

SNSの行動ログに基づくインフルエンサ推定の個人化

荒澤 孔明^{1,a)} 服部 峻¹

概要: ユーザに影響を与える人物(インフルエンサ)の推定は、様々な情報推薦システムにとって重要な基礎技術となる。本稿では、他のユーザの投稿に対する「返信」や「お気に入り登録」といったSNSの行動ログを分析し、インフルエンサをユーザ個々に推定する手法を議論する。またインフルエンサの選定条件を多様化し、適合率を重視した手法や再現率を重視した手法など、複数の推定方式について比較実験を行う。

キーワード: インフルエンサ推定, 行動ログ分析, SNS解析, パーソナライズ, 情報推薦

1. はじめに

影響を与える人物(インフルエンサ)の推定は、消費行動の促進など、様々な推薦システムの重要な基礎技術となる。特に身近な人物の中からインフルエンサを推定する場合、ユーザ個々の分析が伴う。SNSを解析し、ユーザとの親密度や類似度が高い人物を推定する研究[1,2]もあるが、仲が良い事や趣味が合う事はその人物がインフルエンサである事の必要条件とは言い難い。また「刺激的な情報」の発信者を推定する研究[3]もあるが、この手法は、影響を受けている人物を、ユーザに予め選定してもらい、その関連人物を推定するというものであるため、既に影響を受けている人物を推定するタスクとは食い違う。そこで本稿ではユーザ個々のSNSの行動ログを解析し、インフルエンサとしての必要条件を多様に組み合わせる事で、本人が実際に影響を受けている人物を推定する手法を議論する。

Step 1 返信相手とお気に入り登録相手にスコアを付与 (2.3節)

返信相手 u_i	返信スコア $S_{rep}(t \rightarrow i)$	お気に入り相手 u_i	お気に入りスコア $S_{fav}(t \rightarrow i)$
Aさん	0.214	Dさん	0.120
Bさん	0.153	Eさん	0.063
Cさん	0.145	Aさん	0.048

Step 2 インフルエンサ候補を選定しその影響力を算出 (2.4節)

インフルエンサ候補 u_i	影響力 $V(i \rightarrow t)$
Aさん	6種類の算出法を比較検討
Bさん	
Eさん	

Step 3 影響力が閾値以上の候補をインフルエンサと推定する

インフルエンサ	影響力 $V(i \rightarrow t)$
Aさん	閾値以上
Bさん	
Eさん	

図1 ターゲットユーザ u_t に対するインフルエンサの推定フロー

2. インフルエンサ推定

2.1 インフルエンサの必要条件

まず我々は、ある人物 u_i がターゲットユーザ u_t のインフルエンサであるための必要条件として、「ターゲットユーザ u_t はその人物 u_i の言動に対して、度々反応を示している事」という命題を定義した。またその命題を、ターゲットユーザ u_t はその人物 u_i のSNS投稿に対して「返信」や「お気に入り登録」を日常的に行っている事と具現化した。

2.2 インフルエンサ推定の概要

本稿ではこのアイデアに基づき、3つのステップによってターゲットユーザのインフルエンサを推定する(図1)。

第1ステップでは、ターゲットユーザ u_t からこれまで少なくとも1回以上、返信を受けた事のある人物 u_i ($i = 1, 2, \dots$) に対して、ターゲットユーザ u_t からの返信スコア $S_{rep}(t \rightarrow i)$ を算出する。また同様に、ターゲットユーザ u_t からこれまで少なくとも1回以上、お気に入り登録を受けた事のある人物 u_i に対して、ターゲットユーザ u_t からのお気に入りスコア $S_{fav}(t \rightarrow i)$ を算出する。

第2ステップでは、ターゲットユーザ u_t の返信相手及びお気に入り相手から、インフルエンサ候補の人物 u_i を幾人か選定し、候補の人物 u_i がターゲットユーザ u_t に与える影響力 $V(i \rightarrow t)$ を、2つのスコアを用いて算出する。

第3ステップでは、影響力 $V(i \rightarrow t)$ に従って候補をランキングし、その影響力 $V(i \rightarrow t)$ が閾値 $v_t(\epsilon)$ を超える候補をターゲットユーザ u_t のインフルエンサと推定する。但し、閾値 $v_t(\epsilon)$ はターゲットユーザ u_t に依存し、パラメータ $\epsilon = [0, 100]$ によって制御される。これは、候補の人物 u_i がターゲットユーザ u_t に与える影響力 $V(i \rightarrow t)$ を算出しランキングした後、その最小値を0%、最大値を100%として、何%の地点を閾値とするかを表している。

以降2.3節では2つのスコアを、2.4節ではインフルエンサ候補を選定し、その影響力を算出する手法を提案する。

¹ 室蘭工業大学大学院 工学研究科

^{a)} 18096001@mmm.muroran-it.ac.jp

2.3 返信スコア及びお気に入りスコアの算出

本節では、ターゲットユーザ u_t から人物 u_i への返信スコア $S_{rep}(t \rightarrow i)$ 、及びお気に入りスコア $S_{fav}(t \rightarrow i)$ について、それぞれの算出方式を論述する。

2.3.1 返信スコアの算出方式

返信スコアは、ターゲットユーザからこれまで少なくとも1回以上、返信を受けた事のある人物に対して付与される。この時、ある人物 u_i の返信スコア $S_{rep}(t \rightarrow i)$ には、ターゲットユーザ u_t がこれまで行った返信のうち、その相手が人物 u_i であった割合を付与する。以下に示す式中の $C_{rep}(t \rightarrow i)$ は、人物 u_i の投稿に対して、ターゲットユーザ u_t がこれまで行った返信の数であり、 U_{rep}^t は、ターゲットユーザ u_t がこれまで少なくとも1回以上、返信を行った事のある相手 u_k の集合を示している。

$$S_{rep}(t \rightarrow i) = \frac{C_{rep}(t \rightarrow i)}{\sum_{u_k \in U_{rep}^t} C_{rep}(t \rightarrow k)}$$

$$U_{rep}^t = \{u_k \mid \forall k, C_{rep}(t \rightarrow k) \geq 1\}$$

2.3.2 お気に入りスコアの算出方式

お気に入りスコアは、ターゲットユーザからこれまで少なくとも1回以上、お気に入り登録を受けた事のある人物に対して付与される。この時、ある人物 u_i のお気に入りスコア $S_{fav}(t \rightarrow i)$ には、ターゲットユーザ u_t がこれまで行ったお気に入り登録のうち、その相手が人物 u_i であった割合を付与する。以下に示す式中の $C_{fav}(t \rightarrow i)$ は、人物 u_i の投稿に対して、ターゲットユーザ u_t がこれまで行ったお気に入り登録の数であり、 U_{fav}^t は、ターゲットユーザ u_t がこれまで少なくとも1回以上、お気に入り登録を行った事のある相手 u_k の集合を示している。

$$S_{fav}(t \rightarrow i) = \frac{C_{fav}(t \rightarrow i)}{\sum_{u_k \in U_{fav}^t} C_{fav}(t \rightarrow k)}$$

$$U_{fav}^t = \{u_k \mid \forall k, C_{fav}(t \rightarrow k) \geq 1\}$$

2.4 インフルエンサ候補の影響力のランキング手法

続いて本節では、ターゲットユーザのインフルエンサ候補を選定し、ターゲットユーザに与える影響力を算出する事で、候補をランキングする手法を6種類提案する。

2.4.1 返信相手を候補とする手法

この手法は、ターゲットユーザがこれまで少なくとも1回以上、返信を行った事のある相手のみ、インフルエンサ候補として選定するというものであり、手法 **REP** と定義する。この手法では、該当する人物 u_i の影響力 $V(i \rightarrow t)$ として、ターゲットユーザ u_t からの返信スコア $S_{rep}(t \rightarrow i)$ をそのまま付与しランキングする。

$$V(i \rightarrow t) = \begin{cases} S_{rep}(t \rightarrow i) & (u_i \in U_{rep}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

2.4.2 お気に入り相手を候補とする手法

この手法は、ターゲットユーザがこれまで少なくとも1回以上、お気に入り登録を行った事のある相手のみ、インフルエンサ候補として選定するというものであり、手法 **FAV** と定義する。この手法では、該当する人物 u_i の影響力 $V(i \rightarrow t)$ として、ターゲットユーザ u_t からのお気に入りスコア $S_{fav}(t \rightarrow i)$ をそのまま付与しランキングする。

$$V(i \rightarrow t) = \begin{cases} S_{fav}(t \rightarrow i) & (u_i \in U_{fav}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

2.4.3 返信相手かつお気に入り相手を候補とする手法

この手法は、ターゲットユーザがこれまで少なくとも1回以上、返信とお気に入り登録の双方を行った事のある相手のみ、インフルエンサ候補として選定するというものであり、手法 **RaF** と定義する。この手法では、該当する人物 u_i の影響力 $V(i \rightarrow t)$ として、ターゲットユーザ u_t からの返信スコア $S_{rep}(t \rightarrow i)$ とお気に入りスコア $S_{fav}(t \rightarrow i)$ の「積」を付与しランキングする。

$$V(i \rightarrow t) = \begin{cases} S_{rep}(t \rightarrow i) \times S_{fav}(t \rightarrow i) & (u_i \in U_{rep}^t \text{ and } u_i \in U_{fav}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

2.4.4 返信相手またはお気に入り相手を候補とする手法

この手法は、ターゲットユーザがこれまで少なくとも1回以上、返信またはお気に入り登録のどちらかを行った事のある相手のみ、インフルエンサ候補として選定するというものであり、手法 **RoF** と定義する。この手法では、該当する人物 u_i の影響力 $V(i \rightarrow t)$ として、ターゲットユーザ u_t からの返信スコア $S_{rep}(t \rightarrow i)$ の標準得点 $Z_{rep}(t \rightarrow i)$ とお気に入りスコア $S_{fav}(t \rightarrow i)$ の標準得点 $Z_{fav}(t \rightarrow i)$ の「大きい方」を付与しランキングする。

$$V(i \rightarrow t) = \begin{cases} \max\{Z_{rep}(t \rightarrow i), Z_{fav}(t \rightarrow i)\} & (u_i \in U_{rep}^t \text{ or } u_i \in U_{fav}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

但し、標準得点 $Z_{rep}(t \rightarrow i)$ と $Z_{fav}(t \rightarrow i)$ に関しては、返信スコア $S_{rep}(t \rightarrow i)$ とお気に入りスコア $S_{fav}(t \rightarrow i)$ とを統一的に大小比較するために用いている。本稿では、両者の平均 μ_{rep}^t 、 μ_{fav}^t 、並びに分散 σ_{rep}^t 、 σ_{fav}^t を統一するため、該当する人物 u_i の影響力 $V(i \rightarrow t)$ を求める際、 Z 値に変換（平均0、分散1）してから比較している。

$$Z_{rep}(t \rightarrow i) = \frac{S_{rep}(t \rightarrow i) - \mu_{rep}^t}{\sigma_{rep}^t}$$

$$Z_{fav}(t \rightarrow i) = \frac{S_{fav}(t \rightarrow i) - \mu_{fav}^t}{\sigma_{fav}^t}$$

2.4.5 RaF に 2 つのスコアの重みを考慮した手法

この手法は、2.4.3 節の手法 **RaF** に、ターゲットユーザが返信相手とお気に入り相手のどちらを重要人物として重み付けるかという概念を加えたものであり、手法 **RaF(w)** と定義する。インフルエンサ候補の選定法までは手法 **RaF** と同一であるが、この手法では、返信スコア $S_{rep}(t \rightarrow i)$ とお気に入りスコア $S_{fav}(t \rightarrow i)$ を乗算する際、その比率をパラメータ $w_a = [0, 1]$ によって調節できる。

$$V(i \rightarrow t) = \begin{cases} S_{rep}(t \rightarrow i)^{w_a} \times S_{fav}(t \rightarrow i)^{1-w_a} & (u_i \in U_{rep}^t \text{ and } u_i \in U_{fav}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

2.4.6 RoF に 2 つのスコアの重みを考慮した手法

この手法は、2.4.4 節の手法 **RoF** に、ターゲットユーザが返信相手とお気に入り相手のどちらを重要人物として重み付けるかという概念を加えたものであり、手法 **RoF(w)** と定義する。インフルエンサ候補の選定法までは手法 **RoF** と同一であるが、この手法では、返信スコアの標準得点 $Z_{rep}(t \rightarrow i)$ とお気に入りスコアの標準得点 $Z_{fav}(t \rightarrow i)$ を比較する際、どちらを採択しやすくするかを補正值 $z_t(w_o)$ によって調節できる。

$$V(i \rightarrow t) = \begin{cases} \max\{Z_{rep}(t \rightarrow i), Z_{fav}(t \rightarrow i) + z_t(w_o)\} & (u_i \in U_{rep}^t \text{ or } u_i \in U_{fav}^t) \\ 0 & (\text{otherwise}) \end{cases}$$

但し、補正值 $z_t(w_o)$ はターゲットユーザ u_t に依存し、パラメータ $w_o = [-100, 100]$ によって制御される。パラメータ w_o は、補正なしを 0%，負方向の補正限界値を -100%，正方向の補正限界値を 100% とした時、何パーセントの地点を補正值 $z_t(w_o)$ とするかを表している。

3. 評価実験

3.1 実験方法

本実験には、6 名 (表 1) の被験者に協力してもらっており、事前に SNS (Twitter) 内でインフルエンサであると思う人物を 10 名挙げてもらい正解データを作成した。実験手順としてはまず、6 種類の手法 (2.4 節) を用いて、被験者 6 名のインフルエンサを推定する。次に各被験者、また各手法で、正解データとの F 値を評価する。

表 1 被験者の SNS 利用 (行動ログ) の特徴

	返信数/全ての投稿数	お気に入り登録数/日
被験者 1	0.248	0.058
被験者 2	0.823	10.265
被験者 3	0.556	1.553
被験者 4	0.259	1.160
被験者 5	0.226	0.722
被験者 6	0.339	1.289

但し、各手法では幾つかのパラメータを設定する必要がある。これに関して本稿では、手法ごとに、また被験者ごとに、全てのパラメータの組み合わせで実験を行い、6 種類の手法の理論限界値を明らかにする。

3.2 実験結果

表 2 には、行を 6 名の被験者、列を 6 種類の手法とし、各被験者のインフルエンサの推定性能 (F 値) が示されている。まず平均の行に着目すると、最も高い F 値が得られたのは、手法 **RoF(w)** である事が分かる。さらに手法 **RoF** の平均 F 値と手法 **RoF(w)** の平均 F 値との間で片側 t 検定を行った結果、ターゲットユーザが返信相手とお気に入り相手のどちらを重要人物として意識しているかという重みを加味して推定する手法には、そうでない時と比較して、その推定性能を有意水準 0.05 で改善する効果がある事も明らかになった ($p = 0.028$)。従って今後はその重みの調整法についても議論を深めていく必要がある。

3.2.1 被験者 1 に関する考察

ここからは特に目立つ結果が得られた 3 名の被験者について考察する。まず被験者 1 については、どの手法でインフルエンサを推定するかに依って F 値に大きな差が生じるという結果になった。表 2 の平均の列に着目すると、被験者 1 の F 値の標準偏差 (0.315) は、他の被験者と比較して明らかに大きい。また被験者 1 の特徴として、手法 **REP**, **RaF**, **RaF(w)** で推定した時に F 値が著しく低下する事も分かる。ここから被験者 1 は返信相手を重要人物としてほとんど意識していないユーザである事が推測できる。

これに関してスミルノフ・グラブス検定を行った結果、被験者 1 の F 値の標準偏差は異常値として検出されなかった ($p = 0.123$)。これは、この被験者と同様に、返信相手とお気に入り相手のどちらか一方を極端にインフルエンサとして意識しているユーザが少なからず存在する事を意味している。このように様々なタイプのユーザが存在する中、F 値のロバスト性を担保するためには、候補の選出時点で、返信相手のみ (**REP**)、お気に入り相手のみ (**FAV**)、返信かつお気に入り登録の相手のみ (**RaF** と **RaF(w)**) のように、条件を強くするべきではないと考察できる。この考えは、候補を選定する条件を緩和する手法 (**RoF** と **RoF(w)**) の F 値の標準偏差 (0.154 と 0.181) が、条件を厳しくする手法 (**RaF** や **RaF(w)**) の F 値の標準偏差 (0.239 と 0.247) よりも小さい事からも支持できる。

3.2.2 被験者 5 に関する考察

続いて被験者 5 については、どの手法を用いても不十分な F 値となった。最高 F 値を得た手法 **RaF(w)** できえ、2 名しかインフルエンサとして正しく推定できていない。また手法 **RoF(w)** でも、インフルエンサとして 72 名が推定される地点まで閾値を引き下げる事で、辛うじて再現率が 1.0 に達するが、適合率が極めて低くなってしまふ。

表 2 各手法でインフルエンサを推定した場合の F 値（正解数/推定数/候補数）

	REP	FAV	RaF	RoF	RaF(w)	RoF(w)	平均 [標準偏差]
被験者 1	0.057 (1/25/25)	0.667 (7/11/100)	0.087 (1/13/13)	0.621 (9/19/112)	0.087 (1/13/13)	0.667 (7/11/112)	0.364 [0.315]
被験者 2	0.500 (6/14/74)	0.545 (6/12/412)	0.500 (4/6/24)	0.545 (9/23/462)	0.556 (5/8/24)	0.600 (9/20/462)	0.541 [0.038]
被験者 3	0.762 (8/11/36)	0.581 (9/21/379)	0.762 (8/11/29)	0.581 (9/21/386)	0.800 (8/10/29)	0.762 (8/11/386)	0.708 [0.099]
被験者 4	0.500 (6/14/14)	0.545 (9/23/61)	0.522 (6/13/13)	0.529 (9/24/62)	0.526 (5/9/13)	0.571 (8/18/62)	0.532 [0.024]
被験者 5	0.235 (2/7/30)	0.214 (3/18/222)	0.222 (2/8/19)	0.194 (3/21/233)	0.267 (2/5/19)	0.244 (10/72/233)	0.229 [0.025]
被験者 6	0.462 (3/3/28)	0.462 (3/3/553)	0.462 (3/3/20)	0.462 (3/3/561)	0.462 (3/3/20)	0.462 (3/3/561)	0.462 [0.000]
平均 [標準偏差]	0.419 [0.244]	0.502 [0.156]	0.426 [0.239]	0.489 [0.154]	0.450 [0.247]	0.551 [0.181]	

この被験者の場合、そもそも返信スコアとお気に入りスコアの算出時点で、本来インフルエンサであるはずのアカウントが低い水準にいるという問題が発生していた。この原因として、返信スコアやお気に入りスコアを求める際に、ターゲットユーザからある相手への「返信」や「お気に入り登録」の頻度をそのまま用いてしまった事で、その相手の投稿数に依存して算出されてしまうという問題が考えられた。この被験者の正解データを分析すると、企業などの公式アカウントが多く含まれており、頻繁に投稿が行われるアカウントではなかったため、この欠陥を突かれ、2つのスコアが上位に引き上げられなかったと考察できる。

3.2.3 被験者 6 に関する考察

最後に被験者 6 については被験者 1 とは対照的に、どの手法においてもその F 値に差が全く生じなかった。また、いずれの手法でも同一の 3 名がインフルエンサとして推定され、全て適合するという結果であった。この現象は 2 つの特徴が重なり合う事で生じてしまった。

1 つ目は被験者 5 と同様の特徴である。被験者 6 がインフルエンサとして選んだ 10 個のアカウントのうち、7 個が公式アカウントであり、日常的に SNS を利用するユーザよりも投稿数が少なかった。これによって返信スコアやお気に入りスコアを算出する時点で、それら 7 個のアカウントは既に下位に引き下げられていた。2 つ目は、この被験者がインフルエンサとして選んだ残り 3 個のアカウントは、返信スコアのトップ 3 でもあり、お気に入りスコアのトップ 3 でもあったという点である。従ってどの手法を用いてランキングしても、正解データのうち 7 個のアカウントは極端に下位に、残りの 3 個は必ずトップ 3 に固定される事となり、結果として F 値が変動しなかったと考察できる。

4. まとめと今後の課題

本稿では、ユーザに影響を与える人物（インフルエンサ）を推定するために、ユーザの SNS 内での行動に着目した複数の手法を提案し、それらの推定性能を議論した。その結果、ある人物からターゲットユーザへの影響力を算出する際、ターゲットユーザからその人物への返信数とお気に入り登録数を考慮した 2 つのスコアを求め、さらにそれらのスコアの重みを調整しながら計算する事で、最も性能良くインフルエンサを推定できた。その重みはユーザによって変化する事も分かり、今後はその最適化を目指し、ユーザの他の行動を手掛りとしたり、提案手法において、重みの変化による性能の傾向を分析したりし、議論を深める。

また、公式・有名アカウントなど、ユーザとの直接的な関係が薄いアカウントでも、影響力のある発信者として意識されている事が分かり、これらのアカウントは提案手法で再現しきれないという問題も残った。これに関しては、相手のフォロワーが幾らいるか、また単位投稿あたり、不特定多数から平均幾ら返信やお気に入り登録を受け取るかなど、ターゲットユーザに依存せず、相手が固有に持つスコアなども組み合わせる事で、推定性能の向上を図る事ができないかという可能性についても検討していく。

参考文献

- [1] 小寺 暁久, 横山 昌平, 山田 文康, “Twitter におけるユーザ同士の会話に基づいた親密度の評価と時系列的変化の可視化,” DEIM Forum 2015, F8-4 (2015).
- [2] 齋藤 準樹, 湯川 高志, “ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価,” 情報処理学会研究報告, Vol.2011-IFAT-102, No.2 (2011).
- [3] 後藤 清豪, 高田 秀志, “ソーシャルメディア上での行動に基づく「意外な情報」の提供者になり得る人物の推薦手法,” 情報処理学会研究報告, Vol.2010-DBS-151, No.4 (2010).