社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
DE2020-5 (2020-06)

行動ログの機械学習を用いた 他ユーザのSNS投稿に対するお気に入り登録予測

荒澤　孔明† 　服部　　峻†† 　工藤　康生††

†,†† 室蘭工業大学大学院 工学研究科　〒050–8585 北海道室蘭市水元町 27–1
E-mail: †18096001@mmm.muroran-it.ac.jp, ††{hattori,kudo}@csse.muroran-it.ac.jp

あらまし　近年，SNS 上での口コミを利用して商品の宣伝や企業のブランディングなどを行うバイラルマーケティングといった広告戦略に注目が集まってきている．例えば，ある著名人が SNS 上である商品を話題にした際，その投稿を提示しながら，その商品を広告する事で，推薦の説得性を高める事ができる．こうした広告をより高度に行うためには，その SNS 投稿は誰を感化させるものなのかを精確に予測する技術も重要となる．すなわち，SNS 利用者が関心を抱くであろう投稿を特定する方式が求められる．そこで本稿では，SNS でのユーザの行動ログを用いて，あるユーザがある投稿に対してお気に入り登録を行う（関心を持つ）か否かをクラス分類する手法についての諸検討を行う．
キーワード　ランダムフォレスト，パーソナライゼーション，インフルエンサ，情報推薦システム

Bookmarking Forecast for Others' SNS Posts by Machine Learning of Activity Logs

Komei ARASAWA†, Shun HATTORI††, and Yasuo KUDO††

†,†† Graduate School of Engineering, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: †18096001@mmm.muroran-it.ac.jp, ††{hattori,kudo}@csse.muroran-it.ac.jp

Abstract　An advertising strategy called as "Viral Marketing" has public attention. It promotes an item and its company by using posts (reviews) in SNS. For example, when a celebrity wrote a SNS post that has a topic on an item, recommending the item for other users by showing them the post could increase persuasiveness of the recommendation. It is important for the system to exactly forecast person(s) who would be interested in the SNS posts. Therefore, it needs for the system to specify the posts that could induce the interest of a user. This paper proposes a method that analyzes the activity logs of a user in SNS and forecasts whether the user would bookmark a SNS post or not.
Key words　Random Forest, Personalization, Influencer, Recommender System

1. まえがき

近年，SNS 上での口コミを利用して商品の宣伝や企業のブランディングなどを行うバイラルマーケティング [1,2] といった広告戦略に注目が集まってきている．例えば，ある著名人が SNS 上である商品を話題にした際，その投稿を提示しながら，その商品を広告する事で，推薦の説得性を高める事ができる．こうした広告を高度に行うためには，SNS の閲覧者（利用者）が関心を持つであろう投稿を精確に発見する事も重要な基礎技術となる．古典的には，不特定多数が関心を持つであろう SNS 投稿の抽出が目標とされてきた（図 1）．すなわち，抽出すべき投稿には「社会的に影響力のある人物の発言である」という十分条件が求められてきた．そして今日まで，こうした影響力のある人物を推定する研究は盛んに行われてきた [3–5] ．

他方，本稿では，利用者ごとに関心を持つであろう SNS 投稿の抽出に目標を定めている点が従来研究と明確に異なる（図 1）．これによって，抽出される SNS 投稿として，社会的なインフルエンサ（著名人など）の発言だけでなく，個人的なインフルエンサ（友人など）の発言なども加わる事となり，それらの投稿を利用したより多様な情報推薦が期待できる．

以上より，本稿では，ある SNS 投稿に対する関心の有無を「その投稿に対するお気に入り登録の有無」と置き換える事で，あるユーザがある SNS 投稿に対して，お気に入り登録を行うか否かを予測する手法の提案および評価を行う．具体的には，

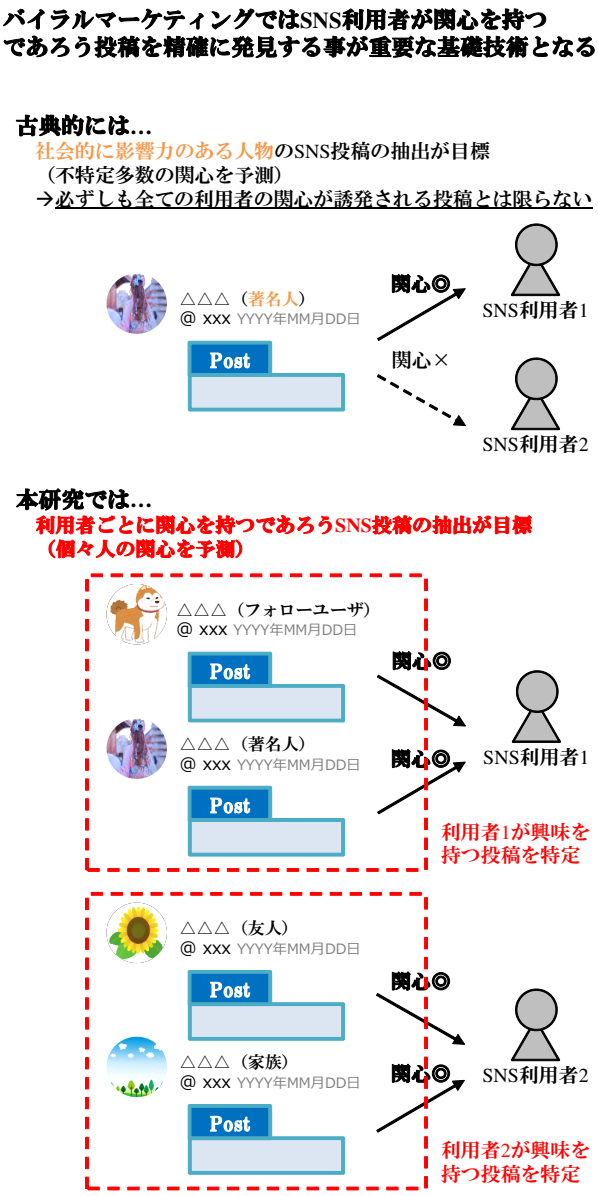

図 1　本稿におけるタスクの従来研究との違い

ユーザがお気に入り登録を行う影響要因（独立変数）を検討し，機械学習の手法を用いて，ユーザごとにお気に入り登録の予測モデルを生成し，その予測性能を評価する．さらに，ユーザのプロファイルや利用状況の違いによって，お気に入り登録を行う要因がどのように異なるのかを分析する．なお，本稿では SNS として Twitter を用いた議論を行う．

2. 独立変数に関する諸検討

本章では，あるユーザがある投稿に対してお気に入り登録を行うか否かを予測するために有用な独立変数についての検討を行う．我々はその影響要因を表 1 に示す 3 種類に大別した．

2.1 投稿内容・形式への魅力

あるユーザがある投稿に対してお気に入り登録を行うための十分条件の 1 つとして，コンテンツそのものが優れている事が挙げられる．我々はこの要因について，表現メディアの違い

表 1　お気に入り登録予測における影響要因

投稿内容・形式への魅力	系統
表現メディアの違い	A-1
話題への関心度	A-2
投稿者の影響力	**系統**
投稿者への関心度	B-1
投稿者へのこれまでの反応	B-2
投稿者の社会的な注目度	B-3
投稿の社会的保証	**系統**
その投稿に対する周囲の反応	C

（表 2 の系統 A-1）と話題への関心度（表 2 の系統 A-2）の側面から検討し，表 2 に示した 9 個の独立変数を設定した．このうち系統 A-2 に該当する「ある被験者 s の文書 D_s の解析に基づくある投稿 p の話題への関心度 $Int(D_s \to p)$」については，以下の通り算出した．まず，被験者 s の解析文書 D_s（以降，蓄積文書群）として，被験者 s の投稿群や被験者 s のお気に入り投稿群，または被験者 s のフォロー相手の投稿群の 3 種類を定めた．そして，その投稿 p に含まれる単語集合 W_p の中で，被験者 s の蓄積文書群 D_s 内に最も出現した単語 w を定め，その出現頻度 $freq_{D_s}(w)$ を，被験者 s からその投稿 p の話題への関心度 $Int(D_s \to p)$ とする．

$$Int(D_s \to p) = \max\{freq_{D_s}(w) | w \in W_p\}$$

2.2 投稿者の影響力

あるユーザがある投稿に対してお気に入り登録を行うための十分条件の 1 つとして，その投稿者がユーザに影響を与えている人物である事が挙げられる．我々はこの要因について，投稿者への関心度（表 2 の系統 B-1），投稿者へのこれまでの反応（表 2 の系統 B-2），投稿者の社会的な注目度（表 2 の系統 B-3）の側面から検討し，表 2 に示した 9 個の独立変数を設定した．このうち系統 B-1 に該当する「ある被験者 s の文書 D_s の解析に基づくその投稿者 u への関心度 $Int(D_s \to u)$」については，以下の通り算出した．まず，被験者 s の解析文書 D_s（以降，蓄積文書群）として，被験者 s の投稿群や被験者 s のお気に入り投稿群の 2 種類を定めた．そして，その投稿者 u のプロフィールコメントに含まれる（その投稿者を表す）単語 w が被験者 s の蓄積文書群 D_s の中でどの程度出現していたかを次式より算出する．ただし，式中の W_u は投稿者 u のプロフィールコメントに含まれる単語集合を示しており，$freq_{D_s}(w)$ は被験者 s の蓄積文書群 D_s 内での単語 w の出現頻度を示している．また，$tfidf_u(w)$ は投稿者 u のプロフィールコメントにおける単語 w の TFIDF 特徴量を示している．

$$Int(D_s \to u) = \sum_{w \in W_u} tfidf_u(w) \times freq_{D_s}(w)$$

なお，TFIDF 分析における TF は，あるユーザのプロフィールコメントにおける単語 w の出現頻度，また DF は，全ユーザ（被験者 100 名，被験者のフォローユーザ，被験者が過去にお気に入り登録を行った相手）のプロフィールコメントにおける単語 w を含むユーザ数として算出している．

表 2　お気に入り登録予測における独立変数の定義（*はダミー変数）

No	系統	変数表記	値域	説明
1	A-1	Txt*	{0,1}	テキストのみで構成される投稿である
2	A-1	TxtRep*	{0,1}	テキストのみで構成される被験者への返信である
3	A-1	Img*	{0,1}	画像のみで投稿される投稿である
4	A-1	ImgRep*	{0,1}	画像のみで構成される被験者への返信である
5	A-1	TxtImg*	{0,1}	テキストと画像で構成される投稿である
6	A-1	TxtImgRep*	{0,1}	テキストと画像で構成される被験者への返信である
7	A-2	IntFav	≥ 0	被験者のお気に入り投稿群の解析に基づくその投稿の話題への関心度
8	A-2	IntTwt	≥ 0	被験者の投稿群の解析に基づくその投稿の話題への関心度
9	A-2	IntFol	≥ 0	被験者のフォロー相手の投稿群の解析に基づくその投稿の話題への関心度
10	B-1	IntPerFav	≥ 0	被験者のお気に入り投稿群の解析に基づくその投稿者への関心度
11	B-1	IntPerTwt	≥ 0	被験者の投稿群の解析に基づくその投稿者への関心度
12	B-2	FavRate	[0,1]	被験者が行ったお気に入り登録の中で相手がその投稿者である割合
13	B-2	RepRate	[0,1]	被験者が行った返信の中で相手がその投稿者である割合
14	B-2	RtwRate	[0,1]	被験者が行ったリツイートの中で相手がその投稿者である割合
15	B-2	RxnRate	[0,1]	その投稿者への総合的な反応　$\max\{\text{FavRate}, \text{RepRate}, \text{RtwRate}\}$
16	B-3	Flw	≥ 0	その投稿の投稿者のフォロワー数
17	B-3	UsFav	≥ 0	その投稿者の平均的な被お気に入り登録数
18	B-3	UsRtw	≥ 0	その投稿者の平均的な被リツイート数
19	C	TarFav	≥ 0	その投稿の被お気に入り登録数
20	C	ChgFav	≥ -1	被お気に入り登録数の変化率　$(\text{TarFav} - \text{UsFav})/\text{UsFav}$
21	C	TarRtw	≥ 0	その投稿の被リツイート数
22	C	ChgRtw	≥ -1	被リツイート数の変化率　$(\text{TarRtw} - \text{UsRtw})/\text{UsRtw}$

2.3　投稿の社会的保証

あるユーザがある投稿に対してお気に入り登録を行うための十分条件の 1 つとして，そのユーザ以外の多くのユーザもその投稿に対して興味を示している事が挙げられる．我々はこの要因について，不特定多数からその投稿への反応に着目し，表 2 の系統 C に示した 4 個の独立変数を設定した．

3.　実 験 方 法

3.1　実 験 手 順

実験手順は以下に示す通りである．

手順 1 被験者が閲覧した SNS 投稿を一定数取得する．

手順 2 被験者がお気に入り登録を行った投稿リストを参照し，各投稿に対して，お気に入り登録を行ったか否か（目的変数）の真偽値を付与する．

手順 3 各投稿に対して独立変数をスコアリングし，ランダムフォレストを用いて目的変数の 2 値分類を行う．

手順 4 3.3 で示す基準を用いてその分類性能を評価する．

本実験では R の caret パッケージを用いて，ランダムフォレストを実装しており，1 つの決定木に学習させる特徴量の個数は，グリッドサーチに基づき最適化した．また本稿では，被験者ごとの予測モデルの評価に，10 分割交差検証を採用した．

3.2　データセット

本実験では，100 名の Twitter 利用者に被験者に協力してもらっており，被験者の平均的な利用状況および被験者周辺のユーザ数については，それぞれ表 3 と表 4 の通りである．Twitter データについては Twitter4j を用いて取得しており，実験で用いたデータ数は表 5 の通りである．ただし，被験者が閲覧した投稿群（タイムライン）は API で取得する事ができない制約があったため，フォローユーザ（表 4 *1）と過去にお気に入り登録を行った事がある相手（表 4 *2）の投稿群から疑似的に被験者が閲覧したであろう投稿群を再現した．最終的な学習データ数とテストデータ数は，表 6 に示されている．

3.3　評 価 尺 度

本実験では，表 7 に示す混合行列を算出したのち，正確率（Accuracy），再現率（Recall），適合率（Precision），特異率（Specificity）を用いて，その予測性能を評価する．

表 3　被験者の Twitter 利用状況

	平均	標準偏差
投稿数/日	7.548	10.013
お気に入り登録数/日	5.178	3.766
返信数/全ての投稿数	0.289	0.245

表 4　被験者周辺のユーザ

	平均人数	標準偏差
フォローユーザ *1	58.700	44.269
過去にお気に入り登録を行った事がある相手 *2	145.310	72.816

表 5　本実験で取得したデータ数

	平均	標準偏差
被験者の投稿数	358.040	88.672
被験者のお気に入り投稿数	344.390	78.925
表 4 *1 の投稿数	190.858	33.525
表 4 *2 の投稿数	193.074	28.408

表 6　学習データとテストデータ

	学習データ		テストデータ	
	FAV	N-FAV	FAV	N-FAV
平均	310.951	311.639	35.439	115.306
標準偏差	70.713	71.148	7.867	26.652

表 7　混 合 行 列

予測＼実際	FAV	N-FAV
FAV	**TP**	**FP**
N-FAV	**FN**	**TN**

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

また，再現率と適合率の調和平均として F 値（F-measure），テストデータにおける FAV と N-FAV の割合を考慮した Balanced-Accuracy を以下の通り評価する．

$$\text{F-measure} = 2 \times \text{Recall} \times \text{Precision}/(\text{Recall} + \text{Precision})$$

$$\text{Balanced-Accuracy} = (\text{Recall} + \text{Specificity})/2$$

4. 実 験 結 果

4.1　予測性能に関する評価

本稿での主要な結果の 1 つとして，実用レベルに近い性能で，個々人のお気に入り登録の予測モデルが生成可能な事を実証した．表 8 では，各種評価尺度を列として，100 名の被験者に対するお気に入り登録予測の結果が要約されている．まず，特異率（Specificity），すなわち真陰性率に着目すると，その他の評価尺度と比較して高く評価されているが，これについては本実験の設計上妥当な結果と判断している．本稿では，約 3 割がお気に入り登録投稿（陽性）であるテストデータに対して予測実験を行っており，すなわち陽性よりも陰性を検出する方が比較的容易なタスクに設計されているため，真陰性率が高くなる傾向は直観と一致する．総合的な評価として，Balanced-Accuracy（特異率と再現率の平均）が F 値（再現率と適合率の調和平均）よりも高い傾向が見られた事についても，同様の解釈である．

他方，本タスクにおいてより重視しなければならない評価尺度は，陽性の検出が関与してくる再現率や適合率であるとも言える．表 8 の再現率（Recall）と適合率（Precision）に着目すると，ともに 7 割程度であり，残りの 3 割は予測漏れや予測誤りであった事が分かる．予測漏れに関しては，お気に入り登録を行うための十分条件が満足してない可能性があり，予測誤りに関しては，独立変数のスコアリング方式が適切でない可能性がある．そのため，今後 2 章での検討事項を再分析する必要がある．ただし，SNS においては，ユーザが気まぐれにお気に入り登録を行うケースがある事も否定できない．したがって，お気に入り登録投稿を漏れなくかつ誤りなく予測するタスクは容易ではないという事を考慮すると，我々はこの予測性能について，極端に悪い結果とは言えず，システムとしてある程度実用可能であると評価している．

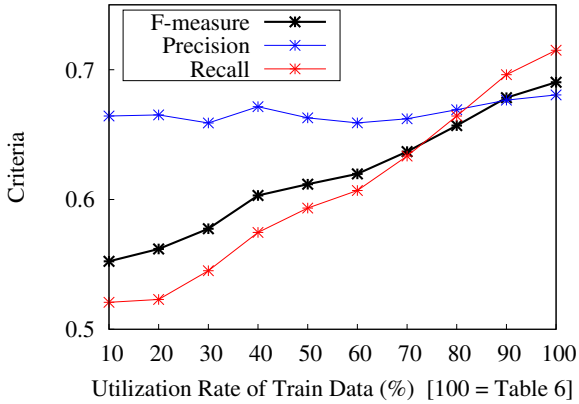

図 2　学習データの利用率と予測性能との関係

4.2　ユーザの SNS 利用状況が予測性能に与える影響

本稿での主要な結果の 1 つとして，フォローユーザ数が多い被験者について，お気に入り登録の予測性能が悪化する課題を明らかにした．表 9 では，ユーザの SNS 利用状況と予測性能との相関関係を評価した．FolU（3 列目）に着目すると，被験者のフォローユーザ数と再現率（Recall）との間には比較的強い負の相関が認められた．この影響で，F 値や Balanced-Accuracy についても悪化させてしまう事が読み取れる．フォローユーザが多い被験者はそれだけ閲覧する投稿も多くなると言える．そのため，投稿の見逃しや，不意なお気に入り登録の確率が高まり，結果として良質な予測モデルが生成できなかったと推測する．

また，被験者が過去にお気に入り登録を行った事がある相手の人数（FavU）と再現率との間にも負の相関関係が認められた．しかしそのトレードオフとして，適合率や特異率を改善する傾向も観測された．よって，過去にお気に入り登録を行った事がある相手が比較的多くても，F 値や Balanced-Accuracy などの予測性能に大きな影響を与えるとまでは断定できなかった．

最後に，単位日当たりの被験者のお気に入り登録数（Fav/D）や単位日当たりの被験者自身の投稿数（TwT/D）などといった，被験者の SNS 利用頻度に関するファクタについては，各種評価尺度との相関は確認されなかった．

4.3　学習データ数が予測性能に与える影響

本稿での主要な結果の 1 つとして，予測モデルの性能は学習データ数に依存し，本稿と同程度の予測性能を得るためには，お気に入り登録を行った投稿が約 220 件必要である事を明らかにした．図 2 は，学習データの利用率（横軸）による，予測性能（縦軸）を示した．ただし，利用率は，表 6 で示した学習データを利用率 100% の基準としている．この図からは，学習データの規模は，適合率には影響を与えていない事が読み取れる．他方，再現率には影響を与えており，その結果 F 値も変化させている．この F 値の変化について，各利用率における平均 F 値

表 8　100 名の被験者のお気に入り登録予測モデルに関する平均性能

	TP	FN	FP	TN	Acc	Spe	Pre	Rec	F	B-Acc
平均	24.887	9.552	11.458	103.848	0.856	0.899	0.681	0.715	0.690	0.807
標準偏差	8.588	5.928	5.294	24.885	0.053	0.041	0.110	0.169	0.128	0.089
最大値	40.000	31.000	37.000	134.000	0.994	1.000	1.000	1.000	0.987	0.996
最小値	2.000	0.000	0.000	18.000	0.681	0.738	0.250	0.143	0.214	0.509
中央値	26.000	9.000	11.000	113.000	0.857	0.903	0.679	0.744	0.701	0.814

表 9　SNS 利用状況と予測性能との相関

	Fav/D	Twt/D	FolU	FlwU	FavU
Acc	0.06	0.02	−0.20	0.08	0.08
Spe	0.11	0.04	0.13	0.08	0.41
Pre	0.09	0.03	−0.08	0.10	0.21
Rec	−0.01	−0.01	−0.39	0.04	−0.24
F	0.04	0.01	−0.29	0.07	−0.07
B-Acc	0.02	0.00	−0.33	0.06	−0.12

Fav/D：単位日当たりの被験者のお気に入り登録数
Twt/D：単位日当たりの被験者自身の投稿数
FolU：被験者のフォローユーザ数
FlwU：被験者のフォロワー数
FavU：被験者が過去にお気に入り登録を行った事がある相手の人数

表 10　各学習データ利用率の平均 F 値に対する分散分析

	自由度	平方和	平均平方	統計量 F	p 値
グループ間	9	2.011	0.223	15.02	≈0
グループ内	990	14.279	0.015		

には差が無いという帰無仮説を立て，一元配置分散分析を行った（表 10）．この結果から，有意水準 0.1% で検定した時，統計量 F に基づく p 値が 0.1 より低いため，帰無仮説は棄却され，F 値は有意な差で学習データ利用率（学習データ数）に依存している事が分かった．加えて，チューキー・クレーマー検定に基づく多重比較を行ったところ，学習データ利用率 100% の平均 F 値と学習データ利用率 60% の平均 F 値との間には有意差が認められた（$p = 0.002$）．したがって，おおよそ表 6 の学習データ数の 70% ほど（正例 220 件，負例 220 件）の利用で，本実験と同程度である 7 割近い F 値のお気に入り登録予測モデルが生成できる事が見通された．

しかしながら，図 2 からは，F 値のピークや収束地点までは読み取る事ができなかった．すなわち，本稿で用意したデータ数では，予測性能の理論限界値を議論するには不十分であり，今後は実験規模の拡張などが課題とされている．

4.4　お気に入り登録予測における各独立変数の重要度

本稿での主要な結果の 1 つとして，お気に入り登録予測をモデリングする際，ほとんどの被験者で投稿の社会的保証（表 2 の C）が重要視されており，その中でさらに，投稿内容・形式への魅力（表 2 の A 系統）が重要視される被験者は 22%，また投稿者の影響力（表 2 の B 系統）が重要視される被験者は 13% 存在する事を明らかにした．以降，各独立変数の平均重要度，また重要度の違いに基づく被験者のタイプ分類を考察する．

表 11　各独立変数のランダムフォレストに基づく重要度

	平均	標準偏差	最大値	最小値	中央値
Txt	17.71	24.30	100.00	0.03	7.26
TxtRep	20.72	29.44	100.00	0.00	6.45
Img	0.55	0.77	8.62	0.00	0.34
ImgRep	0.28	1.16	12.21	0.00	0.00
TxtImg	6.42	8.87	57.14	0.05	2.85
TxtImgRep	0.54	2.29	29.02	0.00	0.00
IntFav	30.79	25.72	100.00	0.09	23.45
IntTwt	8.22	9.92	100.00	0.00	5.22
IntFol	15.09	10.59	71.24	0.06	12.83
IntPerFav	6.11	5.52	54.15	0.35	4.59
IntPerTwt	5.21	4.57	30.70	0.00	3.91
FavRate	35.21	27.20	100.00	0.79	26.32
RepRate	7.32	11.84	100.00	0.00	3.70
RtwRate	3.92	10.08	100.00	0.00	1.20
RxnRate	21.38	18.15	100.00	0.93	16.30
Flw	9.91	7.56	74.27	0.93	7.99
UsFav	8.87	7.37	63.31	0.50	6.49
UsRtw	8.27	7.29	65.82	0.31	5.95
TarFav	38.86	28.66	100.00	0.43	32.54
ChgFav	88.35	22.99	100.00	2.34	100.00
TarRtw	13.72	12.61	65.45	0.15	9.17
ChgRtw	26.48	17.92	93.07	0.49	21.48

4.4.1　平均的な重要度に関する評価

表 11 には各独立変数のランダムフォレストにおける平均的な重要度を示した．まずお気に入り登録を予測するにあたり，平均的に最も重要視された独立変数は，被お気に入り登録数の変化率（ChgFav）であった．標準偏差を考慮しても，大部分の被験者において，ChgFav の重要度が高くモデル化されている事が分かる．その他，被験者のお気に入り投稿群の解析に基づく話題への関心度（IntFav），被験者が行ったお気に入り登録の中で相手がその投稿者である割合（FavRate）などの独立変数も比較的高く評価され，それらの変数を最も重要視した（重要度 100）予測モデルが生成された被験者も少なからず存在した．

4.4.2　各独立変数の重要度の違いによる被験者の分類

ここでは，お気に入り登録の予測モデルにおける各独立変数の重要度は，被験者によってどのようなタイプが存在するのかについて分析する．まず，その概要を把握するために，各被験者の予測モデルに含まれる 22 個の独立変数の重要度を特徴ベクトル（22 次元）として，被験者 100 名に対し，ウォード法による階層クラスタ分析を行った．なお，距離の測度にはユークリッド距離を採用した．

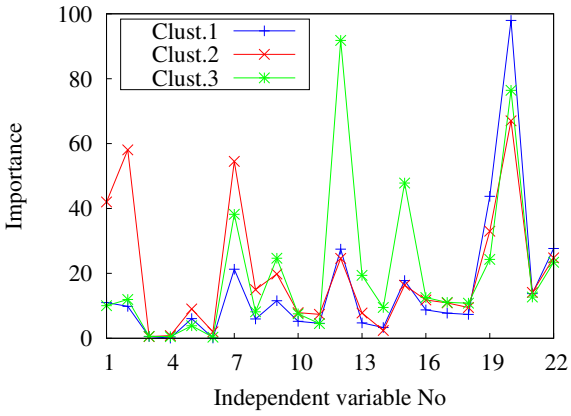

図 3　3 タイプの被験者クラスタにおける各独立変数の平均重要度

結果として，被験者の予測モデルは 3 つ（シルエット分析における最適クラスタ数）に大別できる事が明らかになった．図 3 には，横軸に 22 個の独立変数，縦軸にあるクラスタに該当する被験者の予測モデルにおける平均重要度を示した．以降，各クラスタの特徴を分析していく．

第 1 クラスタでは，ChgFav の重要度のみが突出している．このことから，このクラスタに該当する被験者は，一般的な人物が関心を持つ話題であるならば，その話題への元々の興味や投稿者に影響される事なく，その投稿に対する関心が誘発されるユーザであると言える．すなわち，社会の動向で注意がひきつけられる「社会触発型」のユーザと解釈できる．なお，このクラスタに該当する被験者は全体の 65% であった．

第 2 クラスタでは，ChgFav に加え，テキストのみで構成される投稿である（Txt），テキストのみで構成される被験者への返信である（TxtRep），被験者のお気に入り投稿群の解析に基づくその投稿の話題への関心度（IntFav）といった 3 変数の重要度も高く評価されている．このことから，このクラスタに該当する被験者は，ベースとして「社会触発型」の特徴を持ち，さらにコンテンツに対する魅力や情報のメッセージ性などが強ければ，投稿者に影響される事なく，その投稿に対する関心が誘発されるユーザであると言える．すなわち，社会の動向に加え，コンテンツ自体の魅力などの要因で注意がひきつけられる「コンテンツ触発型」のユーザと解釈できる．なお，このクラスタに該当する被験者は全体の 22% であった．

第 3 クラスタでは，ChgFav に加え，被験者が行ったお気に入り登録の中で相手がその投稿者である割合（FavRate），その投稿者への総合的な反応（RxnRate）といった 2 変数の重要度も高く評価されている．このことから，このクラスタに該当する被験者は，ベースとして「社会触発型」の特徴を持ち，さらに投稿者との良質な関係が満たされていれば，その話題への元々の興味に影響される事なく，その投稿に対する関心が誘発されるユーザであると言える．すなわち，社会の動向に加え，投稿者との関係性などの要因で注意がひきつけられる「投稿者触発型」のユーザと解釈できる．なお，このクラスタに該当する被験者は全体の 13% であった．

5. ま と め

近年，SNS 上でのポジティブな口コミを利用して商品の宣伝や企業のブランディングなどを行うバイラルマーケティングといった広告戦略に注目が集まってきている．こうした広告をより高度に行うためには，SNS の閲覧者が関心を持つであろう投稿（口コミ）を精確に発見する事も重要な基礎技術となる．我々は，従来研究のように，不特定多数に影響を与えるであろう投稿を抽出するだけでなく，利用者ごとに興味を持つ投稿も抽出できる技術の確立を目標としてきた（表 1）．その第 1 歩として本稿では，ユーザの SNS での行動ログを解析し，あるユーザがある投稿に対してお気に入り登録を行う（関心が誘発される）か否かを 2 値分類する手法を提案した．

まず，ユーザがお気に入り登録を行うための十分条件（影響要因）を，投稿内容・形式への魅力（表 2 の A 系統），投稿者の影響力（表 2 の B 系統），投稿の社会的保証（表 2 の C）といった 3 つの側面から検討した．次に，そこで得られた 22 種類の独立変数をランダムフォレストで学習し，ユーザ個々人のお気に入り登録予測モデルを生成し，その予測性能を評価した．そこで得られた本稿での主要な結果は以下の通りである．

- 実用レベルに近い性能で，個々人のお気に入り登録の予測モデルが生成可能な事を実証した．
- フォローユーザ数が多い被験者について，お気に入り登録の予測性能が悪化する課題を明らかにした．
- 予測モデルの性能は学習データ数に依存し，本稿と同程度の予測性能を得るためには，お気に入り登録を行った投稿が約 220 件必要である事を明らかにした．
- お気に入り登録予測をモデリングする際，ほとんど被験者で投稿の社会的保証が重要視されており，その中でさらに，投稿内容・形式への魅力が重要視される被験者は 22%，また投稿者の影響力が重要視される被験者は 13% 存在する事を明らかにした．

文　献

[1] P. Domingos and M. Richardson, “Mining the Network Value of Customers,” Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.57–66 (2001).

[2] M. Richardson and P. Domingos, “Mining Knowledge-Sharing Sites for Viral Marketing,” Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.61–70 (2002).

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy,” Proceedings of International AAAI Conference on Weblogs and Social Media, pp.10–18 (2010).

[4] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” Proceedings of the 19th International Conference on World Wide Web, pp.591–600 (2010).

[5] B. Wang, C. Wang, J. Bu, C. Chen, W. V. Zhang, D. Cai, and X. He, “Whom to Mention: Expand the Diffusion of Tweets by @ Recommendation on Micro-blogging Systems,” Proceedings of the 22nd International Conference on World Wide Web, pp.1331–1340 (2013).