手法, 2010, データマイニングと統計数理研究会第 12 回, pp. 96-100
[4]田中聡, 松本和幸, 吉田稔, 北研二, 情報推薦のための twitter ユーザの性格分析手法, 2016, 人工知能学会第 30 回年次大会
[5]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, 2012, 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