The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

政治ニュース記事クラスタに対する属性ごとのユーザ行動の分析

Analysis of user activity of politics news cluster by user attributes

関 喜史 *1　吉田 光男 *2
Yoshifumi Seki　Mitsuo Yoshida

*1株式会社 Gunosy　*2豊橋技術科学大学
Gunosy Inc.　Toyohashi University of Technology

我々はフィルターバブルやエコーチェンバーのようなユーザのニュース記事の閲覧時に発生するとされる選択的接触行動をユーザ行動ログから定量的に評価することを目指している．この目的のために，これまでに，ユーザ属性ごと行動の違いについて，カテゴリとキーワードを用いて議論してきた．ニュース配信サービスにおいて，政治ニュース記事に対するユーザ属性ごと行動の違いを分析する際に，本稿ではニュース記事のクラスタに着目する．クラスタにおけるクラスタにおけるクリックと Like 行動との相関係数および比率を比較し，ニュースの内容によるユーザ行動の違いを詳細に捉えられることを実験的に示した．

1. はじめに

近年人々はウェブでの情報収集を活発に行なっているが [LINE]，ウェブにおいてはフィルターバブル [Pariser 11]，エコーチェンバー [Jamieson 08] という現象が指摘されている．フィルターバブルは情報がパーソナライズされてしまうことで広い視野を失ってしまうとされる現象であり，エコーチェンバーは自らが好む情報やそれを支持するコミュニティにばかりに接触してしまうあまり，偏った考えがより強化されてしまうとされる減少である．これら 2 つの現象は Brexit やトランプ大統領誕生に大きく影響したとされており，ジャーナリストを中心に問題視されているが [Hooton 16]，これらの現象について定量的な分析はほとんど行われていない．このような背景から我々はニュース配信サービスのユーザ行動ログからこれらの現象についての分析をこれまで行っている．

本研究ではニュース配信サービスにおいて，政治ニュース記事に対するユーザ属性ごと行動の違いを，ニュース記事のクラスタを用いて分析する．以前の研究で我々はユーザの属性別の行動の違いについて，カテゴリとキーワードを用いて議論した [関 18, Seki 18]．若年層の投票率の低下は社会問題化しているなかで [総務]，世代間で政治への関心がどのように異なるのかといアンケートを用いた研究や調査はあるが，実際の行動に基づく研究は多くない [鈴木 12]．本研究ではニュース記事のクラスタを用いることで，ニュースの内容によるユーザ行動の違いをより詳細に捉えることを目指す．

2. データセット

本研究では株式会社 Gunosy が提供するニュース配信サービスであるグノシー *1 における 2018 年 8 月の 1ヶ月間の政治カテゴリにおけるユーザ行動ログを用いる．

本研究ではニュース記事のクリックとニュース記事への *like* を扱う．図 1 にグノシーのサービス画面を示す．ニュース記事リストはタイトルとサムネイル画像からなるセルとして表示されており，ユーザはクリックすることで記事詳細画面で本文を読むことができる．記事詳細画面の下部のナビゲーションバー

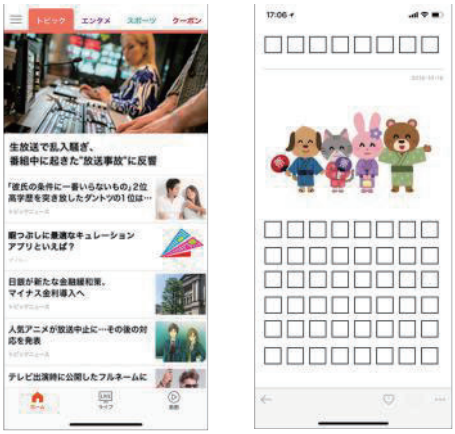

図 1: グノシーのサービス画面: 左の図はニュース記事リスト画面であり，各ニュースリスト記事はタイトルとサムネイル画像からなるセルで構成されている．セルをクリックすると右図の詳細画面に進む．画面下部のツールバーのハート型のボタンが Like ボタンである．Like ボタンはニュース記事本文の末尾にもある．

のハート型のボタンを押すことでニュース記事に *Like* することができる．詳細画面末尾にも同様のボタンがある．2019 年 2 月現在，*Like* ボタンを押してもユーザにメリットのある機能は提供されていないが，いくらかのユーザには活発に利用されている．比較のためには一定以上の行動が必要なため，クリック数が 100 以上の記事に限定している．

本研究ではユーザ属性別の行動の違いに着目するが，ユーザ属性としては性別と年齢を用いる．ユーザはこれらの属性を利用開始時に登録することができる．性別は男性，女性，その他から選べるが，本研究では男性と女性を比較対象に用いる．年齢は 20 歳未満, 20 – 24, 25 – 29, 30 代, 40 代, 50 歳以上から選ぶことができるが，分析で 20 代以下 (Young)，30 代 (Middle)，40 代以上 (Older) の 3 カテゴリに分ける．ユーザは属性を登録せずに利用することもできるが，その場合は一定以上サービスを利用していたユーザを対象に，サービス内での行動から教師あり学習を元にユーザ属性の推定を行っている．分析はこれらの方法によって属性がわかるユーザのみに限定

連絡先: 関 喜史，株式会社 Gunosy，〒 106–6125 東京都港区六本木 6–10–1 六本木ヒルズ森タワー 25 階，03–6455–4562，yoshifumi.seki@gunosy.com

*1 https://gunosy.com

表 1: Each action ratio between demographic attributes.

			all	Politics
number of news articles				1,333
Click ratio	Gender	Male	58.9%	76.2%
		Female	41.1%	23.8%
	Age	Young	34.7%	16.4%
		Middle	30.2%	22.1%
		Older	35.1%	61.5%
Like ratio	Gender	Male	47.7%	78.2%
		Female	52.3%	21.8%
	Age	Young	25.8%	8.8%
		Middle	25.4%	11.0%
		Older	48.7%	80.2%

する．

分析対象となるニュース記事は，一定以上のユーザ行動が必要のためクリック数が 100 以上のものに限定する．このようにして 1,333 件のニュース記事が得られた．それぞれの属性の各行動の比率を表 1　に示す．all はサービス全体での比率を示している．サービス全体に比べて政治カテゴリは男性の年齢層の高いユーザの行動が活発であることがわかる．

3. クラスタリングによる分析

3.1 クラスタリング

記事のクラスタリングはニュース記事のタイトルを用いて行う．まずニュース記事をタイトルを用いてベクトル化する．ベクトルにあたっては Countinuous Bag-of-Words (CBOW) モデルを用いた word2vec を利用する．モデルの作成には過去 2 年間グノシーで配信されたニュース記事データのタイトルを Mecab と mecab-ipadic-NEologd を用いている．同時にそのニュース記事データを用いて inverse document frequency を計算しておく．ニュース記事のベクトル $\boldsymbol{a}$ は以下の式で定義される [米田 17].

$$\boldsymbol{a} := \frac{\sum_{\boldsymbol{w}_i \in W_a} \mathrm{idf}(\boldsymbol{w}_i)\boldsymbol{w}_i}{\left\| \sum_{\boldsymbol{w}_i \in W_a} \mathrm{idf}(\boldsymbol{w}_i)\boldsymbol{w}_i \right\|} \in \mathbb{R}^d,$$

ここで W_a はニュース記事 a のタイトルの単語集合であり，$\boldsymbol{w}_i \in \mathbb{R}^d$ は次元数 d で word2vec にベクトル化された単語である．$idf(\boldsymbol{w}_i)$ は単語 w_i の idf 値を表す．つまりタイトルに出現した単語ベクトルの idf 値による重み付け平均ベクトルをニュース記事ベクトルとしている．

このようにして構築した記事ベクトルを k-means を用いてクラスタリングする．クラスタ数は 10 とした．[*2] 各クラスタの概要と記事数とクリック数，Like 数の分布を表 2 に示す．記事数は総裁選に関する記事が多いクラスタ 6 の記事数が飛び抜けて多い以外は，全体として等しく分布しているが，クリック数と Like 数はバラつきが多い．クリック数，Like 数は政治ゴシップを扱うクラスタ 9 が非常に多く，ユーザから人気のあるクラスタであることがわかる．クラスタ 6 は記事数が他のクラスタの 2.3 倍程度あるため，それと比較するとクリック数，Like 数はそこまで多くない．クリック数と Like 数の傾向はクラスタごとに異なり，例えば沖縄に関係するニュースが多いクラスタ 8 はクリック数は少ないが Like 数は比較すると多い．このようにクラスタごとにクリックと Like の傾向は異なる．

*2　クラスタ数を 2 40 の範疇でエルボー法による分析を行ったが，クラスタ内誤差平方和はクラスタ数 10 以上でなだらかに下がり続けていたため，解釈性を高めるために 10 とした．

表 2: クラスタリングの結果

	概要	記事数	記事数 比率 [%]	クリック数 比率 [%]	Like 数 比率 [%]
1	政権	88	6.6	3.0	2.8
2	選挙・政局	192	14.4	8.0	5.6
3	総裁選 (ゴシップ)	167	12.4	12.4	13.1
4	社会	97	7.3	3.7	2.2
5	事件	108	8.1	3.9	3.6
6	総裁選	283	21.2	12.8	15.1
7	コラム	103	7.7	20.5	14.2
8	沖縄関連	82	6.2	4.8	13.0
9	政治ゴシップ	140	10.5	27.8	27.7
10	外交・海外	73	5.5	2.9	2.6

3.2 各クラスタにおけるユーザ属性間の行動傾向

ここでは各クラスタのユーザ属性間の行動傾向を分析し，議論する．属性間の行動傾向の指標としてピアソンの相関係数 (以下相関係数) と，行動総数の比を用いる．相関係数が高ければ強い相関があり，クラスタ内のニュース記事において行動の属性間の比率がある程度一貫していることがわかる．そのため相関が強ければ行動総数の比から，そのクラスタの特性を議論することができる．一方で相関係数が低ければ，クラスタ内のニュース記事において属性間の行動の比率の一貫性が低いため，行動総数の比で議論することは難しい．まずは相関係数によって，それぞれのクラスタにおける属性間行動の一貫性を確認する．

表 3 にクラスタごとにクリック行動の相関係数と比率を示す．すべての属性間の相関係数は全クラスタにおいて高く，各クラスタとも属性間のクリック傾向には強い相関があることがわかる．表 4 には同様に Like 行動の相関係数と比率を示す．クリックと比較したときに相関係数は小さいが，全体としては強い相関がある．一方でクラスタ別でみると低いクラスタがいくつか存在する．特にクラスタ 4 はすべての属性間で著しく相関係数が低い．また属性ごとに傾向が違い，例えばクラスタ 6 は男女間と middle–older 間では強い相関があるが，young–middle, older–young 間では比較的相関が弱い．今回は相関係数が 0.6 以上のクラスタに絞って，行動総数の比率を議論することにする．

例として男女間のクラスタ間の傾向について議論する．政治カテゴリはもともと男性の行動のほうが多く，男性は女性の 3.2 倍のクリック，3.58 倍の Like を生み出している．クリックの中で男性比率がより高いのはクラスタ 6, 次いでクラスタ 3, そしてクラスタ 10 である．クラスタ 6 とクラスタ 3 は自民党総裁選に関するものであり，クラスタ 10 は外交に関するものである．これらのクラスタはより男性の関心が強いクラスタであるといえる，一方で女性のクリック比率が男性と比較して高いのはクラスタ 8, クラスタ 7, クラスタ 9 であり，それぞれ沖縄，コラム，政治ゴシップのクラスタであり，これらのクラスタは女性の関心が強いクラスタであるといえる．Like をクリックと比較すると，男性の関心の高いクラスタでは，クラスタ 3 がクリックで高い関心の合った他の 2 つに比べて男性の比率が突出していない．またクリックにおいて女性側の関心が高かったクラスタをみると，クラスタ 9 は Like の比率で見れば男性がより Like しているクラスタになっている．一方でクラスタ 8 はクリックよりますます女性側に偏っている．このようにクラスタごとにクリックと Like では行動傾向も異なること

表 3: 各クラスタのクリック行動の属性間の相関と比率

	male-female		young-middle		middle-older		older-young	
	pearson	ratio	peason	ratio	peason	ratio	peason	ratio
all	0.902	3.20	0.992	0.75	0.923	0.36	0.901	3.74
1	0.895	3.60	0.981	0.98	0.949	0.34	0.906	2.98
2	0.930	3.08	0.996	0.86	0.899	0.38	0.870	3.04
3	0.869	4.84	0.913	0.79	0.948	0.28	0.787	4.54
4	0.634	3.44	0.993	0.88	0.811	0.53	0.830	2.14
5	0.880	3.76	0.978	0.75	0.899	0.38	0.864	3.52
6	0.962	5.87	0.995	0.83	0.924	0.23	0.924	5.29
7	0.859	2.61	0.987	0.65	0.929	0.41	0.924	3.70
8	0.987	1.72	1.000	0.80	0.995	0.42	0.995	2.98
9	0.953	2.70	0.999	0.69	0.947	0.40	0.940	3.66
10	0.982	4.36	0.994	0.86	0.995	0.29	0.992	4.02

表 4: 各クラスタの Like 行動の属性間の相関と比率

	male-female		young-middle		middle-older		older-young	
	pearson	ratio	peason	ratio	peason	ratio	peason	ratio
all	0.747	3.58	0.909	0.81	0.845	0.14	0.786	9.09
1	0.719	3.76	0.537	1.05	0.686	0.13	0.840	7.10
2	0.698	3.06	0.767	0.71	0.728	0.19	0.574	7.39
3	0.720	4.93	0.564	0.78	0.632	0.11	0.556	11.65
4	0.263	3.93	0.362	1.2	0.454	0.20	0.359	4.13
5	0.734	3.45	0.610	0.90	0.703	0.18	0.636	6.00
6	0.925	6.51	0.519	0.72	0.909	0.10	0.571	13.90
7	0.769	2.92	0.857	0.72	0.854	0.16	0.818	8.65
8	0.935	1.33	0.975	1.01	0.999	0.22	0.973	4.53
9	0.976	4.79	0.884	0.70	0.650	0.10	0.683	13.72
10	0.945	6.45	0.713	0.76	0.881	0.17	0.796	7.94

が示唆される．

4. まとめ

本稿ではニュース配信サービスにおいて，政治ニュース記事に対するユーザ属性ごと行動の違いを，ニュース記事のクラスタを用いて分析した．我々は以前のユーザの属性別の行動の違いについて，カテゴリとキーワードを用いて議論したが，今回はニュース記事のクラスタを用いることで，ニュースの内容によるユーザ行動の違いをより詳細に捉えることを目指した．ニュース記事タイトルの分散表現を使ってクラスタリングを行い，各クラスタに対する相関係数とクリック行動と Like 行動の比率を求め比較した．クラスタごとに相関の度合いは異なるが大まかには強い相関があり，クラスタの傾向を比較することが可能であること確認し，比率の違いからユーザ属性による話題への関心の違いを議論できることを示した．キーワードベースで行った以前の手法と比較すると，クラスタの解釈については定性的な部分が必要になるが，クラスタ内のデータ量が大きくもてるため，分析が容易になる点，また全体の中でのまとまりを抽出するため，トピックを網羅的に議論できる点で優れている．

今後はユーザ側を属性ベースではなく，興味関心で分類することでユーザの関心の偏りの発見と分析に取り組みたい．また政治ニュースへの関心は社会でのイベントにより変動すると考えられるため，データ量を増やし時間経過による影響についても詳しく調査したい．

参考文献

[Hooton 16] Hooton, C.: Social media echo chambers gifted Donald Trump the presidency. the Independent (2016)

[Jamieson 08] Jamieson, K. H. and Cappella, J. N.: *Echo chamber: Rush Limbaugh and the conservative media establishment*, Oxford University Press (2008)

[LINE] LINE 株式会社：世代間のニュースサービス利用に関する意識調査

[Pariser 11] Pariser, E.: *The Filter Bubble: What the Internet Is Hiding from You*, The Penguin Group (2011)

[Seki 18] Seki, Y. and Yoshida, M.: Analysis of Bias in Gathering Information Between User Attributes in News Application, in *IEEE BigData 2018 Workshop : The 3rd International Workshop on Application of Big Data for Computational Social Science (ABCSS2018)* (2018)

[関 18] 関 喜史：世代による政治ニュース記事の閲覧傾向の違いの分析, 人工知能学会全国大会 (2018)

[総務] 総務省：国政選挙の年代別投票率の推移について

[米田 17] 米田 武, 久保 光証, 関 喜史:ニュース配信システムにおけるパーソナライズの設計と導入, 信学技報, NLC2017-28 (2017)

[鈴木 12] 鈴木万希枝：若年層のニュース消費に関する研究: 情報源接触パタンおよびニュース情報への選択的接触の検討, 三田哲学会 (2012)