Twitterユーザのリツイート情報を用いたトピックの可視化

清水　綾女†　　若林　　啓†　　佐藤　哲司†

† 筑波大学　〒305–8550 茨城県つくば市春日 1-2

E-mail: †s1711519@s.tsukuba.ac.jp, ††kwkaba@slis.tsukuba.ac.jp, †††satoh@ce.slis.tsukuba.ac.jp

あらまし　協調フィルタリングによるユーザの飽きは注目されつつある情報推薦の課題である．論文の推薦などでは確実な推薦が求められる一方，コミュニティの開拓が求められるユーザ推薦ではセレンディピティのある推薦も望まれる．本研究では，ユーザが新しい嗜好を発見できるような推薦情報を提示するシステムを提案する．まず Twitter 上のツイートをトピックモデルによって分類し，ユーザの嗜好を定義する．さらにユーザのリツイートによって嗜好同士が関連付けられると定義し，これらの情報から嗜好をノード，ユーザのリツイートをエッジ，リツイートしたユーザの数を重みとしたグラフを構成することでツイッター上で表現される嗜好の構造を可視化した．

キーワード　Twitter，トピックモデル、可視化

1　はじめに

近年，Web 上で扱われる情報は膨大な数に膨れ上がっており，情報推薦システムの重要性が高まっている．情報を推薦したいユーザと類似の嗜好をもつユーザ群を探し，そのユーザ群が好む情報を推薦する協調フィルタリング [1], [3] は広く使われる情報推薦の手法であり，商品の推薦やユーザの推薦などで用いられている．協調フィルタリングはユーザの好みに対して高い精度を持つ一方，類似の情報ばかりを推薦してしまい，ユーザの飽きを招くという問題が挙げられている [2]．そのため，ユーザにとって魅力的かつ意外性がある出会いを表す概念であるセレンディピティ(serendipity) に着目した情報推薦を試みる動きがある [4], [10]．セレンディピティのある情報推薦のためにはユーザにとって既知の嗜好と未知の嗜好を判別し，未知の嗜好の中でユーザにとって興味深いものを見つける必要がある．そこで本研究では，まず嗜好間の関係を明らかにすることがセレンディピティのある情報推薦において有用であると考え，Twitter のツイートに現れるユーザの嗜好をトピックモデルによって定義し，嗜好をノード，ユーザのリツイートをエッジ，リツイートしたユーザの数を重みとした有向グラフを構成し，その構造を可視化した．

2　先行研究

セレンディピティのある推薦システムとして，これまで多くのアプローチがなされている．徐らはユーザのリツイートのみに現れる嗜好を惹かれた興味として定義し，惹かれた興味から算出した類似度の高いユーザがフォローしているユーザを推薦するシステムを提案した [5]．この手法はユーザのリツイートに出現する惹かれた興味に基づいて推薦するため，ユーザにある程度既知の嗜好を推薦するという点で本研究の目指すシステムと異なる．また，福島らはユーザの行動履歴から興味の偏りを解析し，ユーザの隠れた嗜好を推定した [4]．この手法は行動履歴から作成したプロファイルと記事データに付与されているタグの共起グラフからユーザープロファイルには表れていない嗜好を抽出し，抽出された嗜好と記事データの類似度から推薦リストを作成する．福島らの手法では記事データに付与されたタグの共起グラフをもとに推薦を行っているのに対し，本研究の目指すシステムでは一つのツイートが一つの嗜好を表現すると仮定し，多くのユーザが共通してリツイートする嗜好は関連度が高いというユーザによる嗜好の関連に着目している点でアプローチが異なっている．また，本手法では Twitter 上で表現される嗜好全体の可視化を行うため，興味関心の高いことが予想されるユーザの嗜好に近い嗜好だけでなくユーザにとって遠い嗜好の発見も可能となる．このことによってユーザが通常の情報探索行動では到達しがたい情報の推薦も可能となる．

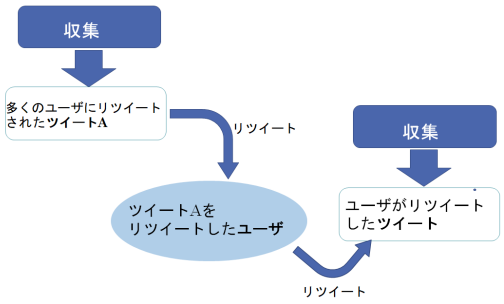

図 1　ツイート収集図

また，ツイートを用いてトピックモデルの学習を行った折本らの研究 [6] があり，トピック分布によってユーザの興味の構造を推察している．本研究は全ツイートを対象としている点とユーザ全体の興味の構造を扱っている点で折本ら [6] と異なる．

3　手　　法

3.1　ツイート収集

日常的な動作の報告など特定の嗜好を表現しない個人的なツイートを調査から排除するため，1000 回以上リツイートされたツイートのみを調査対象として収集 (図 1）した．まず，Twitter

表 1　ツイート検索に用いた 13 語

おすすめ	かっこいい	かわいい	飯	動物
経済	国	雇用	男女	学
政治	スポーツ	スイーツ		

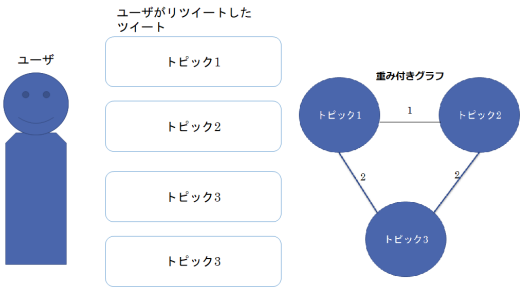

図 2　重み付きトピック関連無向グラフ

の API を用いて表 1 の単語が含まれる 1,000 回以上リツイートされたツイートを収集した．

以下，1,000 回以上リツイートされたツイート群をツイート群 A とする．さらに，ツイート群 A の各ツイートについてリツイートしたユーザ 100 人を収集する．収集したユーザ各々がリツイートしたツイートのうち 1,000 回以上リツイートされたツイートが存在すればリツイートしたユーザと紐づけてツイート群 B として保存する．ツイート群 B の各ツイートについても同様にリツイートしたユーザ 100 人を収集し，収集したユーザ各々がリツイートしたツイートをツイート群 C として保存した．よって，表 1 の語で検索したツイート群 A, ツイート群 A をリツイートしたユーザから得たツイート群 B，ツイート群 B をリツイートしたユーザから得たツイート群 C の 3 つの群をデータとして用いる．

3.2　ツイート分類

次に，収集した全ツイートを Latent Dirichlet Allocation(LDA) の一種であるトピックモデルを用いてトピック分類する．LDA トピックモデルはテキストなどの離散的な情報を扱う確率的生成モデルであり，LDA で文書は各トピックの確率分布モデルにより表される [7]．LDA トピックモデルではトピックの分類数を設定する必要がある．各ツイートはトピックモデルで最も高い確率を示すトピックに属すると仮定し，収集した各ツイートはトピックを表現するものとして扱う．

3.3　グラフ作成

ユーザ毎に，トピックをノード，ユーザ内のトピックの共起をエッジで表現し，共起回数をエッジの重みとしたグラフを図 2 に示す．この図は重み付き無向グラフであり，ユーザが複数のトピックをリツイートすることで表されるトピック同士の繋がりを表す．

次に，ユーザごとに作成された重み付きトピック関連無向グラフを足し合わせ，全トピックの繋がりを求める．ソーシャルネットワークなどの複雑な構造を可視化したときの可読性については様々な議論がなされている [8], [9]．本稿ではトピックの共起を表すエッジを足し合わせ，エッジの本数によって重みづけをした．トピック i，j のエッジの重みをトピック i の出現回数で割った値 C_{ij} を共起割合として定義し，共起割合 $C_{ij} \geqq \theta$ となるような i，j の間に有向エッジを引いた有向グラフを図 3 のように示す．$\theta = 0.2$ のときの引かれたエッジ・エッジで繋がれたノードの数を「エッジ数 (0.2)」「ノード数 (0.2)」のように表した．括弧の中身は θ を示す．採用した θ におけるデータは表 2 内で太字表記した．

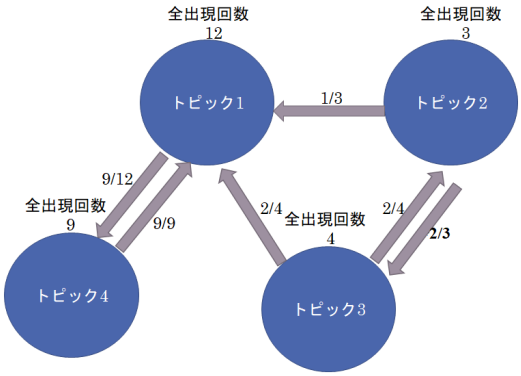

図 3　重み付きトピック関連無向グラフ

図 3 から，トピック 2・トピック 3 はトピック 1 に対して従属性があるがトピック 1 はトピック 2・トピック 3 に対してさほど従属性がないため，トピック 2・トピック 3 を嗜好するユーザの多くはトピック 1 も嗜好するが，トピック 1 を嗜好するユーザが必ずしもトピック 2・トピック 3 を嗜好するとは言えないことが読み取れる．また，トピック 1 とトピック 4 は互いに従属性を持ち合うため，両者を同時に嗜好するユーザが多いことがわかる．

4　結　　果

4.1　デ　ー　タ

ツイートデータは Twitter の API を用いて 2019 年 8 月 16 日から 2019 年 9 月 3 日に収集し，26,031 ユーザがリツイートしたツイート 1,029,492 件 (重複含む) を得た．ユーザのリツイート 1,029,492 件からリツイートによる商品・クーポン等のプレゼントキャンペーンのツイート 86,072 件を削除し，リツイート 943,420 件 (重複含む)，ツイート 89,023 件を分析に使用した．

4.2　ツイート分類

全ツイートを 10，20，40，60，80 トピックに分類した結果が表 2 である．

共起の偏りがないトピックは他のトピックとエッジが引かれずグラフに反映されないため，トピック数 ≧ ノード数となる．トピックの可視化が目的なので，各トピックについてノード数がトピック数と近い値をとるような共起割合を採用した．さらに，過度に密なエッジは可読性を損なうためノード数がトピッ

表 2　トピック数 k を変化させたときのエッジ数

k	エッジ数 (.20)	ノード数 (.20)	エッジ数 (.10)	ノード数 (.10)
10	**18**	**10**	29	10
20	20	20	**25**	**20**
40	4	5	**87**	**40**
60	4	5	99	49
80	1	2	77	54

k	エッジ数 (.08)	ノード数 (.08)	エッジ数 (.07)	ノード数 (.07)
60	**172**	**60**	181	60
80	114	78	**137**	**80**

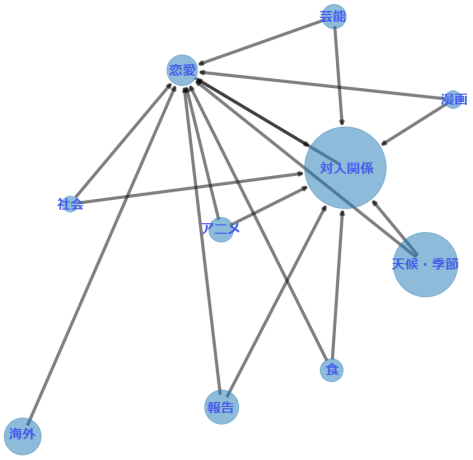

図 4　10 トピックの有向グラフ

図 5　20 トピックの有向グラフ

ク数と近い値をとるもののうち，ノード数よりやや多い程度のエッジ数をとるように共起割合を調整した．

図 4，図 5，図 6，図 7，図 8 はそれぞれ 10 トピック，20 トピック，40 トピック，60 トピック，80 トピックでトピック分類したときの有向グラフである．ノード数が多いグラフについては，可読性を優先しトピック番号のみを表記する．ノードの大きさはトピックに属するツイートがリツイートされた回数を表す．

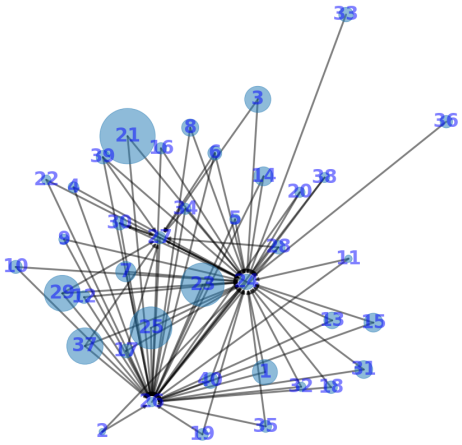

図 6　40 トピックの有向グラフ

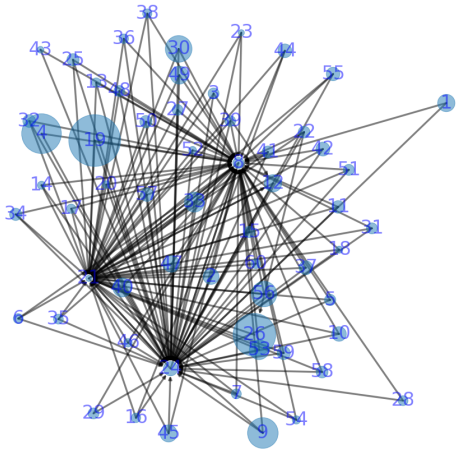

図 7　60 トピックの有向グラフ

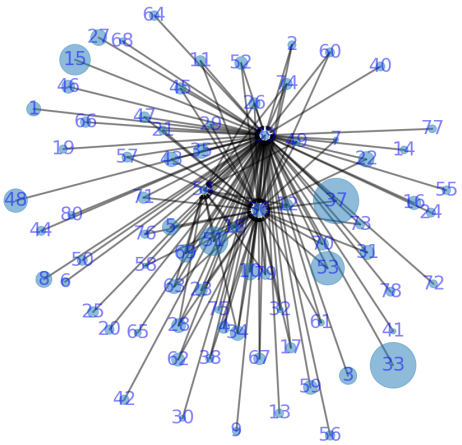

図 8　80 トピックの有向グラフ

表 3 は 10 トピック分類で中心性の高いトピック 2，10 のツイート数，80 トピック分類で中心性の高いトピック 36，39，54 のツイート数を表す．80 トピックで中心性の高い 3 トピックの

表 3　トピック毎のツイート数

トピック番号 (トピック分類数)	属するツイート数
2(10)	20,282
10(10)	20,385
2(10)+10(10)	40,667
36(80)	3,746
39(80)	8,006
54(80)	4,159
36(80)+39(80)+54(80)	15,891

表 4　80 トピック分類における中心性の高いトピックの特徴

トピック番号	全体としての傾向	特徴語
36	季節的な話題	行っ, 今年, 感, 買っ, 誕生
39	人間関係の話題	良い, あなた, 人間, 結果, 会社
54	恋愛関係の話題	女, 映画, 観, おじさん

ツイートのうち，10 トピック分類で中心性の高いトピック 2,10 に含まれているものは 13,119 ツイートである．トピック分類数 K のトピック i に含まれるツイートの集合を $T_{i(K)}$ と表記する．式 (1) は 80 トピックで中心性の高い 3 トピックの 82.6% が 10 トピック分類で中心性の高いトピック 2,10 に含まれていることを示す．

$$\frac{\left[T_{2(10)} \cup T_{10(10)}\right] \cap \left[T_{36(80)} \cup T_{39(80)} \cup T_{54(80)}\right]}{\left[T_{36(80)} \cup T_{39(80)} \cup T_{54(80)}\right]} = 0.826(1)$$

10 トピック分類で中心性の高いトピック 2, トピック 10 が図 7 や図 8 では 3 トピックに分かれていることが読み取れる．(式 1) 及び図 4，図 8 から，トピック分類数の増加によって中心性の高いトピックをより高い精度で抽出できたといえる．表 4 は 80 トピックで分類したグラフで特に中心性の高い 3 トピックそれぞれの特徴である．

季節的な話題，人間関係の話題，恋愛関係の話題はいずれもツイートを理解するための前提知識が多くのユーザに共有されており，多くのコミュニティで共感されやすいため中心性が高くなったと考えられる．しかし，多くのトピックは中心性の高い季節的な話題，人間関係の話題，恋愛関係の話題のいくつかとしか共起割合が高くなっていない．例えば，80 トピック分類でソーシャルゲーム (トピック 26) と舞台・演劇 (トピック 74) は中心性の高い 3 つのトピックに対する関係性が近しいが，両トピック間の共起割合は高くない．このことから，中心性が高いトピックに対してどのような関係性を持つかによって嗜好を分類し，類似の関係をもつ嗜好同士を推薦することでセレンディピティのある推薦を実現できる可能性が示唆される．

5　おわりに

トピックモデルによってツイートに現れる嗜好を分類し，嗜好同士にどのような繋がりがあるのかをグラフによって表現した．これによって中心性の高い嗜好が示され，今後の推薦システムにおける活用が期待される．今後は中心性が高いトピックに対して類似の関係性を持つ嗜好同士に注目したグラフ作成によって類似の関係をもつ嗜好同士推薦システムへの実装を進めていく．また，よりトピック数を増やした場合の可読性を損なわない可視化についても検討していきたい．

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 JP16H02904, JP19K20333）の助成によって行われた．

文　　献

[1] Badrul Sarwar et al. Item-based Collaborative Filtering Recommendation Algorithms. WWW, :2001. 2001, p.285-295.
[2] 土方嘉徳. 嗜好抽出と情報推薦技術. 情報処理学会論文誌. 2007, 48(9), p.959-961.
[3] 岡本一志, 藤井流華. 協調フィルタリング入門 (特集 Web インテリジェンスとインタラクション 2019). 知能と情報. 2019, 31(1), p.5-9.
[4] 福島良典, 大澤幸生. ソーシャルメディアを利用したセレンディピティな情報. 人工知能学会全国大会論文集. 2012, 2012, p.3E1R66-3E1R66.
[5] 徐哲林 ほか. Twitter におけるセレンディピティのあるおすすめユーザの発見. 第 81 回全国大会講演論文集 2019. 2019, 1, p.37-38.
[6] 折本伸之, 渥美雅保. Twitter 連携ニュースフィルタリングのためのトピックモデルを用いたユーザの興味学習に基づくニュース Tweet ランキング. 第 81 回全国大会講演論文集. 2019, 2019(1), p.319-320.
[7] David M. Beli, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research 3. 2003, (993), p.1022.
[8] 越田浩, 細部博史, 脇田建. 大規模社会ネットワークの階層的視覚化手法. コンピュータ ソフトウェア. 2011, 28(2), p.2_202-2_216.
[9] 細部博史. 高次元アプローチによる一般無向グラフの対話的視覚化法. 情報処理学会論文誌. 2005, 46(7), p.1536-1547.
[10] 杉本拓弥ほか. 重みつき完全グラフに基づく異ジャンル間の嗜好傾向表現と そのレコメンデーションシステムへの応用. ファジィシステムシンポジウム講演論文集. 2011, 27, p.237-240.