ニュース記事の共引用関係に基づく類似意見のユーザコミュニティの抽出方法の検討

西上 貴雅 [1]　風間 一洋 [1]　吉田 光男 [2]　土方 嘉徳 [3]

[1] 和歌山大学大学院システム工学研究科　[2] 筑波大学ビジネスサイエンス系　[3] 兵庫県立大学大学院情報科学研究科

はじめに

背景

- ソーシャルメディア上でのニュース記事に対するユーザ反応
  - 意見形成や情報伝播のダイナミクスを理解する上で重要な情報源
- 従来のテキスト解析や統計手法
  - ユーザ間の意見の偏りを把握することが難しい
- ニュース記事の共引用関係（ユーザが言及した記事集合の類似性）の利用
  - 関係の有無でグラフを構築するアプローチ
  - 単純なモジュラリティ最大化　Jaccard係数
    - ↳ 抽出したコミュニティにノイズが混入

目的

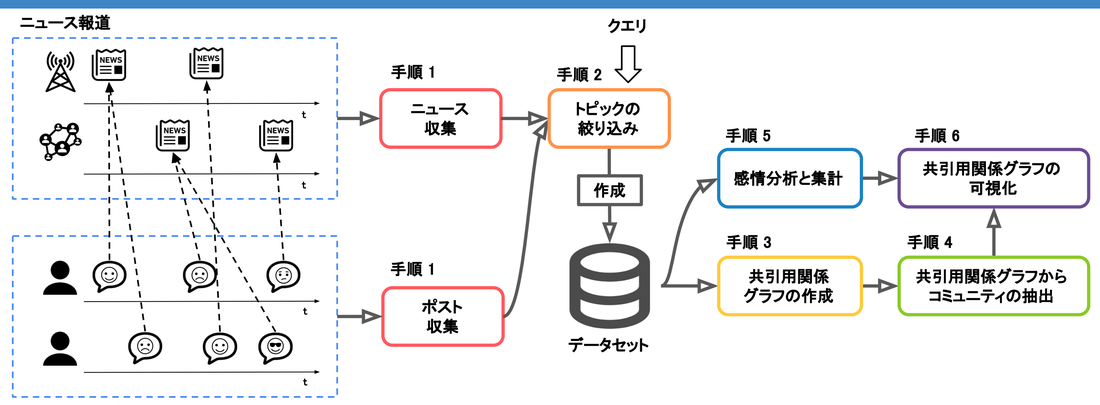
- 意見が類似しているユーザコミュニティを高精度に抽出する手法を提案
  - ニュース記事の共引用関係に基づくグラフ構築手法を広く検討

検討するグラフ構築手法

- **指標**
  - **共引用関係の有無 ☞ 共**
  - **共引用関係の Jaccard係数 ☞ J**
    - ユーザ間の引用ニュース記事集合のJaccard係数
  - **文章の類似度 ☞ 文**
    - 日本語Sentence-Luke[1] でベクトル化 & cos類似度
  - **感情の類似度 ☞ 感**
    - 感情ベクトルのユークリッド距離
- **類似度を算出する粒度**
  - ツイート単位 ☞ **ツ**　ユーザ単位 ☞ **ユ**

ベースライン（共＋J）	ユ（共）
ツ（文）	ユ（文）
ツ（感）	ユ（感）
ツ（共＋文）	ユ（共＋文）
ツ（共＋感）	ユ（共＋感）
ツ（文＋感）	ユ（文＋感）
ツ（共＋文＋感）	ユ（共＋文＋感）

システムの概要



手順1：ニュースとポストの収集

- Twitter API で取得　期間：2022年 1月 1日 ～ 11月 30日
- ニュース記事 ： 1,489,667 件
- ニュース記事に言及したポスト（リポストを含む）： 3,466,472 件

手順2：トピックの絞り込み

- 次のキーワードを全て含むニュース記事とそれらに言及したポストを収集
  - キーワード：「ロシア, ウクライナ, 侵攻」
  - 期間：2022年 1月 25日 ～ 3月 25日（2月24日前後30日間）
- ニュース記事数：　4,432 件　ツイート数 ：　5,647 件

手順3：共引用関係グラフの作成　👈 この部分の検討

手順4：共引用関係グラフからコミュニティの抽出

- Leiden法[2]で共引用関係グラフをクラスタリング

手順5：感情分析と集計

- ポストから中北らの手法[3]を用いて**ユーザの発言部分のみを抽出**
- 難波らのT5を用いた機械学習ベースの感情分析器[4]で感情分析

共引用関係グラフの可視化

ノード

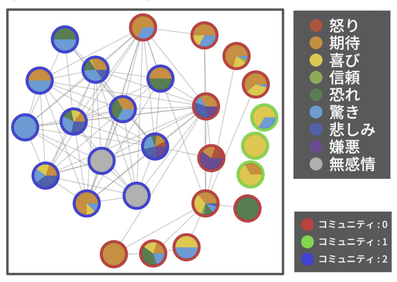
- 外枠 ☞ 所属コミュニティ
- 内側 ☞ 感情の割合
  - 感情 = Plutchikの感情モデル[5]の基本感情8種類 + 無感情
    - ↳ **怒り**、**期待**、**喜び**、**信頼**、**恐れ**、**驚き**、**悲しみ**、**嫌悪**、**無感情**

エッジ

- 類似度などの指標に基づいて有無を決定

配置

- Gephi のForceAtlas2[6] を用いて決定



評価方法

$$Q = \sum_{i \in C} \left\{ \frac{e_{ii}}{2m} - \left(\frac{a_i}{2m} \right)^2 \right\}$$

$$\text{混入度} = \frac{1}{c} \sum_{i=1}^{c} \max_{(u,v) \in V_i \times V_i, u \neq v} d(u,v)^2$$

モジュラリティ：コミュニティ分割の良し悪しを測る指標

- 第1項：クラスタ i に含まれるエッジ数の割合
- 第2項：クラスタ間のエッジ数の割合

混入度：コミュニティを意見ができるだけ類似するように抽出

- **c**：コミュニティ数、**Vi**：コミュニティ i に含まれるノード集合
- **d(u, v)**：ノードuとノードvの感情傾向ベクトルのユークリッド距離

モジュラリティに混入度を組み合わせて評価

- モジュラリティは孤立ノードが少ないほど大きくなる
- 性質が異なるノードは独自のクラスタを形成して欲しい

分析

ノード数 560 個

ツイート単位

		エッジ数	連結成分数	コミュニティ数	孤立ノード数	モジュラリティ	混入度
ベースライン		2796	86	97	73	0.674	12.00
文	モ	146	509	509	496	0.638	2.50
	混	67	548	548	546	0.029	0.00
	平	77678	2	4	1	0.114	15.00
感	モ	4907	138	144	104	0.673	1.65
	混	3396	285	285	229	0.618	0.00
	平	86257	4	7	3	0.104	14.00
共＋文	モ	39	532	533	523	0.766	1.00
	混	13	553	554	551	0.249	0.00
	平	1944	155	165	142	0.579	10.00
共＋感	モ	73	506	506	476	0.897	0.04
	混	58	520	520	496	0.864	0.00
	平	2043	131	142	126	0.566	5.00
文＋感	モ	118	490	490	451	0.935	9.25
	混	20	543	543	529	0.905	0.00
	平	54	524	524	504	0.864	2.00
共＋文＋感	モ	118	490	490	451	0.935	9.25
	混	20	543	543	529	0.905	0.00
	平	41	533	533	519	0.834	2.00

ユーザ単位

		エッジ数	連結成分数	コミュニティ数	孤立ノード数	モジュラリティ	混入度
共		129	490	490	451	0.931	9.25
文	モ	137	514	514	504	0.588	7.00
	混	23	550	550	547	0.480	0.00
	平	78397	2	6	1	0.068	22.00
感	モ	10106	83	83	37	0.783	0.00
	混	10106	83	83	37	0.783	0.00
	平	85676	1	3	0	0.125	22.00
共＋文	モ	32	536	536	524	0.811	3.00
	混	3	557	557	554	0.667	0.00
	平	2033	179	187	166	0.561	18.00
共＋感	モ	236	397	398	337	0.945	0.00
	混	236	397	398	337	0.945	0.00
	平	1915	121	134	109	0.636	18.00
文＋感	モ	119	490	490	451	0.935	13.00
	混	25	542	542	529	0.854	0.00
	平	43	530	530	512	0.847	3.00
共＋文＋感	モ	119	490	490	451	0.935	13.00
	混	25	542	542	529	0.854	0.00
	平	40	533	533	517	0.840	3.00

※ モ：モジュラリティ最大化，混：混入度最小化，平：平均値

孤立ノード率の平均	ツイート単位	ユーザ単位
モ	0.7443	0.6857
混	0.8571	0.7539
平均	**0.3854**	**0.3884**

★ **文章の類似度 vs. 感情の類似度**

- 感情の類似度のほうが混入度が小さい

★ **共引用関係の有無**

- 共引用関係が有るほうが混入度が小さい

★ **モジュラリティ最大化 vs. 混入度最小化 vs. 統計量（平均値）**

- (モ) と (混) の孤立ノード率平均は少なくとも 68%超
  - ↳ 混入度が小さくても過剰に分割している可能性

★ **ツイート単位 vs. ユーザ単位**

- ツイート単位のほうが混入度が小さい

考察

意見の類似性

- 本研究　ポスト文章からユーザの発言を抜き出す
- 文章の類似度
  - 短い文章（特に処理ミス）にセンシティブ
  - 引用部分の抽出ミスが大きいクラスタを形成する傾向大
- 感情の類似度
  - 抽出ミスは無感情となる場合が多い

ツイート単位 vs. ユーザ単位

- ツイート単位は混入度は少ないが孤立ノードが多い
  - 共引用関係でもコメント部が残らない場合は無視されることに起因する可能性あり
- 特定のトピックに絞り込んでいるので、ユーザ単位の類似性で代替可能

閾値設定について

- (モ) や (混) は孤立ノードが多い
  - (平) は孤立ノードは少ないが、妥当か議論の余地あり
- 評価指標に孤立ノードを考慮する必要がある？

統計量による閾値設定の可能性

- ソーシャルメディアにおけるP/N判定は2極化する
  - ↳ サイレントマジョリティは投稿しない
- 平均値はその分布を分断する
  - ↳ 上記と似たようなことが起こっている可能性
- 統計量ベースの閾値設定で計算コストを減らせないか？

まとめ

グラフの構築手法を定量的に分析

- ↳ 混入度をできるだけ小さくしつつ、孤立ノードが少ない構築手法
  - （指標を算出する単位）☞ ツイート単位
  - （用いる指標）☞ 共引用関係 + 感情の類似度
  - （閾値）☞ ※ 議論の余地あり

今後の予定

- 統計量ベースのグラフ構築手法の検討
- 独特の意見を持つ人の行動を分析

関連研究 ・ 参考文献

1. Sonobe Isao. sonoisa/sentence-luke-japanese-base-lite, 2023.
2. Traag, Vincent Antonio et al. “From Louvain to Leiden: guaranteeing well-connected communities.” Scientific Reports 9 (2018): n. pag.
3. 中北雄大, 風間一洋, 吉田光男, 土方嘉徳. 感情とトピックに注目したメディアの報道姿勢の分析. 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2022), B24-2, 2022
4. Nanba, H., Yamamoto, K., Fukuda, S., Shoji, H., Tanishita, M., Kyutoku, Y. and Yamashina, M.: Modeling the Social Acceptability of Technologies Using Twitter Data, Inproceedings of the 2023 IEEE Conference on Systems (2023).
5. Plutchik, R.: The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice, American Scientist, Vol. 89, No. 4, pp. 344–350 (2001).
6. Jacomy, M., Venturini, T., Heymann, S. and Bastian, M.: ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS One 9(6): e98679 (2014).

連絡先

西上 貴雅　和歌山大学大学院 システム工学研究科 システム知能クラスタ 風間研究室　takamasa.nishigami@gmail.com