Vol.2018-MPS-120 No.16
2018/9/26

類似ツイートグラフ構築のための類似度閾値決定法

菅野 健一[1,a)]　伏見 卓恭[1,b)]

概要：Twitter は利用者が多く，ツイートは利用者の率直な本音でもある．ツイートには，アイテムやイベントなどに関するツイートも多数存在しており，アイテムに関する多種多様な意見を収集することができる．しかし，Twitter は文字で表現されているため，アイテムの全体像が把握しづらい．このことから、アイテムの長所・短所の評価について視覚的に見ることでより明確に全体像が把握しやすいと考えられる．本研究では，アイテム名を含むツイートを収集し，類似ツイートをつなげたグラフを構築する．そして，この類似ツイートグラフにおける密結合するサブグラフであるコミュニティから，ユーザニーズに関する表現を抽出する．また，類似ツイートグラフを構築する際に類似度の閾値を適切に設定する必要がある．閾値が低すぎれば，関連しないツイートどうしが結合し，異なる要望表現を持つツイートの連結成分ができてしまう．一方，閾値が高すぎれば，多少の表記ゆれなどにより，本来同じ意味を持つ要望表現のツイートを別の連結成分として分割してしまう．本研究では，各連結成分に表れる要望表現の出現確率に着目して，純度と凝集度を定義し，それらの値が最大値となる類似度を閾値とする方法を提案する．

Appropriate Threshold for Constructing Similar Tweet Graph

Kanno Kenichi[1,a)]　Fushimi Takayasu[1,b)]

1. はじめに

近年，ソーシャルネットワーキングサービス（SNS）の利用者が増え，多くの人が何かしらの SNS を利用している．その中でも Twitter は利用者がとても多く，複数の情報がつぶやき（ツイート）として飛び交う場所である．ツイートにはアイテムやイベントなどに関するものも多数存在しており，それらに関する多種多様な意見を収集することができる．また，若者などのユーザは，フォーラムスレッドなどにわざわざ言葉を選び書き込むより，Twitter でつぶやく方が手間がかからず気楽であり，率直な意見や感想をつぶやく傾向がある．率直な意見はユーザの真意であるため，本当に求めているニーズやアイテムやイベントの評価を抽出できると考えられる．そのため，本研究では口コミサイトなどのフォーラムスレッドではなく Twitter を対象とする．しかし，Twitter は 140 文字以内で表現しなければならないことに加え，ネットスラングや顔文字など様々な表現が用いられていることや，単に時系列順にツイートが表示されていることから，アイテムに関する評判の全体像が把握しづらい．そこで，アイテムに関する長所や短所などの評価をわかりやすく可視化することで，全体像を明確に把握しやすくなると考えられる．

本研究では，TwitterAPI を利用してアイテム名を含むツイートを収集し，類似ツイートをつなげたグラフを構築する．そして，この類似ツイートグラフにおける密結合するサブグラフであるコミュニティから，ユーザニーズに関する表現を抽出することを目的とする．本研究の特徴として，ツイートの文章群を単に表示するのではなく，ツイートをノードとしたグラフを可視化することで，アイテムやイベントに対する評価，ニーズ，改善要望，問題点を視覚的に俯瞰することができる．この可視化結果をもとに，アイテムやイベントの関係者が，改善策を企てるための参考になると考えらえる．

[1] 東京工科大学　コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
[a)] C0115113d1@edu.teu.ac.jp
[b)] fushimity@stf.teu.ac.jp

2. 関連研究

川島らの研究では，Twitter 上から要望を含むツイートの抽出に機械学習のアルゴリズムを適用することで，従来手法と比較してより高い精度での抽出を試みることを目的とし研究を行った [1]．半教師あり学習の手法の一つである「Distant Supervision」を用いて，半自動的に教師データの収集を行った．Distant Supervision を用いた教師データの収集では，予め教師データの判別の手がかりとなる表現を決定しておき，それらの表現的な特徴を含むデータを収集することで半自動的な教師データの収集を可能にした．要望を含む文には「～しろ」「～たい」「～ほしい」といった文末表現が出現することが知られており，合計 19 個の手がかり表現を定義し要望表現辞書とした．そこから，Support Vector Machine を用いて，要望ツイートを要望と not 要望に分けていた．しかし，実験結果の精度は実用レベルまでの向上は得られなく，正解データとなる要望を高い精度で獲得可能な値は異なっている可能性があり，学習データの際に複数のルールなど組み合わせなくてはいけないことや新たな手がかり表現の追加などが課題であった．要望ツイートの設定は詠嘆や命令系など決まった表現で抽出していることから，Twitter などの自由記述では困難であった．そのため，本研究では自由記述の Twitter で対応できるように，要望表現をあらかじめ絞らず，ニーズの対象となるアイテム名のみを設定してツイート群を収集する．また，率直な意見は形容詞で表現されていることが多い．そのことから，収集したツイートを n-gram に分け，名詞-形容詞と分割し類似しているツイートをつなげることで，グラフを構築する点でも異なる．

徳田らの研究では，Twitter で投稿された商品・サービスに対する率直な意見を，サービス提供者にわかりやすい形で届けることを目的とし研究を行った [2]．ネットゲームとソーシャルゲームを対象としツイートの収集を行ったが，一部だけ異なる重複ツイートが含まれてしまうことや，ジャンルを変えると精度が落ちるといった問題点が見られた．また，改善キーワードを予め作りツイート収集を行っていたため改善キーワード語句の追加が課題となっている．

越中らの研究では，自由記述の可視化を目的とし，テキストマイニングによる授業評価アンケートの分析に基づく研究を行った [3]．KH Coder を使い自由記述欄での出現頻出が高い語句を上位 30 個まで選び，出現パターンの似通った語をリンクで結んでネットワークを構築した．本研究では，連結成分のアノテーションをする際に出現頻度が高い語句を抽出するため類似性がある．

和多らの研究では，単語の出現頻度に着目した病院の評判情報の研究を行った [4]．Web 上で公開されている病院

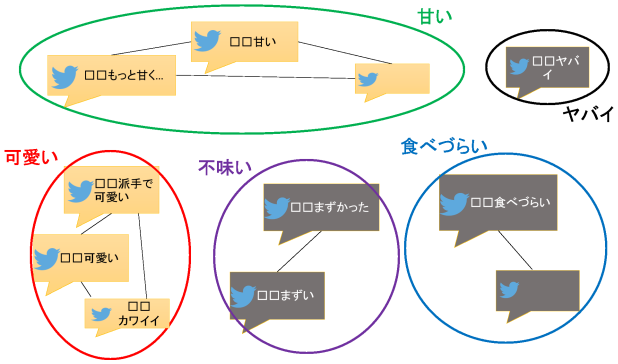

図 1　類似ツイートグラフと要望表現

を対象とし，評判情報の文章群を形態素解析を行い，単語を抽出する．n 個の 2 次元データに対し回帰直線を求め，夏目漱石の小説の文章群と比較して特徴的な名詞・動詞・形容詞・副詞を求めた．表現が自由な Twitter で形態素解析を行うと，品詞の種別が辞書に登録されていない可能性もある．そのため，本研究ではツイートを n-gram に分ける違いがある．また，回帰直線も使用せず比較の仕方も異なる．

3. 提案手法

3.1 類似ツイートグラフ

本研究では，アイテムに関するツイートを収集し，ツイートを文字 n-gram に分割する．Twitter では，ネットスラングや新語，くだけた表現などが多いため，形態素解析により得られる単語より，n-gram が適切であると判断した．分割した n-gram を素性としたベクトルにより各ツイートを表現する．そして，ベクトルのコサイン類似度が高いツイートをつなげることでグラフを構築する（図 1 参照）．提案手法の概要を以下に示す．

(1) アイテム名を含むツイートを収集；
(2) 各ツイートを文字 n-gram に分割；
(3) n-gram を素性としたベクトルを構築；
(4) ベクトル間のコサイン類似度を計算；
(5) 類似度が閾値以上のツイート間にリンクを付与；
(6) 構築したグラフを連結成分に分解；
(7) 各連結成分に有意に多く出現する要望表現によりアノテーションを付与；

あるアイテムに関するツイート集合を V とする．1 件のツイート $u \in V$ を n-gram の頻度ベクトル $\mathbf{x}_u$ で表現し，ツイート $u, v \in V$ 間の類似度をコサイン類似度 $\rho(u,v) = \frac{\mathbf{x}_u^T \mathbf{x}_v}{||\mathbf{x}_u|| ||\mathbf{x}_v||}$ で計算する．類似度が閾値 θ 以上のツイートペアにリンクを付与することで類似ツイートグラフを構築する．すなわち，ツイート集合をノード集合 V，類似ツイート間の関係をリンク集合 $E_\theta = \{(u,v) \in V \times V | \rho(u,v) \geq \theta\}$ としたグラフ $G_\theta = (V, E_\theta)$ を構築する．閾値 θ が大きいと，非常に

類似した字面のツイート間のみにリンクが付与される，一方，閾値を小さくすれば，あまり類似しないツイート間にもリンクが付与される．つぎに，グラフ G_θ を連結成分分解し，類似ツイートからなるサブグラフ群 $\{CC_1, \ldots, CC_K\}$ に分割する．各連結成分 CC_k に対して，要望表現 w の出現したツイートノード数を $c_{k,w}$ とすると，連結成分 CC_k に出現する全要望表現の個数は $a_k = \sum_{w \in W} c_{k,w}$ となる．ここで， W は全ツイートに出現する要望表現の種数を表す．同様に，類似ツイートグラフ全体において要望表現 w が出現した回数は $b_w = \sum_{k=1}^{K} c_{k,w}$ となる．これらより，全要望表現の出現回数は $M = \sum_{k=1}^{K} a_k = \sum_{w \in W} b_w$ であり，周辺分布 $p_k = a_k/M$ と $q_w = b_w/M$ を考えることができる．いま，要望表現 w が連結成分 CC_k にランダムに出現したと仮定すると，2 つの周辺分布から出現回数の期待値を $e_{k,w} = Mp_kq_w$ と計算できる．要望表現 w がランダムではなく，統計的有意に連結成分 CC_k に出現したことを定量化するために，実際の出現頻度 $c_{k,w}$ と期待値 $e_{k,w}$ から Z スコア $z_{k,w}$ を計算する．

$$z_{k,w} = \frac{c_{k,w} - e_{k,w}}{\sqrt{Mp_kq_w(1 - p_kq_w)}} \qquad (1)$$

Z スコアが正で大きな値を示せば，その連結成分に有意に多く出現したことを意味する．本研究では，有意に多く出現した要望表現を用いて各連結成分にアノテーションを付与する．

3.2 類似度閾値決定法

類似ツイートグラフを構築する際に，類似度の閾値を適切に設定する必要がある．閾値が低すぎれば，関連しないツイートどうしが結合し，異なる要望表現を持つツイートの連結成分ができてしまう．一方，閾値が高すぎれば，多少の表記ゆれなどにより，本来同じ意味を持つ要望表現のツイートを別の連結成分として分割してしまう．本研究では，各連結成分に現れる要望表現（形容詞など）の出現確率に着目して，純度（Degree of Purity）と凝集度（Degree of Cohesiveness）を定義し，これらの値が最大となる類似度を閾値とする方法を提案する．

純度は，各連結成分に出現する要望表現の混じり具合を定量化した指標である．前節と同様に，連結成分 CC_k に要望表現 w を含むツイートが出現する回数を $c_{k,w}$，連結成分 CC_k に含まれるすべての要望表現の数を $a_k = \sum_{w \in W} c_{k,w}$ としたとき，連結成分 CC_k の純度を以下のように計算する：

$$P_\theta(CC_k) = \sum_{w \in W} \frac{c_{k,w}}{a_k} \log \frac{c_{k,w}}{a_k}.$$

符号を逆転させたエントロピーを計算しているため，連結成分 CC_k に一部の要望表現だけが偏って存在する場合に値が高くなり，多くの要望表現が混ざっている場合に値が低くなる．そして，すべての連結成分に関して平均をとった値を P_θ とする．

凝集度は，各要望表現の出現する連結成分に関して集中具合を定量化した指標である．前節と同様に，類似ツイートグラフ全体において要望表現 w が出現した回数 $b_w = \sum_{k=1}^{K} c_{k,w}$ としたとき，要望表現 w の凝集度を以下のように計算する：

$$C_\theta(w) = \sum_{k=1}^{K} \frac{c_{k,w}}{b_w} \log \frac{c_{k,w}}{b_w}.$$

凝集度も符号を逆転させたエントロピーを計算しているため，要望表現 w が一部の連結成分のみに存在する場合に値が高くなり，多くの連結成分に点在する場合に値が低くなる．そして，すべての要望表現に関して平均をとった値を C_θ とする．

類似度閾値 θ を下げるにつれ複数の連結成分が合併するため，純度 P_θ の値は低くなる．一方，閾値 θ を下げるにつれ要望表現が少数の連結成分に集まるため，凝集度 C_θ の値は高くなる．両指標がバランスよく高い値を示す類似度を閾値として設定する．いずれの指標の計算においても，連結成分ごとの要望表現の出現回数のみを保持しておき，類似度を下げ連結成分が合併するたびに関連する連結成分のみ出現回数を更新するだけですむため，効率的に計算できる．

4. 評価実験

4.1 データセット

今回実験に使ったデータは 2018 年 6 月に取ったデータを使用している．収集するアイテムの決め方は，話題性があるものや新作・新発売されたモノを中心に積極的に収集した．また，オンラインゲームや携帯ゲームといったモノはユーザのニーズが多いと判断したため，そのようなゲームのツイートも収集した．

収集する際は，キーワードはアイテム名に設定し，10000 件のツイート収集を行っている．しかし，Twitter 上では様々な言い方で表現されていることもあり，アイテム名だけでは 10000 件のツイートが取れない場合は，アイテムの他の言い回しも検索する対象として設定し収集を図った．

今回検索する対象に使ったアイテムは，2018 年 6 月 22 日（金）に発売された「マリオテニスエース」と 2018 年 12 月 7 日（金）発売予定の「大乱闘スマッシュブラザーズ SPECIAL」である．マリオテニスエースは 6 月 26 日にデータを収集し，収集に使用したキーワードは「マリオテニス」の 1 つであり 10000 件のツイートが収集できた．大乱闘スマッシュブラザーズ SPECIAL は 6 月 14 日にデータを収集し，収集に使用したキーワードは「スマブラ」の 1 つであり 10000 件のツイートが収集できた．また，可視化は Cyto Scape を使い可視化を行っている．

5. 実験結果

5.1 「マリオテニス」の実験結果

図 2 は，横軸に類似度閾値，縦軸に純度と凝集度をプロットしたグラフである．赤い線は純度を表し，青い線は凝集度を表している．図 2 を見ると類似度閾値を下げるにつれて純度は減少傾向にあり，凝集度は上昇傾向にあることが確認できる．図 3 は，純度と凝集度を足したものである．図 3 をみると $\theta = 0.55$ 付近で最大値をとることが分かる．したがって，マリオテニスに対しては，類似度閾値 $\theta = 0.55$ が適切であると判断する．

図 4 は，類似度閾値 $\theta = 0.55$ での類似ツイートグラフにおいて，所属ツイートノード数が多い連結成分から順に表示している．Zscore が高い形容詞を各連結成分にアノテーションとして付与した．具体的には，連結成分 1 に出現する形容詞の Zscore は「面白」が 3.27,「楽しい」が 2.76,「おもろい」が 1.98 となっている．連結成分 2 では「熱い」が 4.57 となっている．連結成分 3 では「難しい」が 8.96,「難しく」が 8.41,「難しかっ」が 8.41 となっている．このように，各連結成分に統計的有意に多く存在する形容詞を抽出できており，マリオテニスに関する様々な意見を集約できた．ちなみに類似度閾値を高く設定すると，同じ意味でも複数の連結成分に点在してしまう形容詞が多かった．また，類似度閾値を低く設定すると，1 つの連結成分に雑多な形容詞が集まり各形容詞の Zscore が低くなるため，有意に多く存在する形容詞を抽出できなかった．

5.2 「スマブラ」の実験結果

図 5 は，マリオテニスと同様に，類似度閾値を下げるにつれて純度は減少傾向にあり，凝集度は上昇傾向にあることが確認できる．図 6 は $\theta = 0.40$ 付近で最大値をとることが分かる．したがって，マリオテニスに対しては，類似度閾値 $\theta = 0.40$ が適切であると判断する．

図 7 は，類似度閾値 $\theta = 0.40$ での類似ツイートグラフにおいて，所属ツイートノード数が多い連結成分から順に表示している．Zscore が高い形容詞を各連結成分にアノテーションとして付与した．具体的には，連結成分 1 に出現する形容詞の Zscore は「欲しい」が 2.36,「早く」が 1.53,「楽し」が 0.79 となっている．連結成分 2 では「やば」が 14.23 となっている．連結成分 3 では「熱い」が 7.01 となっていいる．連結成分 4 では.「喜ばしい」が 13.47 となっている．このように，各連結成分に統計的有意に多く存在する形容詞を抽出できており，スマブラに関する様々な意見を集約できた．

6. おわりに

本研究では，アイテムの要望を抽出するために，ユーザの率直な意見が含まれるツイートに着目し，類似ツイートグラフを構築する．この時，類似度の閾値を適切に設定しないと，統計的優位に多く存在する要望表現（形容詞）を抽出することが困難となる．そこで，要望表現の分布に関する純度と凝集度に基づぎ，適切な類似度閾値を決定する手法を提案した．実ツイートを用いた評価実験では，マリオテニスの類似ツイートグラフ構築における類似度閾値は $\theta \simeq 0.55$，スマブラでは $\theta \simeq 0.40$ が望ましいことが分かった．これらの類似ツイートグラフから，アイテムに関する要望を俯瞰することができた．今後の課題として，2-gram 以外での有効性の確認をしていきたい．また，今回の実験に使用したアイテムはどちらもゲームのため，他のジャンルにおいても調べていきたい．

謝辞 本研究は，JSPS 科研費（No.16K16154）の助成を受けたものである．

参考文献

[1] 川島崇秀，佐藤哲司，神門典子 : Twitter からの消費者ニーズの抽出手法に関する提案，DEIM Forum 2016 B5-1.
[2] 徳田勇介 : Twitter からの商品レビュー自動抽出手法の提案 甲南大学 修士論文
[3] 越中康治，高田淑子，木下英俊，安藤明伸，高橋潔，田幡憲一，岡正明，石澤公明 : テキストマイニングによる授業評価アンケートの分析 : 共起ネットワークによる自由記述の可視化の試み，宮城教育大学情報処理センター研究紀要 : COMMUE 22 号 67-74 ページ 2015-03-31
[4] 和多太樹，関隆宏，田中省作，廣川佐千男 : 単語の出現頻度に着目した病院評判情報の分析，社団法人 情報処理学会 研究報告 2005-5-26

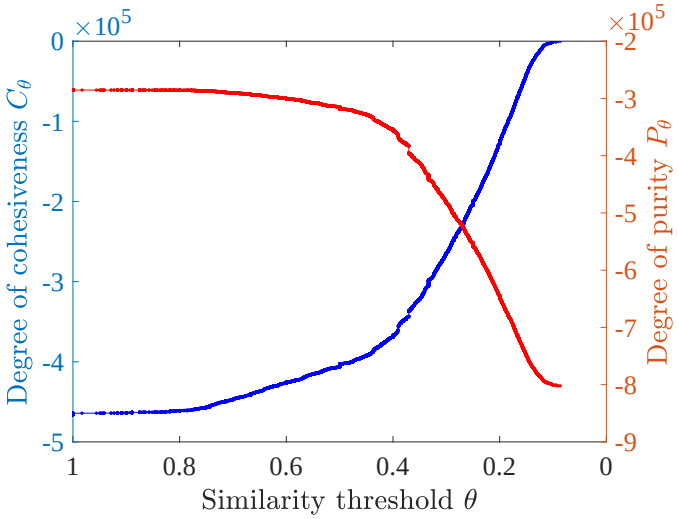

図 2　純度と凝集度の推移

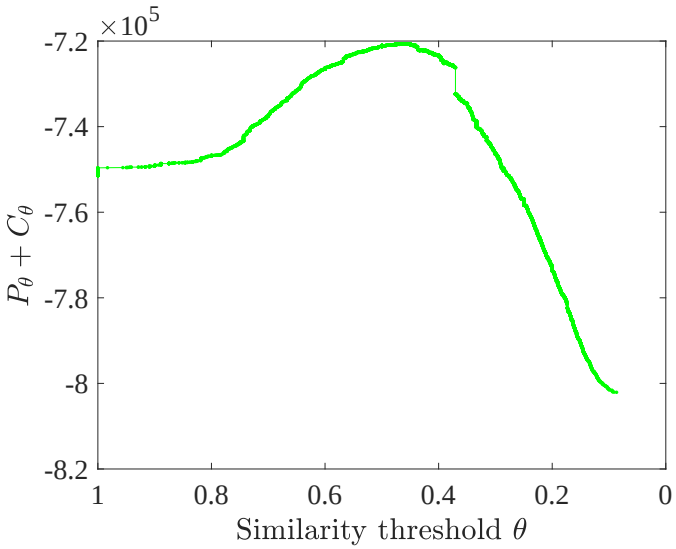

図 3　純度と凝集度の和

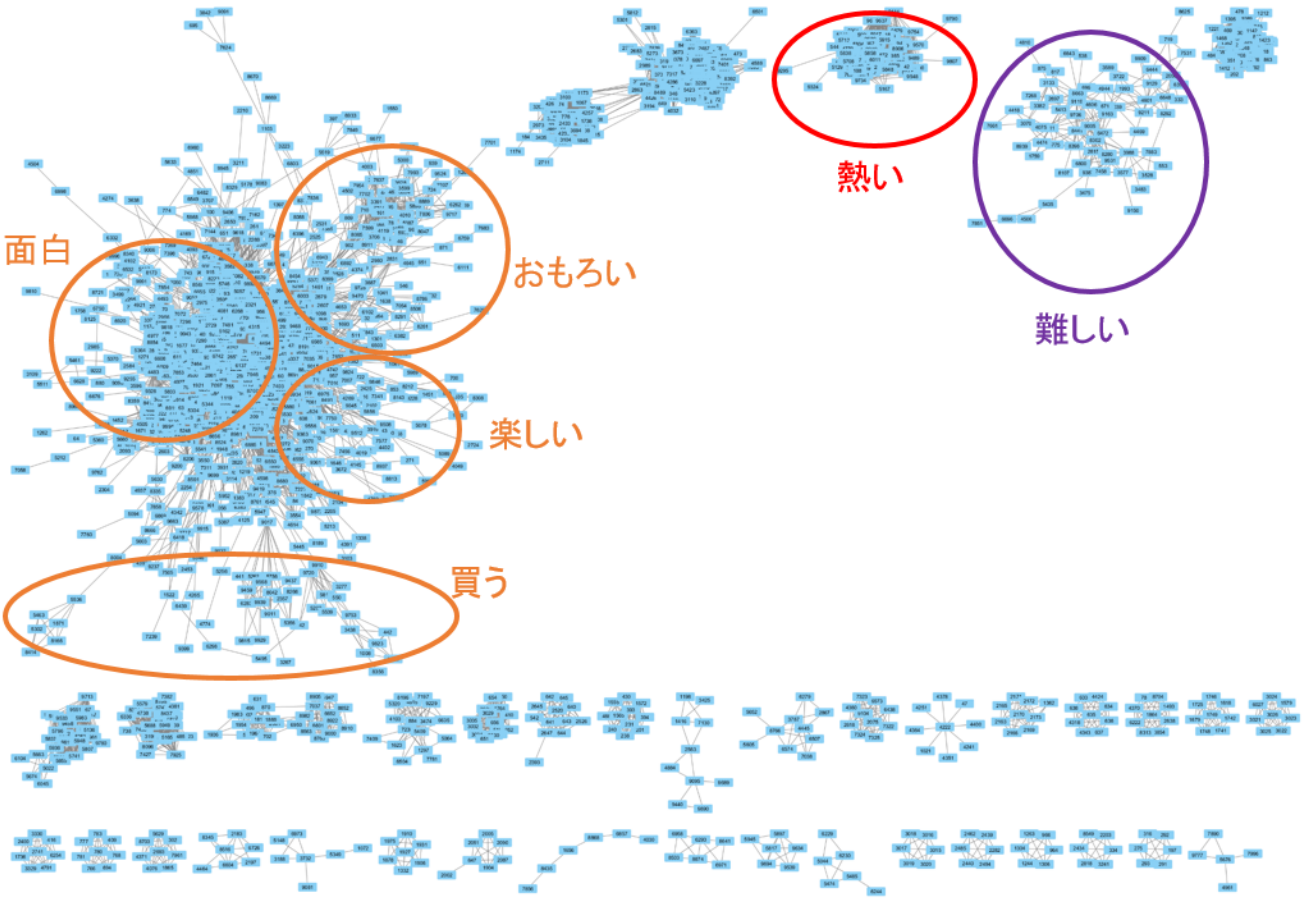

図 4　「マリオテニス」の類似ツイートグラフ $\theta = 0.55$

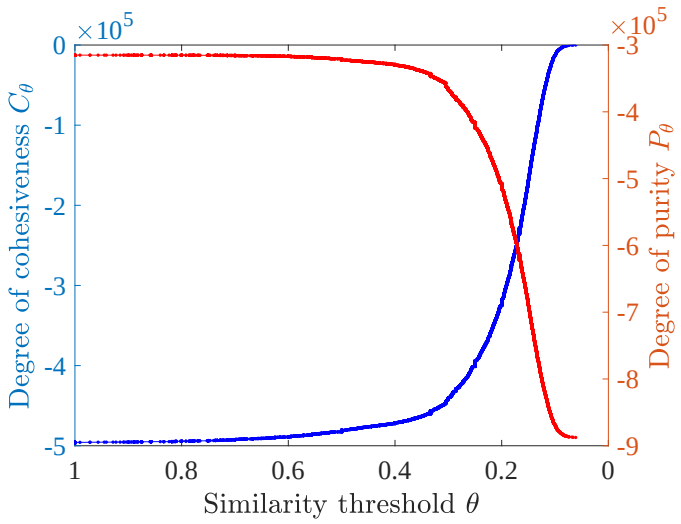

図 5　純度と凝集度の推移

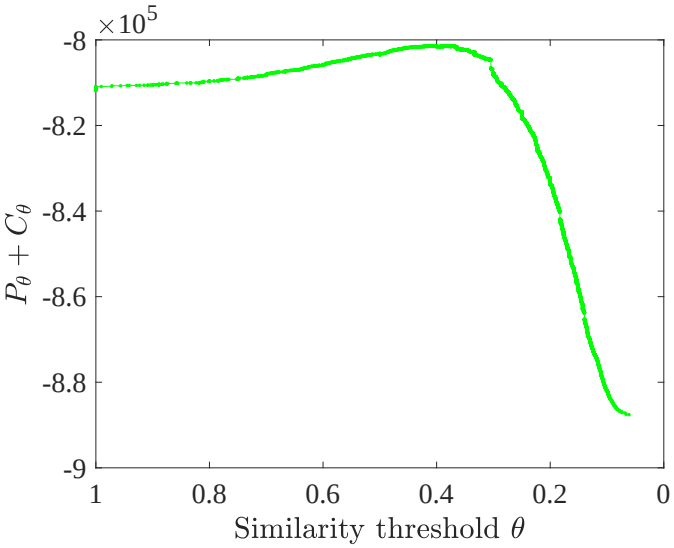

図 6　純度と凝集度の和

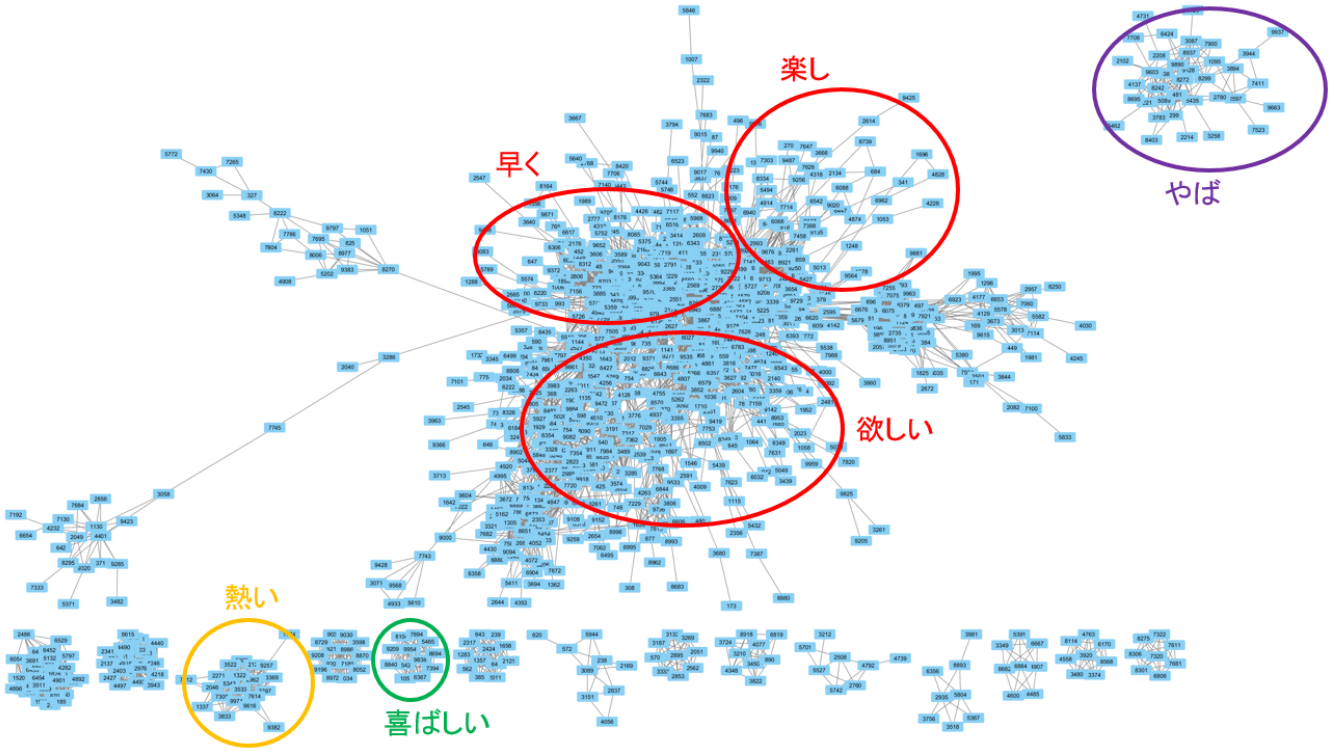

図 7　「スマブラ」の類似ツイートグラフ $\theta = 0.40$