ニュースの多様性に着目した中長期的な行動変容の分析

園田亜斗夢† 関 喜史†† 鳥海不二夫†††

† 東京大学 東京都文京区本郷 7 − 3 − 1
†† 株式会社 Gunosy 東京都港区赤坂 1-12-32 アーク森ビル
††† 東京大学 東京都文京区本郷 7 − 3 − 1
E-mail: †sonoda@crimson.q.t.u-tokyo.ac.jp, ††s.yoshi.mobile.1123@gmail.com, †††tori@sys.t.u-tokyo.ac.jp

あらまし 電子媒体で発信される情報量が増加し，推薦サービスの導入も進んでいる．その中で，過度な推薦により，ユーザに偏った情報のみを提供するフィルターバブルやエコーチェンバーなどの問題が存在するとの指摘もある．我々はこれらのユーザ行動をログデータから定量的に評価することを目指している．これまでに，記事のカテゴリの多様性に基づきユーザの行動変容を議論してきた．本稿では，より長期的なデータに対し分析することで，記事のカテゴリの多様性は増減を繰り返すことを示した．また，記事のカテゴリ情報に加え，記事の人気や参照元の情報を用いることで，ユーザ行動の変化をより詳細に捉えられることを示した．

キーワード online news, diversity changes, オンラインニュース，推薦システム，時系列分析，多様性

1 はじめに

インターネットの登場により新聞社等のメディアの発信するニュースの媒体が新聞から web サイトに拡大し [1]，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．このような状況下で，メディアを通した情報接触において人々は何らかの基準で閲覧する情報を取捨選択せざるを得なくなっていると推測される．広義の推薦システムは，記事を選択する際に必要な時間と労力といった負担を減らし，満足度を向上させるため，多くの分野で導入されている．特に，技術の発達によってユーザの選好に応じて提示情報をパーソナライズするタイプの推薦システムの導入が進んでいる．一方で，フィルターバブル [2] やエコーチェンバー [3] という現象が指摘されている．フィルターバブルは過度の推薦によってユーザに偏った情報のみを提供し広い視野が失われる現象であり，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接触することで，偏った考えがより強化される現象である．「見たいものだけをみる」ことが容易な情報環境下では，異なる意見を持つ他者に対する寛容性の低下やマイノリティに対する偏見の増大，フェイクニュースの無批判な受容などによってこの必要条件を満たすことが難しくなると言った指摘があるが [4]，推薦システムの影響について中長期的なデータを用いて分析した研究は少ない．

本研究では既存のニュース配信サービス上でのユーザ行動の違いを，クリックした記事のカテゴリや人気記事，参照元の情報を用いて長期的なデータを用いて分析する．筆者のこれまでの研究ではユーザの閲覧行動がどう変化するかについて，中期的なデータでカテゴリに対する情報エントロピーを用いて議論した [5] [6]．本研究ではユーザの閲覧行動についてより長期的な分析を行い，カテゴリに対する情報エントロピー以外に人気な記事を読む割合や記事の参照元の情報を用いることで，ユーザ行動の変化をより詳細に捉えることを目指す．

表 1 参照元ごとのクリックに占める割合

参照元	クリックに占める割合
トップ画面	0.384
プッシュ通知	0.349
その他	0.267

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スマートフォンアプリケーションの 2018 年 11 月 1 日から 2019 年 9 月 8 日までの約 10ヶ月間のユーザ行動ログを用いる．この中で，2018 年 11 月の登録ユーザについて分析した．これは既存ユーザではメディアに適合しており変化が小さいと推定されることから，新規ユーザに限定することで条件をそろえるためである．また，閲覧数が一定範囲のユーザを抽出した．

本研究では，クリックしたニュース記事に対し，ユーザがサービス画面のどの位置または通知機能からクリックしたかといった情報や，ニュース記事自体の閲覧数を基にした人気や記事のカテゴリを基にした情報エントロピーなどを扱う．ユーザがサービス画面のどの位置または通知機能からクリックしたかといった情報には，多くのパターンが存在するため，本研究では 3 パターンに集約した．表 1 に参照元ごとのクリックに占める割合を示す．ここで，トップ画面とは対象のアプリケーションを起動した際に表示される画面のことであり，プッシュ通知とは定期的およびニュース性のある事象が起こった際にアプリケーションの通知機能を用いてニュースが表示されるものである．この集計から，サービス内の記事のクリックはトップ画面とプッシュ通知が全体の 7 割以上を占めることがわかる．

つまり，トップ画面とプッシュ通知に含まれる記事から多くのクリックはされており，トップ画面とプッシュ通知には推薦システムが導入されているため，多くのクリックは推薦された

記事から選ばれていると言える．

図 1 に記事ごとのクリック数を示す．ここから，多くのユーザにクリックされやすい記事は記事全体に対し少数であることがわかる．分析期間にクリックされた記事は約 90 万記事であるが，クリック数が上位の 1,316 記事で全クリック数のうち 25%を占める．本研究ではこのクリック数上位 1,316 記事を人気記事と定義する．

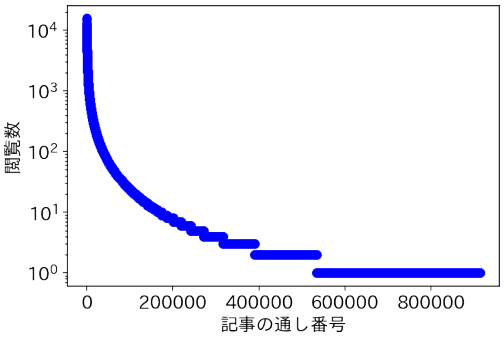

図 1　記事ごとのクリック数

3　情報エントロピーによる分析

3.1　情報エントロピー

本研究ではこれまでの研究た [5] と同様に，クリックした記事の多様性の評価の一つとしてカテゴリの集中度を用いる．カテゴリの集中度はカテゴリエントロピーと呼ぶこととし，情報量エントロピーを用いて次のように求められる．

$$H(u) = -\sum_i p_i \cdot logp_i$$

ただし，p_i は各集合におけるカテゴリ c_i の存在確率で，閲覧数 N のとき，$p_i = c_i/N$ と表されるものである．

カテゴリエントロピーが大きなユーザは，様々なカテゴリに所属する記事を読んでいるユーザであり，カテゴリエントロピーの小さなユーザは，特定のカテゴリに所属する記事を集中的に読んでいるユーザである．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，読んでいる記事のカテゴリの集中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，閲覧記事は特定のクラスタに集中するため，多様性が失われ，カテゴリエントロピーは低下すると考えられる．

3.2　カテゴリ

本研究では，記事をカテゴリによって評価するが，対象サービスではカテゴリを付与しており，文章情報等から教師あり学習とルールベースを組み合わせたもので実現されている．分析では，これらのカテゴリを著者が自身の知識を元に集約したものを用いた．これは，既存のカテゴリの粒度が様々であり，「スポーツ」というカテゴリがある一方で，サッカーには「サッカー」,「サッカー日本代表」,「国内サッカー」,「海外サッカー」と詳細な分類がされるものもあり，同等に扱うことは不適切であると考えられるためである．

表 2　全クリックにおける集約したカテゴリの占める割合

カテゴリ	クリックに占める割合 [%]
社会・政治・経済	42.3
エンタメ	38.5
スポーツ	9.83
コラム	5.05
モノ	1.86
動物	0.557
食	0.403
サブカルチャー	0.0767
テクノロジー	0.0048
不明	1.52

表 2 にカテゴリごとのクリックに占める割合を示す．ここから，カテゴリごとのクリック数は異なり，一部のカテゴリに多くのクリックが集中していることが分かる．

3.3　カテゴリエントロピーの時系列的な変化

ここではカテゴリエントロピーの時系列的な変化を分析し，議論する．分析期間中継続していたユーザに限定し，週ごとに読んだ記事のカテゴリエントロピーを計算した．図 2 に各週ごとに全ユーザのカテゴリエントロピーの平均と出稿された記事全体のカテゴリエントロピー，人気記事のカテゴリエントロピー，プッシュ通知とトップ画面からクリックされた記事のエントロピーを示す．記事全体のカテゴリエントロピーに比べ，

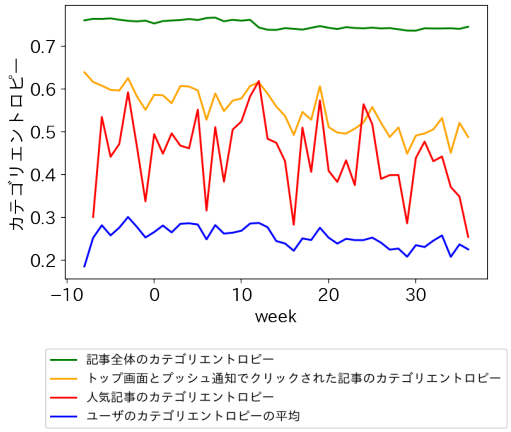

図 2　カテゴリエントロピーの変化

人気記事のカテゴリエントロピーと全ユーザのカテゴリエントロピーの平均，プッシュ通知とトップ画面からクリックされた記事のエントロピーは週ごとに大きく増減する．全ユーザのカテゴリエントロピーの平均と記事全体のカテゴリエントロピーの相関係数は 0.53 であり中程度の相関がある．一方，全ユーザのカテゴリエントロピーの平均と人気記事のカテゴリエントロピーの相関係数は 0.71 であり記事全体に対する相関より強い相関がある．また，全ユーザのカテゴリエントロピーの平均とプッシュ通知とトップ画面からクリックされた記事のエントロピーの相関係数は 0.63 であり記事全体に対する相関より強い相関がある．

このことから人気記事のカテゴリエントロピーの影響や記事

の参照元の影響が大きく，ユーザのクリックした記事のカテゴリの多様性だけを時系列的に比較しても，ユーザの興味の変容を捉えることは難しい．そこで，ユーザの行動を詳細に評価するため，ユーザごとのクリックに含まれる人気記事の割合や参照元の割合を用い評価し，議論する．

4　人気記事と参照元による分析

4.1　人気記事と参照元の割合変化

2 章で定義したように，クリック数が上位の記事を人気記事として扱う．全クリック数のうち上位 10%を占める場合，上位 25%を占める場合，上位 50%を占める場合と人気記事の閾値を変化させたとき，含まれる記事数はそれぞれ 350 記事，1,316 記事，5,978 記事であった．それぞれの定義に対し，横軸に各ユーザの期間中のクリックした記事に対する人気記事の占める割合をとり，縦軸にユーザ数を取ったヒストグラムを図 3 に示す．この図から人気記事の定義が上位 10%から上位 50%と広くなるに従い，人気記事の割合が大きいユーザ数が増えることが分かる．また，人気記事の定義が上位 10%，上位 25%のときは分布が左に偏っており，人気記事だけ読んでいるユーザは少ないことが分かる．このことから，人気記事は多くのユーザが閲覧し少数の記事で多くのクリック数を獲得するが，各ユーザは人気記事だけ読むのではなくそれ以外の記事も読んでいることが分かる．

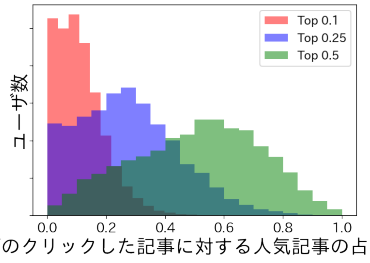

図 3　人気記事の割合ごとのユーザ数

以降では人気記事とは，記事をクリック数の多い順に並べ全クリックのうち上位 25%を占めるクリック数上位の記事として議論する．図 4 は縦軸に人気記事のユーザのクリックに占める割合の平均についてとり，横軸に週を取ったものである．この図から，期間の経過に従い人気記事の割合は減少していくことが分かる．このことから，期間の経過に従い人気記事の影響を受けにくくなっていることが分かる．

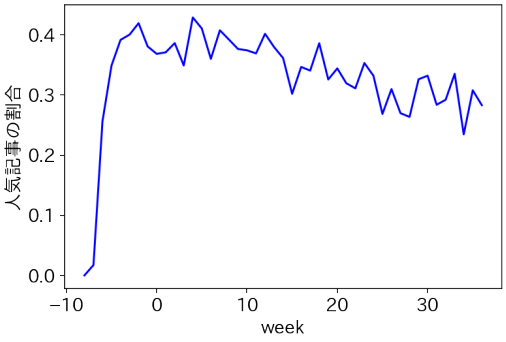

図 4　人気記事のクリックに占める割合の変化

図 5 は縦軸にプッシュ通知のクリックに占める割合をとり，横軸に週を取ったものである．期間の経過に従いプッシュ通知の記事から選択する割合は増加している．これは，サービス利用開始後初期は探索的に多くの記事を読むが，期間の経過に従いサービスの利用方法が定まり探索的な行動が減り，プッシュ通知のクリックの割合が相対的に大きくなることによると考えられる．また，推薦システムの初期段階では，興味関心を推定するための情報が少ないことにより有効な推薦が期待できないというコールドスタート問題 [7] が，期間の経過により解決されプッシュ通知される記事のパーソナライズの精度が向上していることも要因の一つとして推察される．

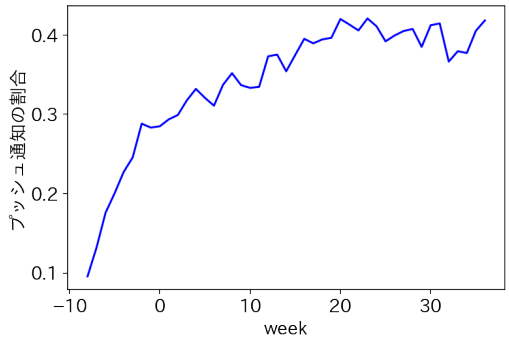

図 5　プッシュ通知のクリックに占める割合の変化

図 6 は縦軸にトップ画面のクリックに占める割合をとり，横軸に週を取ったものである．期間の経過に従いトップ画面の記事から選択する割合は減少している．この結果からも，期間の経過に従いサービスの利用方法が定まり探索的な行動が減っていることが推察される．

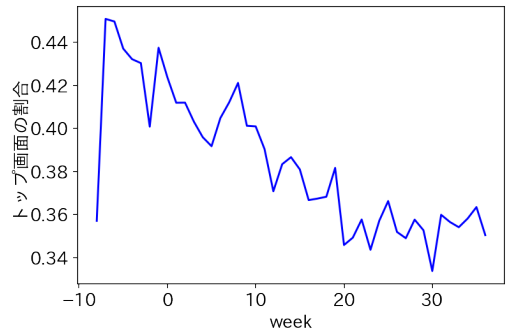

図 6　トップ画面のクリックに占める割合の変化

4.2　人気記事の特徴

表 3 にカテゴリごとの人気記事のクリックに占める割合を示す．表 2 と比較すると，社会・政治・経済とスポーツの占める割合は大きく，他のカテゴリの占める割合は小さいことが分かる．これは社会・政治・経済とスポーツの記事は幅広い属性のユーザに受け入れられやすいことから人気になりやすいためと考えられる．

表 4 に人気記事に対する参照元ごとのクリックに占める割合を示す．ここから，人気記事は 9 割以上がトップ画面かプッシュ通知経由ということが分かり，この値は通常の記事と比較すると大きい．つまり，人気な記事はトップ画面やプッシュ通

表 3　人気記事のクリックにおける各カテゴリの占める割合

カテゴリ	クリックに占める割合 [%]
社会・政治・経済	50.6
エンタメ	32.6
スポーツ	11.0
コラム	2.42
モノ	0.446
サブカルチャー	0.120
食	0.100
不明	2.49

表 4　人気記事の参照元ごとのクリックに占める割合

参照元	クリックに占める割合
トップ画面	0.238
プッシュ通知	0.690
その他	0.072

表 5　人気記事のクリックに占める割合が大きいユーザと小さいユーザの参照元ごとのクリックに占める割合

参照元	クリックに占める割合 (人気記事の割合の大きいユーザ群)	クリックに占める割合 (人気記事の割合の小さいユーザ群)
トップ画面	0.324	0.400
プッシュ通知	0.442	0.351
その他	0.234	0.249

知に含まれやすく，またトップ画面やプッシュ通知に含まれた結果人気記事となるという相互関係が成り立っていることがうかがえる．

4.3　人気記事の割合の大小による分析

人気記事の割合でソートし，人気記事の割合が上位 10,000 ユーザ（人気記事の割合の大きいユーザ群）と下位 10,000 ユーザ (人気記事の割合の小さいユーザ群) を抽出し，それぞれの特徴を分析する．表 5 に各ユーザ群ごとの参照元ごとのクリックに占める割合を示す．人気記事の割合の大きいユーザ群の方がプッシュ通知の割合が大きいことが分かる．

人気記事の割合の大きいユーザ群と人気記事の割合の小さいユーザ群の閲覧数に対し t 検定を行うと有意水準 1%で人気記事の割合の小さいユーザ群の方が閲覧数が多いことが分かった．また，カテゴリエントロピーの平均はそれぞれ 0.752，0.762 で，t 検定を行うと有意水準 5%で人気記事の割合の小さいユーザ群の方がカテゴリエントロピーが大きいことがわかった．しかし，それぞれの群で閲覧数とカテゴリエントロピーの相関係数は 0.27，0.20 と中程度の相関があることから，人気記事の割合の小さいユーザ群の方がカテゴリエントロピーが大きいのは閲覧数の影響の可能性もある．そこで，閲覧数の上限と下限を設定し，閲覧数の平均が一致するようにユーザを抽出した．このとき元の半数以上のユーザが残るように操作した．このとき，カテゴリエントロピーの平均はそれぞれ 0.756，0.771 で，t 検定を行うと有意水準 5%で人気記事の割合の小さいユーザ群の方がカテゴリエントロピーが大きいことがわかった．このことから，人気記事の割合が小さいユーザはより多様なカテゴリの記事をクリックしていることが分かった．

5　離 脱 予 測

5.1　離脱の定義

ここではこれまでに分析した記事やユーザの特徴を用いて，ユーザがサービスを離脱するか否かを予測することができるか検証する．サービスの離脱予測はサービス提供者にとって重要な情報であるが，ここまでの分析で用いた特徴量を用いて予測精度を向上させることができるのであれば，それらの特徴量はユーザ行動を特徴付ける変数としての意義を物と言える．

人気記事の割合やカテゴリエントロピーの値は週ごとに変化するため，入会週を固定し分析する必要がある．そこで，2018 年 11 月 5 日〜11 日に入会したユーザを抽出し，2019 年 1 月 1 日の週時点で継続しているユーザに限定し，それまでの情報を用いて離脱を予測する．モデリングにおいては離脱と継続のユーザ数は離脱したユーザ数に合わせ，ダウンサンプリングする．離脱は 2019 年 1 月 28 日以降のログデータが存在しないユーザを離脱として定義する．

5.2　結　　果

週ごとのクリック数，人気記事の割合，カテゴリエントロピー，プッシュ通知の割合，カテゴリの割合を特徴量として予測すると正解率が 62.3%となった．一方，週ごとのクリック数のみで予測した場合は正解率が 55.3%となった．このことから，週ごとのクリック数，人気記事の割合，カテゴリエントロピー，プッシュ通知の割合，カテゴリの割合を特徴量として用いた場合はクリック数のみで予測した場合より予測精度が向上することがわかった．つまり，人気記事の割合，カテゴリエントロピー，プッシュ通知の割合，カテゴリの割合はユーザのサービス継続率に影響を与えていることが分かることから，これらの指標を用いてユーザ行動を分析することは意味があると言える．

6　結　　論

株式会社 Gunosy が提供するニュース配信スマートフォンアプリケーションのユーザ行動について，クリックした記事のカテゴリや人気記事，参照元の情報を用いて長期的なデータを用いて分析した．筆者のこれまでの研究では，ユーザ行動について中期的なデータを用い記事のカテゴリのエントロピーに基づいて議論したが，今回はより長期的なデータに対し人気記事や参照元の情報を追加で用いることで，ユーザ行動の変化をより詳細にとらえることを目指した．カテゴリエントロピーによる分析では，ユーザのカテゴリエントロピーの変化は長期的にも安定することなく増減を繰り返し，人気記事の影響を受けることを示した．また，ユーザのクリックに占める人気記事の割合や参照元の占める割合の時系列的な変化を分析することで，サービス利用開始後初期は探索的に多くの記事を読むが，期間の経過に従いサービスの利用方法が定まり探索的な行動が減ることが確認できた．

今後は，閲覧記事だけでなく表示されたが読まれなかった記

事と表示されなかった記事についても分析を進めていきたい．また，記事のカテゴリをクリックしたユーザに基づき分類することで，押されやすい記事の分析や興味の偏りの分析を行いたい．

文　献

[1] 総務省. 平成 28 年情報通信メディアの利用時間と情報行動に関する調査報告書. http://www.soumu.go.jp/main_content/000492877.pdf.

[2] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[3] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[4] 小林哲郎 et al. ソーシャルメディアと分断化する社会的リアリティ (特集 twitter とソーシャルメディア). **人工知能学会誌**, 27(1):51–58, 2012.

[5] Atom Sonoda, Fujio Toriumi, Hiroto Nakajima, and Miyabi Gouji. Analysis and modeling of behavioral changes in a news service. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 73–80. IEEE, 2018.

[6] Atom Sonoda, Yoshifumi Seki, and Fujio Toriumi. Analysis of factors that affect users ' behavioral changes in news service. In *IEEE/WIC/ACM International Conference on Web Intelligence-Volume 24800*, pages 35–42. ACM, 2019.

[7] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260. ACM, 2002.