3M-2

Twitterユーザのクラスタリング

上野　彰　　深海　悟
大阪工業大学大学院　情報科学研究科　　大阪工業大学　情報科学部

1　研究背景・研究目的

近年，SNS(Social Networking Service)のひとつとして Twitter が注目されている．Twitter は 140 文字以内の短文（ツイート）を投稿し，それを他のユーザが購読するサービスである．他のユーザの事をフォロワー(Follower)と呼ぶ．

Twitter のフォロワーとなるにはフォロー（follow）するだけで，相手の認証が不要な点が特徴である．フォロワーになる際には「友人」や「同じ趣味」といった何らかのキーワードに基づいてフォロワーとなる傾向がある．

しかし，Twitter にはユーザ情報の検索が存在せず，ユーザをフォローする際には Twitter の ID を知っておく必要がある．この理由のひとつとして，同一の属性の集合が作成されていない事が挙げられる．例えば，「同一の趣味の人を探す」等の場合，自分がフォローしているユーザがフォローする人物のプロフィールを手動で確認するしかない．そのため，漠然としたキーワード等からユーザを探す際には非常に煩雑な作業が必要であった．

そこで本研究では，Twitter における同一の属性（キーワード）を持つクラスタを作成，調査を行う．クラスタの作成には Twitter 内でユーザが任意に作成出来る「リスト」機能を用いる．クラスタを作成することで，Twitter におけるユーザに対するキーワード検索機能の実装が可能となる．

2　類似研究

Twitter を用いた研究は既に行われており，Akshay らは，ユーザがどのように Twitter を利用しているかを研究している[1]．また，風間らは，ツイートが災害時にどのように伝搬していくのかを分析したものがある[2]．これらの研究は Twitter がどのように利用されるかを調査したものであり，我々のように，Twitter ユーザのクラスタを作成し分析を行う研究は少ない．

3　提案・手法

Twitter のクラスタを作成するために，本提案では Twitter の機能にある「リスト機能」を用いる．リスト機能はユーザが任意で作成することができ，フォローの関係でないユーザのツイートをグループ化し，購読出来る．リストにはリスト名，リスト ID，作成者のユーザ ID と登録しているユーザの ID 等が記載されている．

本提案では作成されたリストに登録されているユーザが同じ属性であると推定することで，ひとつのリスト自体が小さなクラスタと考える．また，同一のリスト名が複数検出される可能性は高いが，それらも本提案では，ひとつのクラスタとして考える．

4　実験データの収集方法

本提案を検証するために，Twitter からリスト一覧を取得する．リストはユーザ ID からしか検索できないため，ユーザ ID から登録されているリストを抽出し，リストに記載されるユーザ ID から更にそのユーザが登録されているリストを抽出する．これを，再帰的に探索することで，リスト一覧を取得した．

探索した結果，2012 年 12 月 28 日から 2013 年 10 月 25 日の約 10 ヶ月の期間にてユーザ 2,758,581 人，全リスト件数は 1,058,422 件の情報を取得した．また，リスト情報から抽出したクラスタ数は 39,3852 個が取得でき，取得できたクラスタのうち，リストの数が多かったリスト名の上位 20 件を表 1 に示す．

表 1 ：出現数の多いクラスタ

クラスタ名	リスト数	クラスタ名	リスト数
news	31301	design	4288
music	12961	friends	3962
sports	10638	amigos	3873
celebs	6814	social-media	3553
my-favstar-fm-list	5478	ニュース	3508
noticias	5311	most-liked	3079
politics	5135	entertainment	3067
media	5101	business	3051
celebrities	4393	list	2822
tech	4380	famosos	2456

5　提案手法の調査

本章では 4 章で作成したクラスタから，大阪工業大学を示す oit クラスタを作成した．oit クラスタはリストの合計 34 個，ユーザ数は合計で

1279 人であった．作成した oit クラスタが大阪工業大学を示すものかの検証を行う．

作成した oit クラスタ内の大阪工業大学が関係するユーザ数を計測した．計測は自己紹介欄に oit という文字列や所属学部学科等の大阪工業大学を連想出来るキーワードが記載されていないかを手作業で分析した．その結果，1279 人のユーザ中，343 人存在し，全体の 26.82%であった．つまり，oit クラスタには更に複数のクラスタが存在する事が考えられる．そこで，抽出した oit クラスタの中から，更に大阪工業大学を指す oit クラスタを生成するために，リスト間の重複ユーザをデータとしてクラスタリングを行った．

最短距離法を用いたデンドログラムを図１に示す．デンドログラム作成にはリスト間の重複ユーザ数をデータとして利用している．ユーザ数は 2 つのリストの両方に同じユーザ ID があるかを 34 個のリスト全てに総当りで実施した．この時，計測されたユーザ数を全てのユーザ数で正規化した値をクラスタリング作成時のデータとして入力している．

結果より，図１の破線で囲った部分に oit ユーザが多く存在する可能性がある事がわかった．つまり，他のリストに同じユーザが多く存在すれば，それらのリストは同じ oit の意味を持つリストの可能性が高い．

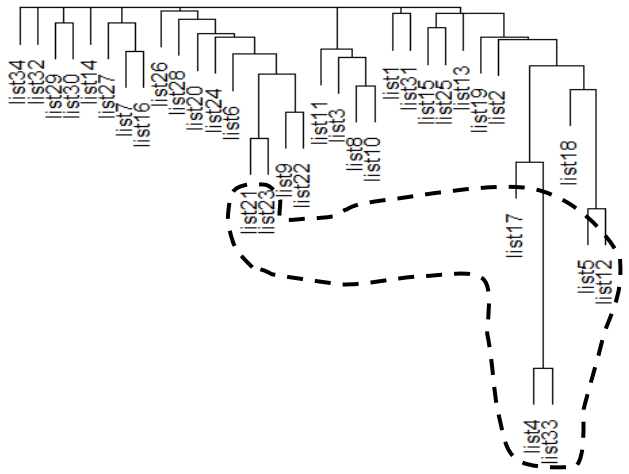

図１　oit クラスタのデンドログラム

６　検証

５章の調査結果をシンプソン係数を用いて検証する．シンプソン係数はひとつの集合からふたつのキーワードの集合を生成し，キーワードの関係の強さを定量的に計測する．値は 0-1 の範囲で表現され，値が大きいほど関係が強いとなる．以下に式を示す．

$$\mathrm{simpson} = (|X \cap Y|) / \min(|X|, |Y|)$$

シンプソン係数を本研究に適用する概念を図２に示す．図２の黒三角はユーザデータを示しており，ユーザデータにはユーザ ID と登録されているリストの名前と ID が保持されている．本検証ではこのユーザデータのリスト ID を用いてシンプソン係数を求める．つまり，34 個のリスト間の関係を求める．

結果，他の全てのリストとのシンプソン係数が 0 であるリストが 2 つ検出された．つまり，2 つのリストに所属するユーザは，他のリストと一切の関係の無いユーザである．これらは図１で示す list32 及び list34 であり，登録ユーザ数は合計で 288 人であった．

また，図１に示すクラスタリング結果とシンプソン係数を比べると，図１の点線で囲った部分の係数は高い事がわかった．つまり，oit クラスタの中を更に細かなクラスタに分ける事で，精度の高いクラスタリングができたと考えられる．

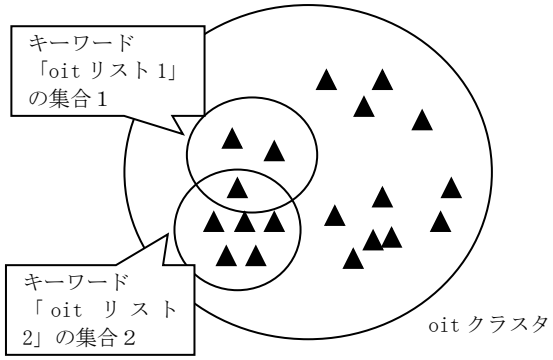

図２　oit クラスタの概念図

７　まとめ

本研究では，Twitter の検索機能を実装するために，Twitter 内のクラスタリングを実施した．クラスタリングには Twitter でユーザが任意に作成するリストを使った．約 10 ヶ月で収集したデータに本提案を適用した．その検証は oit リストを対象にし，シンプソン係数を用いる事でより精度の高いクラスタリングが実施できた．今後はより精度の向上を目指す．

参考文献

[1] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng, Why We Twitter: Understanding Microblogging Usage and Communities August 12, (2007)
[2]風間 洋一　Twitter における情報伝播
人工知能学会誌 Vol.27 No.1 pp.35-42(2012)