The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ネットワーク構造に基づく新聞記事の分類による読者の行動分析

Analysis of readers by Classification of Articles in a News Service from Network Structures

園田 亜斗夢 *1　鳥海　不二夫 *1　中島 寛人 *2　郷治 雅 *2
Atom Sonoda　Fujio Toriumi　Hiroto Nakajima　Miyabi Gouji

*1東京大学　*2日本経済新聞社
The University of Tokyo　Nikkei Inc.

Information is transmitted through websites, and the immediate reaction to various information is required. Hence, the efforts for readers to select information themselves have increased, which leads to the further improvement of recommendation services that can reduce such burdens. On the other hand, it is pointed out that filter bubbles that only provide biased information to users are generated due to the redundant recommendation. In this research, we analyzed behavioral changes prior to recommendation by clustering, and showed that behavior changes after average number of browsing number and article type change during the period.

1. はじめに

新聞社等のメディアの発信するニュースの主な媒体が新聞から web サイトに拡大するに従い，記事の速報性の向上や記事数の増加が進んでいる． これに伴い，ユーザが記事を選択する際に必要な時間と労力が増大していると推測される．このような負担を減らし，満足度を向上させるため，ユーザの嗜好に応じた推薦サービスは多くの分野で導入されている。一方で，過度の推薦によってユーザに偏った情報のみを提供するフィルターバブルが発生しているとの指摘もある [Pariser 2011]．

記事の推薦は，ユーザの選択行動を変容させる可能性がある [Nguyen 2014]．本研究では，フィルターバブルの発生を抑制し，多様な意見に触れる機会を提供できるような推薦システムの開発のために，推薦システム開発に先立ち，既存のニュースサイト上でどのような場合にユーザは行動を変容させるのかについて分析する．新聞社の 1 つである日本経済新聞社の web サービス日経電子版を用いて，筆者らが提案した記事の分類手法 [園田 2018] に基づき，分類された記事の情報を用いて，ユーザの行動が変容する過程について分析する．実際の推薦システムでは，ユーザの行動履歴を閲覧した記事のクラスタの割合を分析し，それに基づき記事を推薦し，読む記事を変化させることで読者の行動変容を促すことを目指す．本研究では，推薦システムの開発に向け，閲覧記事の多様性が増加したユーザの行動変容要因を明らかにする．

2. 記事のクラスタリング

2.1 分析に用いるデータ

本研究では，2017 年 5 月 21 日から 8 月 20 日までの 3ヶ月間の日本経済新聞社の web サービス日経電子版のデータを用いる．この間に読まれた記事は全部で約 60 万記事であり，会員数は約 200 万人であった．

この中で，一定程度以上読まれている記事のみを扱うため，今回は読者が 100 人以下の記事は対象から除いた．これは，将来的な推薦システムの開発には，ニュースという即時生の高いコンテンツであるという性質上，24 時間以内の記事を優先して推薦するため，定期的な更新が必要となり，計算時間の観点からデータを減らす必要があることと，推薦ではフィルターバブルを抑制するという目的から全体のユーザ数に対し極端に読者の少ない特殊な記事を推薦することは考えていないためである．分析対象となるデータはこれらの処理を行った約 7 万記事である．

連絡先: 園田　亜斗夢，東京大学，東京都文京区本郷 7 － 3 － 1，090-1364-4850，sonoda@crimson.q.t.u-tokyo.ac.jp

2.2 関連記事ネットワークの構築

ある 2 つの記事を読んだユーザが複数人いた場合，2 つの記事は共通した興味を引く内容を有していると考えられる．つまり，記事を読んだユーザの重複度から記事の類似性を求めることができる．そこで，ユーザの重複度の高い記事同士をリンクで結ぶことで，記事ネットワークの構築を行う．2 つの記事 a_i，a_j のユーザ群 U_i，U_j の重複率は Simpson 係数を用いて次のように求められる．

$$Sim(a_i, a_j) = \frac{|U_i \cdot U_j|}{min(|U_i|, |U_j|)} \quad (1)$$

ここでは，前述の類似度が閾値 0.62 以上の記事の間にリンクを張ることで，重みあり無向ネットワークを構築した．このとき，モジュラリティQ=0.936，リンクは 38002 件，ノードは 17233 件であった．

なお，このような類似度を測る指標としては，Simpson 係数のほかに Jaccard 係数，Dice 係数などがあるが，共起を用いた関係性の強さを表現するための指標としては Simpson 係数が適切であるとされている [松尾 2005].

2.3 記事のクラスタリング

ここで, 上記のネットワークに対してクラスタリングを行った．クラスタリング手法には Modurality を基準とする Louvain 法 [Blondel 2008] を用いた．クラスタリングにより，2470 件のクラスタが得られた．これらのいずれかのクラスタに含まれる記事は，図 1 のようになっている．横軸に記事の投稿日，縦軸に記事数をとったものである．記事数が多い日と少ない日があるのは，休日は記事が減るためである．この図から，対象期間内に発表された記事が多く含まれていることがわかる．一方，期間の後半では，発表された後の日数が少なく，閲覧数が適切に蓄積されていないことから，クラスタリング結果に含まれる記事数が減少していることがわかる．このことから，この先の行動分析では，クラスタリング結果に含まれる記事数が十分ある，クラスタリング対象期間の前半と中盤を比較する．

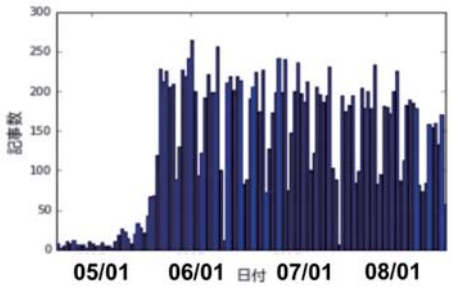

図 1: いずれかのクラスタに含まれる記事数の推移

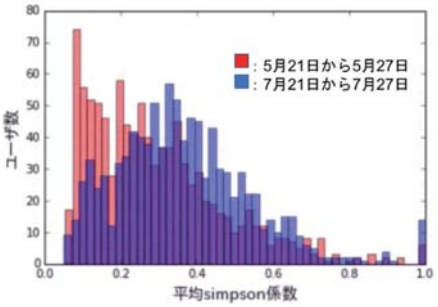

図 2: クラスタ間平均類似度の変化

3. 読者の行動変容の分析

読者の行動を，読んだ記事の種類，と定義して，ある一定の期間経過後，その行動が変化したかどうか評価する．特に，本研究では閲覧行動の多様性の変化を行動変容として分析する．

3.1 閲覧行動の多様性の評価

読者ごとの閲覧記事を，前章のクラスタリングのどのクラスタに属するかにより評価し，読んだ記事のクラスタの類似度により閲覧行動の多様性を評価する．クラスタ間の類似度は，クラスタに含まれる記事を読んだユーザに基づいて，2 つのクラスタ c_i，c_j のユーザ群 U_i，U_j から Simpson 係数を用いて次のように求められる．

$$Sim(c_i, c_j) = \frac{|U_i \cdot U_j|}{min(|U_i|, |U_j|)} \tag{2}$$

これにより，クラスタ毎の類似度がわかる．これを用いて，ユーザ u の閲覧性の多様性は，u が読んだ全ての記事 n 件について，2 つの記事 a_i，a_j の属するクラスタ $c(a_i)$，$c(a_j)$ から以下の式で表される，クラスタ間平均類似度により評価する．

$$f(u) = \frac{\sum_{i,j,i\leq j} Sim(c(a_i), c(a_j))}{{}_{n+1}\mathrm{C}_2} \tag{3}$$

一定期間ごとに閲覧行動の多様性を評価し，多様性に変化が見られた読者について，その特徴を分析する．一般に，フィルターバブルとは推薦システムによりユーザがその人の観点に合わない情報から隔離され，ユーザ自身の興味があると判断された範囲に集約されていくことである．したがって，読んでいる記事のクラスタの類似度でその影響を測ることができると期待される．例えば，もしフィルターバブルによって観点に合う情報のみと接するようになれば，閲覧記事は類似したものが多くなるため，多様性が失われ，$f(u)$ の値は増加すると考えられる．

3.2 実験

分析では，5 月 21〜27 日（期間 1）と 7 月 21〜27 日（期間 2）の行動を比較した．これは，前章で確認した通り，クラスタリング対象期間の後半では，クラスタに含まれる記事が減少するためである．この期間に，それぞれ 10〜100 記事読んでいるユーザを選択し，その中のクラスタに含まれる記事を 10 以上読んでいるユーザから 1000 ユーザをランダムサンプリングする．これらの読者について，前節で定義したクラスタ間平均類似度を求める．

3.3 結果と考察

クラスタ間平均類似度は図 2 のようになった．期間 1 の平均類似度の平均は 0.292, 期間 2 の平均は 0.364 となり，有意水準 1 %で有意に平均類似度は上昇していた．クラスタ間平均類似度は，ユーザの興味に基づいてクラスタリングされた記事のクラスタの類似度によって，ユーザの読んでいる記事の興味の幅を測るものである．つまり，クラスタ間平均類似度が大きいということは類似した記事を読んでいるということで，クラスタ間平均類似度が上昇した場合は興味が狭まり，低下した場合は興味の幅が広がったと言える．よって，この分析から，既存の表示システムでも，期間が進むにつれ，興味が偏っていくと考えられる．これには，ランキングやあなたへのおすすめという推薦システムが導入されていることが影響していると推測される．

4. 読者の行動変容を導く特徴の分析

ここでは，前章で得られた期間 1，2 のクラスタ間平均類似度について，ユーザごとの変化を分析する．前章で述べたとおり，全体ではクラスタ間平均類似度は上昇し，つまり，多様性は減少していた．一方，クラスタ間平均類似度が低下し，多様性が増加していた読者もいた．そこで，多様性増加グループと多様性減少グループを比較する．

4.1 対象ユーザと比較する特徴量

多様性増加グループは 1000 読者中 259 ユーザであった．一方，多様性減少グループとして，多様性が減少した読者から増加グループと同数の上位 259 ユーザを選んだ．これらの多様性減少グループに含まれるユーザは，類似度が 68.7 %以上上昇している特に類似度が上昇しているユーザ群で構成される．これらの多様性減少グループと多様性増加グループの間で，期間中の閲覧数，クラスタエントロピーと，行動評価の指標となる期間 1，2 のクラスタ間平均類似度と，ユーザ属性として年代，入会年，性別を比較した．まず，前章で定義したクラスタ間平均類似度の変化と，期間中に読んだ記事数，クラスタの種類の平均情報量（クラスタエントロピー）の関係を分析する．なお，クラスタエントロピーとは，クラスタ間の類似度は考慮せず，読んだクラスタの種類の多様性を測る指標で，p_i を各ユーザにおけるクラスタ c_i の存在確率として以下のように表される．

$$H(u) = -\sum_i p_i \cdot logp_i \tag{4}$$

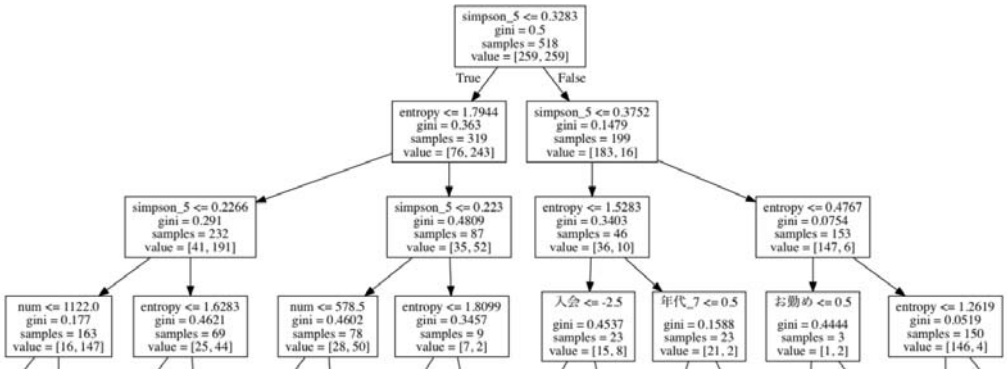

図 3: 期間 1 の平均類似度を含めた決定木分析

表 1: 多様性の変化別の特徴量の平均

多様性	増加	減少	有意水準
閲覧数	536	486	P<0.05
クラスタエントロピー	1.47	1.60	P<0.01
期間 1 の平均類似度	0.416	0.179	P<0.01
期間 2 の平均類似度	0.322	0.409	P<0.01

4.2 結果

各指標の結果を表 1 に示す．表 1 より，多様性増加グループのユーザは記事の閲覧数が多いことがわかることから，多くの記事を読んだユーザは興味が広がると考えられる．一方，多様性増加グループのクラスタエントロピーは低い．このことから，単純に異なるクラスタの記事を読むだけでは，3.1 節で定義したクラスタ間平均類似度で測れる多様性は増加しないことがわかる．これは，クラスタにはそれぞれ類似度があり，類似度の高いクラスタと低いクラスタでは意味が異なるためと，クラスタに含まれる記事の数に差異があるためと考えられる．

期間 1，2 のクラスタ間平均類似度に関しては，平均類似度が低下した多様性増加グループは期間 1 がもともと高く，平均類似度が上昇した多様性減少グループは期間 2 が低いという結果が得られた．

ユーザ属性に関しては，性別と入会年には有意差が認められず，年代については，会員数の少ない 70 代と 20 代以下以外は有意差が見られなかった．

4.3 決定木分析

上記の多様性増加・減少グループについて，特徴量として前節と同様の期間中の閲覧数，クラスタエントロピーと，行動評価の指標となる期間 1 のクラスタ間平均類似度と，ユーザ属性として年代，入会年，性別を用いて決定木分析を行った．ただし，期間 2 の平均類似度に関しては，実際の推薦では，推薦した後に推薦の影響を測定するために用いるもので，推薦の前には測定不可能なため除いた．交差検定を行った結果は，77.2 %の精度で予測できた．この時の決定木を図 3 に示す．value の 1 項目は多様性増加グループ，2 項目は多様性減少グループである．この図から，期間 1 のクラスタ間平均類似度の影響が大きいことがわかる．これは，前節で考察したように，期間 1 のクラスタ間平均類似度が低いものは上昇し，高いものは低下するという傾向があるということだと考えられる．

そこで，期間 1 の平均類似度以外の特徴量の予測への寄与を調べるため，期間 1 の平均類似度を除いた分析もしたところ，精度 64.1 %で予測できた．この時の決定木を図 4 に示す．これより，中心の二つの分岐に注目すると，クラスタエントロピーが 1.5889 より低く閲覧数が 531 より多いものは多様性増加グループの割合が多く，クラスタエントロピーが 1.5889 から 1.8593 の間にあり，閲覧数が 1028 以下のものは多様性減少グループの割合が多いことがわかり，大きくは前節の考察と一致し，閲覧数が多いと多様性が増加し，興味の幅が広がりやすいと言える．

表 2: 多様性増加グループに所属するユーザが多く閲覧したクラスタ

多様性増加グループ [%]	多様性減少グループ [%]	クラスタの主な内容
72.6	57.5	短信
66.8	51.4	市場速報
59.1	44.8	米国政治＋医療
30.1	14.7	経済指標
30.1	14.7	科学技術

4.4 特徴的なクラスタの分析

次に，特定のクラスタの記事を読んだ読者は多様性に変化が生じるという仮定のもと，多様性増加・減少グループの一方に多く含まれるクラスタについて分析する．表 2，3 に各グループに多く含まれるクラスタを抜粋した．多様性増加グループに多いクラスタには，国内の政治問題は多くは含まれておらず，米国政治についてもトランプ大統領に関するものであった．また，科学技術や経済情報，市場情報が多く含まれた．一方，多様性減少グループに多いクラスタには政治問題や，企業に関する情報が多く含まれていた．このことから，政治に興味があることと，興味の幅を広げることには，負の相関関係が存在する可能性が示唆された．

4.5 考察

多様性増加グループは少ないが，そのようなグループのユーザは他のユーザに比べて，多くの記事を読んでいることがわかった．その上で，読んでいるクラスタの種類が多いことは興味を広げることにはつながらないことがわかった．これは，[園田 2018] のクラスタリング手法では，クラスタ間に類似度の高いクラスタとそうではないクラスタが生成され，また，それぞれのクラスタに含まれる記事の数に差異があることによる

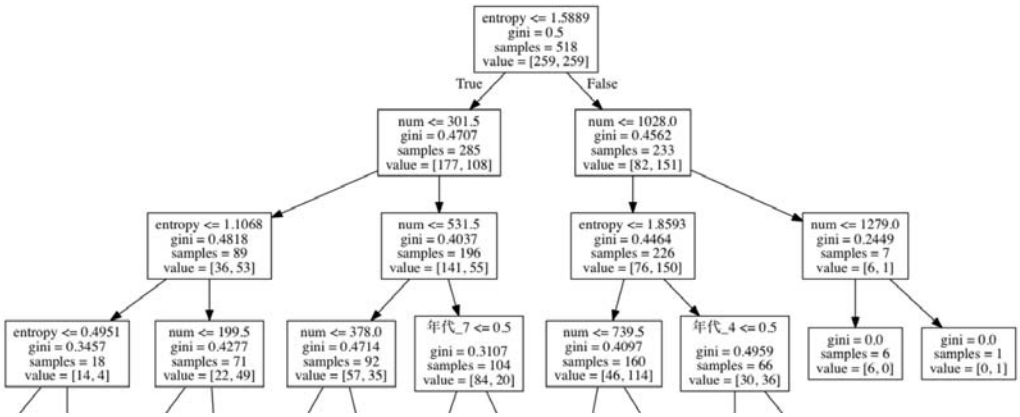

図 4: 期間 1 の平均類似度を除いた決定木分析

表 3: 多様性減少グループに所属するユーザが多く閲覧したクラスタ

多様性増加グループ [%]	多様性減少グループ [%]	クラスタの主な内容
29.3	43.6	IT 企業
19.3	34.7	ヤフー
18.1	33.6	アジア政治＋交通
16.2	32.4	米国政治
8.1	22.4	加計学園問題

と考えられ，推薦では，クラスタの類似度等も考慮する必要があると考えられる．また，ユーザ属性と閲覧数，ユーザエントロピーのみでその後の興味が広がるかどうかを一定程度予測することができることがわかった．しかしながら，期間の開始時点でのクラスタ間平均類似度の情報がある場合は 77%と比較的高い精度で予測できるが，この情報がない場合は，64%と精度が下がることから，ユーザの行動の変化を説明する際に，年齢・性別・職業のようなユーザ属性や閲覧数だけで説明できる部分は限定的であるといえる．このことから，ユーザ属性等に関わらず，行動変容を起こさせることも可能であると期待される．

また，多様性増加グループと減少グループには閲覧割合の異なるクラスタがあり，その中でも政治に関するクラスタなどは，多様性減少グループに所属するユーザが多く読んでいることがわかった．これは，政治に関する記事は同時に関連する記事が多く出されることで，話題に興味を持ったユーザが続けて関連する記事を読み，ほかの話題に興味が向きづらくなるのではないかと推測される．

5. 結論

本研究では，日経電子版の記事について，ユーザの閲覧記録に基づいて，類似度ネットワークを構築し，ユーザの興味に基づいた記事をクラスタリングし，それぞれのクラスタ間の類似度の平均によりユーザの行動の多様性を評価し，2 つの期間の行動を比較し，行動変容を導く特徴について分析した．これらの結果から，閲覧数やクラスタの種類，ユーザの年齢が行動変容に影響を与えていることが確認できた．しかしながら，それらの影響は限定的で，表層的な属性では行動変容の可能性はさほど限定されないこともわかった．また，多様性増加グループと減少グループでは，読んでいるクラスタの内容も異なることがわかった．

今後の課題は，行動変容を導く要因のさらなる分析，これらの結果を利用した推薦システムの開発などが挙げられる．本研究結果から，ユーザの興味は一定期間を経て，多くの記事を読むことで変化するということが確認された．そこで，推薦システムの開発においては，多くの記事が読まれ，クラスタ間の距離を考慮した多様な記事を推薦することで，フィルターバブルの発生を抑制し，多様な意見に触れる機会を提供できるような推薦システムの開発が可能であるか実験することで，行動変容を導く要因のさらなる分析が可能になると考えられる．

参考文献

[Pariser 2011] Pariser, Eli. The filter bubble: What the Internet is hid-ing from you. Penguin UK, 2011.

[Nguyen 2014] Nguyen, Tien T., et al. ”Exploring the filter bubble: the effect of using recommender systems on content diversity.” Proceedings of the 23rd international conference on World wide web. ACM, 2014.

[園田 2018] 園田　亜斗夢, 鳥海 不二夫, 中島 寛人，郷治　雅，ネットワーク構造に基づく新聞記事の分類，第 15 回社会システム部会研究会，2018.

[松尾 2005] 松尾豊, et al. ”Web 上の情報からの人間関係ネットワークの抽出.” 人工知能学会論文誌 20.1 (2005): 46-56.

[Blondel 2008] Blondel, Vincent D., et al. ”Fast unfolding of communi-ties in large networks.” Journal of statistical mechanics: theory and experiment 2008.10 (2008)