3B3-2

ソーシャルブックマークとしてのTwitterリスト機能の応用

Application of Twitter List Function as Social Bookmarks

榊　剛史*1　　松尾　豊*1
Takeshi Sakaki　　Yutaka Matsuo

*1東京大学
The university of Tokyo

近年, マイクロブログサービスである Twitter が脚光を浴びており, ビジネス・研究分野において, 様々な分析やサービスが発表されている. 本論文では Twitter のリスト機能に注目する. リスト機能は, ユーザーを整理するための機能であるが, 同時にユーザーへのソーシャルブックマーク（ＳＢＭ）と捉えることができる. 既存のＳＢＭ分析の手法を適用することで,Twitter ユーザーの属性抽出や特徴語抽出を行う.

1.　はじめに

近年, ブログや SNS をはじめとするソーシャルメディアサービスが注目を浴びている. ソーシャルメディアとは, 一般のユーザーが情報を作り出し, 発信していくメディアのことである.

その中でも特にマイクロブログサービスの 1 つである Twitter が世界的に急成長を遂げている. 現在, 登録ユーザーは 1 億 500 万人,1 日あたりの新規ユーザー獲得数が 30 万人,2009 年〜2010 年の成長率は 700% 超えており, 最も大きな Web サービスの 1 つであると言える.

Twitter のサービス開始当初から,Twitter を活用した研究については様々な試みがなされてきたが, ユーザー数の増加に伴い, ここ 1 年で Twitter を対象とした研究の数は急速に増してきている. Twitter 研究会や各研究会のでの特集など,Twitter に注目する学術コミュニティにも増加している. Java らは Twitter をソーシャルネットワークと捉え, ソーシャルネットワーク分析の手法を用いて Twitter のネットワーク分析を行った [Java 07]. 榊らは Twitter のリアルタイム性を利用し, リアルタイムな地震検出手法を提案している [Sakaki 10].

本研究では,Twitter のリスト機能に注目する. リスト機能とは, 各ユーザーが自分のリンクのあるユーザー (Twitter では, リンクすることを follow,follow しているユーザーを friends と呼ぶ) を整理するための機能である. 具体的には, 自分の Follow しているユーザーをいくつかのグループにグルーピングし, 各グループに名前をつける機能である.1 つ 1 つのグループを**リスト**, リストにつけた名前を**リスト名**と呼ぶ.

このリスト機能は, 前述の通りユーザーを整理する機能であるが, 視点を変えれば, 図 1 のように**「自分の friends にリスト名でタグ付けをしている」**と考えることができる. つまり, ユーザーによるユーザーへのソーシャルブックマーク機能であると考えられる. このような考え方を用いて, ユーザーの属性分析を行うサービスも存在している*1.

ソーシャルブックマークの研究については, ソーシャルブックマークにおける Folksonomy を Actor-Concept-Instance による 3 層構造を持つグラフと定義し、Folksonomy を Ontology に拡張する可能性について言及した Peter Mika による研究 [Mika 05] や, Folksonomy におけるユーザーのリンクを用い,

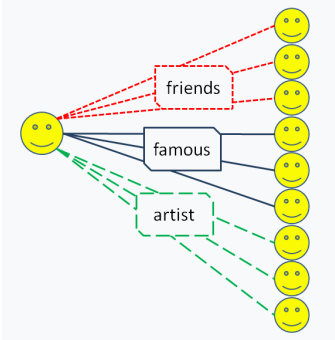

図 1: ユーザーによるユーザへのＳＢＭ

そのネットワークを活用する新しいソーシャルブックマークサービスを提案した大向らによる研究が挙げられる [Ohmukai 05].

本研究ではソーシャルブックマークを対象とした研究に用いられる手法を用いて, 既存研究とは違った角度から Twitter を分析することを目指す. 具体的には, リスト名によるユーザーの属性判別, 同一リスト名に共通する特徴語の抽出を行った。

2 章では, データの整備および全世界と日本でのリスト機能の使われ方の違いについて分析する. 3 章ではリスト名を使ったユーザー属性の判別の手法について言及する. また,4 章では同じ名前を持つリストを代表するようなキーワードの抽出, 6 章では, 議論と今後の研究について述べる.

2.　リスト機能についての分析

本章では, 分析対象とするデータの収集方法について説明するとともに, データの集計値からリストの使われ方について分析する.

2.1　データの整備

本研究で用いるデータは以下のような手順により収集した.

1. 2009 年 12 月 22 日〜29 日に Tweet を投稿したユーザーから 60 万人を抽出
2. ユーザーがリストを作成している場合, そのリスト名をすべて収集

連絡先: 榊剛史, 東京大学工学系研究科, 東京都文京区弥生 2-11-16 工学部 9 号館，03-5841-1161，sakaki@biz-model.t.u-tokyo.ac.jp

*1　http://www.mustexist.com/list_tags/

表 2: 高頻度に使用されるリスト名

順位	世界	日本
1	friends 71297	bot 16035
2	news 47185	news 8037
3	music 42466	conversationlist 6044
4	celebs 27255	music 4315
5	conversationist	list 3193
6	amigos 17164	friends 2623
7	bot 16106	famous 2155
8	celebrities 15660	met 11185
9	sports 14542	info 1395
10	media 13971	who-i-met 1186

3. 収集した各リストについて, リストに含まれるメンバーをすべて収集

2.2 リスト機能の使われ方の違い

取得したデータについて, 各統計情報により, 日本と全世界でのリスト機能の使われ方を比較する.

表 1 は全世界および日本でのリスト所有比率や 1 ユーザーあたりのリスト数になどの単純統計量である. これを見ると, いずれの統計量も若干の差はあるものの, ユーザー 1 人あたりのリスト所有比率や 1 ユーザーあたりのリスト数には大きな違いはない. これより, リスト機能活用の度合いにおいては, 日本と全世界ではそれほど大きな差は無いと考えられる.

表 2 は, 全世界および日本での各リストのリスト名の集計し, 出現頻度が高い順に 10 件抜き出したものである. これを見ると, 日本では"bot"が非常に多いことが分かる. また,"list"や"info"という意味の無いリスト名をつける傾向が見て取れる. なお,Twitter における bot とは, プログラムにより自動的に呟きを投稿しているアカウントを意味している. また"met"や"who-i-met"など自分の直接の知り合いを表すリスト名が上位にランクされているのも日本の特徴である.

3. リスト名を用いたユーザーの属性判別

ソーシャルブックマークでは, ユーザーによってつけられたタグによってブログ記事やニュースの投稿を分類し, 検索を容易にする. それと同様に, 本章ではユーザーによってつけられたリスト名によってユーザーを分類し, 検索することが可能であるかを検証する.

まずユーザー u_i にリスト名 l_m のリストに含まれる回数を以下のように定義する.

$$N(u_i, l_m) \tag{1}$$

ここで, リスト名 l_m が与えられた場合, ある回数以上リスト l_m に含まれるユーザーは, l_m というキーワードに関連が深いと考えられる. そこで l_m に関連するユーザーを以下のように定義する.

$$Rel(l_m) = U(u_i | N(u_i, l_m) > n_{th}) \tag{2}$$

n_{th} : 閾値

この定義に基づき, あるキーワードに関連するユーザーの抽出実験を行った. 実験の概要は以下の通りである.

表 3: リストによるユーザー分類実験

リスト名	精度	再現率	F 値
politics	0.518(29/56)	0.707(29/41)	0.598
bot	1.000(100/100)	-	-

- 用いたリスト名 : "politics", "bot"
- 対象リスト : 日本語ユーザーのリストのみ
- "politics"の正解データ : 現役国会議員のアカウント
- "politics"の閾値 $n_{th} = 5$:F 値が最も高くなるように設定
- "bot"の正解データ : 正解データの準備が困難なため, 抽出した上位 100 ユーザーについて bot かどうかを人手により判定

実験結果は表 3 の通り.

実験結果に対する考察について, まず"bot"については 100%の精度が得られている. これは bot の定義が明確でリスト化されやすい,"bot"というリスト名のリストが非常に多いことに起因していると考えられる.

次に"politics"についてであるが,"bot"と比較すると精度はそれほど高くない. ただし, 実際に誤って判別された例を検証すると, オバマ大統領 (@BarackObama) や片山さつき元国会議員 (@katayama_s), 自民党広報部 (@jimin_koho) など, 語検出のほとんどが実際には政治家, または政治関係の団体であった. そのため精度の低さは,「現役国会議員のみ」という正解データの起因するものであるといえる. また, 再現率は 70%であるが, これは閾値を低く設定すれば 95%となり, ほぼすべての現役国会議員のアカウントを抽出することができる.

このようにリスト名を使うことで, ある程度の精度でユーザー分類が行うことができた.

4. リスト名による分類を用いた特徴語の抽出

本章では, リスト名によるユーザー分類を用いて, リスト名に関連する語を抽出することを試みる.

手順は以下の通り.

1. $Ret(l_m)$ より,$N(u_i, l_m)$ の上位 n_u ユーザーを選ぶ
2. 各ユーザーについて, 直近の呟き n_t 件を取得する.
3. $Ret(l_m)$ に属するユーザーの呟き集合を $Tw(l_m)$ とする
4. 複数のリスト名について 1.,2.,3. を行う
5. ある語の w の呟き集合 $Tw(l_m)$ への出現回数を $tf_w(m)$, 語 w を含む呟き集合の数を df_w とし, 各語の tfidf 値を算出する.
6. リスト名 l_m 毎に,tfidf 値の高い語を抽出し, それを l_m の特徴語とする.

本実験では,$n_u = 20$, $n_t = 200$ とし, またリスト名としては, 表 5 の語を用いた.

抽出された語は表 4 の通りである. 細かく中身について見ていくと, まず表 4 を見ると,"news","media"においては「日経

表 1: リスト機能の使用に関する単純統計量

項目	世界	日本
クロールユーザー数	107907	11653
収集リスト数	314891	32454
リスト所有比率	0.524(601259/314891)	0.616(32454/52709)
リスト数/1 ユーザー	2.92(314891/107907)	2.78(32454/11653)
メンバー数/1 リスト	10.5(3295519/314891)	9.49(307942/32454)

表 4: 抽出された特徴語 その 1

news	media	art	artist	web	iphone	friend	friends
ストーカー	今朝の空	sound,silence	美術店	sideA	KENTHE390	ヨゴレ	元女房
キプロス	小名浜	ハーモニカ	正平	人の口コミ	ShinjiHori	クリノッペ	小指
殺人事件	デカルト	剣さん	妖怪マンガ	圧縮	HMDT	差押	8tracks
エイシス	開花状況	8 小節	Galleries	Intel	林会長	supply	auth
mokei	桜の開花状況	sound	原久路	南武線	KENTHE	スペーサー	^ ω ^
橋下	蜃気楼	メルヴィル	仏教	nakatanigo	NSArray	本予算	リリパ
子会社化	カタクリ	ムッシュ	ベルメール	飼い主	3.0.	TAJO	9auth
日経平均	エリア化	綾小路	Moore	E メール	インスタンス (	差押禁止	道聴塗説
世界記録	スコア速報	f	Gogh	リアリティ	iPhoneiPad	RT@GIFT_Japan	1_again
歩道橋事故	小名浜 (福島	フランス座	ギャラリー	IntelCore	installed	wRT@simachoko	ゆとりちゃん
高地	冬の天気分布	焼肉屋	テンパ	田口さん	AirTags	予算書	REST
明石歩道橋	会見オープン	小節	楽部	IntelCorei	terminal	梅子	al1_again

表 5: 特徴語抽出に用いたリスト名

art , artist , bot, bots, conversationlist, creator, famous, friend, friends,info, iphone, list, media, met, music, news, pixiv, politics, web, who-i-met

平均」「歩道橋事故」「開花状況」「高気圧の中心」など, 時事関係の語が多く取れているのが分かる.

また”art”,”artist”では,「ハーモニカ」「8 小節」「美術店」「Gogh」など, 音楽, 美術に関する語が取得できている.

次に,”web”では「Intel」「E メール」などの IT 関係の語が, ”iphone”では「iPhone/iPad」「Airtags」といった iphone に関連する語が取得できている.

また.”friend”,”friends”では, ユーザー名が多く見受けられる. これは,Reply,Retweet などを頻繁に行っているために, ユーザー名がつぶやきに含まれやすいためではないかと推測される.

このように, リスト名によって分類したユーザーのつぶやきを取得することで, リスト名に特徴的な語を抽出することが可能であると考えられる.

5. 結論

本論文では,Twitter のリスト機能をユーザーによるユーザへのソーシャルブックマークの一種と捉え, ソーシャルブックマークの分析でよく使われる Folksonomy の手法を用いて, ユーザーの分類および特徴語の抽出を行った.

その結果,1. ユーザーの分類,2. 特徴語の抽出, いずれにおいても既存手法の適用可能性を示した. ただし, あくまで今回は試行的な論文であり, 整備された正解データを準備することができなかったため, 数値評価は十分であるとは言えない.

今後の研究においては, 十分な数値評価を行った上で, 他のソーシャルブックマークの手法を用いてユーザの分類精度の向上やキーワードによるユーザーの検索, リスト名の構造化, 特徴語の抽出などさまざまな問題に取り組んでいくことを考えている.

参考文献

[Java 07] Java, A., Song, X., Finin, T., and Tseng, B.: Why we twitter: understanding microblogging usage and communities, in *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, New York, NY, USA (2007), ACM

[Mika 05] Mika, P.: Ontologies are us: A unified model of social networks and semantics, *Proc. orf 4th International Semantic Web Conference* (2005)

[Ohmukai 05] Ohmukai, I., Hamasaki, M., and Takeda, H.: A Proposal of Community-based Folksonomy with RDF Metadata, *Proc. orf 4th International Semantic Web Conference* (2005)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proc. 18th International World Wide Web Conference (WWW2010)* (2010)