1C-01

非負値行列因子分解とサポートベクタ回帰モデルに基づいた共感された質問記事における特徴抽出手法の提案

輪島　幸治†　古川利博††　佐藤哲司†††

† 筑波大学図書館情報メディア研究科　†† 東京理科大学工学部情報工学科
††† 筑波大学図書館情報メディア系

1. はじめに

近年，情報通信技術の進歩により，CGM(Consumer Generated Media) やオンラインコミュニティなどインターネットにおける個人を主体に情報発信を行うソーシャルメディアが台頭してきている．ソーシャルメディアでは，個人の胸中や購入した商品やサービスの所感などを自由に情報発信できる．発信した情報は，情報収集など多種多様に利用されている一方で，発信者に対する返信は，共感されない場合，返信が行われない場合も少なくない．本研究では，オンラインコミュニティを対象に，返信が数多く行われた質問記事における特徴を抽出し，かつそれらの性質の評価をする．提案手法の概要を図 1 に示す．

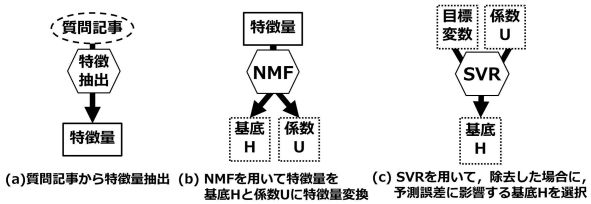

図 1　提案手法の概要

提案手法は，非負値行列因子分解 (NMF) による特徴量変換，サポートベクタ回帰モデル (SVR) を用いた基底選択および特徴量選択で構成されている．質問記事から得られる特徴量を特徴量変換し，変換特徴量である基底 H を得る．そして各基底の重みである係数 U の値を説明変数として，説明変数を変化させ，非線形回帰手法で評価する．そして，返信数に対する予測誤差を算出し，影響が大きい基底を評価する．評価では，文献 [1] で用いた既存研究の日本語テキスト情報の特徴量を用いる．特徴量は，表層情報，トピックを用いた話題，語種，基本語，意味属性，言語表現，文末表現，品詞，固有表現，評価表現の 10 種類，31 個の特徴量である．次元数は 2,071 次元である．

Feature Extraction in Question Article of Empathy based on Non-negative Matrix Factorization and Support Vector Regression

Koji Wajima†,Toshihiro Furukawa††,Tetsuji Satoh†††

†Graduate School of Library, Information and Media Studies, University of Tsukuba

††Dept.Information and Computer Technology., Science University of Tokyo

†††Faculty of Library, Information and Media Science, University of Tsukuba

ところで，ソーシャルメディアでは，インターネットの情報は，企業の評判分析など多様な分野で活用されている[1]．オンラインコミュニティにおける返信では，ポジティブあるいはネガティブに共感した場合に行われ，マクロ的[2] な秩序現象が創発される場合もある．したがって，購買行動などにも影響が大きく，SIPS[3]など共感に基づいた生活者消費行動モデルなども，提案されている．本研究では，感性に基づく曖昧な要素である共感を特徴量変換と非線形回帰手法を用いることで，評価する．

2. 非負値行列因子分解

本研究では，特徴量変換手法に，非負値行列因子分解 (NMF: Non-negative Matrix Factorization) [2] を用いる．NMF は観測行列 Y を基底行列 H と係数行列 U の積に分解するアルゴリズムである．詳細は，文献 [2] にゆずり，ここでは，概略を述べるに留めることにする．NMF を式 (1) に示す．

$$\left(y_{i,j}\right)_{NK} \simeq \Sigma_{m=1}^{M} h_{j,m} u_{m,i} \tag{1}$$

式 (1) は NMF による観測行列 Y の行列分解である．$\left(y_{i,j}\right)_{NK}$ は観測行列 Y を表す．$h_{j,m}$ は基底行列 H の成分，$u_{m,i}$ は係数行列 U の成分である．行列分解される行列 H,U は，一意に決まらず，一般に分解誤差が発生する．NMF では，分解誤差に相当する乖離度規準を定義し，規準に基づく更新式の反復で最適解を求める．本研究では，一般化 Kullback-Leibler ダイバージェンスを用いた．更新式は収束するまで繰り返し適用する．

3. サポートベクタ回帰モデル

本研究では，非線形回帰手法に，文献 [3] などで用いられているサポートベクタ回帰モデル (SVR:Support Vector Regression) を用いる．SVR では，入力空間から，カーネル関数を用いて高次元の特徴空間へ写像する．そして，特徴空間で線形回帰を行う非線形回帰手法である．SVR は，汎化能力が高い回帰モデルであることが知られている．SVR の回帰関数を式 (2) に示す．

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + bias \tag{2}$$

1 Oracle Help Center - Cloud Documentation : https://docs.oracle.com/en/cloud/saas/index.html

2 現象に対する視野が大きいさま．巨視的．

3 SIPS : http://www.dentsu.co.jp/sips/index.html

$K(x_i, x)$ は入力 x_i を特徴空間へ写像するカーネル関数である．本研究では，カーネル関数に RBF カーネルを用いた．α_i,α_i^*,$bias$ などの詳細は文献 [3] などを参照していただきたい．

4. 評価実験

本研究の実装は，NMF およびサポートベクタ回帰モデルのアルゴリズムの実装，予測誤差の算出には scikit-learn[4][5] を用いた．コミュニティはゼネラル・メディアやクラス・メディア [4] など，メディアの特性が大きく影響する．本研究では，2 種類のオンラインコミュニティを用いる．コミュニティ1 には，Apple Inc.[6] が提供している Apple サポートコミュニティに 2008 年 10 月 1 日から 2014 年 1 月 24 日に投稿された質問記事 10,391 件を用いる．コミュニティ2 には，Stack Exchange, Inc.[7] が提供している Stack Exchange Data Dump のうち，2018 年 5 月 5 日までに投稿された ja.stackoverflow.com の質問記事 35,945 件を用いる．閲覧数に対する返信数の相関係数を算出した結果，コミュニティ1 では 0.42，コミュニティ2 では 0.76 という相関係数が得られた．また，コミュニティの利用者の性質から，コミュニティ1 は社会全般の人々を対象としたゼネラル・メディア，コミュニティ2 は特定の集団などを対象とするクラス・メディアであると推定した．相関係数と推定したメディアの性質から，本研究においては，コミュニティ2 における質問記事を共感された質問記事であると定義した．

まず，基底評価に先立ち，2 種類のコミュニティのデータを統合し，特徴量抽出および特徴量変換を行った．本研究における特徴量の次元数は 2,071 次元である．ゆえに，NMF を適用する観測行列 Y は，46,336 行 2,071 列の長方行列である．

次に，基底評価では質問記事に対する返信数を目標変数として，MAE および RMSE の予測誤差に影響が大きい基底を評価した．基底を評価した結果を図 2 に示す．

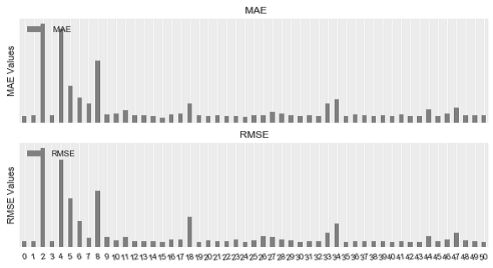

図 2　基底評価の結果

結果，基底 2 を除去した場合に，返信数の予測誤差に影響が大きいことが明らかになった．次に，基底 2 に対する寄与率上位 100 個の特徴量を選択し，選択特徴量で，共感された質問記事の文書分類を行った．Precision，Recall，F-measure の評価指標で評価した結果を表 1 に示す．また，偽陽性率 (FPR) と真陽性率 (TPR) を用いた ROC カーブを図 (3) に示す．

4 scikit-learn : https://scikit-learn.org/stable/
5 scikit-learn : https://github.com/scikit-learn
6 Apple : https://www.apple.com
7 Stack Exchange: Hot Questions : https://stackexchange.com/

表 1 から，4 種類の分類器のうち，MLP において，Precision 基準で 0.91，F-measure 基準においては 0.93 という優れた分類精度が得られた．また，AdaBoost においても，F-measure 基準で 0.91 の十分な分類精度が得られた．加えて，図 (3) から，各分類器における FPR のスレッショルドの最適化で，TPR は向上することも明らかになった．コミュニティにおける返信は利用者が質問記事に共感し，返信という形式で，コミュニティに参加したことに相当する．結果から，利用者の性質および相関係数を基に対象としたコミュニティ2 の質問記事に対して，適切な分類が行えた．したがって，特徴量選択した集合が，感性に基づく曖昧な要素である共感に相当する特徴であると言える．

表 1　返信が多い共感された質問記事の分類結果

Classification methods	Precision	Recall	F-measure
AdaBoost	0.88	0.94	0.91
RandomForest	0.84	0.97	0.90
MLP	0.91	0.95	0.93
KNeighbors	0.85	0.93	0.89

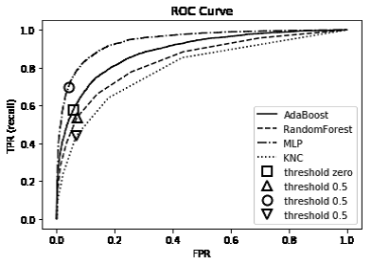

図 3　Receiver Operating Characteristics (ROC)

5. まとめ

本稿では，NMF と SVR を用いて，オンラインコミュニティにおける共感された質問記事を評価した．結果，提案手法で得られた選択特徴量で，文書分類において優れた分類精度が得られた．共感など利用者の感性に基づく要素は，情報化社会において，企業やブランドの生涯顧客価値を高めていく過程で，重要な要素の一つである．今後の課題は，選択特徴量数や複数基底の評価を行い，提案手法の分類精度を向上させたい．

謝辞

本研究に連関した研究に関して，有益な御教授ならびにシミュレーションやデータ整理などをお手伝い頂いた研究室の関係諸氏に感謝の意を示します．また，本研究に連関した研究協力および研究助成頂いた皆様に御礼申し上げます．

文献

[1] 輪島幸治, 木暮啓, 古川利博, 佐藤哲司. 可読性に基づいた日本語テキスト情報の特徴量評価. 第 10 回データ工学と情報マネジメントに関するフォーラム DEIM2018, Mar 2018.
[2] 亀岡弘和. 非負値行列因子分解. 計測と制御, Vol. 51, No. 9, pp. 835–844, sep 2012.
[3] 小林正幸, 小西康夫, 藤田貞雄, 石垣博行. サポートベクタ回帰モデルを用いた超音波モータの位置決め制御. 精密工学会誌論文集, Vol. 72, No. 5, pp. 596–601, 2006.
[4] 亀井昭宏. 電通広告事典. 電通, 2008.