Vol.2015-ICS-178 No.11
2015/3/3

Twitter上での発話履歴の時系列パターンに基づく特定発話行動予測手法の検討

阿部 秀尚[1,a)]

概要：Twitter をはじめとする SNS では，ユーザの日々の興味や関心がテキストやその発話行動として可視化されている．そのなかで，企業をはじめとして，興味・関心のあるユーザを特定する要求が高まりつつある．しかしながら，単に発信情報を受信するユーザが真に興味・関心を持っているかどうかを的確に判断することは困難が伴う．本研究では，Twitter 上でリツイートと呼ばれるユーザが受信した情報を再発信する行動に着目し，過去の発話履歴の内容に基づいて再発信行動の予測を行うモデル構築手法の開発を目指す．本稿では，インターネット上で大規模な通信販売を行うサイトの発信したテキストとそのテキストを受信するフォロワーについて，再発信されたテキストとフォロワーの日々の発話テキストから得られた特徴語の比較を行う．さらに，行動予測モデル構築のため，フォロワーの発話テキストから得られた特徴語の使用頻度に基づく評価指標の時系列パターン抽出の可能性について検討する．

An Analysis for Developing User Behavior Analytic Model Construction Method on Twitter

Abstract: According to popularization of many SNS such as Twitter, enterprise users and some other people want to reach interested users more efficiently. However, it is difficult to detect more deep interests of the users without considering the users' behavior. In this study, I focus on a characteristic behavior of the users on Twitter, called retweet by followers. By taking followers' tweet history, a method constructing analytic models based on temporal patterns of term evaluation indices is described in this paper. The method combines ordinary used characteristic term extraction and a temporal pattern extraction of the terms usages. In the experiment, both of the characteristic terms in the retweeted text and the followers' tweet history is shown on the three major e-commerce accounts in Japan. Based on the results, some temporal patterns of the term importance indices are also shown to discuss the feasibility for predicting the retweet behavior of the followers.

1. はじめに

近年，Twitter をはじめとする SNS（Social Media Services) の普及によって，SNS ユーザー（以降，単にユーザと呼ぶ）の日常の興味・関心に基づく発話や行動がメディア上に可視化されるようになってきた．これに伴い，企業や政治家などは，より自身に興味・関心が近いユーザにより効率的に考えを伝え，さらに多くの関心のあるユーザとの接触を求めるようになってきている．このため，ネットワーク分析手法を用いた興味・関心の近いユーザの検出手法 [1] やテキストマイニングに基づく分析手法が数多く開発されてきた [2]．しかしながら，ユーザの行動や発話の履歴を考慮せずに次の行動を的確にとらえることは困難が伴い，行動予測をより的確に行える手法の開発が求められている．

本研究では，ユーザが情報拡散を行う再発信行動である Twitter 上でのリツイート (retweet) と呼ばれる行動に着目し，ユーザの発話履歴の内容からこの再発信行動を予測するモデルの構築手法の開発を行う．このため，本稿では，リツイートされたテキストに含まれる特徴語群とフォロワーと呼ばれる特定アカウントからの情報に興味があると思われるユーザの発話履歴に含まれる特徴語群の差異について検討を行う．また，発話履歴に含まれる特徴語の使用頻度に基づき算出される評価指標群について，時系列パターンの生成の可能性について示す．これらの結果をもと

[1] 文教大学情報学部情報システム学科
Dept. of Infomation Systems, Bunkyo University, Namgegaya 1100, Chigasaki, Kanagawa 253–8550, Japan
[a)] hidenao@shonan.bunkyo.ac.jp

に，発話履歴の時系列パターンに基づく情報再発信行動を予測するモデル構築手法の開発について考察する．

以降では，2 節において，提案手法を述べる．次に，語句を評価し，特徴語として選定するための評価指標群の定義について，3 節で述べる．4 章に示す特徴語の抽出実験では，日本国内で多数のフォロワーを有するインターネット通販サイトの 3 アカウントについて，リツイートされたテキストに含まれる特徴語群とリツイートしたフォロワーの発話履歴からの特徴語群をそれぞれ抽出する．さらに，1 つのアカウントのフォロワーの発話履歴について，重要語評価指標として用いられる指標の時系列パターンの生成を行う．最後に，5 節でまとめを述べる．

2. 発話履歴の時系列パターンに基づく情報再発信行動を予測するモデル構築手法

本手法では，ユーザの特定の発話行動が過去の発話内容から影響を受けていることを仮定し，過去の発話内容から得られる各種の特徴に基づき特定の発話行動が出現することを予測するモデルを構築する．

テキスト分析においては，過去の発話内容の特徴を得る手法として，テキスト集合からの特徴語抽出が良く知られている．しかしながら，特徴語抽出で得られた特徴語群は，過去のどの時点で出現するか，またその出現傾向などの情報を一切持たない．このため，阿部は時間軸方向の変化パターン（以降，時系列パターン）に着目し，複数の語句の評価指標の時系列パターンから語句の出現などを予測するモデルを構築する手法の開発を行ってきた [3]．

一方，より多くの特徴によって説明することで従属する変数の予測が正確になる，という観点からみると，従来の特徴語の出現に基づく特徴と語句の評価指標の時系列パターンのいずれが予測に有用であるかは事前に自明ではない [4]．このため，本手法では，ユーザの発話履歴を用いることによって，履歴中の内容をより正確に特徴づける特徴として，特徴語の出現と使用されている語句の属する時系列パターンの双方を用いることが必要である．さらに，行動という点については，発話の回数や間隔についての時系列パターンも同様に特定の行動に関連付けられることが考えられる．

以上の手法の開発について検討するため，特定の行動を行ったユーザの発話内容を代表する特徴語群，および特徴語の使用傾向の変化を示す時系列パターンについて，実データでの抽出を行う．

3. 語句の出現頻度に基づく評価指標群

テキスト中の語句 [*1] は，1 つ以上の単語の連なりとして表される．前後関係にある単語 w_i と単語 w_j については，必ず $i < j$ であり，この 2 つの単語から成る語句 $term_x$ は $term_x =< w_i, w_j >$ と表す．語句のこのような性質から，テキスト中の語句はすべて前後の順序関係がある系列データとみることができ，語句は部分系列として考えることができる．

以上から，自然言語処理における語句の重要度指標のほか，系列パターン評価指標を語句を評価する指標として，次のように定義する．

3.1 テキスト中の語句の重要度評価指標

語句の重要度を測るため，自然言語処理やテキストマイニングでは種々の重要度評価指標が開発されてきた．これらの指標の基準となるのは主に語句の出現頻度である．語句の出現頻度は単語が出現した回数を 1 つの文書内であっても重複して数え上げる**語頻度** (TF: term frequency)，語句の含まれる文書数にあたる**文書頻度** (DF: document frequency) の 2 つの出現頻度計測基準が一般に用いられる．

表 1 に L 個の単語から成る語句 $term_i =< w_1, \cdots, w_L >$ $(L \leq 1)$ の代表的な評価指標を示す．TFIDF は TF および DF の双方を考慮し，TF に対象文書全体 $|D|$ と DF の割合を重みとした重要度指標として，重要語句抽出に最もよく用いられる．さらに，2 つの出現頻度計測基準ごと，単純な割合から，出現頻度に基づく性質を測るための指標が定義できる．

表 1 語句に対する出現頻度計測基準毎の重要度評価指標.

	頻度計測基準	
	語句を含む文書数 $DF = \mid D_{\in term_i} \mid$	語句の延べ出現頻度 $TF = \sum_j freq(term_i, d_j)$
支持度	$DF / \mid D \mid$	$TF / \sum TF_{term_i}$
オッズ	$DF / (DF - \mid D \mid)$	$TF / (TF - \sum TF_{term_i})$
自己情報量	$(DF / \mid D \mid) \log_2 (DF / \mid D \mid)$	$(TF / \sum TF_{term_i}) \log_2 (TF / \sum TF_{term_i})$
Jaccard係数	$\frac{DF}{DF(w_1 \cup \ldots \cup w_L)}$	$\frac{TF}{TF(w_1 \cup \ldots \cup w_L)}$
TFIDF	$TF * \log(\mid D \mid / DF)$	

3.2 系列パターン評価指標による語句の評価指標

系列パターンの評価指標は，アイテム集合 I に属するアイテム $i \in I$ の列である系列データ $s_i =< i_1, \cdots, i_m >$ から成る系列データセット $D = \{s_i\}$ における部分系列 α の出現頻度 $freq(\alpha, D)$ に基づき，系列パターンの様々な性質を計量化する指標である．系列データセット D における部分系列 $\alpha =< i_1, \cdots, i_j >$ $(j \leq m)$ の出現頻度の数え上げには，語句の重要語評価指標と同様に二種類の出現頻度基準が代表的である．

各文書内での重複も考慮した頻度基準 TF と重複を考慮しない頻度基準 DF を用いて，順序を持たないアイテム集

[*1] 語句とは，単語あるいは複数の単語から成る句を指す．

合の評価指標群 [5]，および系列パターン内のアイテムを考慮した評価指標群 [6] を基に確信度 (Confidence) に基づく指標を定義する．ここで α を語句 $term_i$，各アイテムを語句 $term_i$ 中の単語 w_l とする．

表 2 部分系列 α に対する出現頻度計測基準毎の系列パターン評価指標．

	頻度計測基準	
	語句を含む文書数 $DF = \mid D_{\in term_i} \mid$	語句の延べ出現頻度 $TF = \sum_j freq(term_i, d_j)$
先頭確信度 (H-Conf)	$\frac{DF}{DF(w_1, D)}$	$\frac{TF}{TF(w_1, D)}$
最大確信度 (MaxConf)	$\max\left(\frac{DF}{DF(w_l, D)}\right)$	$\max\left(\frac{TF}{TF(w_l, D)}\right)$
全体確信度 (AllConf)	$\frac{DF}{\max(DF(w_l, D))}$	$\frac{TF}{\max(TF(w_l, D))}$
系列全体確信度 (SeqAllConf)	$\frac{DF}{\max(DF(\beta_{\subset term_i}, D))}$	$\frac{TF}{\max(TF(\beta_{\subset term_i}, D))}$

表 2 に示すように，語句の系列パターンとしてのアイテム間の順序関係を考慮すると，α の出現頻度と α の部分系列 β の出現頻度との割合である確信度について 8 種類以上の指標が定義できる．

4. インターネット通販アカウントとそのフォルワーの発話行動分析

本節では，Twitter 社の API[7] を通じて得られるツイートと呼ばれる各ユーザが発信した 140 文字以内のテキストを対象に，有名なアカウントの発信したツイートのうちリツイートと呼ばれる再発信されたツイートの特徴と再発信を行ったユーザのそれ以前の期間でのツイートの特徴分析を行う．ここで，ユーザの一般的な興味をより広くとらえるため，再発信の行動を「リツイートされたツイートに含まれる特徴語をツイートすること」と定義する．

実験の手順は，以下のとおりである．

(1) ある期間で有名なアカウントのリツイートされたツイートに含まれる特徴語を抽出
(2) 1. の特徴語を含むツイートを行ったユーザを列挙
(3) 1. の期間よりも前に 2. で列挙したユーザが発したツイートを収集
(4) 3. で収集したツイートに含まれる特徴語とそれらの出現頻度に基づく評価指標群の時系列パターンを抽出

以上の結果から，本研究で開発を目指す発話履歴の時系列パターンに基づく情報再発信行動を予測するモデル構築手法への適用について検討する．

本実験では，収集したテキスト集合から，特徴的な語句となりうる候補語句を得るため，自然言語処理で用いられる自動用語抽出手法を各テキスト集合に適用する．特徴語の候補となる語句の抽出には，中川らの開発した FLR スコアに基づく自動用語抽出手法 [8] を利用した．その際，FLR スコア算出に必要とされる名詞の同定は MeCab[9] および同時に配布される IPA 辞書 (mecab-ipadic-2.7.0-20070801) を利用し，形態素解析結果から名詞の同定を行った．FLR スコアの算出の結果から，$FLR(term_i, D) > 1.0$ となる語句の内，文字数が 2 以上のものを対象として特徴語の候補とした．

4.1 再発信されたテキストに含まれる特徴語の抽出

本実験では，セブンネットショッピング (7_netshopping)，Amazon.co.jp(AmazonJP)，楽天市場 (RakutenJP) の各公式アカウントから発信されたツイートの内，リツイートと呼ばれる再発信をされたツイートの集合について，特徴語抽出を行う．

ツイートは，2015 年 1 月 15 日から 2015 年 1 月 20 日にかけて，各アカウントが発信したツイートを対象とした．各アカウントの発信したツイートのうち，リツイートされたツイート数，FLR スコアによる特徴語の候補語句数は表 3 のとおりである．

表 3 各アカウントから 2015 年 1 月 15 日から 2015 年 1 月 20 日に発せられたツイートの内，リツイートされたツイート数，および FLR スコアによる候補語句数．

アカウント	$\mid D \mid$	候補語句数
セブンネットショッピング	76	369
Amazon.co.jp	193	857
楽天市場	127	540

次に，表 4 に各アカウント毎の TFIDF の値に基づく上位 10 位までの語句と各語句の支持度および先頭確信度を示す．なお，支持度と先頭確信度は文頻度による頻度の数え上げ基準による．

表 4 に示す結果より，各アカウントによって発せられたツイートの内，特にフォロワーがリツイートした特徴的が語句群が得られた．これらの特徴語群は，各アカウントが発したツイート内の語句にとどまっており，フォロワーであるユーザの興味の一部と合致したものと考えられる．そのため，より広くユーザの興味に関連した語句を得るため，これらの語句を含むツイートを再発信したユーザの過去の発話内容について，特徴語や語句の使用頻度の変化についても調査する必要がある，と考えられる．

4.2 再発信を行ったユーザが発信したテキストの特徴語と時系列パターン

Twitter では，各アカウントのフォロワーと呼ばれる，常に各アカウントの発信する情報を得ているユーザを得ることができる．フォロワーの発信したツイートのテキストには，ユーザの興味による様々な語句が含まれ，興味を反映

表 4 各アカウント毎の TFIDF 値に基づく上位 10 位までの語句と各語句の支持度および先頭確信度（文頻度基準）.

セブンネットショッピング				Amazon.co.jp				楽天市場			
語句	TFIDF	支持度(DF)	先頭確信度	語句	TFIDF	支持度(DF)	先頭確信度	語句	TFIDF	支持度(DF)	先頭確信度
限定	41.39	0.26	1.00	/	78.36	0.34	1.00	楽天 ポイント	84.18	0.23	0.58
:	36.92	0.16	1.00	タイム セール	73.57	0.22	1.00	楽天	74.57	0.39	1.00
セブン	33.47	0.29	1.00	OFF	63.38	0.20	1.00	応募	51.21	0.26	1.00
特典	32.95	0.22	1.00	!	58.52	0.16	1.00	フォロワー	51.21	0.26	1.00
予約	32.53	0.43	1.00	%	55.98	0.15	1.00	♪	46.70	0.38	1.00
予約 受付	31.95	0.37	0.85	人気	53.11	0.14	1.00	フォロー 解除	44.47	0.26	1.00
/	30.02	0.33	1.00	PC	52.78	0.06	1.00	当選 確率	44.47	0.26	1.00
月	30.02	0.17	1.00	限定	52.14	0.10	1.00	完了	44.47	0.26	1.00
発売	29.21	0.20	1.00	受付	52.12	0.12	1.00	/	44.47	0.26	1.00
アカチャンホンポ	27.06	0.14	1.00	チェック	51.10	0.13	1.00	下	44.31	0.23	1.00

しているものと考えることができる．そのため，各アカウントの発信したツイートをリツイートと呼ばれる再発信する行動についても，それ以前の普段から発信しているテキストに関連する特徴語が含まれているものと推定する．

以上を確認するため，本節では，前節で述べたインターネット通販に関する有名アカウントのフォロワーについて，リツイートよりも以前のツイート内容の特徴語抽出と時間経過による変化パターンを評価指標の時系列パターンとして抽出する．ここでは，2015 年 1 月 15 日から 2015 年 1 月 20 日の各有名アカウントのツイートを再発信したユーザ [*2] について，2015 年 1 月 1 日から 2015 年 1 月 14 日に発信されたツイートを取得した．

図 1 にセブンネットショッピングのフォロワーのうち，表 3 に示した語句を含むツイートを 2015 年 1 月 15 日から 1 月 20 日までに行った上，2015 年 1 月 1 日から 1 月 14 日の間にツイートのあったユーザ数を示す．

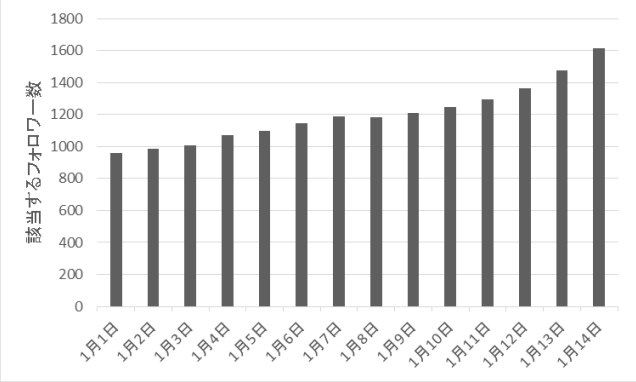

図 1 表 3 に示した語句を含むツイートを 2015 年 1 月 15 日から 1 月 20 日までに行った上，2015 年 1 月 1 日から 1 月 14 日の間にツイートのあったセブンネットショッピングのフォロワー数.

これらのユーザのツイートから得たテキストに対し，FLR スコアによる語句の抽出を行った結果，271,234 個の候補語句が得られた．次に，FLR スコア上位 1000 個までの特徴語について，取得日ごとのテキスト集合について各語句の重要語評価指標および系列パターン評価指標の計算を行った．表 5 に 1 月 1 日と 1 月 14 日のテキスト集合それぞれで TFIDF による評価値が上位 10 位までとなった語句を示す．

[*2] TwitterAPI の制限のため，ランダムに取得した 5000 ユーザのうち，該当するユーザである．

表 5 セブンネットショッピングのフォロワーの 1 月 1 日と 1 月 14 日のテキスト集合における TFIDF 評価値と語句（上位 10 位まで）.

1月1日		1月14日	
語句	TFIDF	語句	TFIDF
フォロー	9644.49	RT	14476.75
RT	8056.19	フォロー	12019.62
人	6355.03	人	10623.82
日	5086.92	日	7808.90
無料	4992.20	無料	7361.78
相互	4766.66	相互	5757.38
方法	3855.28	楽天	5049.00
支援	3320.98	方法	5009.66
年	3236.76	ww	4923.89
月	3160.53	!!	4546.57

表 5 から，TFIDF による評価値の順位が取得日の違うテキスト集合で変化することが分かる．このことから，各語句の評価指標値は，取得日の異なるテキスト集合において変化していると言える．また，これら上位に順位づけられる語句は，表 4 にあげるような語句とは明らかに異なることから，日頃発せられているツイートの内容はリツイートの内容ととは異なる．ただし，一部ボットと呼ばれるプログラムによる自動リツイートや定期的にツイートする機能のあるアカウントも含まれるため，これらの判定と併せて，人間のユーザが行う行動の特徴を語句の差異や語句が属する時系列パターンの違いから分類予測を行っていく必要がある．

5. おわりに

本稿では，ユーザが情報拡散を行う再発信行動である Twitter 上でのリツイート (retweet) と呼ばれる行動に着目し，リツイートされたテキストに含まれる特徴語群とフォロワーと呼ばれる特定アカウントからの情報に興味がある

と思われるユーザの発話履歴に含まれる特徴語群の差異について検討を行った．この結果，各アカウントが発信したツイートのうち，再発信されたツイートに含まれる語句とユーザの発話内容とは差異があることが確認された．また，日時の差異によって，同一の語句と評価指標の組み合わせでも，値はもちろん相対的な順位も変化することが確認できた．

今後は，以上の検討を基にユーザの発話履歴から各評価指標の時系列パターンの抽出を行い，再発信行動の説明変数として時系列パターンの有無を用いた予測モデルの構築を行っていく．また，語句の使用頻度の変化によるパターンだけではなく，発話間隔など行動そのものの変化についても時系列パターンを得て，従来の特徴語による特徴づけと合わせた予測モデルの構築手法として開発を進めていく．

謝辞 本研究は，日本学術振興会・科学研究費補助金（基盤研究 (C) 24500175）の助成によるものである．

参考文献

[1] 今森大地，田島敬史：Twitter におけるフォローに関する影響力に基づくハブ度の推定，第 6 回データ工学と情報マネジメントに関するフォーラム（第 12 回日本データベース学会年次大会），B7-1 (2014).
[2] 奥村 学：マイクロブログマイニングの現在，電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション (第 3 回集合知シンポジウム)，Vol. 111, No.427, pp.19-24 (2012).
[3] Abe, H.: Analysis for Finding Innovative Concepts Based on Temporal Patterns of Terms in Documents, Theory and Applications for Advanced Text Mining, Shigeaki Sakurai (ed.), pp.37-50 (2012).
[4] 阿部秀尚，津本周作：重複系列の発生パターンに関する時系列マイニングとその医療応用，人工知能学会誌，Vol. 27, No. 2, pp.154-161 (2012).
[5] Wu, T., Chen, Y., and Han, J.: Association Mining in Large Databases: A Re-examination of Its Measures, In Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 621-628 (2007)
[6] 岩沼宏治：テキスト系列マイニングにおける有用性尺度について，人工知能学会誌，Vol.27, No.2, pp.136-145, (2012)
[7] Twitter Wep API 1.1: `https://dev.twitter.com/docs/api/1.1`
[8] Nakagawa, H.: Automatic term recognition based on statistics of compound nouns. Terminology, Vol.6, No.2, pp.195-210 (2000)
[9] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: `https://code.google.com/p/mecab/`