クリックとインプレッションに基づいたニュース記事の分類

Classification of news articles based on impressions and clicks

園田 亜斗夢 [*1]　関 喜史 [*2]　鳥海 不二夫 [*1]
Atom Sonoda　Yoshifumi Seki　Fujio Toriumi
[*1] 東京大学　The University of Tokyo
[*2] 株式会社 Gunosy　Gunosy Inc.

要旨: フィルターバブルやエコーチェンバーといった現象が社会課題となっている．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，記事のカテゴリやタイトルの言語情報から得られるベクトルの多様性に基づきユーザの行動変容を議論してきたが，フィルターバブル等の課題の本質的な理解のためには，ユーザの興味に基づいた記事の理解が重要であると考えている．本稿では，情報の分類手法として，ユーザの利用ログに基づく情報の分類法をニュース記事に対し適用した．また，単純なクリックログだけでなくインプレッションログを利用することで，より正確に記事の分析を行えることを示した．

キーワード: オンラインニュース，トピック分類，インプレッション

Abstract: The phenomenon of filter bubbles and echo chambers has become a social issue. Our goal is to quantitatively evaluate these behaviors from log data. So far, we have discussed users' behavioral changes based on the diversity of article categories or vectors obtained from linguistic information of the titles. However, we believe that it is important to understand articles based on users' interests in order to essentially understand issues such as filter bubbles. In this paper, we applied the information classification method based on the user's usage log to news domain. We also showed that using the impression logs as well as click logs allows for more accurate analysis of the news articles.

Keywords: online news, topic classification, impression logs

1　はじめに

偏ったアイテムのみを消費することで特定の主張を真実として捉える状態と定義できるフィルターバブル [Pariser 11] やエコーチェンバー [Jamieson 08] といった現象が社会課題となっている．これらの問題を解決することは，社会性，事業性両面で重要な意味を持つ．例えば，大統領選挙における情報の流通過程では，支持する政党の主張のみを真実として捉えることで，陰謀論やそれに伴う暴動などが問題になっている．また，反トラスト法に反するような対抗措置なども正当化されているなど，法令遵守の観点でも深刻な状況である．事業的には，多様な主張

を持ったユーザに支持されることで収益力を高めることができ，また，ユーザが多様な意見を受容できる状態にあることは，新たな情報を発信した際に興味を集めやすくメディアの事業展開の可能性が広がる．

著者らはこれまでにも既存のニュース配信サービス上でのユーザ行動について，既存のカテゴリやタイトルのテキスト情報に基づき情報エントロピーを用いて多様性について議論した [Sonoda 18][Sonoda 19]．一方で，エコーチェンバーやフィルターバブルといった問題を解決するという観点では，支持する主張以外の情報に対する受容性を評価し，受容性が低い状態から受容性を高めるために必要な条件を検討することは重要である．しかし，ニュースが提供する情報について，特定の主張の情報のみに接しているかどうかを評価する際に，既存のカテゴリやテキスト情報のみを用いることは検証できる内容に限界がある．なぜなら，同様のカテゴリや内容の記事でも，記者のスタンスや文調によって読者の興味を集めるかどうかは変わり，発信者の性別によっても変わる場合もあるからである．

そこで，本研究では既存のカテゴリやテキスト情報を用いずに，ユーザの興味を反映した情報の分類方法として，ユーザの利用ログに基づく情報の分類法をニュース記事に対し適用する．また，単純なクリックログだけでなくインプレッション（表示）ログを利用し，記事が表示されたかどうかの影響を考慮した分析を行う．さらに，得られたネットワーク，クラスタを分析し，本手法によって構築されたネットワーク構造と分類された情報群の特徴を明らかにする．

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 12 月 5 日から 1 週間のユーザ行動ログとニュース記事の表示履歴を用いる．分析対象のデータは，対象期間の 1ヶ月前の一定期間中に入会したユーザに限定し抽出した．これは，入会時期の影響を受けないようにするためである．対象データには約 2,400 万件のクリックログが含まれる．対象のアプリケーション上でユーザはニュース記事の一覧を縦方向にスクロールし，気になったニュース記事をクリックすることで本文を読むことができる．

本研究では，クリックしたニュース記事とユーザに表示されたニュース記事の履歴（インプレッションログ）に対し，ニュース記事ごとのユーザの重複率を基にした分析を行う．

3　ネットワーク構造を用いたニュース記事の分類手法

3.1　二部グラフによるネットワークの構築

本研究では，記事の分類には馬場らが提案した [Baba 15, Uchida 17]Tweet を分類する手法を基に，言語情報を利用することなしに，ユーザの行動ログ，つまり記事を閲覧したか否かと表示されたか否かのみに注目し，ニュース記事を分類する手法を提案する

ある二つの記事を同時に閲覧したユーザが複数人いた場合，二つの記事は共通した興味を持たれる内容を有していると考えられる．つまり，記事を閲覧した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．そこで，閲覧したユーザの重複度から記事をクラスタリングすることで，ユーザの興味に基づいたクラスタを得ることが期待される．それぞれのクラスタに含まれる記事の内容から，アプリケーションを利用するユーザの興味の特徴を明らかにできると考えられる．このように，ユーザの重複度のみを利用してクラスタリングを行うことで，記事の既存のカテゴリや言語的なクラスタリングでは得られない「興味を持つユーザの類似度」によって記事を分類することが可能である．

二つの記事 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．

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

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 [Matsuo 05]．

3.2 記事ネットワーク

ここでは，前述の類似度 $Sim(a_i, a_j)$ が閾値 th 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．実際には，ユーザのクリックログとインプレッションログの二つのデータを用いて，ログ数が上位の記事についてそれぞれ類似度 $Sim(a_i, a_j)$ を算出した．ログ数が上位の記事ペアに限定したのは，計算量が $O(N^2)$ のオーダーで増えるためであるが，記事ペアを限定するに当たり全ログ数の 85%以上が含まれるように閾値を設定した．これにより，期間中にクリックされたユニークな記事のうち 10%ほどがネットワークに含まれた．これは，記事のクリック数や表示回数は人気なものに偏りが存在することによる．

クリックはユーザの興味を反映していると考えられる．インプレッションは，ユーザに表示されたログである．ニュースアプリケーションは，ニュースというコンテンツの特性上，アイテムの入れ替わりが激しい．また，本研究の対象アプリケーションはカテゴリごとに表示タブ（画面）が分かれている．このような環境下では，インプレッションの類似度が大きいということは，同時刻や同様の時間帯にアクセスしている可能性が高いことや，同様のカテゴリに含まれる記事であることを意味する．ユーザが記事をクリックするには，ユーザに記事が表示されていなければならない．つまり，インプレッションの類似度が高い記事はインプレッションの類似度が低い記事に比べ，クリックの類似度が高くなる可能性が高いという仮説が立つ．そこで，クリックログの類似度からインプレッションログの類似度を引いた類似度を算出し，新たな類似度を定義した．具体的には，クリックの類似度を Sim_{click}，インプレッションの類似度を Sim_{imp} とすると次のように求められる．

$$Sim_r = Sim_{click} - r \cdot Sim_{imp} \quad (2)$$

このとき，r はインプレッションの類似度 Sim_{imp} をどれだけ考慮するかのパラメータである．本実験では，$r = 1$ のときと，対象記事群のクリック率の中央値 k を用いて $r = k$ としたときの二通りを試した．

ネットワークを構築するにあたり，関係性の少ない二つの記事，すなわち Simpson 係数の小さいリンクの影響を排除するため，前述の類似度 $Sim(a_i, a_j)$ が上位 N_{th}% のリンクのみを抽出した．ここで，N_{th} の取り方によってネットワークの構造が変化する．また，ほかの記事とリンクで結ばれていない記事すなわち独立したノードは，今回の分析対象から除外した．本研究では，$10^{-7} \leq N_{th} \leq 1.0$ で変化させた．

3.3 クラスタリング

ここで，上記の得られたネットワークについてコミュニティ抽出を行い，記事の類似性に基づいたクラスタを獲得する．そのために，上述の得られた重みつきネットワークに対し，ネットワーククラスタリングの手法を適用する．クラスタリング手法としては，モジュラリティを基準とする Louvain 法 [Blondel 08] を用いた．モジュラリティとは各クラスタの結合度合いを表す指標であり，コミュニティ間のリンクが疎であるほど高い値を出す指標であり，モジュラリティを最大化することで，結合度の高いコミュニティを抽出する．

4 表示と選択の関係

本研究で適用する情報の分類手法は，閲覧したユーザの重複度が高い記事は関係のある情報であるという仮定に基づいている．一方で，アプリケーションの設計上，表示されていない記事

はユーザに気づかれず，クリックが行われることはない．ニュースアプリケーションは，ニュースというコンテンツの特性上，アイテムの入れ替わりが激しく，数時間で表示される記事は更新されている．つまり，クリックログの類似度はインプレッションの影響を受けていることが推察される．

そこで，クリックの類似度とインプレッションの類似度の関係をネットワークの特徴とクラスタに含まれる記事の内容によって評価する．

4.1 ネットワークの特徴

クリックの類似度 Sim_{click}，インプレッションの類似度 Sim_{imp}，インプレッションを考慮したクリックの類似度 $Sim_{r=1}$，$Sim_{r=k}$ の四つの類似度について，それぞれ，類似度の抽出パラメータ N_{th} を $10^{-7} \leq N_{th} \leq 1.0$ で変化させ，ネットワーク構造の変化を確認した．

4.1.1 モジュラリティ

ここでは，ネットワークにおけるクラスタ化を評価するために，モジュラリティQ を用いた．これは，適切なパラメータを選び，モジュラリティQ が大きい，すなわち結合度の高いコミュニティが抽出できたとき，記事の分類が適切に行われたと考えらえれるためである．

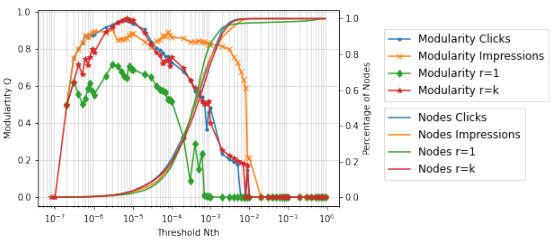

図 1: N_{th} による各類似度のモジュラリティQ の変化と対象のネットワークに含まれるノード数の変化．N_{th} を大きくすることで，含まれるノード数は増加するが，モジュラリティQ は $N_{th} = 10^{-5}$ 付近で最大値を取ることがわかる．

各パラメータにおけるモジュラリティQ の変化を図 1 に示す．これより，どの類似度についても，$N_{th} = 10^{-5}$ 付近で最大値を取ることがわかる．特にクリックの類似度ネットワークは，$N_{th} = 6 \cdot 10^{-6}$ のとき $Q = 0.960$ と最大値となり，また，類似度 $Sim_{r=k}$ のネットワークについても $Q = 0.959$ となった．一方で，インプレッション類似度 Sim_{imp} ネットワークは $10^{-6} \leq N_{th} \leq 10^{-2}$ の範囲で，ほぼ一定のモジュラリティQ の値となり，最大値がわかりにくい分布となった．$N_{th} = 6 \cdot 10^{-6}$ のとき，抽出されたネットワークに含まれる記事数は，$N_{th} = 1.0$ のときに含まれる記事数の 2.3%であった．

また，本研究の応用先であるユーザ行動の評価のためには分類される記事の網羅性も重要である．そこで，ノード数の減少が起きない最大の N_{th} について調査したところ，$N_{th} = 3 \cdot 10^{-2}$ のとき，$N_{th} = 1.0$ のときに含まれる記事数と同等となった．このとき，モジュラリティQ は 0 と低く，クラスタ化が弱いネットワークであり，このネットワークから，提案手法を用いて記事の分類を行うことは困難であると考えられる．

4.1.2 類似度

クリックの類似度に与えるインプレッションの影響を評価するために，クリックの類似度 Sim_{click} とインプレッションの類似度 Sim_{imp} について，相関分析を行った．全ての記事ペアのクリックの類似度 Sim_{click} とインプレッションの類似度 Sim_{imp} について，相関係数を求めたところ，0.717 となった．次に，クリックの類似度 Sim_{click} を目的変数，インプレッションの類似度 Sim_{imp} を説明変数として単回帰分析を行うと，

$$Sim_{click} = 0.271 \cdot Sim_{imp} - 0.019 \qquad (3)$$

という回帰直線が得られた．

各類似度におけるリンクの数の分布を図 2 に示す．全体として，インプレッションの類似度 Sim_{imp} が大きくなるほどクリックの類似度 Sim_{click} が大きくなるという関係性が見て取れる．一方で，インプレッションの類似度 Sim_{imp}

がそれほど大きくなくとも，クリックの類似度 Sim_{click} が大きいという特徴的なリンクも存在することがわかる．

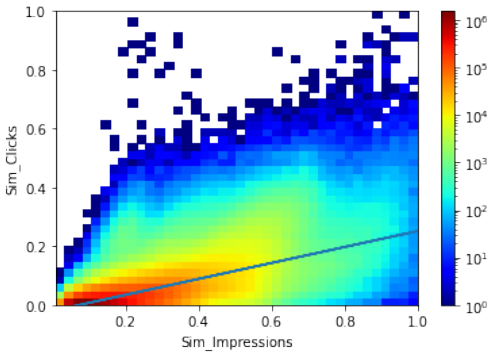

図 2: クリックの類似度 Sim_{click} とインプレッションの類似度 Sim_{imp} の分布．カラーバーはリンクの数の密度を表す．

ここで，クリックの類似度が $Sim_{click} \geq 0.75$ のリンクについて抽出しリンクで繋がれた記事ペアのそれぞれのクリック数を確認すると，クリック数が多い人気な少数の記事とクリック数が少ない多様な記事のペアであった．各記事ペアについて，クリック数が多い記事群と少ない記事群に分けた際に，クリック数の中央値は対象記事全体の中央値と比較して，クリック数が多い記事群の中央値は 10 倍以上であり，少ない記事群は同程度であった．つまり，クリック数の多い人気記事は異なる興味の記事を結びつける効果がある．今後，人気記事で紐付けられたクリック数が少数の記事同士の関係性を分析することで，興味の変容に与える人気記事の効果をより詳細に捉えられることが期待される．

4.2 クラスタに含まれる情報

ここでは，$N_{th} = 6 \cdot 10^{-6}$ のときと $N_{th} = 3 \cdot 10^{-2}$ のときに得られるクラスタに含まれる記事の内容について議論する．

$N_{th} = 6 \cdot 10^{-6}$ のとき，クリックの類似度 Sim_{click} とインプレッションを考慮したクリックの類似度 $Sim_{r=k}$ から得られたクラスタは，それぞれ 38 クラスタあった．内容を目視確認すると，類似度 Sim_{click} と類似度 $Sim_{r=k}$ で，同じ ID の記事がそれぞれのクラスタにまとめられていることが確認できた．リンク数の制限を非常に強く行った場合，ネットワークに含まれる記事数が非常に少なくなり，クリックの類似度 Sim_{click} とインプレッションを考慮したクリックの類似度 $Sim_{r=k}$ のそれぞれのネットワーク構造の差異が小さくなることで，そこから得られるクラスタも同じようなものが得られていると考えられる．

表 1: $N_{th} = 6 \cdot 10^{-6}$ のときの $Sim_{r=k}$ のクラスタの内容

No.	トピック
0	大相撲の暴行事件の人気記事 1 件により紐付けられた雑多な記事
1	大相撲の暴行事件に関する情報とそれに関する著名人のコメント
2	Twitter の画像つき投稿で構成されるまとめ記事
3	在日米軍や犯罪等に関わる社会性の高いニュース
4	自動車会社会長の退任に関する情報

一方で，インプレッションを考慮したクリックの類似度のパラメータが $r = 1$ の $Sim_{r=1}$ の場合は，クリックの類似度 Sim_{click} や $Sim_{r=k}$ とは異なる結果が得られた．得られたクラスタは，8 クラスタあった．内容を目視確認すると，クリックの類似度 Sim_{click} や $Sim_{r=k}$ から得られたクラスタよりよくまとめられていた．そのうち記事数が十分にあったクラスタについて内容を確認したところ，表 1 のようなトピックのニュースがまとめられていた．No.2 のようなクラスタは，タイトルは全て異なるが，まとめ記事という観点で共通していた．また，No.1,3,4 のクラスタは内容は類似したニュースであるが，タイトルに含まれる事件の登場人物は異なったりしていたが，類似したニュースであった．このようなクラスタは，既存のカテゴリやタイトルの言語情報だけではまとめることは難しいと判断できる．ただし，No.0 のクラスタは，大相撲の暴行事件に関する速報記事のクリック数が非常に大きく，その記事に接続されるまとまり

のない記事が集約されていた．このような分類を防ぐために，クリック数が非常に大きい記事はネットワークの構築の対象から外すことは今後の課題である．

$N_{th} = 3 \cdot 10^{-3}$ のとき，類似度 Sim_{click} と類似度 $Sim_{r=k}$ から得られたクラスタは，それぞれ 5 クラスタあった．内容を目視確認すると，$N_{th} = 6 \cdot 10^{-6}$ のときと異なり，類似度間で共通したクラスタが作成されるというような対応関係が見られなくなっていた．また，対象記事の約半数が含まれる大きいクラスタが存在し，解析を行うにあたってクラスタの粒度が十分ではなかった．この傾向は，他の類似度の場合も同様であった．リンク数の制限が弱く，ネットワークに含まれる記事数の減少がないとき，4.1.1 項で確認したように，モジュラリティQ の値が低く，うまくコミュニティ抽出が行えていないことが確認できた．

以上より，モジュラリティを最大化するようにリンク数の制限を非常に強く行ったほうが記事の分類が上手く行えていることが確認できた．その際，ネットワークを構築するリンクには類似度 $Sim_{r=1}$ を用いたほうが良いこともわかった．

5 結論

本研究では，ニュースアプリケーション上でのクリックログとインプレッションログを用いたネットワークの関係性について分析を行い，また，二部グラフとコミュニティ分類の手法を用いてニュースアプリケーション内で提供されるニュース記事の分類を行った．

記事ペア間のクリックの類似度とインプレッションの類似度の間には強い相関がある一方で，インプレッションの類似度が小さくてもクリックの類似度が大きいリンクが存在し，そのようなリンクはクリック数を多く集めている記事に接続していることが確認された．このことから人気な記事は，普段確認するタブが異なり目に触れる記事が違うユーザでも共通してクリックされていると推察され，興味が異なるユーザから広く興味を集めていると考えられる．

また，得られたネットワークに対しコミュニティ抽出を行ったところ，カテゴリをまたいだり，言語情報の類似度は低いような情報群がクラスタに分類できていることが確認された．つまり，ユーザの重複度のみを利用してクラスタリングを行うことで，記事の既存のカテゴリや言語的なクラスタリングでは得られない「興味を持つユーザの類似度」によって記事を分類することが可能であるといえる．

本手法を応用することで，ユーザの興味を測定し，行動の変容を分析することが可能になると期待される．具体的には，本手法を拡張し，クラスタに含まれる記事の網羅性の向上と追加された記事のクラスタリングの対応を行い，ユーザの興味の変容を評価することが今後の課題である．

参考文献

[Baba 15] Baba, S., Toriumi, F., Sakaki, T., Shinoda, K., Kurihara, S., Kazama, K., and Noda, I.: Classification method for shared information on twitter without text data, in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1173–1178ACM (2015)

[Blondel 08] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of statistical mechanics: theory and experiment*, Vol. 2008, No. 10, p. P10008 (2008)

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

[Matsuo 05] Matsuo, Y., Tomobe, H., Nakashima, H., Ishizuka, M., et al.: Social network extraction from the web information, *Transactions of the Japanese Society*

for Artificial Intelligence vol. 20, pp. 46–56 (2005)

[Pariser 11] Pariser, E.: *The filter bubble: What the Internet is hiding from you*, Penguin UK (2011)

[Sonoda 18] Sonoda, A., Toriumi, F., Nakajima, H., and Gouji, M.: Analysis and Modeling of Behavioral Changes in a News Service, in *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 73–80IEEE (2018)

[Sonoda 19] Sonoda, A., Seki, Y., and Toriumi, F.: Analysis of Factors that affect Users' Behavioral Changes in News Service, in *IEEE/WIC/ACM International Conference on Web Intelligence-Volume 24800*, pp. 35–42ACM (2019)

[Uchida 17] Uchida, K., Toriumi, F., and Sakaki, T.: Evaluation of retweet clustering method classification method using retweets on Twitter without text data, in *Proceedings of the International Conference on Web Intelligence*, pp. 187–194ACM (2017)