

マイクロブログにおける情報伝搬距離に着目した ユーザプロファイリング

島田 諭[†] 山口裕太郎^{††} 佐藤 哲司[†]

[†] 筑波大学大学院図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学情報学群知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

E-mail: [†]{sat,yamaguchi,satoh}@ce.slis.tsukuba.ac.jp

あらまし ソーシャルメディアやマイクロブログにおけるユーザ間での情報伝搬を有向グラフとみなしたとき、あるユーザのシードからの距離は、新しい情報や評価の定まっていない情報への態度、情報伝搬における役割といった、ユーザの総合的な特性を反映していると考えられ、トピックやコミュニティに依存しないユーザプロファイリングに有用と考えられる。本論文では、Twitter 上で1ヶ月間に投稿された約3億件の日本語ツイートをを用い、約300万人のユーザの特性を分析する。

キーワード Twitter, クラスターリング, グラフマイニング

User Profiling based on Information Propagation Distance in Microblogs

Satoshi SHIMADA[†], Yutaro YAMAGUCHI^{††}, and Tetsuji SATOH[†]

[†] Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, Ibaraki, 305-8550 Japan

^{††} College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
1-2 Kasuga, Tsukuba, Ibaraki, 305-8550 Japan

E-mail: [†]{sat,yamaguchi,satoh}@ce.slis.tsukuba.ac.jp

Abstract In microblogs, information propagation between users is able to be modeled as a directed graph. A user's distance from seed in the graph is shown the user's general characteristics, e.g., attitudes to novel information, roles in information propagation, etc. Thus, it's thought that the distance in information propagation is effective to profiling users depended from particular topics or communities. In this paper, we analyze 0.3 billion tweets posted by 3 million users in a month.

Key words Twitter, clustering, graph mining

1. はじめに

ソーシャルメディアやマイクロブログにおけるユーザ間での情報伝搬を有向グラフとみなしたとき、あるユーザのシードからの距離は、新しい情報や評価の定まっていない情報への態度、情報伝搬における役割といった、ユーザの総合的な特性を反映していると考えられ、トピックやコミュニティに依存しないユーザプロファイリングに有用と考えられる。

本論文では、Twitter 上で1ヶ月間に投稿された約3億件の日本語ツイートをを用い、約300万人のユーザの特性を分析する。

2. 関連研究

現在、Twitter を用いた情報抽出やユーザプロファイリングでは、他のユーザのツイートを引用するリツイート（以下、RT

と記す）や、特定の話題に関するツイートであることを明示するハッシュタグを用いることが多い。しかし、RT やハッシュタグは、投稿時に厳密な書式を用いて記述する必要があり、これらの機能を活用できるユーザの比率は必然的に低くなると考えられる。多様なユーザを網羅できるユーザプロファイリングを実現するには、より多くのユーザが実際に使用している機能や書式を用いることが欠かせない。

Kwak らは、Twitter における RT による記事間のつながりをツリー構造とみなす RT ツリーを提案し、情報伝搬におけるシードからの距離とユーザの関係を分析している [1]。シードからの距離が最大4となる RT ツリーが最多で、最長では18となる RT ツリーもあり、一連の RT ツリーには、100人から1,000人のユーザが含まれることを報告している。

Boyd らは、ツイートにおける、返信（リプライ）、RT、URL

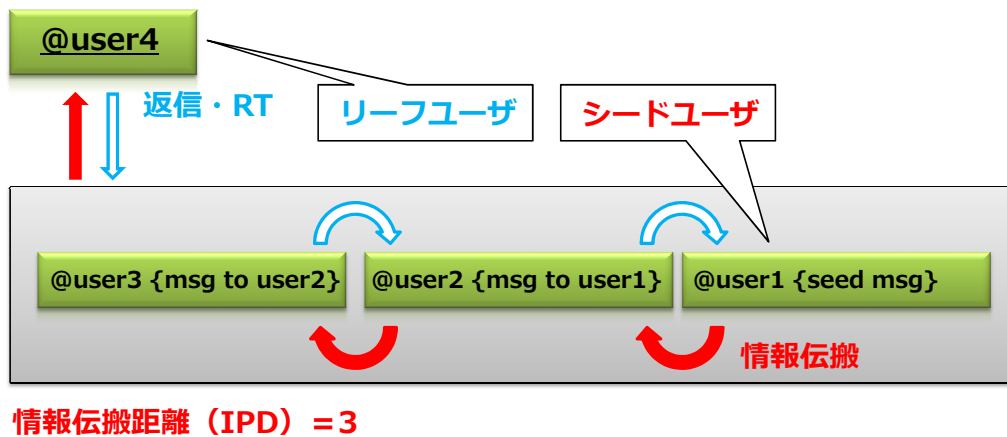


図 1 レスポンスツリーにおけるレスポンスと情報伝搬

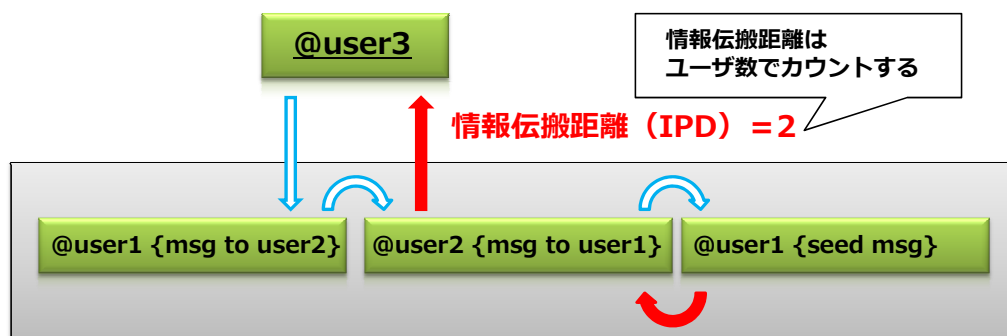


図 2 レスポンスツリーにおける情報伝搬

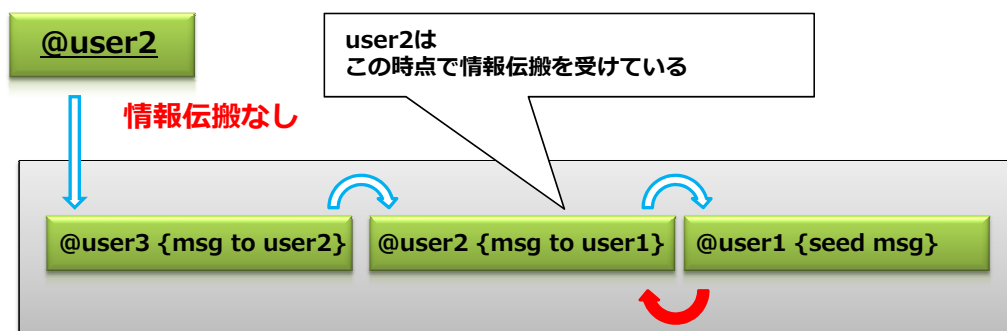


図 3 レスポンスツリーにおける情報伝搬

やハッシュタグを含むツイートなどの比率を報告している [2]。半年間にわたり、パブリックタイムラインを 5 分ごとに収集し、ランダムにサンプリングした 72 万ツイート、43 万ユーザについて分析した結果、返信は 36%、「RT」「retweet」「via」など、RT を意図したツイートは全体の 3% で、ハッシュタグを含むツイートは 5%、URL を含むツイートは 22% だったことを報告している。また、返信と RT が厳密には区別されずに使われていること、および、「thx @user」「via @user」「r @user」といった、Twitter のシステム上で返信や RT として扱われない非公式の書式での返信や RT が広く使われていることを報告している。

3. アプローチ

Boyd らの調査 [2] で明らかになっているように、Twitter で

は、非公式な書式を含む返信および RT を用いるユーザが多い。このため、Twitter 上での情報伝搬を扱う際に、公式の書式で記述された RT だけに着目すると、ユーザの網羅性が下がるという問題がある。

本研究では、Kwak らの RT ツリー [1] を拡張し、非公式な書式を含む返信および RT（以下、レスポンスと総称する）を情報伝搬の経路とみなす、レスポンスツリーを定義する。レスポンスツリーにおけるレスポンスと情報伝搬の流れを図 1 に示す。ユーザ *user4* が「(RT) @user3 [msg to user2] @user2 [msg to user1] @user1 [seed msg]」というツイートを投稿したとする。「@」で始まる文字列はユーザ名を表す。ユーザ名によって分割される、ユーザ名以外の文字列をメッセージと称する。メッセージは、メッセージの直前に記されたユーザから、メッセージの直後に示されたユーザへのレスポンスである。た

表 1 収集されたデータの概要

ツイート数		305,265,021	
ユニークユーザ数		3,394,610	
返信	(LIKE '%@%')	117,738,390	38.6%
2 文字目以降にのみ「@」を含むツイート	(LIKE '%_@%_')	35,706,813	11.7%
うち, RT	(LIKE '%RT%')	(21,182,362	6.9%)
うち, 非公式 RT	(LIKE '%_ RT%@_ %')	(6,789,638	2.2%)
うち, QT	(LIKE '%_ QT%@_ %')	(741,931	0.2%)
計		153,445,203	50.3%
URL を含むツイート	(LIKE '%http://_ %')	43,897,802	14.4%
うち, 「@」を含むツイート	(LIKE '%@%_')	(15,356,280	5.0%)
ハッシュタグを含むツイート	(LIKE '%#_ %')	19,449,245	6.4%
うち, 「@」を含むツイート	(LIKE '%@%_')	(5,999,687	2.0%)

表 2 ツイートの収集方法

収集期間	2011 年 11 月 16 日 0:00:00 から 2011 年 12 月 15 日 23:59:59 まで
収集条件	
lang	ja (日本語)
geocode	35.002054, 134.997758, 2000km (日本の全域)

表 3 分析対象のユーザ

from_user	返信または RT したユーザ	2,818,223
to_user	返信または RT されたユーザ	4,137,923
	うち, 収集条件内のユーザ	2,398,314
	うち, 収集条件外のユーザ	1,739,609

だし, ツイートの末尾のメッセージには, レスポンスの相手となるユーザは存在しない. このとき, *user4* は, *user3* および *user2* を介して, *user1* が最初に発信した情報の伝搬を受けたとみなす. 情報伝搬の発信源となっている *user1* をシードユーザ, 伝搬した情報の最終的な受信者となっている *user4* をリーフユーザと称する. また, シードユーザが発信したメッセージをシードメッセージと称する. 図 2 に示すように, 経由した異なりユーザ数を情報伝搬距離 (IPD) とする. また, 図 3 に示すように, レスponsツリーにリーフユーザが含まれている場合は, 情報伝搬がなかったとみなす.

ユーザの情報伝搬距離は, 新しい情報や評価の定まっていな情報への態度, 情報伝搬における役割といった, ユーザの総合的な特性を反映していると考えられる. このため, 情報伝搬距離は, トピックやコミュニティに依存しないユーザプロファイリングに有用と考えられる.

4. 分 析

本論文では, まず, Twitter に投稿された日本語ツイートを網羅的に収集し, Twitter 上の返信, RT, ハッシュタグなどの機能が, ユーザにどのように使用されているかを明らかにする. 次に, 非公式な書式を含む返信および RT を用いた情報伝搬距離の算出を試み, ユーザプロファイリングにおける有用性を検討する.

4.1 分析に用いるデータ

Twitter の Search API^(注1)を用いて収集した日本語ツイートを分析に用いる. 収集方法を表 2 に示す. 可能な限り網羅的にツイートを収集するため, 日本の全域が該当する位置情報^(注2)を検索条件とした. Twitter では, ユーザがフリーテキストで記入する「location」と, 投稿に用いられるクライアントソフトウェアが GPS 等に基づいて自動的に付与する「geocode」^(注3)の 2 種類の位置情報があり, ツイートごとに記録される. 実際には「geocode」が記録されたツイートは少なく, 「location」も具体的な地名でなく「地球のどこか」「自分の部屋」といった曖昧な内容であることが多い. ただし, Twitter の Search API では, 日本語で記入された「location」が特定の地名と一致しなければ, デフォルトでは東京とみなされているとみられる. このため, 上記のような曖昧な「location」を持つツイートも, 上述の検索条件によって取得できる. なお, ユーザのフォローに関する情報, お気に入り, およびツイートを非公開に設定しているユーザのツイートは, この方法では取得できない. これらの情報の収集には, 時間あたりのリクエスト数やユーザによる許諾など制約が多いため, 網羅的な収集は困難である. このため, 本研究では, 公開されているツイートのみを用いることとした.

4.2 返信, RT, URL, ハッシュタグの比率

収集されたデータの概要を表 1 に示す. 収集されたツイート数は, 同じ期間にパブリックタイムラインで収集可能な件数の約 700 倍であり, 実際に同じ期間にパブリックタイムラインで収集された日本語ツイートの件数の約 3,000 倍に相当する. ツイート数およびユニークユーザ数から, 1ヶ月間のユーザ 1 人あたりのツイート数は平均 100 であることがわかる.

収集されたデータにおける, 返信および RT の比率は, 50.3%だった. なお, 2 文字目以降にのみ「@」を含むツイートは, ユーザ名より前にコメントが挿入されたツイート, および RT, 非公式 RT, QT を含む. 括弧内は, 集計に用いた SQL 文における正規表現である. このうち, 「RT」「QT」の文字列を含むツイートは, 全体の 9.3%で, Boyd らの調査結果 [2] と比較し, 約 2 倍多いことがわかった. 一方, URL やハッシュ

(注1): <http://search.twitter.com/search.json>

(注2): 兵庫県西脇市を中心とする半径 2,000km の円内.

(注3): 検索結果において緯度と経度が返される.

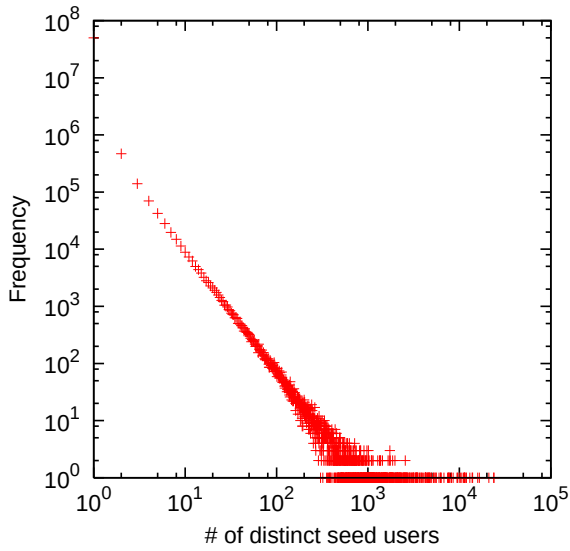


図 4 シードメッセージごとの異なりシードユーザ数の分布

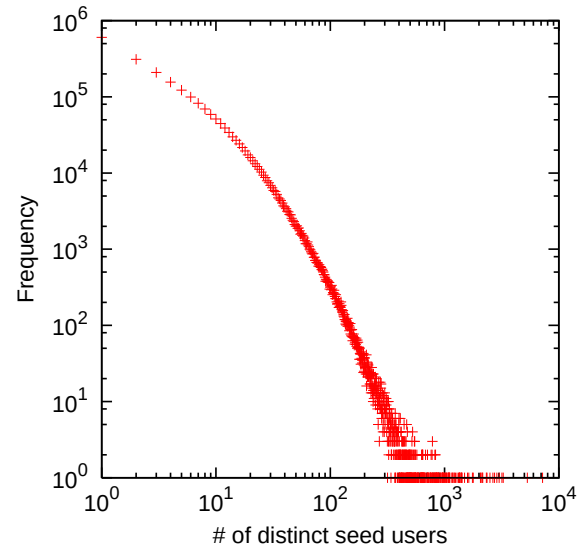


図 6 ユーザごとの異なりシードユーザ数の分布

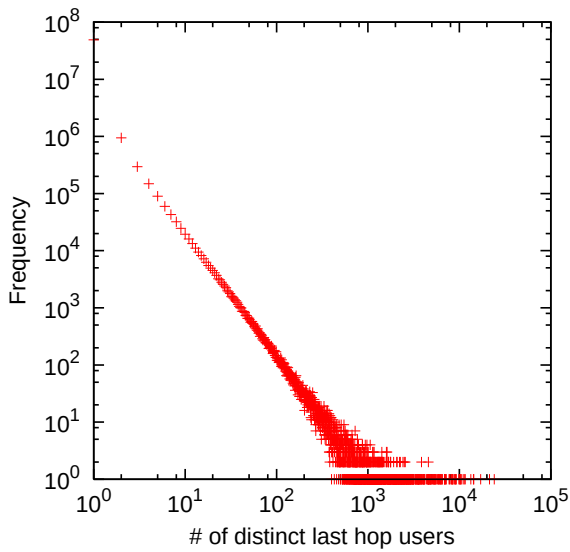


図 5 シードメッセージごとの異なりリーフユーザ数の分布

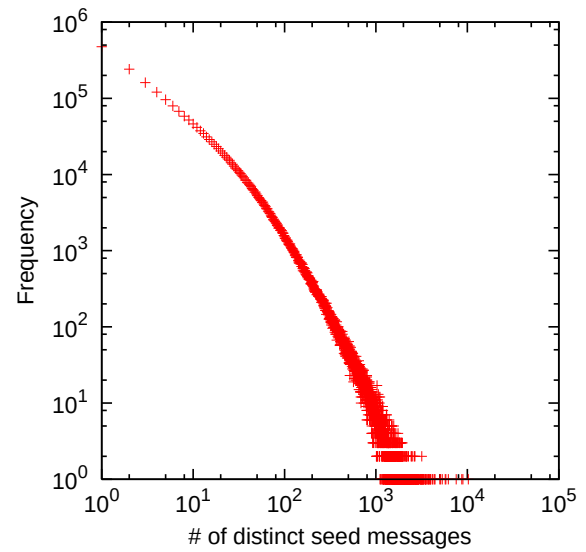


図 7 ユーザごとの異なりシードメッセージ数の分布

タグを含むツイートの比率は、Boyd らの調査結果と同程度となった。

次に、ユーザごとの分析結果を示す。分析対象のユーザを表 3 に示す。返信または RT を 1 回以上投稿したユーザは、全体の 84.4% だった。このうち、RT を 1 回以上投稿したユーザは 66.7% だった。これらの結果から、多くのユーザは、Boyd らの調査結果 [2] と同様、返信と RT を厳密に区別せず、広い意味で「他者への反応」として投稿していることがわかった。

URL を含むツイートを 1 回以上投稿したユーザは 67.4%、ハッシュタグを含むツイートを 1 回以上投稿したユーザは 39.4% だった。しかし、前述したように、URL やハッシュタグを含むツイートの件数は非常に少ない。さらに、URL を含み、かつ「@」を含むツイートは、URL を含むツイートの 35.0% (データ全体の 5.0%)、ハッシュタグを含み、かつ「@」を含むツイートは、ハッシュタグを含むツイートの 30.8% (データ全体の 2.0%) だった。このため、返信や RT に URL やハッシュタグが含まれることは多いが、自ら URL やハッシュタグ

表 4 異なりシードユーザ数が多いシードメッセージ (上位 10 件)

ユーザ数	メッセージ
23,769	おやすみー
23,910	おかえりー
20,972	おはよー
11,922	ありがとうございます！
16,565	おはようございます！
15,837	おやすみなさい！
13,016	おやすみ
13,945	おやすみなさい
7,580	大丈夫？
7,416	ありがとう！

を使用するユーザはあまり多くないと考えられる。

4.3 シードメッセージ、シードユーザ、リーフユーザ

収集したデータを用いて、レスポンスツリーにおけるシードメッセージ、シードユーザ、リーフユーザを抽出した。その結果、65,070,480 件のシードメッセージ候補が抽出され、文字列

表 5 異なりリーフユーザ数が多いシードメッセージ（上位 5 件）

ユーザ数	メッセージ
24,027	店頭での蒸しが足りずあと 15 分はかかると言われたスライム。「ご自宅で水少しかけてラップして、5 分位レンジで温めれば食べれますよ」と教えてもらって捕獲したのに…言われた通りにしたらメラゾーマ受けてスライムが死んだ。 http://t.co/AF5aST3q
21,342	1 1 月 2 9 日（火）午前 9 時 1 0 分 旅先でだいすけ君が急逝しました。あまりに突然のことで、我々番組スタッフも心の整理が出来ずにあります。だいすけ君を愛し、応援してくれた皆様、今まで本当にありがとうございました。だいすけ君のご冥福を心よりお祈り申し上げます。
17,070	ドコモ iPhone 参入を否定された日経について「東スポ並み」と言われ複雑。うちなら「ドコモ iPhone 獲得のため霊媒師使いジョブズ氏と会談」ですよ
14,364	身長－体重が 1 1 0 以上の人は公式リツイート（これはあんまりいないだろう）
14,249	【自転車乗る人へ】 明日から、自転車の取り締まりが厳罰化されます。信号無視、一時停止無視、傘差し運転、携帯メール、二人乗りは 2 回、飲酒運転は 1 回でアウト。違反したら赤切符が切られ、罰金 1 0 万円…！みなさんお気をつけて…

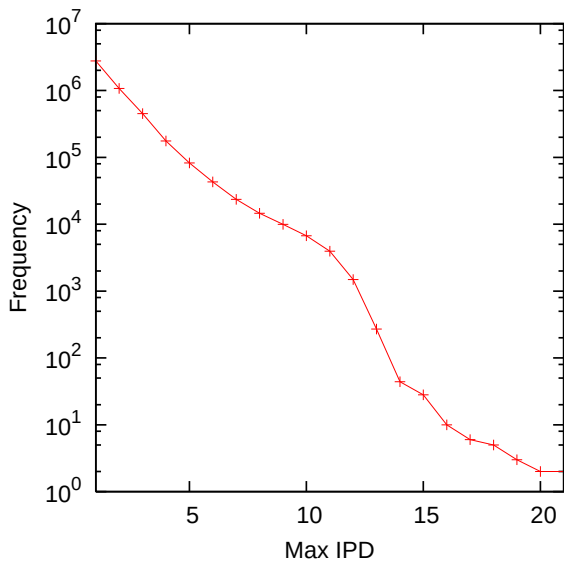


図 8 リーフユーザごとの最長 IPD の分布

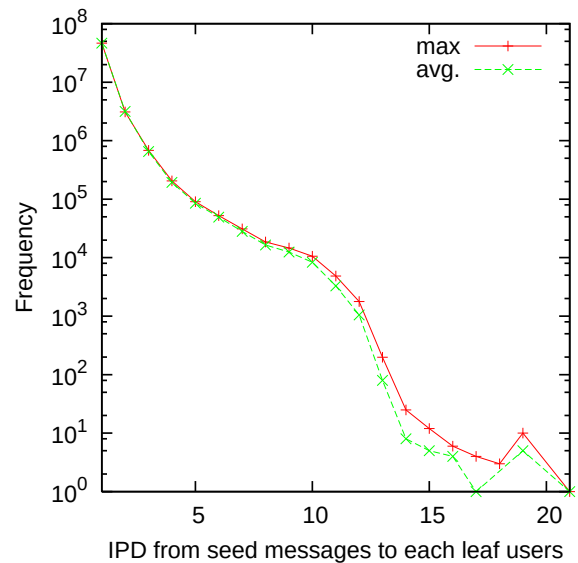


図 9 シードメッセージごとの最長 IPD および平均 IPD の分布

が完全一致するメッセージの重複を除いた 50,745,948 件をシードメッセージとした。シードメッセージごとの異なりシードユーザ数の分布を図 4 に、シードメッセージごとの異なりリーフユーザ数の分布を図 5 に示す。いずれもべき分布を示し、異なりシードユーザ数の最大は 48,655、異なりリーフユーザ数の最大は 24,027 だった。

ユーザごとの異なりシードユーザ数の分布を図 6 に、異なりシードメッセージ数の分布を図 7 に示す。異なりシードユーザ数の最大は 7,220、異なりシードメッセージ数の最大は 10,246 だった。

異なりシードユーザ数が多いシードメッセージの上位 10 件を表 4、異なりリーフユーザ数が多いシードメッセージの上位 5 件を表 5 に示す。「おやすみー」など、あいさつのメッセージにおいて、異なりシードユーザ数が多い、すなわち、多くのユーザがこのような内容を投稿し、それに対し他のユーザからのレスポンスがあることがわかる。また、時事的な話題において、異なりリーフユーザ数が多い、すなわち、このような内容の情報が多くのユーザに伝搬していることがわかる。

4.4 情報伝搬距離 (IPD)

リーフユーザごとの最長 IPD の分布を図 8 に示す。シード

メッセージごとの最長 IPD および平均 IPD の分布を図 9 に示す。IPD の最大は 21 だった。公式の書式で記述された RT のみを用いる Kwak らの RT ツリー [1] では、距離 4 をピークとする分布となったことが報告されている。これに対し、レスポンスツリーを用いて抽出したシードメッセージでは、最長 IPD および平均 IPD が 1 となるシードメッセージが最も多い。また、IPD が 10 程度となるシードメッセージも多い。本論文におけるレスポンスツリーでは、RT ツリーに対し、RT と返信の混在を許す拡張を行ったことで、RT ツリーにおいて 4 だったピークが、10 程度にまで延伸されたと考えられる。また、主にリーフユーザの返信からなる、IPD が非常に短いシードメッセージと、RT の連鎖や、RT に対する返信などを含む、IPD が中程度のシードメッセージとに二極化していることが示唆される。

4.5 k-means 法による探索的クラスタリング

レスポンスツリーによる情報伝搬距離 (IPD) のユーザプロファイリングにおける有用性を検討するため、本論文では、IPD に基づく指標を素性とした、k-means 法による探索的なクラスタリング [3] を試みる。

クラスタリングに用いる素性を表 6 に示す。シードメッセー

表 6 クラスタリングに用いる素性

素性名	説明
$rstw(n)$	リーフユーザの IPD が n となるレスポンスにおけるシードメッセージの重みの総和
$rstw_all$	リーフユーザがレスポンスしたシードメッセージの重みの総和
to_users	リーフユーザがレスポンスした異なりシードユーザ数
msg_ids	リーフユーザがレスポンスした異なりシードメッセージ数
rs_total	リーフユーザの延べレスポンス回数

表 7 クラスタリングに用いる素性セット

素性セット	用いる素性
A	$rstw(n)$
E	$rs_total, rstw_all$
D	to_users, msg_ids
B	$to_users, msg_ids, rs_total$
F	$to_users, msg_ids, rs_total, rstw(n)$
C	$to_users, msg_ids, rs_total, rstw_all$

ジには、多くのシードユーザによって投稿される一般的な内容から、少数のシードユーザのみによって投稿される特徴的な内容までがある。本論文では、情報検索において一般に知られる tfidf 法 [4] を援用し、シードメッセージの重み付けを行う。

シードメッセージ M_i の重み w_i は式 (1) を用いて算出する。

$$w_i = \log_2\left(\frac{rs_total_i}{rs_total_{max}} + 1\right) \times \frac{1}{\log_2(to_users_i + 1)} \quad (1)$$

ここで、 rs_total_{max} は、対象データ中のすべてのシードメッセージにおけるレスポンス回数の最大値、 rs_total_i はシードメッセージ M_i に対するレスポンス回数、 to_users_i はシードメッセージ M_i の異なりシードユーザ数とする。

素性 $rstw(n)$ は、リーフユーザがレスポンスしたシードメッセージの重みを、IPD ごとに総和した値である。ここで、 n は、対象データ中における IPD の最大値とする。素性 $rstw(n)$ は、 $rstw(1)$ から $rstw(n)$ までの n 個の値からなる。

クラスタリングに用いる素性セットを表 7 に示す。素性セット A は、IPD およびシードメッセージの重みを含むセットである。素性セット E および C は、IPD を含まず、シードメッセージの重みを含むセットである。素性セット D および B は、IPD およびシードメッセージの重みのいずれも含まないセットである。なお、素性セット F は、全素性を含むセットである。

6 組の素性セットを用い、クラスタ数を 8 および 27 に設定し、12 通りのクラスタリングを行った。クラスタ数は、クラスタリング結果を人間が探索的に解釈する際に、3 個の潜在変数を仮定し、これらの変数が 2 値をとる場合 ($2^3 = 8$)、および 3 値をとる場合 ($3^3 = 27$) について分析することを想定して設定した。それぞれ、各クラスタに属するリーフユーザ数の降順でクラスタ ID をソートし、分割結果を比較する。

まず、クラスタ数 8 におけるクラスタリング結果を図 10 に示す。素性セット A における最大クラスタ (Cluster1) と最小クラスタ (Cluster8) の差は 10^6 である。他の 5 組の素性セットでは、 10^5 程度であるのと比較し、クラスタのサイズの差が大きい。このことから、IPD に基づく指標のみからなる素性セット A は、他の素性よりも鮮鋭な分割ができることが示唆さ

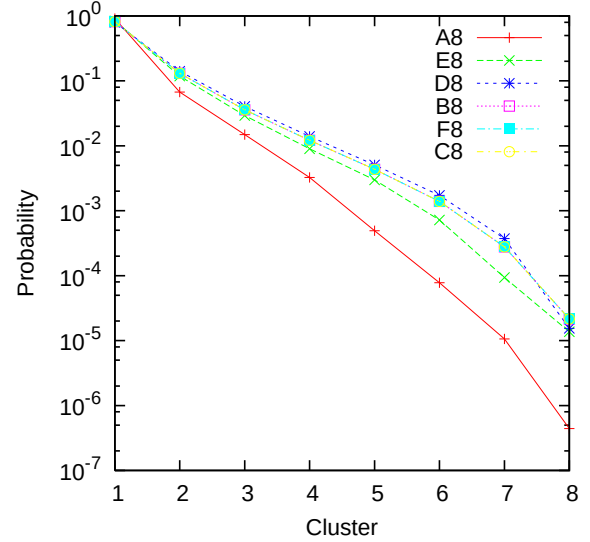


図 10 クラスタ数 8 におけるクラスタリング結果

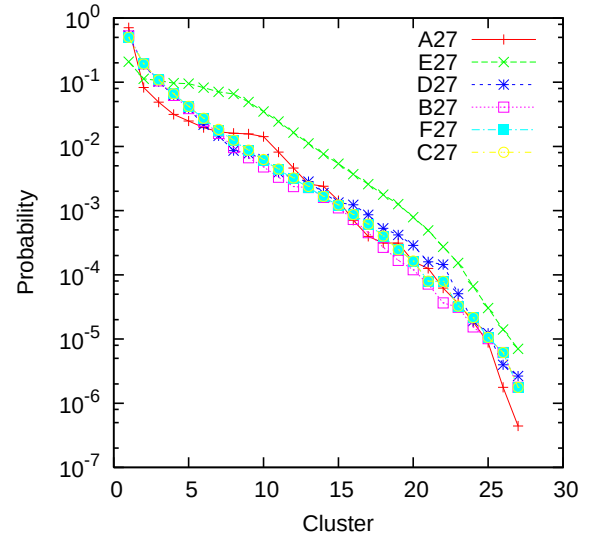


図 11 クラスタ数 27 におけるクラスタリング結果

れる。

次に、クラスタ数 27 におけるクラスタリング結果を図 11 に示す。最大クラスタ (Cluster1) と最小クラスタ (Cluster27) の差は、素性セット A は、素性セット E を除く他の素性セットと同程度である。このことから、IPD に基づく指標のみからなる素性セット A による鮮鋭な分割は、クラスタ数が少ないときのみ可能であることが示唆される。

なお、クラスタリングには R^(注4) (version 2.14.0) を用いた。

(注4) : <http://www.r-project.org/>

素性セット A を用いたクラスタ数 27 のクラスタリングの実行時間は、分析に用いた環境^(注5)において 226.0 秒だった。

5. おわりに

本論文では、まず、Twitter に投稿された日本語ツイートを網羅的に収集し、Twitter 上の返信、RT、ハッシュタグなどの機能が、ユーザにどのように使用されているかを明らかにした。その結果、ユーザ全体の 84.4% は、返信や RT をしたことがあり、Twitter を利用する上で他のユーザとの「つながり」を重視するユーザが大部分を占めていることがわかった。ただし、公式の書式での返信や RT よりも、非公式の書式での返信や RT を用いるユーザが多いことがわかった。また、公式、非公式を問わず「RT」と記述したツイートをしたことのあるユーザは、全体の 56.3%、さらに、ハッシュタグを使用したことのあるユーザは全体の 39.4% にとどまっており、厳密な使い分けをするユーザは少数であることがわかった。

次に、非公式な書式を含む返信および RT を用いた情報伝搬距離の算出を試み、ユーザプロファイリングにおける有用性を検討した。

今後の課題としては、より長期間のデータを用いての分析や、URL やハッシュタグを含むツイートと、それ以外のツイートにおける情報伝搬上の違いなど、詳細な分析に取り組む予定である。

謝 辞

本研究の一部は科研費 (21500091) の助成を受けたものである。ここに記して謝意を示す。

文 献

- [1] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a social network or a news media?, *Proceedings of the 19th international conference on World wide web*, WWW '10, New York, NY, USA, ACM, pp. 591–600 (2010).
- [2] Boyd, D., Golder, S. and Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, Washington, DC, USA, IEEE Computer Society, pp. 1–10 (2010).
- [3] 新納浩幸: R で学ぶクラスタ解析, オーム社 (2007).
- [4] 北 研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム, 共立出版 (2002).

(注5) : Xeon E5440 (2.80GHz) , CentOS 5.7 (x86_64)