非負値行列因子分解を用いた文章の書き手判別

福井 諒太[1]，増井 隆治[2]，久富 望[3]，曽我部 舞奈[4]，段 正楠[5]，大関 真之[6*]

[1]三重県立津高等学校，[2]京都大学工学部，[3]京都大学大学院情報学研究科，[4]京都大学大学院医学研究科，[5]京都大学農学部，[6]東北大学大学院情報科学研究科

要旨

企業は利益を上げるために，消費者の情報を分析する必要がある．ゆえに分析の効率化は重要であると言える．そこで本研究では，コンピュータにおいて機械学習を行い，レビューの自動解析に挑んだ．この機械学習の手法の一つに非負値行列因子分解がある．まず負の値でない数字で構成されている行列を非負値行列という．そして非負値行列を二つの非負値行列の積に分解することを非負値行列因子分解という．この研究では，文章を構成する単語の出現回数をもとに作った非負値行列において，非負値行列因子分解を適用，テストを行うことで正答率を算出した．結果，非負値行列因子分解は，人間の認識と似た判別ができる可能性が示唆された．

目的

文章を書く人間には年齢や性別といった特徴がある．人は顔文字や感嘆符の多い感情豊かな文章を見ると，女性の文章でないかと推測できる．そのように文章の書き手の特徴の差異が，文章の違いとして現れるならば，文章の書き手がどんな人か判別できる．さらに，コンピュータによって文章の書き手の特徴を判別することで社会に貢献できる点がある．ネットショッピングを運営する企業は，購入者の商品に対する意見であるレビューを収集している．このレビューを自動で解析することができ，そしてレビューの書き手の情報をマーケティングに利用できるのだ．そこで本研究は，コンピューターによる文章の書き手の判別という課題に挑んだ．

序論

人間が自然に行う学習と同等のことをコンピューター上で実現しようとする試みを機械学習という．機械学習は大量のデータを収集し，それをさまざまの手法を用いて解析することで行う．そしてその機械学習の手法の一つに，非負値行列因子分解（Non-negative matrix factorization，以下，NMF と略す）がある．行列とは，数字を長方形状に並べたものである．特に行列を構成する値が負の値でないとき，これを非負値行列という．そして非負値行列を二つの非負値行列に分解することを NMF という．先行研究 (1) として，NMF を白黒の顔画像の輝度の情報で作った非負値行列に適用したものがある（図 1）．

このとき分解された二つの行列は目や鼻といった顔に特徴的な構造を表すものと，もとの顔画像を復元するための構造の組み合わせ方を示すものの二つに分かれる．そのため，レビューに NMF を適用するとき共通する特徴と，各年代における特徴の組み合わせ方が得られることが期待できる．またこのとき，分解には二つのメリットがある．一つは行列が非負値で構成されており，組み合わせ，つまり足し算で，元画像を復元しようとするため，非常に単純な操作となるという点であり，二つ目は特徴的な構造の組み合わせるという手順が，人間の認知の仕組みと似ているという点である．ゆえに人間の直感と類似した結果が期待できる．実際、Lee らの論文では様々な機械学習の手法での顔画像の表現を比較していた．人間が顔を認識するとき，特徴的な構造として用いるのは目や鼻のようなパーツである．そして，目や鼻といった部分構造を得られたのは NMF だけであった．

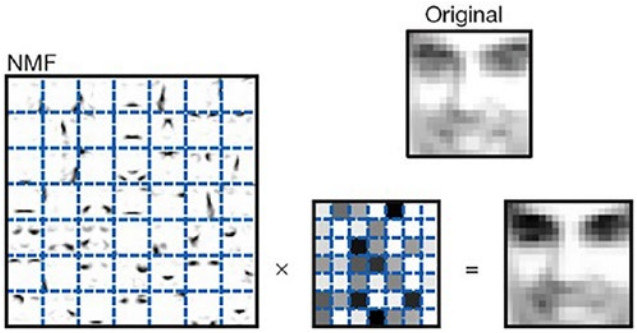

図 1. NMF による顔画像の表現 (1).

方法

今回，対象とした文章は楽天グループ，楽天市場に集められた「マキタ　コードレス掃除機　充電式クリーナ」に関するレビューである．この商品はレビューの数が多く，様々な年齢性別の書き手が多いため，解析の対象に選んだ．まず書き手が，30 代の男性，30 代の女性，40 代の男性，40 代の女性，50 代の男性，50 代の女性であるレビューをそれぞれ 150 件ずつ，計 6 × 150=900 件のレビューを学習データとして集める．そして形態素解析エンジン MeCab を使い，それぞれのレビューを構成する単語の品詞情報を調べる．ここで，「小学生から高校生までの男女を対象とした作文の分析の結果，形容詞 1 個に対する動詞の数は男 5.40，女 4.34 である．」「1000 字の文章中の強意副詞の出現頻度平均は男性 2.8 回，女性 3.8 回である」という記述 (2) を基に，特徴的な品詞を名詞，代名詞，形容詞，副詞，終助詞，記号に絞った．ただ，該当する全ての単語を用いると，データ量が過剰になる．そこで全レビューにおいて 3 回以上出現した上記の品詞の単語のみをリスト化し，特徴量として用いる．そして，それぞれのリストに含まれる単語の出現回数を調べる．行列は図 2 に示すように，行列の一つの列は一つのレビューを表し，各行はリストの単語の出現回数となるように作る．

また，行列を構成する値は単語の出現回数を数えたものであるから，構成する数字はすべて非負値である．そしてこの行列

* 内容に関する連絡先：mohzeki@smapip.is.tohoku.ac.jp

図 2. 行列の作り方

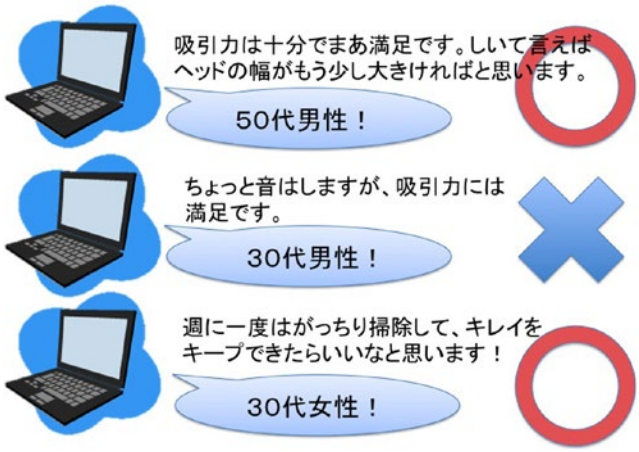

図 4. NMF によるレビュー判別の例

に NMF を適用することによって，女性は形容詞を多く使う，50 代の人は固い言葉を好んで使う，といったそれぞれの年齢性別の特徴が抽出できることが期待される．次に正答率のテストを行う．新たに，同じ 6 つの書き手の特徴をもつレビューを，30 件ずつテストデータとして集め，同様の手順で非負値行列を作る．そして，学習時と同じ共通する特徴をもとに分解し，組み合わせ方の情報を持つ行列に注目する．分析には最小二乗誤差法を用いた．各年代における特徴の組み合わせ方の行列の中から，二乗和誤差が一番小さい行列が答えであると判断させた．今回は性別のみの判別，年齢のみの判別，年齢性別の両方の判別の 3 つで正答率を調べた．

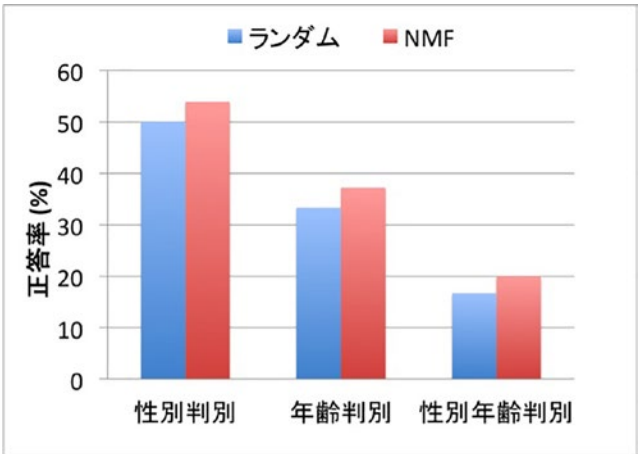

図 3. レビュー判別の正答率

結果

性別の場合，ランダムで判別を行うと正答率は 50% となる．なぜならば選択肢が「男性」「女性」の 2 通り存在するためである．同様に年齢の場合，「30 代」「40 代」「50 代」と 3 通りの選択肢があるため，年齢性別の場合は選択肢が 2 × 3 ＝ 6 通り存在するから，ランダム判別での，それぞれ正答率は 33.33%，16.67% となる．NMF の性別の判別は，正答率が 53.89% と 3.89% 高かった．また年齢判別は 37.22%，年齢性別両方の判別は 20.00% と，それぞれ確率より 3.89%，3.33% 高くなった．

考察

NMF が誤った判別を行った文章は，中性的な文章など人間が見ても判別しにくいレビューが多い印象を受けた．しかし，感嘆符や形容詞が多い「週に一度はがっちり掃除して，キレイをキープできたらいいなと思います！」（30 代女性）の文章や，「しいて言えば」のような語を用い，硬い印象を受ける「吸引力は十分でまあ満足です．しいて言えばヘッドの幅がもう少し大きければと思います」（50 代男性）のような特徴的な文章は，判別ができた．このことから NMF は人間の認識と似た判別ができた可能性が高い．

今後の展望

今回は文章を構成する品詞にのみ注目して解析を行った．しかし，日本語百科大辞典によると女性の文章には男性の文章より，体言止めや「はなびらがはらはらと散りかかって…」のように連用形で止める，連用形止めなどの回数も多いとされる．そこで文章構造の情報を解析し，判定に使用すると精度が上がる可能性がある．また基本的に現役世代は平日の昼頃は働いている．ゆえに，レビューをあまり投稿しないと思われる．そこで曜日ごと，一時間に区切った時刻をリスト化された単語と同様に横に並べる．そしてレビューの投稿時刻に合致するものに「1」しないものに「0」としたならば，非負値行列として扱えるため，判別の精度が上がる可能性がある．

参考文献

1. Lee, D. D. & H. S. Seung. Learning of the parts of objects by non-negative matrix factorization. Nature 401:788–791. (1999).
2. 金田一晴彦, 柴田武, 林大(編). 日本語百科大辞典 [縮刷版]. pp. 560–561. 大修館書店, 東京.