D-024

類似ツイートグラフに基づくユーザニーズの可視化手法
A Visualization Method of User Demands Based on Similar Tweets Graph

菅野 健一 †
Kenichi Kanno
伏見 卓恭 †
Takayasu Fushimi

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 を使っていることが分かる．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ことが分かる．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．

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

†東京工科大学コンピュータサイエンス学部

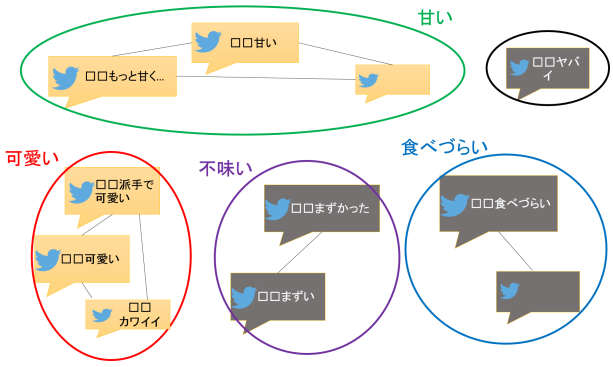


図 1: 類似ツイートグラフと要望表現

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．

3. 提案手法

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 $\{C_1, \ldots, C_K\}$ に分割する．各連結成分 C_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 C_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 C_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 $e_{k,w} = M p_k q_w$ 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 C_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．

$$z_{k,w} = \frac{c_{k,w} - e_{k,w}}{\sqrt{M p_k q_w (1 - p_k q_w)}} \quad (1)$$

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．

4. 評価実験

提案手法により抽出，アノテートした要望表現が適切なものかどうか，実例により定性的に評価する．本稿では，n-gram の $n = 2$ とし，すなわち，2-gram により各ツイートを表現し，要望表現として形容詞全般を用いる．形容詞には，ユーザの率直な感想が含まれており，要望に直結する場合が多いと感じたからである．また，先行研究 [1] のようにあらかじめ要望表現を限定しないため，多様な要望を抽出できると考えられる．

4.1. 実験データ

実験データは，2018 年 4 月から検索を始めた．一ヶ月ごとに新しいデータを取り直し，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 11 日 (月) に発売された「コカ・コーラクリア」と 2018 年 6 月 22 日 (金) に発売された「マリオテニスエース」である．コカ・コーラクリアは 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「コカ・コーラクリア」「コカコーラ・クリア」「コカコーラクリア」の 3 つであり 7444 件ツイートが収集できた．マリオテニスエースも同じ 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．

4.2. 実験結果

図 2 は，類似度閾値 $\theta = 0.9, 0.8, 0.7$ 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．表 1 に，図 2 の各連結成分に対するアノテーションワードを示す．アノテーションワードを抽出できなかった連結成分は省略している．

6 月 26 日に収集したコカコーラクリアの結果では，類似度閾値 $\theta = 0.9$ での類似ツイートグラフでは,「まずい」「くそまずい」と二文字

程度の違いでも別の連結成分となってしまった．また、検索キーワードが「コカ・コーラクリア」「コカコーラ・クリア」「コカコーラクリア」と 3 つで検索したことから,「コカ・コーラクリアうまい」「コカコーラクリアうまい」といった「・」の位置が違うだけで同じ意味のものも別の連結成分となってしまい，細かすぎる結果となってしまった．表 1(a) の類似度閾値 $\theta = 0.9$ におけるアノテーションワードを見ると，C_2 の「まずい」は「コカ・コーラクリア」と一緒に出現する．一方，C_4 の「まずい」は「コカコーラクリア」と一緒に出現する．すなわち，閾値が高すぎるため，同じ意味を持つ要望表現であるにもかかわらず,「・」の位置の違いにより別連結成分となってしまった．

$\theta = 0.8$ での類似ツイートグラフでは，1 つの連結成分のサイズが $\theta = 0.9$ のときより大きくなったものの,「コカコーラクリアまっず」や「コカ・コーラクリアまっず」のように 2 つの連結成分に分かれてしまった．$\theta = 0.7$ での類似ツイートグラフでは，最大連結成分はサイズが大きくなり複数の要望表現を含む結果となった．しかし，コカ・コーラクリアの「・」の場所が違う場合でも同じ連結成分に含まれるようになった．以上のことから，コカ・コーラクリアでは $0.7 < \theta < 0.8$ が適切であると考えられる．

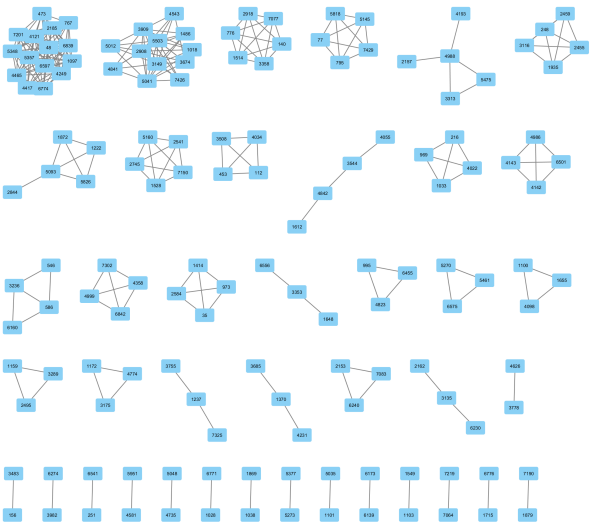
5. おわりに

本研究では，コカ・コーラクリアの類似ツイートグラフ構築における類似度閾値は $\theta \simeq 0.8$ が望ましいことが分かった．しかし，類似の要望表現を有する連結成分が存在するため，今後は適正な閾値を自動で計算する手法の提案に着手していきたい．

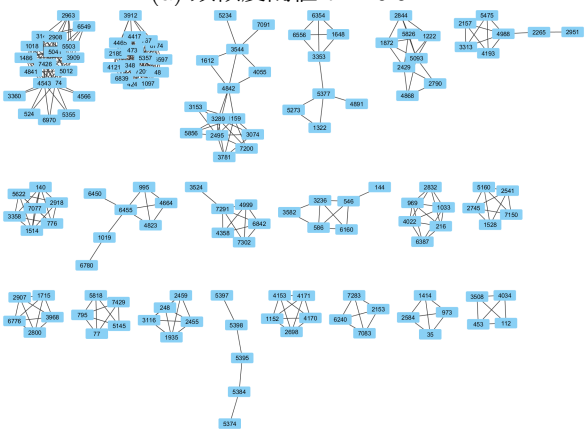
謝辞 本研究は，JSPS 科研費 (No.16K16154) の助成を受けたものである．

参考文献

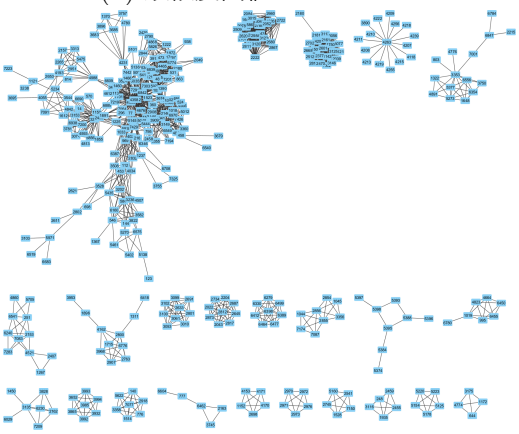
[1] 川島崇秀，佐藤哲司，神門典子：Twitter からの消費者ニーズの抽出手法に関する提案，DEIM Forum 2016 B5-1.



(a) 類似度閾値 $\theta = 0.9$



(b) 類似度閾値 $\theta = 0.8$



(c) 類似度閾値 $\theta = 0.7$

図 2: 「コカコーラクリア」に対する類似ツイートグラフ

表 1: 「コカコーラクリア」に対するアノテーションワード

(a) 類似度閾値 $\theta = 0.9$

C_k	w	$z_{k,w}$	$c_{k,w}$
C_2	まずい	3.9	10
C_4	まずい	2.7	5
C_7	早く	6.6	5
C_8	うまい	4.2	5
C_9	ない	4.4	4
	黒い	4.4	4
C_{13}	うま	3.4	1
	うまー	3.4	1
	うまい	1.4	2
C_{16}	うまい	3.3	3
C_{17}	なつかし	6.9	3
C_{19}	まずい	1.7	2
C_{21}	不味い	6.9	3
C_{22}	美味しく	7.0	2
C_{24}	美味しい	6.9	3

(b) 類似度閾値 $\theta = 0.8$

C_k	w	$z_{k,w}$	$c_{k,w}$
C_1	まずい	5.2	12
	まずっ	2.1	1
	まずぃ	2.1	1
C_3	甘	3.2	1
	うま	3.2	1
	うまー	3.2	1
	うまい	2.4	4
C_5	美味しく	6.3	8
C_7	うまく	4.6	2
	うまい	3.3	5
C_8	うまい	5.1	7
C_{11}	甘い	9.2	1
C_{12}	まずい	3.8	5
C_{15}	早く	8.8	5
C_{16}	美味しく	4.9	5
C_{18}	ない	6.0	4
	黒い	6.0	4
C_{25}	なつかし	9.0	3
C_{26}	まずい	2.4	2
C_{28}	不味い	9.0	3
C_{30}	美味しい	6.1	3
C_{31}	美味しく	9.0	3
C_{32}	美味しい	6.1	3
C_{34}	美味しかっ	9.0	3

(c) 類似度閾値 $\theta = 0.7$

C_k	w	$z_{k,w}$	$c_{k,w}$
C_1	まずい	4.4	23
	うまい	2.5	14
	甘	1.4	2
	うま	1.4	2
	不味い	1.1	8
	なう	1.0	1
	薄い	1.0	1
	まずっ	1.0	1
	わろ	1.0	1
	ひどい	1.0	1
C_3	詳しく	12.1	1
C_5	美味しく	5.9	11
	おいしく	5.7	4
C_7	美味しく	7.4	11
C_{14}	うまく	6.2	2
	うまい	4.4	5
C_{15}	美味しい	9.7	8
C_{20}	早く	11.8	5
C_{21}	ない	7.3	5
C_{23}	黒い	8.2	4
	ない	4.2	4
C_{24}	不味かっ	5.9	1
	不味い	4.3	3
C_{26}	まずい	2.9	2
C_{29}	美味しい	11.9	4
C_{30}	甘い	8.5	1
C_{32}	不味	12.0	2
C_{35}	なつかし	12.0	3
C_{37}	不味い	5.1	3
C_{40}	美味しい	5.9	3
C_{41}	美味い	12.0	3
C_{43}	美味しかっ	12.0	3
C_{45}	ない	5.6	3