AKIBA Lab

Akihabara

CyberAgent

秋葉原ラボ技術報告

2

Volume.

Akihabara Lab

巻頭言「新しい時代に向けて」

秋葉原ラボ 研究室長 福田 一郎

技術報告 vol.2

「平成最後の○○」という言葉を頻繁に耳にする．平成という時代が終わり，新しい時代がはじまるという風潮である．改元を 5 月に控えているが，この文章を書いている時点では，まだ次の元号は発表されていない．何という元号になるのかは気になるところであるし，システム開発を多少なりとも知っている身からしてみると，早く発表してほしいという人の気持ちもよく分かる．新元号が発表されれば，なおさら新しい時代の到来を感じるようになることだろう．

よく考えてみると，改元が直接的に影響するのは日本国内だけであり，国内においても多くの場面で西暦を使用しているため，生活に大きな変化があるということはないであろう．しかし，実質的な変化がなくとも何かしらの区切りがあると，その前後で状況を比較したり，気持ちや気分が切り替わるような感じがしたりするのは不思議なものである．

私達はインターネットおよびソフトウェアを中心とする技術領域を担当しているが，その進歩と変化は他の技術領域に比べて速いとよく言われる．変化についていくのがやっとのところもあるが，その変化は突然に出てきたように思われるものであっても，実際は関連技術や基礎的なものから発展してきた結果であることは間違いない．

秋葉原ラボも，技術の変化を敏感に察知し，追従し，結果的に少しでも技術を発展させ，サービスおよび業界に良い影響を与えていけるように日々研鑽しているところだ．改元のように明確な区切りが生まれるわけではないであろうが，振り返ってみるとここが変化点であったなと感じることはある．私の場合は Hadoop 関連技術を多く知るきっかけとなった第 1 回の Hadoop Conference Japan 2009 や，秋葉原ラボを設立した 2011 年 4 月 1 日などが典型的な変化点となった．

今回が Vol.2 となる秋葉原ラボ技術報告でも，前回と同様に，秋葉原ラボでの最近の取り組みとその成果をまとめている．今回掲載している内容が技術発展の大きな変化点に直接にはなっていないかもしれないが，今後の発展・変化につながるものになっていくことを望んでいる．そして今年が新たな発展の年になることを期待している．

目次

AbemaTV[a]は当社とテレビ朝日とが共同で展開しているリニア配信型の国内向けインターネット動画サービスである．ニュース，ドラマ，バラエティ，アニメやスポーツなど幅広いジャンルの専門チャンネルを有し，約 20 チャンネルが 24 時間編成されている．2016 年 4 月の開局以降，新規チャンネルの開設や統廃合を含めた継続的な機能拡張がなされている．

視聴にあたっては，インターネットブラウザもしくはスマートフォンやタブレット，テレビデバイスに用意された専用アプリケーションを用いる．いずれの場合も無料で利用できて必ずしもユーザ登録を必要としない点や，特にスマートフォンやタブレットでは番組表のフリック操作のみでチャンネルおよび番組を選択できる点が特徴である．

またオンデマンド配信型のサービスとして Abema ビデオ[b]も用意されており，リニア配信で見逃した番組の視聴や，アーカイブされた動画の検索および推薦などのニーズに応えている．AbemaTV の番組表から直接にアーカイブされた動画にアクセスできることはもちろん，タイムシフト再生や倍速再生のシームレスな利用も可能であり，独特なユーザ体験を提供している．なお Abema ビデオは月額課金のサブスクリプション型の動画サービスに分類できるが，一部の機能は無料でも利用できる．

AbemaTV の利用ログも当社の他のメディアサービスと同様に，秋葉原ラボのログ集計・分析基盤 Patriot に集約されている．加えてレポーティングを容易にするために整備した中間テーブル群を Google BigQuery[c]に構築しており，データ分析や可視化に役立てている．

本技術報告では，秋葉原ラボで実施された分析やシステムの実装について，以下の 3 つを取り上げる．1.1 章では，AbemaTV のユーザの利用ログに基づいた統計モデリングにより，サービス利用のアクティブ度の定量化を試みている．1.2 章では，Abema ビデオで提供している推薦システムの概要と実装について述べている．1.3 章では，AbemaTV を Web メディアの一つとする観点から，主にニュース番組に焦点を当てて社会的意義について論じている．

チャンネル改編を含む継続的な機能拡張に伴うユーザ体験の変化や，リニアとオンデマンドとを横断した利用，および多様なチャンネルのフリック操作によるザッピングなど，いずれも AbemaTV に特徴的な機能に着目した分析および実装がなされている点にご注目いただきたい．

[a] https://abema.tv
[b] https://abema.tv/video
[c] https://cloud.google.com/bigquery/?hl=ja

Akihabara Lab

1.1 AbemaTV におけるユーザアクティブ度のモデル化

松田 和己

概　要　インターネットサービスにおけるユーザの利用態度の把握は，サービス運用において重要となる．本稿では，インターネットテレビ局 AbemaTV を対象に，ユーザのアクティブ度の定量化を行うことでサービス運用への寄与を目指す．先行研究では，multivariate adaptive regression splines（MARS）を用いた定量化方法が提案されている．しかしながら AbemaTV では，時期によってさまざまな番組が配信され，また各種機能が逐次リリースされるため，一度構築したモデルで一貫した評価を行うことは難しい．本稿では，異なる期間を対象にしたモデル化を実施することで，より汎用的な定量化方法を検討する．

Keywords　AbemaTV，MARS，アクティブ度，行動分析

1　はじめに

AbemaTV[*1]は，当社とテレビ朝日が共同で展開するインターネットテレビ局で，番組表に基づくリニア配信が特徴である．AbemaTV では，オリジナル番組の他に，ドラマやアニメ，スポーツ，将棋などのさまざまなコンテンツを約 20 あるチャンネルで 24 時間無料で提供しており，それらの番組はスマートフォン，タブレット，パソコン，テレビで視聴可能で，据置型のテレビで放送する通常のテレビ局と異なり多様な視聴体験を提供している．2016 年の開局以降，チャンネルの増減を伴った番組の改編，対象デバイスの拡大やデバイス専用アプリケーションの提供，終了後の番組を視聴できる「Abema ビデオ」を始めとした各種機能のリリースが進み続けており，ユースケースが多様化している．

一般にインターネットサービスは，ユーザにとって導入コストが小さい一方で，利用中のサービス離脱が発生しやすいことが知られており，ユーザの継続的な利用を実現することがサービスの成功と直結する．そのため，ユーザの行動データを分析し，継続して利用しているユーザと離脱ユーザの差異を明らかにすることや，サービス事業者の意図通りには使われていない機能を発見することによって，UX（ユーザエクスペリエンス）を向上させることが重要となる．また，分析結果を実際のサービス改善に活用する上で，人間が分析結果を解釈できること，特に，どのような行動や使い方が継続的な利用に対して正もしくは負の効果があるかを理解しやすいことが求められることが多い．

これに対して，和田らは multivariate adaptive regression splines（MARS）[1] を用いた分析を行った [2, 3]．これらの報告では，ユーザごとにある 1 日の視聴行動と属性情報に加えてその直近 7 日間のサービス利用日数を説明変数とし，直後の 7 日間の利用日数を目的変数とした継続・離脱分析を行い，継続利用しているユーザほど高いとされるアクティブ度を定量化することで，継続しやす

[*1] https://abema.tv

いユーザと離脱しやすいユーザの把握とそれに寄与するユーザ行動を明らかにした．しかしながら，これらの報告で扱われているデータの対象期間はAbemaTV 開局直後であり，時期によって提供されるサービス内容やユーザの利用傾向が異なることを考慮できていない．実際に，AbemaTV ではチャンネルの新規開設や統廃合，番組の改編や放送時間の変更，新しい機能のリリース，UI（ユーザインタフェース）の変更などに伴って，提供される体験が変化し続けている．

本稿では，変化する利用態度を考慮することで分析対象とする時期に依存しないアクティブ度表現の構築と適切な変数選択手法を検討するため，ユーザに提供される体験が異なる期間を対象にした分析をそれぞれ行い，結果を考察する [4].

2　手法

本稿で用いた手法の概要について説明する．ここで，サービスに対するアクティブ度は，高いユーザほどサービスを熱心に利用しており，低いユーザほどサービスをあまり利用していないということを表す指標とする．

まず，ユーザに提供される視聴体験がそれぞれ異なっていると考えられるいくつかの日を選択し，和田らの先行研究と同様の手法でそれらの日のユーザのアクティブ度をモデル化する．具体的には，あるユーザのある日のアクティブ度を，翌 7 日間のサービス利用日数を通して，特定の属性や行動が寄与しているかどうかを分析する．

このように，異なる期間のユーザのアクティブ度をそれぞれモデル化し時期による要素の違いを考察することで，よりロバストで汎用的な指標を検討したい．

3　モデル

特定の日のユーザのアクティブ度をモデル化する方法について述べる．ある日のあるユーザについて，目的変数である翌 7 日間のサービス利用日数 Y と，当日および直近 7 日間の行動情報からなる説明変数ベクトル $\boldsymbol{x}$ に関する n 組のデータ $\{(Y_i, \boldsymbol{x}_i);\ i = 1, \cdots, n\}$ が得られたとする．ここで，説明変数ベクトルの各要素は属性や行動の有無や程度を表す数値であり，属性や行動の有無についてはカテゴリカルな因子，視聴時間などの連続量についてはそのままの値とする．

目的変数 Y は，当日のサービスに対するアクティブ度 p $(0 \leq p \leq 1)$ を用いた二項分布

$$Y \sim \mathrm{Binom}(7, p)$$

に従うと仮定する．このとき先行研究と同様に，アクティブ度 p に対して MARS モデルを用いて，

$$\mathrm{logit}(p) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{x})$$

とモデル化する．ここで，$\mathrm{logit}(\cdot)$ はロジット関数，β_m は基底関数の重みを調節するパラメータ，M は基底関数の個数，$h_m(\cdot)$ は基底関数である．MARS モデルにおける基底関数は hinge 関数からなる．hinge 関数は，1 つの変数 x に対する 1 つの境界点を c としたとき，$h^+(x) = \max(0, x - c)$ や $h^-(x) = \max(0, c - x)$ と表現される．パラメータの推定，変数選択の詳細については，文献 [3] を参照されたい．結果として，アクティブ度 p に対して寄与するいくつかの説明変数とそれぞれの境界点がモデル選択基準を用いて選ばれる．また，本稿では，統計解析向けプログラミング言語 R[*2]の earth[*3]パッケージを用いて分析を行った．

4　データ

本稿では，2017 年 10 月 23 日，2018 年 2 月 26 日，2018 年 11 月 5 日，2018 年 11 月 12 日の AbemaTV ユーザのデータを利用した．ここで，曜日の差を極力排除するため，これらの日付は全て平日の月曜日を選択した．このとき，これらの日に AbemaTV を利用したユーザの中からランダムサ

[*2] https://www.r-project.org/
[*3] https://CRAN.R-project.org/package=earth

ンプリングを行い学習データとした．また，抽出されたユーザについて上記 4 日間の行動データの他に，前後 7 日ずつの利用日数も取得している．なお，抽出対象とした 2017 年 10 月以降に，72 時間ホンネテレビ[*4]の放送，アニメチャンネル全面リニューアル[*5]，動画ダウンロード機能[*6]，追っかけ再生機能[*7]のリリースがあり，ユーザ層およびユースケースの多様化が続いていると考えられる．

また，抽出対象となるユーザについて，サービスを本質的には利用していなかったユーザを除外するため，任意の番組を 30 秒以上視聴したユーザに限定した．同じく，サービスの利用日数を取得する際も 30 秒以上の番組視聴を条件としている．本分析で扱う目的変数と説明変数については，表 1.1.1 に示すように，ある特定の日の翌 7 日間の利用日数を目的変数とし，ある特定の日の中での各種行動と直近 7 日間の利用日数を説明変数とした．

5 結果

集計対象とした 4 日間それぞれに MARS モデルを適用した結果を表 1.1.2 に示す．

結果から，分析対象とした 4 日間に共通して，直近 7 日間の利用日数が強く影響していることがわかる．また，視聴時間については，多い分にはあまり影響がなく，少ない場合はマイナスに働いている．アニメ，韓流，麻雀を視聴していることはある程度のボリュームでプラスの効果があり，それぞれ習慣的に視聴するジャンルであることが示唆されている．

次に期間によって差がでている点に着目して考察する．まず 2017 年 10 月 23 日については，他の 3 つの対象期間への結果と共通して含まれる要素によって構成されており，今回の対象期間の中では標準的な期間と考えられる．

2018 年 2 月 26 日は，将棋を視聴していることがプラスに働いている．これは翌 7 日間（2018 年 2 月 27 日〜2018 年 3 月 5 日）に将棋の公式戦生中継番組が 3 つ編成されていたことによるものだと考えられる．

続いて 2018 年 11 月 5 日は，パソコンでサービスを利用していることが特にマイナスに効いている．これについては明確な理由を見いだせていないが，パソコンでの視聴では他のデバイスに比べてユーザ識別子が変更されやすい（シークレットウィンドウの利用や Cookie の削除などによる）ために，翌 7 日間の利用日数が 0 日となりやすく，それによりマイナスに効く量として変数が選択されているものと考えられる．

2018 年 11 月 12 日は，格闘を視聴していることが大きくマイナスに効き，加えてオンデマンドの視聴のみであることもマイナスに効いている．この日は平日でありながら，ボクシングの世界戦の生中継特番が放送されていた．一般に特番は視聴者が集まりやすい一方で，その番組のみを特に視聴するユーザの割合が多くなるので，翌 7 日間の利用日数としては相対的に低くなる傾向がある．また特番の放送終了後にオンデマンド視聴をするユーザも多い．したがって，格闘視聴とオンデマンドのみの視聴がマイナスに働いたと考えられる．

以上により，アクティブ度に影響するユーザの属性や行動について，期間によって共通の要素と特有の要素とを見出すことができた．特有の要素のほとんどは特番などの一時的かつ不定期な編成が要因と考えられるので，汎用的な指標を目指す観点からは変数として利用しないほうが適切であろう．たとえば，各日のモデルに対して，複数のモデルからさらに精度の高いモデルを構築する手法であるモデル平均化法などを応用することなどが考えられる．

[*4] https://abema.tv/channels/abema-special/slots/AkkG6KZAFkUszf
[*5] https://abematimes.com/posts/3882858
[*6] https://www.cyberagent.co.jp/news/detail/id=21536
[*7] https://www.cyberagent.co.jp/news/detail/id=21583

表 1.1.1　目的変数と説明変数の詳細

種別	期間	内容	範囲
目的変数	翌 7 日間	利用日数	0 ～ 7
説明変数	直近 7 日間	利用日数	0 ～ 7
	当日	新規登録	新規 or 既存
		デバイス	スマートフォンのみ or タブレットのみ or パソコンのみ or テレビのみ or 複数
		総視聴時間 (hour)	0 ～
		視聴ジャンル	バラエティ or ドラマ or アニメ or リアリティショー or 韓流 or スポーツ or 格闘 or 釣り or 将棋 or 麻雀
		視聴方法	リニア視聴のみ or オンデマンド視聴のみ or 併用

表 1.1.2　選択された変数および境界点と係数パラメータの推定値

2017 年 10 月 23 日		2018 年 2 月 26 日		2018 年 11 月 5 日		2018 年 11 月 12 日	
変数	係数	変数	係数	変数	係数	変数	係数
切片	1.30	切片	1.32	切片	1.37	切片	1.39
直近 7 日利用 $>$ 6 日	0.64	直近 7 日利用 $>$ 6 日	0.65	直近 7 日利用 $>$ 6 日	0.70	直近 7 日利用 $>$ 5 日	0.51
直近 7 日利用 $<$ 6 日	−0.44	直近 7 日利用 $<$ 6 日	−0.43	直近 7 日利用 $<$ 6 日	−0.47	直近 7 日利用 $<$ 5 日	−0.50
視聴時間 $>$ 2h	0.03	視聴時間 $>$ 2h	0.02	視聴時間 $>$ 2h	0.04	視聴時間 $>$ 5h	−0.04
視聴時間 $<$ 2h	−0.29	視聴時間 $<$ 2h	−0.27	視聴時間 $<$ 2h	−0.22	視聴時間 $<$ 5h	−0.18
アニメ視聴あり	0.19	アニメ視聴あり	0.20	アニメ視聴あり	0.16	アニメ視聴あり	0.17
韓流視聴あり	0.30	韓流視聴あり	0.27	韓流視聴あり	0.25	韓流視聴あり	0.25
麻雀視聴あり	0.28	麻雀視聴あり	0.24	麻雀視聴あり	0.29	麻雀視聴あり	0.31
		将棋視聴あり	0.37	パソコンのみ	−0.23	格闘視聴あり	−0.52
						オンデマンドのみ	−0.17

6　まとめ

本稿では，インターネットテレビ局 AbemaTV のユーザのアクティブ度を異なる期間でモデル化し比較した．その結果, 配信される番組や提供サービスが変化し続ける中では，同じ曜日を対象にしたとしても日ごとに共通の変数と固有の変数があることがわかった．今後はモデル平均化法による改善アプローチによって，汎用的かつロバストなアクティブ度の表現手法の確立を目指す．これにより異なる期間のアクティブ度比較や時系列変化を捉えることを可能にし，それに寄与する行動を明らかにしたい．このアプローチの結果については稿を改めて報告する．

参考文献

[1] Friedman, J. H.: Multivariate adaptive regression splines, *The annals of statistics*, pp. 1–67, 1991.

[2] 和田計也，福田一郎：インターネットテレビにおけるユーザの視聴行動分析，日本マーケティング学会カンファレンス・プロシーディングス，Vol. 5，pp. 62–65，2016.

[3] 和田計也:インターネットテレビサービス AbemaTV におけるユーザアクティビティ分析，CyberAgent 秋葉原ラボ技術報告，CyberAgent，Vol. 1，2018.

[4] 松田和己，和田計也，福田一郎：インターネットテレビ局 AbemaTV における番組視聴データを用いたユーザー行動分析，統計関連学会連合大会報告集，2018.

Akihabara Lab

1.2 AbemaTV を支える推薦システム

前田 英行，福田 鉄也

概　要　膨大なコンテンツを提供するサービスにおいて，ユーザが興味をもつコンテンツを発見するのに推薦システムは有用である．近年では e-commerce や動画サイトなどをはじめとする多くのインターネットサービスに推薦システムが導入されている．本章ではインターネットテレビ局 AbemaTV が提供する動画サービスにおいて実運用している推薦システムのシステムアーキテクチャを紹介し，実運用上で課題となる膨大なデータに対する計算時間の削減や，ユーザの行動履歴に対する応答性などの課題をどのように解決しているか述べる．

Keywords　機械学習，推薦システム

1　はじめに

近年 YouTube[*1]，Netflix[*2]，Hulu[*3]などのインターネット上で動画を楽しむサービスが一般的となっている．当社もテレビ朝日と共同で，オリジナルのコンテンツの他，スポーツ，ドラマ，音楽，麻雀，将棋，ゲームなど約 20 チャンネルの番組をすべて無料で楽しめるインターネットテレビ局 AbemaTV[*4]を提供している．

AbemaTV では，番組表に基づくリニア配信に加えて，Netflix，Hulu などのような SVOD（Subscription Video on Demand）型の Abema ビデオというサービスも提供している．SVOD 型の動画サービスでは，ユーザ自身が視聴したい動画を膨大なコンテンツの中から発見できる必要がある．動画のようなマルチメディアデータでは，興味のあるコンテンツの特徴を言葉で表現しづらい場合があり，ユーザによる検索だけでは膨大なコンテンツを十分に提供することが難しい．このような場面ではユーザが視聴したいコンテンツをサービス提供者側から推薦（レコメンデーション）する機能によりユーザへの視聴を促進することが重要となる．

近年，推薦分野の研究では精度向上を目的とした推薦アルゴリズムは多く登場してきた．しかし，実運用上で耐えられる推薦機能を提供するには精度を追求するアルゴリズムだけではなく膨大な量のデータにおける計算時間の削減やユーザの直近の行動をリアルタイムに反映させることなども重要となってくる．

本章では実際に AbemaTV で運用を行っている推薦システムのアーキテクチャについて紹介し，推薦システムを実運用する上での課題をどのように解決しているか紹介する．秋葉原ラボで開発した推薦システムは，推薦候補の抽出とリランキン

[*1] https://www.youtube.com/
[*2] https://www.netflix.com/
[*3] https://www.happyon.jp/
[*4] https://abema.tv/

グのシステムを分割することで，精度を保ちつつ実用的な処理時間を実現している．以降 2 節では AbemaTV の推薦機能の概要を説明し，3 節ではシステムアーキテクチャについての詳細を説明する．

2　AbemaTV における推薦機能

この節では AbemaTV で提供している主な推薦機能と AbemaTV の推薦システムについての概要を説明する．

AbemaTV において推薦機能は Abema ビデオに導入されている．Abema ビデオで提供している推薦機能は主に以下の 2 点である．

- 番組から番組への推薦
  - ある番組を起点にその番組の関連番組を推薦する機能
  - Abema ビデオの番組ページなどで提供される
- ユーザから番組への推薦
  - 個々のユーザの直近の視聴履歴を元に番組推薦する機能
  - Abema ビデオのトップページなどで提供される

“番組から番組への推薦” はユーザがある番組を視聴中もしくは番組詳細ページを閲覧中にその番組の関連番組を推薦することで同じような番組を観たいユーザに見つけやすくする目的があり，一方で “ユーザから番組への推薦” は特定の番組によらずユーザの過去の視聴履歴の傾向からユーザの興味がある番組を推薦することが目的である．このような違いがあるため 2 つの推薦機能が提供されるページが異なる．

3　AbemaTV における推薦システム

この節では AbemaTV の推薦システムについて説明する．

AbemaTV の推薦システムは推薦候補を生成する処理とリランキングを行う 2 つの処理から構成される．

- 推薦候補生成処理
  - オフラインで計算された番組の類似度に基づき推薦候補を絞る
- リランキング処理
  - 推薦候補生成処理で生成された推薦候補からユーザとの関連度スコアを付与し推薦結果を並び替える

推薦候補を絞った上でリランキング処理を行うことで，オンラインでの計算時間をおさえつつ複雑な機械学習モデルを用いた推薦を実現している．番組の類似度計算はオフラインのバッチ処理で網羅的に行う．これによりすべての番組を推薦システムの対象にする．リランキング処理ではユーザからのリクエストごとに機械学習を用いた順序づけを行うことで，パーソナライズされた推薦を実現している．

また，処理を 2 段階にすることで，さまざまな変更に柔軟に対応できるシステムアーキテクチャを構成している．秋葉原ラボでは，各処理のアルゴリズムの精度向上などの改善には継続的に取り組んでおり，その成果を実システムに適用することは重要である．また，要件に応じたビジネスロジックの挿入などにも柔軟に対応できる必要がある．

全体のシステム構成を図 1.2.1 に示す．推薦機能は Recommendation API を介して提供される．Recommendation API は，データ処理基盤（Patriot）やストリーム処理エンジン（Zero）などの他システムと連携し推薦リストを生成する．番組の類似度計算やリランキングのための機械学習モデルの生成は，Patriot 上で行っている．Patriot は Hadoop を用いた秋葉原ラボで開発されたデータ処理基盤で，ユーザの行動ログなどのデータが蓄積されており，蓄積されたデータに対して Spark などの並列分散処理を適用できる [1]．また，“ユーザから番組への推薦” では，Zero を用いることでユーザの視聴行動をリアルタイムに反映した推薦を実現している．Zero は秋葉原ラボで開発されたストリーム処理エンジンであり，行動ログなどのイベントデータをリアルタイムに処理できる [2]．

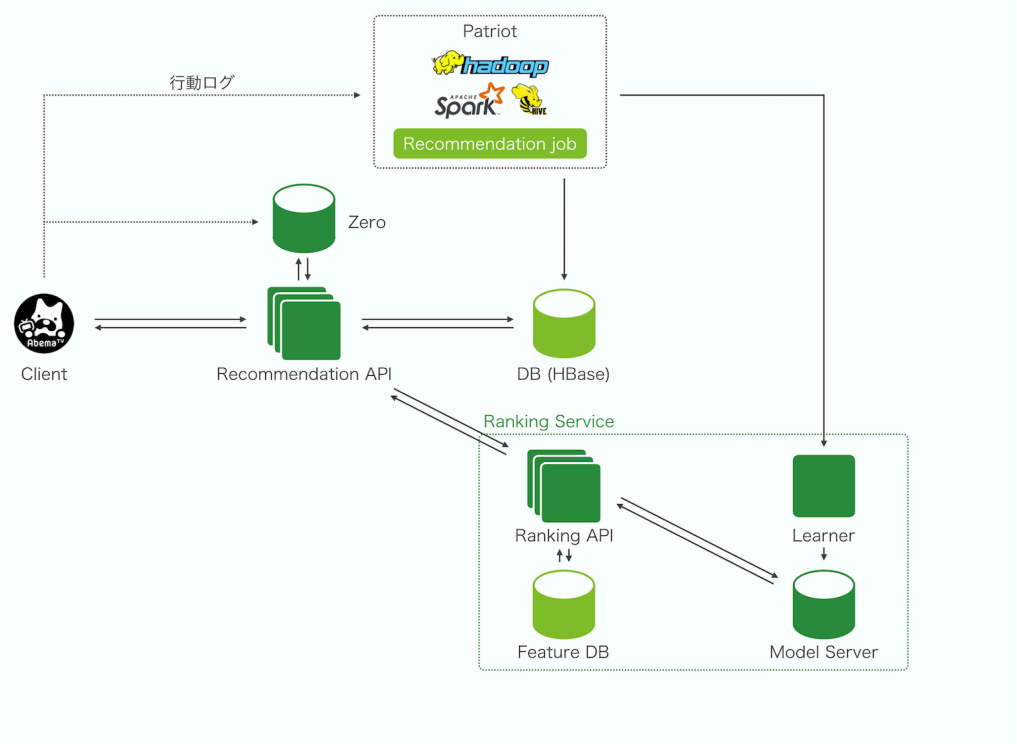

図 1.2.1 システム構成図

Ranking Service は機械学習によるリランキング機能を提供している．

以降，推薦候補生成処理とリランキング処理の流れについて説明する．

3.1 推薦候補生成処理

推薦候補の生成は Zero や Patriot に転送されたユーザの視聴履歴などの行動ログデータをもとに行う．Patriot 側ではユーザの視聴履歴のログを元に類似度計算ジョブを実行する．類似度計算にはユーザの視聴履歴ログを Implicit Feedback とした協調フィルタリング [3] を用いている．協調フィルタリングのアルゴリズムは重み付き Matrix Factorization を使用している．

AbemaTV の推薦システムでは視聴履歴ログとして AbemaTV と Abema ビデオの両方の視聴ログを使用している．両方のログを使用するメリットとしては AbemaTV しか視聴していないユーザに対して Abema ビデオの推薦をすることが可能となることが挙げられる．また AbemaTV では番組予約の機能があるため，ユーザがすでに番組予約をしていた場合は視聴情報にさらに重み付けをしている．

重み付き Matrix Factorization により各番組の特徴を表す次元圧縮された特徴ベクトルを得ることができる．この特徴ベクトルを元に各番組の類似番組の上位 N 件を取得する探索処理を実行する．この上位 N 件は番組推薦候補として HBase に保存される．なお，この N は数十と設定するが，推薦結果を AbemaTV のサービス側で何件表示させるかなどのサービス要件で変わる．

Recommendation API はクライアントからのリクエストに応じて 2 節で述べた 2 種類の推薦機能のいずれかを実行する．2 節より “番組から番組への推薦” に該当する処理は Recommendation API では起点となる番組の類似番組を HBase から取得する．この時点で HBase には事前に計算された推薦候補の結果が保存されているので HBase から結果を取得するだけで推薦候補結果を取得できる．

“ユーザから番組への推薦” に該当する処理は Recommendation API でユーザの直近の視聴履歴データを Zero から取得し，この視聴履歴を元に

類似する番組を HBase から取得し，集約して推薦候補として利用する．“ユーザから番組への推薦”の推薦候補は HBase に事前に結果を保存していない．この理由としてはユーザの直近の行動に対するリアルタイム性を反映させたい点とユーザ数は番組数に対して圧倒的に数が多いので，全ユーザの推薦候補を事前計算するのは非効率だからである．このアーキテクチャは YouTube や Amazon での例を参考にしている [4, 5].

HBase から取得した推薦候補は，後述する Ranking Service を用いてリランキング処理を行い，その結果を最終的な推薦結果としてクライアント側に返却する．

3.2 リランキング処理

リランキング処理は推薦候補生成処理で生成された推薦候補の結果からユーザとの関連度の高い番組をなるべくランキングの上位に推薦することを目的とする．AbemaTV の推薦では，ユーザが推薦候補の番組を視聴する可能性があるかどうかという二値分類器のモデルを学習し，その予測確率をスコアとして順位づけを行う．

図 1.2.1 において，リランキング処理は Ranking Service で実行される．リランキング処理に関わる各コンポーネントの役割は以下のようになる．

- ユーザと番組の関連度合いを計算する機械学習モデルの生成（Learner）
- ユーザと番組の関連度合いの推論を行う（Model Server）
- Recommendation API より推薦候補結果を受け取り Feature DB より特徴量を付与し Model Server へ関連度スコアの計算をリクエストする（Ranking API）
- 推論時に使用する特徴量を保存する DB（Feature DB）[*5]

リランキング処理で行われるモデル作成の処理について説明する．Learner は，Patriot からユーザの行動ログや番組のメタ情報などを取得して学習データを作成し，オフラインで二値分類の機械学習モデルを生成する．二値分類器のアルゴリズムは Gradient Boosting Decision Tree を使用している． 学習されたモデルは Model Server へデプロイされる．

次にこのモデルを使用した推論時の処理についての説明を行う．Ranking Service は，Recommendation API から推薦候補とユーザ ID を入力として推論を行う．まず，Ranking API は，ユーザの特徴量と推薦候補となっている各番組に対しての特徴量を Feature DB から取得する．特徴量はたとえばユーザのデモグラフィック情報や番組情報を表すメタ情報などがある．次に，この特徴量を元に Model Server へ関連度スコアの計算をリクエストする．Model Server は各特徴量に対してユーザと番組との関連度スコアを計算し，推薦候補の結果に付与して Recommendation API に返却する．Recommendation API ではこのスコアの高い順に並び替えを行いクライアント側へ返却を行う．

4 まとめ

本稿では，AbemaTV における推薦システムについて説明し，実運用上の課題をどのように解決しているか述べた．AbemaTV では新しいユーザが日々増加しており，大規模な推薦システムを精度高く提供するために今後もシステム面の改善や機械学習アルゴリズムの改善を行っていく必要がある．また今回紹介した推薦機能以外の新しい推薦機能も今後検討していく必要があると考えている．

参考文献

[1] 梅田 永介，飯島 賢志，斎藤 貴文: データ処理基盤 Patriot における継続的改善の取り組み，CyberAgent 秋葉原ラボ技術報告，サイバーエージェント，vol. 1, pp. 4–11 (2018)
[2] 佐藤 栄一: リアルタイムのデータ活用を支援

[*5] 特徴量のデータは Patriot からユーザのデモグラフィック情報や番組のメタ情報を取得し保存する

するストリーム処理エンジンの開発，CyberAgent 秋葉原ラボ技術報告，サイバーエージェント，vol. 1, pp. 17–25 (2018)

[3] Hu, Y., Koren, Y. and Volinsky, C.: Collaborative filtering for implicit feedback datasets, Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pp. 263–272 (2008)

[4] Linden, G., Smith, B. and York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, pp. 76–80 (2003)

[5] Davidson, J. et al.: The YouTube Video Recommendation System, Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 293–296 (2010)

Akihabara Lab

1.3 分断化した社会における AbemaTV の役割に関する検討

高野 雅典

概　要　AbemaTV はテレビのような受動的なユーザ体験（受け身視聴）とコンテンツ品質を実現しつつ，テレビへの接触が少ない若年層の利用者が多いという特徴をもつ．したがって従来のマスメディアが担ってきた報道の役割を AbemaTV のニュースコンテンツによって補間できる可能性がある．本稿では AbemaTV のメディアとしての価値向上を目指した 2 つの研究を紹介する．一つはユーザの接触コンテンツの多様性やその変化を評価する枠組みを提案するものであり，もう一つは AbemaTV のザッピングというテレビに似たユーザ体験がユーザの関心に与える影響を調査するものである．両研究結果は AbemaTV のユーザが接触するコンテンツの多様性を増加させる可能性，ニュースへの関心が薄いユーザへの関心喚起ができる可能性を示唆する．

Keywords　ニュースメディア, 社会的分断, 選択的接触

1　はじめに

インターネットやスマートフォンの普及により，多くの人が膨大な情報源にアクセスできるようになった．新聞・テレビニュース・雑誌などの既存のメディアに加え，ニュースサイト・個人のブログ・まとめサイト・ソーシャルメディアでの専門家・一般の人の発言などである．情報源の多様化によって，情報の受け手は情報源を選別することが可能になった．

このような自身の選好に合わせて情報を選択すること（選択的接触）によって受け取る情報に偏りが生じる [14, 20, 23]．選択的接触は社会構造を分断する（社会的分断）という問題が指摘されている．社会が分断化すると自分が所属するクラスタと触れ合うことが多く，自分の所属しないクラスタと触れ合う機会は少なくなる．たとえば政治思想・知識である．リベラルな政治態度の人はリベラルな情報源に接触しがちであり，保守的な政治態度をもつ人は保守的な情報源に接触しがちになる [6]．政治に関心がないと政治情報への接触機会が少ないため [14]，政治知識の有無に関しても分断が発生する [15, 20]．

政治的・社会的な問題に関する態度・知識の分断は特に重要な課題の一つである．社会的分断は同じ立場の情報への偏ったアクセスをより加速するため，意見・主張の極端化が発生しがちである．極端な意見をもつと異なった意見の他者との相互理解や建設的な議論が難しくなる [19]．また誤った情報が蔓延しやすいという問題もある [11]．それはたとえ誤っていても自分自身の意見に近い情報であるため信じてしまいやすく，分断されているため訂正情報が届きにくいからである [2]．

現在では多くの Web メディアで提供される推薦機能（ニュースサイトでの記事，ソーシャルメディアでのアカウントなど）は，偏った情報源へのアクセスをさらに促進し，社会的分断はさらに加速されることが懸念される．推薦機能はユーザのアクセス履歴などの行動情報から，そのユーザが好む

情報を提示することだからである．このようなシステムによっていつの間にか自身が望む情報しか見えなくなってしまうことはフィルターバブルと呼ばれる [13].

従来はテレビなどのマスメディアが人の政治知識の乖離を防ぐ役割を果たしていた [22]．以下の点においてニュースへの関心のないユーザに対しても比較的情報を届けやすいからである．

- 受動的なメディアである．ニュース記事のように能動的にテキストを読む必要がないため関心が低いユーザでも情報を得やすい．
- ニュースコンテンツは特定のトピックスではなく様々なトピックを扱うパッケージになっている．
- ニュース制作会社によってコンテンツが制作されているため，個人のソーシャルメディアの発言やまとめサイトに比べて平均的な品質が高い．

このようなテレビニュースのような受動的でパッケージ化されたメディアによる情報は知識の平均化をもたらす [17, 20]．しかし，近年の日本では若年層は Web メディアへの接触が頻度が高く，一方でテレビ・新聞などのマスメディアへの接触は少ないため，この役割を果たしづらくなっている [24].

したがって社会的分断を緩和するには従来のマスメディアが担ってきた役割を Web メディアが補間する必要がある．たとえば，政治や国際情勢に関心を持たない人に対して，政治知識を促進するにはさりげない情報提示が有効である．Kobayashi ら [9] は日本のポータルサイト Yahoo!ニュースのトップページに掲載されるトピックス見出しを閲覧するだけで，政治に関する知識の学習効果があることを示した．Epstein & Robertson [3] は検索エンジンの偏った検索結果は選挙の結果を左右しうることを示した．

まとめると社会的分断を緩和するためには，1) テレビのような受動的なパッケージ化されたメディアによる情報は知識の平均化に有用，2) 若年層のメディア接触には Web メディアが有効，3) Web メディアのさりげない情報提示は政治などへの関心のない層への知識向上に有用と言える．

インターネットテレビ局 AbemaTV[*1]は Web メディアでありながらテレビのようなユーザ体験を提供するサービスであり，また若年層の利用者が多い [1] ため，1, 2 を満たす．またチャンネルをテレビと同様に順番に変更するというユーザインタフェースによって，意図しないコンテンツへの接触機会があり 3 を満たす可能性がある．秋葉原ラボでは AbemaTV が放送するニュースコンテンツによって社会的分断を緩和することを目指して，計算社会科学・社会心理学・政治コミュニケーションの研究者と共同研究を進めている．本稿ではそのための準備として実施した以下の枠組みの開発（a）と行動傾向の基礎的な調査（b）を紹介する．a) ユーザの接触コンテンツの多様性の定量化 [7, 8]，b) ニュースの「チラ見」がニュース視聴の動機に与える影響 [21] の 2 つを紹介する．a は東京大学大学院 工学系研究科の鳥海不二夫先生，西口真央研究員，垣内弘太氏との共同研究である．b は立命館大学 情報理工学部 小川祐樹先生との共同研究である．

2 ユーザの接触コンテンツの多様性の定量化

前節で述べたように人は情報源に選択的に接触し，その結果，接触する情報源が偏る [14, 20, 23]．これらは質問紙調査による接触メディアの傾向 [14, 20] やソーシャルメディア上のフォロー関係 [23] によって確かめられている．一方で実際の行動履歴からメディアやコンテンツに対する接触を検証した研究は少ない（日本の事例では [9, 16] がある）．

本研究 [7, 8] では，ユーザの情報接触とその影響に関する基礎的な研究として視聴するコンテンツの多様性を定量化する枠組みを構築することを

[*1] https://abema.tv/

目的として実施した．

2.1　提案手法と結果

接触コンテンツの多様性の定量化

まず接触コンテンツ（AbemaTV では視聴番組）の多様性の定量化について述べる．最もシンプルな指標化の方針はコンテンツのメタデータ（カテゴリやタグ）に基づくものである．しかしメタデータは分析したい粒度（国際情勢，ヨーロッパの動向，イギリスの EU 離脱，各政党など）で付与されているとは限らず，適切なタグ付けは流行や社会情勢にも左右されうる．そこで本研究ではユーザの行動履歴から評価粒度を調節可能な接触コンテンツの多様性を定義し，定量化する．

ユーザの行動履歴からコンテンツを特徴づけるために，ユーザと接触コンテンツからなる行列を考える．各ユーザが接触したコンテンツには 1，そうでないコンテンツには 0 が入ったユーザ数 × コンテンツ数の行列である．

作成した行列の各列をコンテンツベクトルとして，それらのコサイン類似度を用いてコンテンツのクラスタリングを行う．ユーザが接触したコンテンツのクラスタの頻度の情報エントロピー（Cluster Diversity Entropy; CDE）が本研究で提案する接触コンテンツ多様性である（図 1.3.1）．CDE はクラスタリングのクラスタ数 K_d を変えることで任意の粒度で接触コンテンツの多様性を計算できる．

なお実際の計算ではコンテンツベクトルは次元縮約を行った．それは各ユーザが接触するコンテンツは全コンテンツのごく一部であるため，コンテンツ間の類似度を計算する際に一致する要素が非常に少なくなってしまい，扱いにくいからである．そこで LINE (2nd) [18] を用いてベクトルの次元縮約をした（図 1.3.2）．LINE (2nd) とはネットワークのノードを 2 次ノード（最短経路が 2 のノード）まで考慮してベクトル化するデータ縮約手法である．コンテンツどうしはユーザを介して接続される二部グラフである（図 1.3.2 左）ため，LINE (2nd) を使うことでユーザの類似度まで考慮してコンテンツのベクトルを作ることができる．

接触コンテンツの多様性変化の評価

前節で構築した多様性指標 CDE を用いて AbemaTV ユーザの接触コンテンツの多様性の変化を評価・分析する．多様性の変化は図 1.3.3 に示すように時系列で 2 つの期間（Start Block，End Block）で CDE の差によって評価する．ユーザの CDE の差 $CDE_{end\_block} - CDE_{start\_block}$ が正であれば，そのユーザの接触コンテンツの多様性が増したということである．ここでは AbemaTV の利用初期データ[*2]を除去し，その後に視聴した 10 コンテンツを Start Block，データセットにおける最後の 10 コンテンツを End Block とした．使用したデータセットの期間は 2017 年 6 月 26 日から 2017 年 7 月 2 日である．分析対象ユーザ数は 1 ヵ月以上 AbemaTV を利用していたユーザから 891 人のデータを用いた．コンテンツ数は 20,878 である．

図 1.3.4 にクラスタ数 K_d を変えたときの CDE の差を示す．ここでは階層的クラスタリングを利用して段階的に K_d を変えた．K_d が小さいとき（大きな少数のクラスタで CDE を評価したとき）は CDE の差は正になり，K_d が大きくなる（小さな多くのクラスタで CDE を評価する）と CDE の差は負になった．すなわち，接触コンテンツの多様性は巨視的視点では増加し，微視的視点で評価すると減少した．

2.2　まとめ

本研究では人が視聴するコンテンツの多様性変化を評価する枠組みを提案した．その枠組の有効性を検証するために，その枠組で「AbemaTV」におけるユーザの番組視聴の多様性変化について分析した．その結果，視聴番組の多様性は微視的視点では減少する（たとえばいくつかのニュースコンテンツのうち特定の番組のみ視聴する）ものの，巨視的視点では増加する（たとえばドラマ中心に

[*2] 新規登録日と 2 日目以降で最初に見た 10 コンテンツ

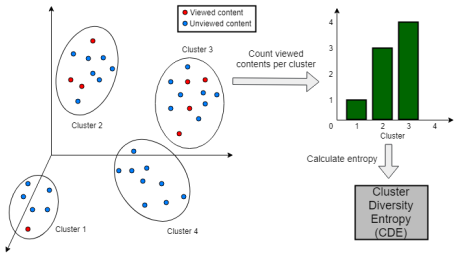

図 1.3.1 コンテンツのベクトル表現からの多様性指標 CDE の計算方法

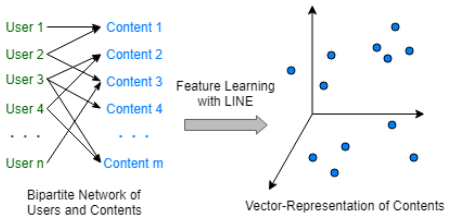

図 1.3.2 コンテンツのベクトル表現の作り方

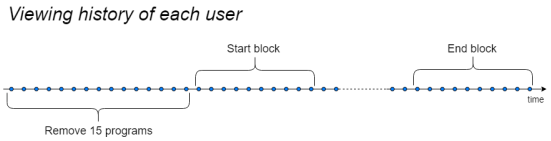

図 1.3.3 接触コンテンツ多様性の変化はスタートブロックとエンドブロックを比較して評価する.

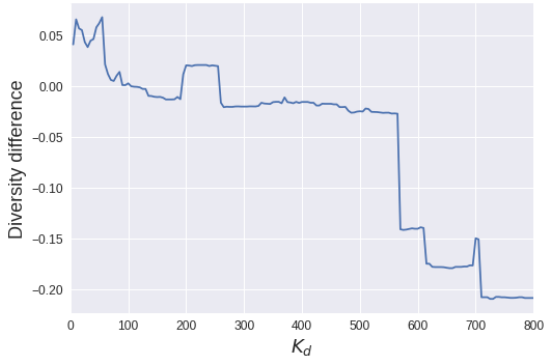

図 1.3.4 クラスタ数 K_d による CDE の差の変化

見ていたがニュース番組も見るようになる）ことがわかった.

この枠組みを用いると任意の粒度で多様性の変化を評価できる. たとえば, 政治に関する社会的分断の文脈では, 政治への関心の有無・知識量・政治態度（リベラル／保守）に影響するニュースコンテンツはどの粒度での多様性を必要とするかについてなどの検討ができる.

3 ニュースの「チラ見」がニュース視聴の動機に与える影響

Kobayashi ら [9], Epstein ら [3] が示したようにさりげない情報提示によって人の知識・関心・行動に影響を与えることができる. 本節ではニュースコンテンツの画面を「チラ見」することによるニュースチャンネルの継続視聴への影響を評価した.

AbemaTV では通常のテレビと同様に順番にチャンネルを変えながら視聴する番組を選ぶ（ザッピングと呼ばれる; 図 1.3.5）が, その際にニュース番組が目に入ることがある（「チラ見」と呼ぶ）. この普段の「チラ見」がニュースに関する知識や視聴動機を高めている可能性がある. 本節ではこの仮説を確かめるために次の研究課題（Research Question; RQ）を検証する（図 1.3.6）.

- **RQ**: 大きな政治・社会的ニュース視聴がきっ

図 1.3.5　AbemaTV のチャンネル変更の様子のイメージ図（ニュース番組からバラエティ番組）

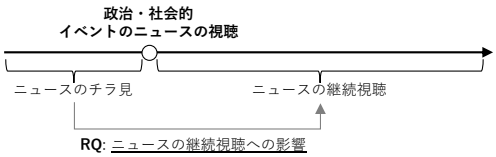

図 1.3.6　本研究で取り組む研究課題のイメージ図

表 1.3.1　分析対象の政治・社会的イベント．時差のため米合衆国のニュースについてはその日の翌日とした．

番号	イベント名	日付
0	米大統領選投開票	2016/11/09
1	米大統領就任式	2017/01/21
2	テロ等準備罪裁決	2017/06/15
3	小池都知事当選	2017/07/02
4	北朝鮮ミサイル発射	2017/08/26
5	北朝鮮ミサイル横断	2017/08/29
6	衆院解散表明	2017/09/25
7	衆院解散	2017/09/28
8	衆院選挙公示	2017/10/10
9	衆院投開票	2017/10/22

かけでニュース番組を見るようになったとき，その効果の持続性にニュースのチラ見は貢献するか？

3.1　データセット

本研究では図 1.3.6 のように「政治・社会的イベント」の事前のユーザの振る舞いが当日やそれ以降に与える影響について分析した．対象とした「政治・社会的イベント」ニュースは表 1.3.1 の 10 件である．分析対象のユーザは表 1.3.1 の前にほとんどニュースを見ていなかった（興味をもっていなかったと推察される）ユーザとした．具体的には対象日の 1 日前から 7 日前までさかのぼった 7 日間を「事前期間」として，以下の 4 つの条件を満たすユーザを分析対象ユーザとした．

- 事前期間よりも前に AbemaTV を利用していた
- 事前期間にある程度（合計 60 分以上）の AbemaTV の番組視聴をした
- 事前期間でのニュース視聴（一度に 30 秒以上視聴）をしなかった
- 対象日にニュース番組を 5 分以上視聴した

一度で 30 秒未満の視聴を「チラ見」と呼ぶ．また利用端末の OS は iOS と Android のユーザに限定した．

政治・社会的イベント前 7 日間でのユーザの振る舞いを表す変数として，事前期間における AbemaTV での総視聴時間 u, 事前期間におけるニュースチラ見数 c（一度の視聴が 30 秒以下），事前期間におけるニュースチャンネル以外の番組チラ見数 c'（ザッピング頻度），対象日当日のニュース視聴時間 s（分），対象日当日のニュースチャンネル以外の番組視聴時間 s'（分）とした．ザッピング行動を考慮するのは事前のニュースチラ見 c の効果とザッピング行動 c' によってニュースチャンネルにたどり着いた効果を分けて知りたいからである．また利用端末の OS によってアプリケーションの挙動が多少異なるため，OS を表す変数 o（iOS であれば 1，Android であれば 0）も分析の際には考慮した．本研究で着目するのは c と c' で，それら

以外は共変量調整のための変数である．

3.2 分析結果

政治・社会的イベントがきっかけでニュース視聴をした際に，その後のニュース視聴の継続について事前ニュース視聴 c が与えた影響を次の生存時間分析モデルによって分析した．ニュース視聴を 7 日間で 30 分以上していることを生存とした．すなわち「ニュース視聴をやめた日」とは「ニュース番組視聴時間が次の 7 日間の合計で 30 分以下になった最初の日」であり，イベント発生日からその日までが生存時間である．期間は最大 90 日とした（それ以上は生存と扱った）．

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{1.3.1}$$

$$\begin{aligned}\log \mu = {} & \beta_0 + \beta_1 \log_{10}(u+1) \\ & + \beta_2 \log_{10}(c+1) + \beta_3 \log_{10}(c'+1) \\ & + \beta_4 \log_{10}(s+1) + \beta_5 \log_{10}(s'+1) \\ & + \beta_6 o + \sum_{i=1}^{9} \beta_{7,i} e_i\end{aligned} \tag{1.3.2}$$

ここで $S(t)$ は生存関数で，生存時間（継続してニュースを視聴した日数）が t を超える確率を表す．$\beta_{7,i}e_i$ は「米大統領投開票」（$i = 0$）を基準としてそれ以外のイベント i（$i \geq 1$）や日程の効果を示す調整のための変数である．生存時間の分布には指数分布，ワイブル分布，対数正規分布，対数ロジスティック分布の 4 種類を検討し，全イベントに対して最も AIC が小さかった対数正規分布を採用した．

このモデルを用いて分析した．対象ユーザ数は 49,658 である．その結果を表 1.3.2 に示す．ニュースチラ見回数（c）は継続したニュース視聴に対して正の影響があった（$\beta_2 > 0$）．一方でザッピング頻度（c'）は有意に負の影響があった（$\beta_3 < 0$）．すなわちチャンネル変更などの過程でニュース番組の画面が目に入ること（c が大）は，普段ほとんどニュースを見ていないユーザの継続的なニュース視聴を促した．それに対しユーザのザッピングをしやすい傾向は継続的なニュース視聴を抑制した．

また事前期間での AbemaTV の番組視聴時間が 60 分未満のユーザにおいてはニュースチラ見回数・ザッピング頻度ともに有意に負の影響があった．

3.3 議論

本研究ではインターネットテレビ局 AbemaTV において「政治・社会ニュースに関心のないユーザ（AbemaTV においてニュースを見ていない）」のニュースへの関心を高める方策を検討するために，ニュース番組の画面を「チラ見」していると，ニュース番組をほとんど見ていなくてもニュース視聴への動機を高めるという仮説を立て検証した．

その結果，AbemaTV をある程度利用しているユーザにおいて，普段ほとんどニュース番組を見ていなくてもニュースの画面をチラ見をしていることは，政治・社会的イベント発生時のニュース番組視聴をした場合に，その後の継続的なニュース視聴も促進することがわかった．したがってチラ見をしやすいユーザインタフェースは政治・社会ニュースに関心のないユーザに関心を持たせる効果があると言える．対して，AbemaTV をあまり利用していなかったユーザについてはニュース番組のチラ見はニュース番組の継続視聴を抑制する結果になった．この理由については今後検証が必要である．

一方でザッピング行動傾向はニュースの継続利用は抑制する結果になった．チラ見によってニュース視聴を促進されたユーザについてはザッピングによってそのイベントのニュースに接触する機会が多いためだと考えられる．対して，ザッピング行動の傾向は継続性を低める傾向にあった．一般にニュース番組よりもエンターテイメント性の高い他のコンテンツの方が興味を引きやすいと考えられる．そのため，複数のチャンネルから見たいものを選ぶザッピング行動傾向が高いと，普段はニュースを視聴しづらかったのだと思われる．

ある政治・社会的イベントに興味をもつ・理解するためには，それの基盤となる知識が重要であろう．図 1.3.5 に示すように AbemaTV を含む

表 1.3.2　式 1.3.1 の回帰係数

Explanation Value	Coefficient	Standard Error	t-value	p-value
β_1 $(\log_{10}(u+1))$	-0.1852	0.01848	-10.024	2.0×10^{-16}
β_2 $(\log_{10}(c+1))$	0.0620	0.02774	2.2360	2.54×10^{-2}
β_3 $(\log_{10}(c'+1))$	-0.2532	0.01456	-17.392	2.0×10^{-16}
β_4 $(\log_{10}(s+1))$	-0.0145	0.00556	-2.6090	9.07×10^{-3}
β_5 $(\log_{10}(s'+1))$	0.5185	0.01602	32.359	2.0×10^{-16}
β_6 (o)	-0.0617	0.01377	-4.4830	7.34×10^{-6}
$\beta_{7,1}$（米大統領就任式）	-0.2828	0.02738	-10.328	2.0×10^{-16}
$\beta_{7,2}$（テロ等準備罪裁決）	0.2822	0.03726	7.5740	3.62×10^{-14}
$\beta_{7,3}$（小池都知事当選）	0.2151	0.03045	7.0630	1.63×10^{-12}
$\beta_{7,4}$（北朝鮮ミサイル発射）	0.2476	0.03641	6.8000	1.04×10^{-11}
$\beta_{7,5}$（北朝鮮ミサイル横断）	0.0202	0.02796	0.7240	4.69×10^{-1}
$\beta_{7,6}$（衆院解散表明）	0.3687	0.03617	10.1930	2.0×10^{-16}
$\beta_{7,7}$（衆院解散）	0.2394	0.03629	6.5960	4.22×10^{-11}
$\beta_{7,8}$（衆院選挙公示）	0.1843	0.03796	4.8560	1.20×10^{-6}
$\beta_{7,9}$（衆院投開票）	-0.0532	0.02850	-1.8660	6.20×10^{-2}
β_0 (Intercept)	0.5499	0.08365	6.5740	4.91×10^{-11}
$\log \sigma$	0.3519	0.00317	110.91	2.0×10^{-16}

ニュース番組は画面上に重要な情報をテロップで表示していることが多い．Kobayashi ら [9] が示したようにポータルサイトではトピックス見出しを閲覧するだけで，個別の見出しをクリックして記事自体を読まなくても政治に関する知識の学習効果がある．同様にニュース画面を「チラ見」していることで政治・社会ニュースに関する知識をユーザが獲得していた可能性がある．実際にそういった知識が獲得されたかどうかについては今後質問紙調査によって検証する予定である．

4　おわりに

本稿では AbemaTV が放送するニュースコンテンツによって社会的分断を緩和することを目指して行ってきた 2 つの準備的研究を紹介した．一つはユーザの接触コンテンツの多様性やその変化を定量化する枠組みを提案したものであり，もう一つは AbemaTV のザッピングというユーザ体験がユーザの関心に与える影響を調査したものである．両研究結果は AbemaTV のユーザが接触するコンテンツの多様性を増加させる可能性，ニュースへの関心が薄いユーザへの関心喚起ができる可能性を示唆する．それらはニュース番組においては，政治的姿勢や知識の分断を緩和し，健全な政治的・社会的議論の実現に寄与することが期待できる．今後はこれらの研究と社会調査を組み合わせ，ユーザの政治知識・関心・姿勢への効果を計測し，AbemaTV のメディアとしての社会的価値をより高める方策について検討する予定である（立命館大学 小川祐樹先生，神奈川大学 高史明先生との共同研究）．

本稿で紹介した研究以外にも豊橋技術科学大学 情報・知能工学系の吉田光男先生，早野宙氏と AbemaTV のコンテンツ視聴と情報探索行動の関連 [4]，立命館大学 小川祐樹先生，西朋里氏，神奈川大学 高史明先生と AbemaTV におけるニュース種別によるユーザの意見表出の違いについて研究を実施しており，多角的な視点で AbemaTV の社会的影響について調査を進めている．

このようなメディアにおけるユーザの情報接触とその影響をユーザの実際の行動履歴から調査した研究は少ない．それはメディアを提供する企業でなければユーザの行動履歴データを持っていないため多くの研究者にとって研究対象にすることが難しいからである．このように行動履歴に基づく接触コンテンツの分析はコンテンツを提供する事業者自身が研究することが有効である．日本企業の事例としてはヤフー社（国立情報学研究所との共同研究） [9] や Gunosy 社 [16] が挙げられる．こういった取り組みは事業者が自身が提供するコンテンツやサービスの社会に与えうる影響を知る

ためにも重要であろう.

参考文献

[1] CyberAgent, Inc. CyberAgent 2Q FY2018 Presentation Material, 2018.

[2] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 113, No. 3, pp. 554–9, jan 2016.

[3] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112, No. 33, pp. E4512–21, aug 2015.

[4] Hiroshi Hayano, Masanori Takano, Soichiro Morishita, Mitsuo Yoshida, and Kyoji Umemura. Analysis of the Influence of Internet TV Station on Wikipedia Page Views. In *The 3rd International Workshop on Application of Big Data for Computational Social Science (workshop at IEEE BigData2018)*, 2018.

[5] Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer, and Olga Slivkó. Does Wikipedia Matter? The Effect of Wikipedia on Tourist Choices. *ZEW Discussion Paper*, Vol. 15, No. 089, 2017.

[6] Shanto Iyengar and Kyu S Hahn. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, Vol. 59, No. 1, pp. 19–39, mar 2009.

[7] Kota Kakiuchi, Mao Nishiguchi, Fujio Toriumi, Masanori Takano, and Kazuya Wada. What influences people to broaden their horizons? In *IEEE/WIC/ACM International Conference on Web Intelligence (WI2018)*, 2018.

[8] Kota Kakiuchi, Fujio Toriumi, Masanori Takano, Kazuya Wada, and Ichiro Fukuda. Influence of Selective Exposure to Viewing Contents Diversity. In *The 10th International Conference on Social Informatics (SocInfo)*, 2018.

[9] Tetsuro Kobayashi and Kazunori Inamasu. The Knowledge Leveling Effect of Portal Sites. *Communication Research*, Vol. 42, No. 4, pp. 482–502, jun 2015.

[10] Adam D I Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 111, No. 24, pp. 8788–90, jun 2014.

[11] Steven Kull, Clay Ramsay, and Evan Lewis. Misperceptions, the Media, and the Iraq War. *Political Science Quarterly*, Vol. 118, No. 4, pp. 569–598, dec 2003.

[12] Kevin Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, Vol. 39, No. 3, pp. 629–649, sep 2017.

[13] Eli Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin, UK, 2011.

[14] Markus Prior. News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout. *American Journal of Political Science*, Vol. 49, No. 3, pp. 577–592, jul 2005.

[15] Markus Prior. *Post-broadcast democracy : how media choice increases inequality in po-*

litical involvement and polarizes elections. Cambridge University Press, 2007.

[16] Yoshifumi Seki and Mitsuo Yoshida. 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)*, Seattle, WA, USA, 2018.

[17] Adam Shehata. Active or Passive Learning From Television? Political Information Opportunities and Knowledge Gaps During Election Campaigns. *Journal of Elections, Public Opinion & Parties*, Vol. 23, No. 2, pp. 200–222, may 2013.

[18] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, pp. 1067–1077, New York, New York, USA, 2015. ACM Press.

[19] S. Yardi and D. Boyd. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society*, Vol. 30, No. 5, pp. 316–327, oct 2010.

[20] 稲増一憲, 三浦麻子. 「自由」なメディアの陥穽. 社会心理学研究, Vol. 31, No. 3, pp. 172–183, mar 2016.

[21] 高野雅典, 森下壮一郎, 小川祐樹. インターネットテレビ局 AbemaTV におけるニュースの「チラ見」がニュース視聴の動機に与える影響. 日本行動計量学会 第 46 回大会, 2018.

[22] 小林哲郎. マスメディアが世論形成に果たす役割とその揺らぎ. 放送メディア研究, No. 13, pp. 105–128, 2016.

[23] 小林哲郎. ソーシャルメディアと分断化する社会的リアリティ. 人工知能学会誌, Vol. 27, No. 1, pp. 51–58, 2012.

[24] 保高隆之. 情報過多時代の人々のメディア選択. 放送研究と調査, Vol. 12, 2018.

当社のメディア事業には，ユーザがコンテンツを投稿できるサービスが数多く属している．中でも Ameba の中心サービスである Ameba ブログは，2004 年からの 14 年以上に亘り 24 億を超える投稿を受け入れてきた日本最大級のブログサービスである．Ameba ブログは誰でも自由にブログの作成が可能であり，文章に加えて画像・動画・音声などの多様なコンテンツを用いてユーザの経験・知見・意見・創作を全世界に容易く公開できる．

ブログサービスという開かれたメディアが抱える最も重要な責務の 1 つに，「自由」と「品質」の両立がある．Ameba ブログはあまねく人々の自由な自己表現の場としての役割を託されて開発と運営が行われているが，無制限に投稿を認めることはサービス全体の品質低下につながるためである．たとえば，アフィリエイトリンクが記事の大半を占めるような記事を大量に投稿する行為（いわゆるアフィリエイトスパム）への対処が行われなければ，他ユーザの記事が埋没してしまう．ユーザが精魂込めて作り上げたコンテンツへのアクセスが阻害されるような状況では，もはやブログサービスは自己表現の場という役割を喪失するのである．

秋葉原ラボでは Ameba ブログの自由と品質を保つために，主に次に挙げる 2 つの取り組みを進めている．1 つ目は，不適切な投稿が行われることを未然に防ぎ，不適切な可能性がある投稿についても迅速に確認できるフィルタリングシステムの整備である．このために，どのようなタイプの不適切投稿が存在するかの類型化と，各タイプに適したフィルタの考案と開発を行っている．2 つ目は，実際に投稿されているコンテンツの状況を確認するアノテーションの枠組みの整備である．アノテーションとは，コンテンツを人手で確認し，たとえば不適切か否かといった情報を付加する処理を指す．アノテーションを効率的・持続的に実施するための下地を整えることは，前述のフィルタを精緻化するため，そして現状の Ameba ブログ全体の品質を俯瞰するために不可欠である．

本技術報告では以上の 2 つの取組みについて，各章に分けて紹介する．2.1 章では，不適切投稿のフィルタリングについて詳説する．具体的には，スパム分類などの知見や，フィルタリングの枠組みと手法を解説する．続く 2.2 章では，アノテーションシステムについて詳説する．ここでは，アノテーションの一連のプロセスを自動化するデータフローや，さまざまなユースケースに対応できるインタフェースを紹介する．

Akihabara Lab

2.1 アメブロを護る: 不適切コンテンツの傾向分析と検知手法の検討

角田 孝昭，數見 拓朗

概　要　本稿では，秋葉原ラボで開発しているコンテンツフィルタの手法と実装や，実運用の過程で得られた知見を紹介する．Ameba ブログのように自由な投稿を認めているサービスには，さまざまな種類の不適切投稿が行われる．本稿では，この中から特にコピーコンテンツスパムを取り上げ，具体的なフィルタリング手法と実装方法について説明する．実装したフィルタは総合監視基盤システム Orion（オライオン）上で動作しており，さまざまなサービスの健全性を支えている．我々の取り組みは，Ameba ブログ上の不適切コンテンツを減少させると同時に，同様のサービスを運営する他社との共有を通じて Web 全体の健全化を促進している．

Keywords　機械学習，スパム，コピーコンテンツ，Hashing，SimHash

1　はじめに

当社のメディア事業では，ユーザのコンテンツ投稿を受け入れるサービスを数多く展開している．たとえば，Ameba ブログ（以下，アメブロと略する）ではユーザが自由にブログを作成し，文章をはじめとするコンテンツを投稿・公開できる．また，アメブロや AbemaTV などのサービスでは，記事や番組など各コンテンツへのコメントをテキスト形式で投稿できる．くわえて，ユーザ間のコミュニケーションを主目的としているサービスも存在する．たとえば，ピグパーティでは仮想空間上でのテキストチャットによるコミュニケーションを行える．

アメブロのようなメディアサービスを運営する上で重要な課題となるのは，サービス健全性の堅守である．特にアメブロでは簡単にブログを作成できるしくみにより投稿の門戸を全世界に開いているが，これは悪意をもつユーザに対しても不適切コンテンツを投稿する機会を与える危険性と表裏一体である．そのため，不適切コンテンツからの防御の成否は，サービスの健全性に直結する．くわえて，不適切コンテンツが大量に投稿されれば，検索エンジン上の表示順位低下や，ストレージ・ネットワークなどの資源が無用に消費されるといった問題にもつながる．

健全性を脅かすような不適切な投稿には，さまざまな種類が存在する．たとえば，アメブロを不正に利用した事例として，アフィリエイト報酬の獲得を企図して大量のアフィリエイト記事を投稿したケースが存在する（アフィリエイトスパム）．このような迷惑行為により，一部の新着記事ページや検索ページなどで一般ユーザの記事が埋もれてしまう問題が発生した．また，コメントを受け入れているサービスやチャットサービスでは，コメント対象やチャット相手への誹謗中傷の他，未成年からの個人情報聞き出しや脅迫・犯行予告といった深刻な犯罪と関連したものも考えうる．

秋葉原ラボではメディア事業が抱えるサービスを健全に保つため，これらの投稿を抽出するため

のフィルタと，フィルタに基づいた総合監視基盤システム Orion（オライオン）を開発・運用している．さまざまな不適切投稿に備えたフィルタ群を，さまざまなサービスに接続可能な Orion 上で動作させることで，あらゆるサービスの不適切投稿を迅速に発見し対応できるしくみを実現している．

本稿は，次のような展開で我々の取り組みを紹介する．まず，2 節では不適切コンテンツの種類とそれらの割合を示し，対応方針について議論する．次に，3 節では，不適切コンテンツの中でもコピーコンテンツスパムと呼ばれる投稿を取り上げ，我々が検出に用いている手法と実装方法を紹介する．続いて，4 節では総合監視基盤システム Orion の概要を説明する．5 節では，これらの取り組みの成果を客観的な指標から評価する．最後に，6 節で本稿を総括し，今後の課題を述べる．

2 不適切コンテンツの類型化

ユーザのコンテンツ投稿を受け入れるサービス運営者にとって，実際にどのような不適切コンテンツが投稿されているのかを把握することは重要である．不適切コンテンツの投稿傾向を把握できれば，これらへ対処する優先度や手法の決定が可能になるためである．そのため，アメブロでは不適切コンテンツの類型化を行い，不適切コンテンツを含めてどのようなコンテンツが投稿されているかを俯瞰するための調査を継続的に実施している．

本節では，特にスパムと呼ばれる不適切コンテンツについて説明し，その類型化を試みる．続いて，これらスパムへの対策について議論する．

2.1 スパム（スプログ）

本稿では，サービス利用規約の観点から望ましくない不適切コンテンツの中でも，特に**スパム**と呼ばれるコンテンツに注目する．ブログ上のスパムは**スプログ**とも呼ばれ，たとえば次のような目的をもって，大量に投稿されるコンテンツを指す．

- 当該ブログや特定の外部 Web サイトの，サービス内の各種ランキングや検索エンジンにおける順位を操作する目的
- アフィリエイト広告の表示数やクリック数を不当に増加させる目的

不適切コンテンツの中でもスパムを取り上げる主な理由は，スパムは次に示す 2 つの特徴をもつためである．

1. 投稿の検出が比較的難しい．不適切コンテンツにはスパムの他，誹謗中傷・法令に反する表現を含むコンテンツも存在するが，これらは特定の文字列を含むものが多いため，キーワードマッチで検知しやすいものが多い[*1]．一方で，スパムに用いられている表現は一般的なコンテンツとの差異が小さい場合も多いため，単純なキーワードマッチによる弁別は困難である．
2. 投稿による影響が大きい．スパムは大量に投稿されるため，その規模からランキング・トレンド・推薦などのサービス内の広範な機能に影響を及ぼす恐れがある．同時に，スパム割合が増加することにより，サービス全体の価値を低下させる要因にもなる．

2.2 スパムの種類と割合

我々は，長期にわたるアメブロ運営過程で得られたさまざまなスパムに基づき，それらを表 2.1.1 のように類型化した．類型化にあたっては，高橋らによる分類 [高橋他, 2009] や Google の公開しているウェブマスター向けガイドライン[*2]を参考にしている．なお，スパム投稿は表 2.1.1 の分類で見たときには複数に該当する場合もある．たとえば，多くのコピーコンテンツスパムはコンテンツの剽窃が主目的ではなく，剽窃したコンテンツにより閲覧数を増やすことで別サイトの SEO 効果を不正に上げたりアフィリエイト報酬を不正に得たりすることが目的である．そのため，コピー

[*1] ただし，隠語や伏せ字などを用いることで，キーワードマッチによる検知を回避しようとする投稿も存在する．我々は，これらの投稿に対して対処できるよう，単純な文字列の一致によらない手法も取り入れている．

[*2] https://support.google.com/webmasters/answer/35769?hl=ja（参照: 2019 年 1 月）

表 2.1.1　スパム（スプログ）の分類とデータセットにおける割合

スパムの種類	説明	割合
発リンクスパム	リンク先サイトに SEO 効果を出すことを主目的として作成されたコンテンツ	34.6%
コピーコンテンツスパム	他のエントリやサイトからコンテンツを盗用し，同一のまま，またはわずかなリライトのみを加えて投稿されたコンテンツ	22.4%
アフィリエイトスパム	オリジナルなコンテンツがほとんどなく，アフィリエイトリンクが主となっているコンテンツ	17.7%
ドアウェイスパム	リンク誘導を主目的として作成されたコンテンツ	15.1%
ワードサラダスパム	一定の規則に基づいて言葉や文章を組み合わせることで生成した，一見文法的には正しくても意味が通っていないコンテンツ	1.4%
キーワードスタッフィングスパム	文章とは関係なくキーワードを多数詰め込み，多数の検索ワードにヒットすることを企図して作成されたコンテンツ	0.8%
リダイレクトスパム	訪問ユーザを強制的に他サイトにリダイレクトするコンテンツ	0.2%
アダルトスパム	卑猥なコンテンツ	0.1%
ノンクレジットスパム	広告コンテンツにも関わらず，[PR]［広告］などの明記がないコンテンツ	0.1%

コンテンツスパムは発リンクスパムやアフィリエイトスパムであることも多い．

スパムの性質により，有効な検知手法は異なったものとなる．たとえば，発リンクスパムは，エントリ内やブログ内での外部へのリンク数を見たり，逆にブログ全体から特定のドメイン・ページへのリンク数を見たりすることで検出しやすい．また，アフィリエイトスパムに対しては，特定のドメインを含む URL へのリンク数に基づくルールベースや，リンク数などを素性に取り入れた機械学習による手法が有効である．一方で，アダルトスパムに対しては，特定のキーワードが含まれているかを見るルールベースや，アダルトスパム特有の単語の出現パターンを学習した機械学習モデルによる判別が効果的である．

続いて，アメブロにおける各スパムの割合を，表 2.1.1 の「割合」列に示す．スパム種別の割合は，次の手順によるアノテーションを行ったデータから算出した．まず，アノテーション対象として，ある特定の期間における 55,852 件の投稿をランダムサンプリングにより抽出した．次に，3 人のアノテータにより，投稿が不適切か否かの判別を行った．最後に，1 人でも不適切としたエントリについて，別のアノテータが表 2.1.1 の分類にしたがっていずれのスパムであるかを分類した．表 2.1.1 の結果に示すように，発リンクスパムが最も多く（34.6%），次いでコピーコンテンツスパム（22.4%），アフィリエイトスパム（17.7%）の順に多かった．複数のブログサイトを対象とした高橋らの調査 [高橋他, 2009] と比較すると，コピーコンテンツスパム[*3]・アフィリエイトスパムが上位に来ている点で共通している．一方で，高橋らのデータではアダルトスパムが比較的多いことを報告しているが（第 3 位）[高橋他, 2009]，本データでは比較的少ない件数であった．この理由は，アメブロがスパムであるかを問わずアダルトコンテンツの投稿を禁止しているためと考えている．

3　コピーコンテンツスパムの検出

本節では，多様なスパムの中でも特にコピーコンテンツスパムを取り上げ，その検出手法を紹介する．コピーコンテンツスパムを取り上げる理由は，比較的多く見られるスパムである一方で検出が難しいためである．コピーコンテンツスパムの割合は，我々の調査（表 2.1.1）では第 2 位であり，

[*3] 高橋らは「引用型」スパムとしている [高橋他, 2009].

高橋らの調査では第 1 位であった [高橋他, 2009]. しかし，コピーコンテンツスパムのエントリの大半は「剽窃であること」を除けば一見して問題のないコンテンツであるため，単語の出現パターンなどの特徴を利用した機械学習による分類器は有効ではない[*4].

そこで，本節では「剽窃であること」を検出することで，コピーコンテンツスパムであるかを判定する手法について説明する．なお，本節の説明は數見・内藤の報告 [數見・内藤, 2015] に基づいている．

3.1 検出手法

我々は，ユーザが投稿した記事が剽窃であるかの判別を，大量の記事集合から類似する記事を抽出するタスクとして考える．このタスクを解くにあたっては，テキスト間の類似度を高速に計算する必要がある．アメブロでは一日あたり数十万件の記事が投稿されており，これらの記事毎に大量の候補との間での類似度計算が求められるためである．

我々は記事間の類似度を高速に計算するため，SimHash [Charikar, 2002] に注目した．SimHash は，データをある方法でハッシュ化し，類似度の計算をハッシュ値間の低コストなビット演算に帰着させることで高速化を実現している．理論的には，ランダムに振る舞うハッシュ関数を用いて類似度を確率変数で表し，類似度の計算を経験的確率の計算で置き換えている．

SimHash によるコサイン類似度の近似

Charikar が提案した SimHash は, 近似的なコサイン類似度を高速に求める手法である [Charikar, 2002]．重複コンテンツ抽出のアルゴリズムとしても採用事例が多く，Web サイト [Manku et al., 2007, Pi et al., 2009] やショートメッセージ [Kumar & Govindarajulu, 2013] の重複検出に用いられている．

まず，類似度を求めたいデータが d 次元ベクトル $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ で表されるとする．SimHash では次のようなハッシュ関数 $h(\boldsymbol{x})$ を考える．

$$h(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{r} \cdot \boldsymbol{x} \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1.1}$$

ここで $\boldsymbol{r} \in \mathbb{R}^d$ は各成分が平均 0，分散 1 の正規乱数からなるベクトルである．このハッシュ変換を施した $h(\boldsymbol{x})$ と $h(\boldsymbol{y})$ が一致する確率は，ベクトル $\boldsymbol{x}, \boldsymbol{y}$ のなす角 θ を用いて次のように書ける[*5].

$$P(h(\boldsymbol{x}) = h(\boldsymbol{y})) = \frac{\pi - \theta}{\pi} = 1 - \frac{\theta}{\pi} \tag{2.1.2}$$

これを θ について解けば，θ は $h(\boldsymbol{x})$ と $h(\boldsymbol{y})$ の一致確率から計算できることが分かる．

$$\theta = \pi \left(1 - P(h(\boldsymbol{x}) = h(\boldsymbol{y}))\right) \tag{2.1.3}$$

SimHash では，$P(h(\boldsymbol{x}) = h(\boldsymbol{y}))$ を経験的確率から推定することで θ を近似的に求める．経験的確率を求めるため，相異なる $\boldsymbol{r}$ の下での f 個のハッシュ関数 $h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_f(\boldsymbol{x})$ による変換を施した f ビットからなるビットベクトル $H(\boldsymbol{x})$ を考える．

$$H(\boldsymbol{x}) = \{h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_f(\boldsymbol{x})\} \tag{2.1.4}$$

$H(\boldsymbol{x})$ を用いると，経験的確率 $\hat{P}(h(\boldsymbol{x}) = h(\boldsymbol{y}))$ は次のように求められる．

$$\begin{aligned} &\hat{P}(h(\boldsymbol{x}) = h(\boldsymbol{y})) \\ &= 1 - \hat{P}(h(\boldsymbol{x}) \neq h(\boldsymbol{y})) \\ &= 1 - \frac{H(\boldsymbol{x}), H(\boldsymbol{y}) \text{ 間で一致しないビット数}}{f} \end{aligned} \tag{2.1.5}$$

$H(\boldsymbol{x}), H(\boldsymbol{y})$ 間で一致しないビット数とは，すなわち $H(\boldsymbol{x}), H(\boldsymbol{y})$ のハミング距離に相当する．これ

[*4] アメブロのコピーコンテンツスパムに限った場合，剽窃元の記事はニュースサイトが多いことから，スパム特有の単語出現パターンは多くの非スパム記事とは異なった傾向にはなっている（時事用語やニュース記事独特の記法が多く出現する）．しかし，ニュースサイトから適切に引用した非スパム記事も存在するため，単語出現パターンを根拠にした判定は誤検知を起こす可能性が高い．

[*5] 式 2.1.2 は，幾何的には d 次元空間をランダムな超平面で 2 つに分割した際に $\boldsymbol{x}, \boldsymbol{y}$ が同じ側にある確率としてとらえることもできる．式 2.1.2 の証明は Goemans & Williamson が与えているが [Goemans & Williamson, 1995], Ravichandran らによる解説 [Ravichandran et al., 2005] も参考になる．

を式 2.1.3 に代入することで

$$\theta \approx \pi \left(1 - \frac{H(\boldsymbol{x}), H(\boldsymbol{y}) \text{ 間のハミング距離}}{f}\right) \tag{2.1.6}$$

を得る．式 2.1.6 は，結局 θ の推定値 $\hat{\theta}$ が

$$\hat{\theta} \propto -\big(H(\boldsymbol{x}), H(\boldsymbol{y}) \text{ 間のハミング距離}\big) \tag{2.1.7}$$

であることを示している．よって，実用上は $H(\boldsymbol{x}), H(\boldsymbol{y})$ のハミング距離をそのまま $\boldsymbol{x}, \boldsymbol{y}$ 間の距離（類似していない度合い）として用いる場合が多い．我々の実装においてもハミング距離を利用している．

さらに Manku らは，以上に示す SimHash の空間計算量を削減する手法を提案している [Manku et al., 2007]．計算方法の詳細は Manku らの論文 2 節 [Manku et al., 2007] に譲るが，この手法では各次元ごとに適当なハッシュ変換を適用した値[*6]に注目することで，実際に相異なる f 個の乱数ベクトル $\boldsymbol{r}$ を生成することなく変換後のビットベクトル $H(\boldsymbol{x})$ を得られる．我々の実装も Manku らの提案手法を用いている．

テキストのハッシュ化

以上で説明した SimHash を用いるため，我々は記事テキストの文字 bi-gram を利用した．テキストに注目した理由は，コピーコンテンツスパムの多くはテキスト検索のヒットを企図していることから，剽窃コンテンツの対象はテキストであることが多いためである．また，単語ではなく文字に注目した理由は，コピーコンテンツスパムのように表層的な類似度を測る場合は文字を単位とする方がより適切なためである．Manku らの手法 [Manku et al., 2007] を利用すれば，文字 bi-gram をハッシュ化した値から記事全体のハッシュ値を生成できる．

3.2　実装

本節では，アメブロに投稿されるコピーコンテンツスパムを SimHash により検出するための実装について説明する．まず，実装の中心である，大量のハッシュ値データベースから特定のハッシュ値と類似するレコードを高速に抽出するしくみについて説明する．続いて，このしくみを用いたシステム全体の構成を示す．

ハッシュ値のデータベースへの格納

3.1 節では類似度の計算をハミング距離の計算で近似できることを示したが，クエリとなるハッシュ値に対して大量のハッシュ値との間でハミング距離を計算する必要がある場合は，依然として計算コストが高い．ここで目的がコピーコンテンツスパムの検出であることを考えると，類似度が低いハッシュ値については関心がない．そのため，あらかじめハッシュ値の一部ビットが一致するレコードのみを対象とすることで，ハミング距離の計算が必要となるレコードの限定が可能になる．ここでは，ハッシュ値のデータベース上への格納方法を工夫することで，ハッシュ値の一部が一致するレコードを抽出する手法について説明する．

Manku らは，ハッシュ値の一部ビットが一致するレコードを高速に抽出するため，ハッシュ値のビットを複数の変換規則で並び替えて，変換規則ごとに別個のテーブルに格納する手法を提案している [Manku et al., 2007]．各テーブルは，変換後ハッシュ値の先頭 p ビットをキーにもつレコードをソートされた状態で保持する．このキーを利用すれば，クエリとなるハッシュ値を各テーブルと対応する規則で変換し，各テーブルで変換後ハッシュ値の先頭 p ビットが一致するレコードを高速に抽出できる．p の長さは絞り込める量とレコード容量とのトレードオフとなるが，適切な p を選択すれば，一致するキーの数が元の全レコード数

[*6] 各次元（素性）ごとに適当なハッシュ変換を施すため，各次元に対応する入力が一意であれば必ずしも全次元数 d が定まっていなくてもよい．たとえば，入力として数値を入力（`hash(1)`）とするのではなく，文字列を入力（`hash("a")`）としてもよい．この特徴は，3.1 節のように文字列を入力とする場合と極めて相性が良い．すなわち，入力されうる全単語（vocabulary）が定まっていない状況においてもハッシュ値を求めることができるため，あらゆるテキストを入力とできる利点を有する．

ハッシュ値
00000001
00100010
01010101
11111111

→

テーブル T_1（左 1 ビット回転）	
000	00000010
010	01000100
101	10101010
111	11111111

テーブル T_2（左 2 ビット回転）	
000	00000100
100	10001000
010	01010101
111	11111111

テーブル T_3（左 3 ビット回転）	
000	00001000
000	00010001
101	10101010
111	11111111

...

図 2.1.1 ハッシュ値のビットを複数の方法で並び替えて別個のテーブルに格納する例．先頭 $p = 3$ ビットをキーとしている

00100011 →

テーブル T_1（クエリ: 01000110）	
000	00000010
010	01000100
101	10101010
111	11111111

テーブル T_2（クエリ: 10001100）	
000	00000100
100	10001000
010	01010101
111	11111111

テーブル T_3（クエリ: 00011001）	
000	00001000
000	00010001
101	10101010
111	11111111

...

図 2.1.2 ハッシュ値の一部が一致するレコードを抽出する例．下線部が，先頭 $p = 3$ ビットの一致を示している

の $1/2^p$ となることが期待できるため大きく絞り込める．なお，用意すべきテーブル数についての詳細は Manku らの論文 3.1.2 節 [Manku et al., 2007] を参照されたい．

具体的なデータの格納方法を，ハッシュ値が $f = 8$ ビットであり，キーとなる先頭ビット数が $p = 3$ である場合を例として，図 2.1.1 を用いて説明する．まず，ハッシュ値を格納するためのテーブル T_i を複数用意する．次に，各テーブルごとにハッシュ値を並び替える規則を割り当てる．この規則は，他テーブルの規則と重複しない全単射な写像であれば何でも良く，たとえば図 2.1.1 ではビット回転[*7]を例に説明している．最後に，抽出対象としたいハッシュ値の集合を各テーブルの変換規則にしたがって変換し，先頭 p ビットをキーにもつレコードとして格納し，ソートすることで完了となる．

次に，図 2.1.1 で用意したテーブルを用い，実際にクエリとなるハッシュ値が与えられた際に一部ビットが一致するレコードを抽出する手法を図 2.1.2 を用いて説明する．まず，与えられたハッシュ値を各テーブルと対応する規則で変換する．次に，各テーブルにおいて，変換後ハッシュ値と先頭 p ビットが一致しているものを，キーを参照して抽出すればよい．

Apache HBase 上での表現

実際に以上で説明したハッシュ値を変換したデータを格納するため，我々のシステムでは Apache HBase（以下，HBase と呼ぶ）を利用している．HBase は高い性能とスケーラビリティを備えたカラム指向型データベースであり，その性能から採用実績も多い．秋葉原ラボにおいても，コピーコンテンツスパムのデータベースにとどまらず，さまざまなシステムを支える基盤として機能している．

[*7] ビットシフトにおいて，あふれた桁を逆の端に持ってくるような演算．たとえば，11000001 を左に 1 ビット回転すると 10000011 となる．ビットローテーション，循環シフトとも呼ばれる．

[*8] https://hbase.apache.org/

HBase[*8] でテーブル毎の先頭 p ビットが一致するハッシュ値を検出できる構造を作るには Rowkey と Qualifier を利用すればよい．具体的には，これらを以下のように設定する．

Rowkey　テーブル番号 i と変換後のハッシュ値の先頭 p ビットを結合したバイナリとする．

Qualifier　変換後のハッシュ値全体とする．

Value　必要に応じ，記事の属性情報などを保存してもよい．

類似ハッシュ値を検索する際は，確認したい投稿のハッシュ値をテーブルの数だけ変換し，その先頭 p ビットとテーブル番号 i を組み合わせたバイナリが一致する Rowkey のレコードを取得すればよい．該当するレコードを取得後は，それぞれの Qualifier を参照すればハッシュ値全体を得られる．

システム構成

アメブロに投稿されるコピーコンテンツスパムを検出するシステムの概略図を図 2.1.3 に示す．処理は，コピー元記事データベースを用意するバッチ処理と，投稿記事のコピーコンテンツスパム判定を行うオンライン処理の 2 つに分かれている．

バッチ処理では，コピー元となる記事の候補をハッシュ値に変換し，3.2 節で説明した手順でデータベースへ格納する．コピー元記事の候補には，アメブロ内部の記事に加え，いくつかの他サイトからの記事も利用している．この理由は，コピーコンテンツスパムはニュースサイトなど他サイト上の記事を剽窃の対象とする場合も多いためである．

オンライン処理は，投稿の度に行われるスパム判定処理である．入力された投稿記事をハッシュ値に変換し，データベース上のハッシュ値と照合する．この結果，類似するハッシュ値が見つかった場合，コピーコンテンツスパムの疑いがあると判定して結果を返す．

4　Orion: 総合監視基盤システム

3 節で説明した手法を含め，各種コンテンツフィルタは秋葉原ラボで開発している総合監視基盤システム Orion（オライオン）上で動作している．Orion は，自動フィルタと人間のオペレータが効率的に協働できるしくみや，多様なサービス上のさまざまな文書・画像・動画投稿を同様の画面で監視できる Web UI を備える監視システムである [Fujisaka, 2018]．本節では Orion の概要を説明する．

Orion において，特にフィルタリングの観点から見た主要な特長は，次の 2 点に集約できる．

まず，フィルタとオペレータの 2 段階チェックにより，確実な判定を行っている点がある．各種フィルタによって不適切と判定されたエントリは，必ずしも真に不適切とは限らない場合がある．たとえば，3 節で説明したコピーコンテンツスパムの抽出手法は経験的確率に基づく近似から類似度を求めているため，低確率ながらオリジナルのコンテンツをコピーと誤って判定してしまう可能性がある．そのため，各フィルタで不適切とされたエントリは，オペレータが各サービスの規約にのっとって真に不適切なコンテンツと判定した場合に限って非表示などの対応が行われる[*9]．このしくみにより，自動フィルタにより大量のコンテンツを確認するコストを軽減しながら，オペレータによる確認作業を経ることで確実な対応を実現している．

次に，サービスやコンテンツに応じてフィルタを柔軟に使い分けられるしくみによって，高い汎用性を実現している点がある．監視すべきコンテンツや，監視の観点・基準には，サービス間でさまざまな共通点・非共通点が存在する．たとえば，サイバーエージェントが運営するサービスではいずれも誹謗中傷や犯罪と関連する投稿を禁じているため，どのサービスでもこのような不適切投稿を抽出する需要が存在する．一方で，コピーコンテンツスパムが投稿されるのはブログが中心であ

[*9] ただし，極端なスパムについてはオペレータによる確認を経ずに非表示状態となる場合がある．具体例として，極端な連続投稿や，特定のテンプレートを利用したスパムがある．

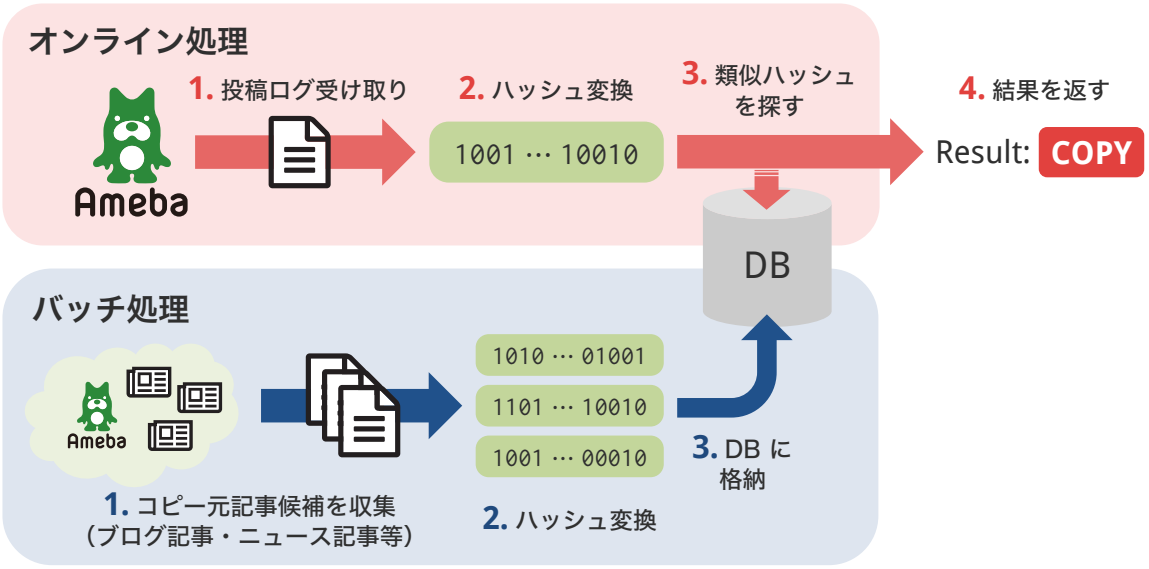

図 2.1.3　コピーコンテンツスパムを検出するシステムの概略図

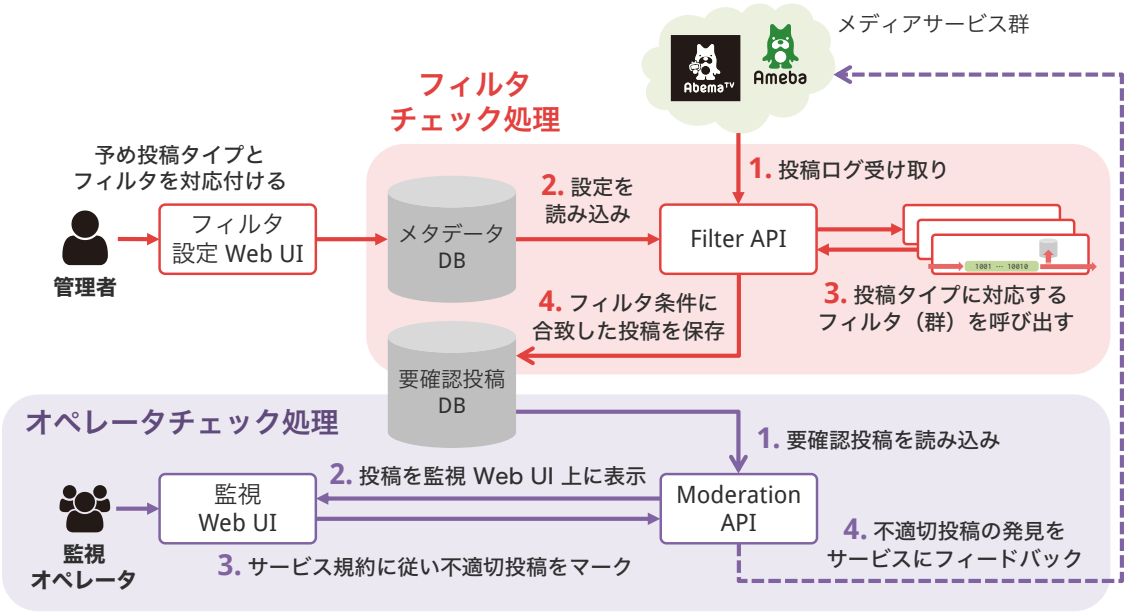

図 2.1.4　Orion の概略図

るため，コピーコンテンツスパムフィルタはブログからの需要が特に高い．こう言ったさまざまな需要に応えるため，Orion は特定のサービスにおける特定のコンテンツと，フィルタ（群）を多対多関係で対応付けられるしくみを備えている．

Orion の処理の流れを，図 2.1.4 に示す概略図を用いて説明する．

ステップ 1: フィルタチェック処理

ユーザがサービスに投稿すると，サービス内部から Orion へ投稿データが転送される．Orion は，あらかじめ対応付けられたフィルタ（群）により転送された投稿をチェックする．フィルタのスコアが特定の閾値を上回った場合，投稿データは不適切なコンテンツの可能性があるとしてデータベースへ保存される．

ステップ 2: オペレータチェック処理

オペレータは Orion の Web インタフェース

を通じて，データベースに保存された不適切コンテンツの可能性がある投稿を確認する．オペレータが確認した結果，真に不適切である投稿が発見された場合，Orion からサービスの API を呼び出すことにより不適切コンテンツが発見されたことをフィードバックする．

Orion の処理の流れにおいて，サービスが関係するのはフィルタの呼び出し処理と判定結果の受け取り処理のみである．そのため，各サービスがこれらの処理を実装すれば，すぐに監視を始められる汎用性を備えている．

5　スパム対策効果の確認

本節では，以上に示した手法をはじめとするスパム対策の効果を，客観的指標により確認する．まず，サービス上で不適切コンテンツが占める度合いを表す指標（Spam Index）を定義し，続けてこの指標の推移を示す．

5.1　Spam Index の定義

我々はサービスの健全度を客観的に評価するため，Google Search Console[*10]を利用する．Google Search Console では，Google が目視によりブログをスパムなどの不適切コンテンツとして評価したリストを確認できる[*11]．

Google Search Console 上の不適切コンテンツのブログリストを用い，サービスの不適切コンテンツの割合 Spam Index（SI）を次の式 2.1.8 のように計算する．

$$\mathrm{SI}_t = \frac{S_t}{S_t + H_t} \tag{2.1.8}$$

SI_t は時刻 t における不適切コンテンツの割合であり，S_t は時刻 t での Google Search Console 上の不適切コンテンツのブログリストに存在するアカウント数を示す．H_t は，リストに存在しない全アカウント数を示す．

5.2　Spam Index の推移

2016 年 5 月から 2018 年 10 月における SI の推移を図 2.1.5 に示す．なお，図中では 2016 年 5 月における SI が 100 となるよう正規化している．

図 2.1.5 より，SI は調査を始めた 2016 年 5 月からおおむね低下傾向にあり，スパム対策の効果が客観的指標からも確認できる．特に（1）2016 年中期から 2016 年末と（2）2018 年中期から 2018 年後期に掛けて SI が大きく減少している．期間（1）では，ある特徴に基づくルールベースフィルタを複数追加することで，スパムアカウントを大きく減らすことができた．期間（2）では，キーワードスタッフィングスパム・海外から投稿されるスパムへの対策を強化した結果，さらなるスパムの削減を実現している．

6　おわりに

本稿では，アメブロにおける不適切コンテンツの中でもスパムと呼ばれるコンテンツに注目し，その分類と割合を紹介した．その中からコピーコンテンツスパムを取り上げ，検出手法を紹介した．また，これらのフィルタを組み込んだ総合監視基盤システム Orion の概要を説明した．最後に，スパム対策の効果を Spam Index の観点から確認した．

本稿で取り上げたコピーコンテンツスパム検出の課題の一つとして，援用するハッシュ化手法の精緻化がある．SimHash はコサイン類似度を近似する手法であるのに対し，Jaccard 係数を近似する手法として minwise hashing [Broder et al., 1997], b-bit minwise hashing [Li & König, 2010], Odd Sketches [Mitzenmacher et al., 2014] などが提案されている．中でも Odd Sketches は，b-bit minwise hashing と比較して類似度が高いデータ間においても頑健であることが示されている．そのため，コピーコンテンツスパムの検出タスクにも適していると考えている．

また，技術的側面以外からの課題として，不適切

[*10] https://search.google.com/search-console/about?hl=ja（参照: 2019 年 1 月）
[*11] https://support.google.com/webmasters/answer/9044175?hl=ja（参照: 2019 年 1 月）

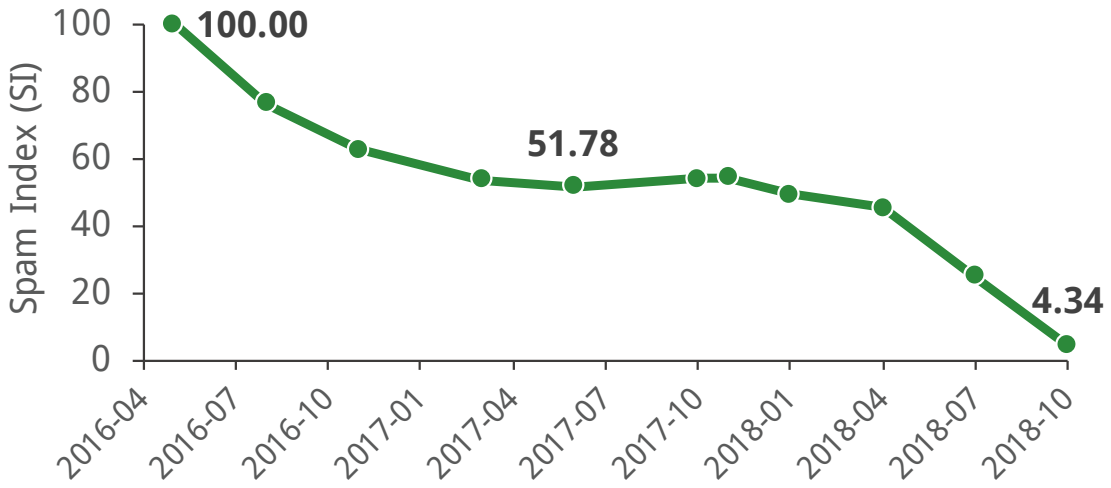

図 2.1.5 Spam Index（SI）の推移．2016 年 5 月における SI が 100 となるよう正規化している

コンテンツへの対処を共通の目的として掲げる企業間連携の活性化がある．不適切コンテンツへの対処は，サイバーエージェントにとどまらず，同様のホスティングサービスを運営する企業間で共通する課題である．たとえば，アメブロである特徴をもつスパムが観察された場合，同様の特徴をもつエントリが別のブログサービスにも投稿されている場合がある．この場合，スパムの態様や対処方法を企業間で共有することで，他社サービスにおけるスパムの対処も促進できる．現在アメブロでは，このような課題解決を目的として設立されたプロジェクトである Advanced Hosting Meetup[*12]への参画を通じ，Web 全体の健全化を進めている．本プロジェクトでは，Google ウェブマスター向け公式ブログの記事[*13]で紹介されているように，単独のホスティングサービスでは実現できない取り組みを行っている．我々は，引き続き企業間連携を推し進めることで，アメブロなどの自社サービスにとどまることなく Web 全体の健全化を促したいと考えている．

参考文献

[Broder et al., 1997] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, Geoffrey Zweig. “Syntactic Clustering of the Web.” *Computer Networks*, Vol. 29, pp. 1157–1166, 1997.

[Charikar, 2002] Moses S. Charikar. “Similarity Estimation Techniques from Rounding Algorithms.” in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC-2002)*, pp. 380–388, 2002.

[Fujisaka, 2018] Yusuke Fujisaka. “Orion: An Integrated Multimedia Content Moderation System for Web Services.” in *ACM International Conference on Multimedia Retrieval (ICMR-2018), Industrial Session*, 2018.

[Goemans & Williamson, 1995] Michel X. Goemans, David P. Williamson. “Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming.” *Journal of the ACM*, Vol. 42, No. 6, pp. 1115–1145,

[*12] https://webmaster-ja.googleblog.com/2016/01/advanced-hosting-meetup.html（参照: 2019 年 1 月）
[*13] https://webmaster-ja.googleblog.com/2017/07/advanced-hosting-meetup-2017.html（参照: 2019 年 1 月）

1995.

[Kumar & Govindarajulu, 2013] J. Prasanna Kumar, P. Govindarajulu. “Near-Duplicate Web Page Detection: An Efficient Approach Using Clustering, Sentence Feature and Fingerprinting.” *International Journal of Computational Intelligence Systems*, Vol. 6, pp. 1–13, 2013.

[Li & König, 2010] Ping Li, Arnd Christian König. “b-Bit Minwise Hashing.” in *Proceedings of the 19th International World Wide Web Conference (WWW-2010)*, pp. 671–680, 2010.

[Manku et al., 2007] Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma. “Detecting near-duplicates for web crawling.” in *Proceedings of the 16th International World Wide Web Conference (WWW-2007)*, pp. 141–149, 2007.

[Mitzenmacher et al., 2014] Michael Mitzenmacher, Rasmus Pagh, Ninh Pham. “Efficient Estimation for High Similarities Using Odd Sketches.” in *Proceedings of the 23rd International World Wide Web Conference (WWW-2014)*, pp. 109–118, 2014.

[Pi et al., 2009] Bingfeng Pi, Shunkai Fu, Weilei Wang, Song Han, Roboo Inc, P R China. “SimHash-based Effective and Efficient Detecting of Near-Duplicate Short Messages.” in *Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT-2009)*, pp. 20–25, 2009.

[Ravichandran et al., 2005] Deepak Ravichandran, Patrick Pantel, Eduard H. Hovy. “Randomized Algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering.” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pp. 622–629, 2005.

[高橋他, 2009] 高橋哲朗, 雄也野田, 岩倉友哉. スプログの調査と実システムにおける判別手法. 言語処理学会第 15 回年次大会, 2009 年.

[數見・内藤, 2015] 數見拓朗, 内藤遥. 実システムへ適用可能な引用型スプログの検知手法の検証. 人工知能学会合同研究会 第 9 回データ指向構成マイニングとシミュレーション研究会（SIG-DOCMAS-2015）, 2015 年.

Akihabara Lab

2.2 アノテーションを支える Orion Annotator の紹介

上辻 慶典

概　要　秋葉原ラボではコンテンツ推薦やスパム検知などのシステムで機械学習を活用している．推薦システムではアクセスログのようなユーザのフィードバックを学習データとして用いるが，スパム検知などでは人手でアノテーション作業を行い学習データとすることが多い．そこで，機械学習に求められる大規模のデータに対するアノテーション作業を効率化するため，我々はアノテーションシステム Orion Annotator を開発した．Orion Annotator は当社データ基盤との連携やストレスなく操作できるアノテーション画面，さまざまなタスクをサポートするカスタマイズ性を備えている．本稿では Orion Annotator 開発の背景や，その機能について紹介する．

Keywords　機械学習, アノテーション

1　背景

秋葉原ラボではコンテンツ推薦やスパム検知などのシステムで機械学習を活用している．こういった機械学習で高い精度を達成するためには，豊富で正確な学習データが必要である．推薦などでは学習データとして購買履歴のようなユーザのフィードバックを利用できるが，スパム検知のようにアノテーションによって学習データを人力で作成しなければならない機械学習タスクも多い．

たとえば Ameba ブログでは高精度なスパムフィルタの構築とサービス品質の把握のため，ブログ記事に対してスパム投稿か否か，またスパム投稿であればその種類を付与するアノテーションを実施してきた（2.1 章参照）．このアノテーションは，Orion Annotator が導入されるまでは次の手順を月ごとに行っていた．

1. Ameba ブログのスパム対策チームがアノテーション対象となるブログ記事（エントリ）を抽出し，アノテーション作業管理者に URL のリストを e メールで送信する．
2. アノテーション作業管理者がアノテーション作業者にエントリを分配する．
3. アノテーション作業者がエントリがスパム投稿かを判定し，スプレッドシート上にアノテーション結果を入力する．
4. アノテーション作業管理者がアノテーション結果を集計し，スパム対策チームに送信する．

しかし，上記のフローによるアノテーションでは，次の課題があった．

- 人手で入力されたスプレッドシート形式のファイルでアノテーション結果が納品されており，分析や機械学習の前処理に工数を要していた．そのため，より構造化・規格化されたデータ形式が必要だった．
- エントリの送信やアノテーション結果の受領が e メールで行われており，アノテーション結果を利用したレポーティングやモデル学習の自動化が難しかった．

- アノテーション作業者にアノテーション専用のインタフェースが用意されず，ブログ記事の確認やアノテーションの記入に煩雑な操作を要していた．

そこで，これらの課題を解決するためのシステムとして我々は Orion Annotator を開発した．Orion Annotator はエントリの抽出やアノテーション済みデータの出力を自動化するデータフローと，アノテーション作業のためのインタフェースを提供するシステムである．自動化されたデータフローによって人を介さないアノテーションデータの利活用を実現し，操作性に優れたインタフェースによって効率的なアノテーション作業を可能としている．さらに，インタフェースは高いカスタマイズ性を備えており，画像分類や情報抽出など多くのアノテーション需要を満たすアノテーション基盤として利用できる．

本稿では，まず 2 節で既存のアノテーション支援ツールと比較し Orion Annotator の特徴を述べる．次に 3 節で Orion Annotator の提供するデータフローについて，4 節でインタフェースについて説明する．

2　既存ツールとの比較

既存のアノテーション支援ツールの概観を通じ，Orion Annotator を開発するに至った背景を説明する．機械学習において学習データのアノテーションは重要なタスクであり，アノテーションのためさまざまなツールが開発されてきた．

たとえば画像・動画メディアに対するアノテーションでは LabelImg[*1]，自然言語処理の分野では brat [1] などがある．これらのプロダクトは画像やテキストといった特定の分野に特化しているため，必要なアノテーションタスクの要件を満たせない場合がある．たとえば Ameba ブログのスパム判別アノテーションの場合，テキストではなく実際のブログ記事ページを表示させてアノテーションを行う必要があり，既存のツールでは対応できなかった．

AWS SageMaker Ground Truth[*2]のように SaaS として提供され，インタフェースのカスタマイズ機能を備えたプロダクトもある．Orion Annotator は当社のシステムとシームレスに連携する点で優れており，データフローを統一的に扱うことで新たな実装を要さずアノテーションを開始できる．

3　Orion Annotator のデータフロー

Orion Annotator はアノテーション対象の抽出からアノテーション済みデータの出力までを担うシステムである．アノテーション作業は手動で行われるが，それを取り巻く入出力を自動化することで継続的なデータ整備を支えている．Orion Annotator の扱うデータフローの概要を次の図 2.2.1 に示す．本節ではデータフローの各段階について順に説明する．

3.1　エントリのインポート

データフローの起点はアノテーション対象（エントリ）のインポートである．エントリは当社のデータ処理基盤である Patriot [2] からインポートされるほか，限られた件数のエントリに対してアノテーションする場合は CSV ファイルからインポートすることもできる．Ameba ブログのスパム判別アノテーションの場合は，Patriot に保存されているブログ記事から，月ごとに一定数をランダムサンプリングしインポートしている．

Orion Annotator へのデータのインポートは Hive クエリで記述され，サンプリングのロジックや前処理を柔軟に設定することが可能である．したがって，Orion Annotator に新たなデータソースとの連携を実装することなく，Patriot に蓄積されたさまざまなサービスのデータを用いたアノテーションを開始できる．

[*1] https://github.com/tzutalin/labelImg
[*2] https://aws.amazon.com/jp/sagemaker/groundtruth/

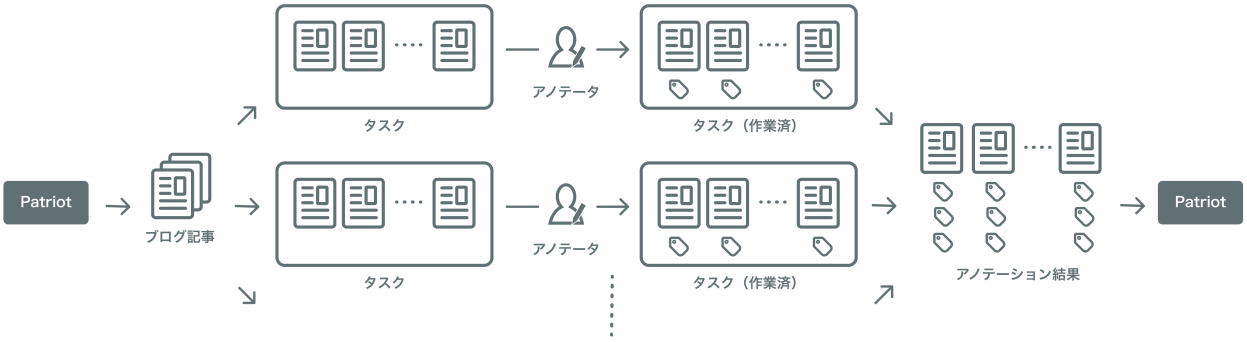

図 2.2.1 Orion Annotator のデータフロー

3.2 アノテーション作業者への割り当て

Orion Annotator はエントリのインポートと同時にアノテーション作業者への割り当てを行う．エントリは，あらかじめアノテーション管理者が規定した作業者ごとのエントリ件数などの設定にしたがって分配され，タスクという形で作業者に割り当てられる．タスクが割り当てられ次第，作業者はアノテーションを開始できる．

なお，管理者は 1 つのエントリが複数のタスクに分配されるように設定することも可能である．この設定によって同じエントリに対して複数の作業者がアノテーションを付与できる．

3.3 アノテーションのエクスポート

アノテーション結果を分析や機械学習から利用するため，日次バッチで Patriot にエクスポートする．エクスポートする際はタスクの完了を待つことなく，アノテーションが完了しているエントリをすべてエクスポートの対象とする．したがって，アノテーションデータを逐次的に機械学習モデルへ反映させることができる．

Patriot へと出力されたデータは，当社のデータ可視化ツールである Patriot FDC [3] や Apache Zeppelin[*3]などでレポーティングや分析ができる．Ameba ブログのスパム判別アノテーションでは Patriot FDC のレポーティング機能を利用し，アノテーション傾向の変化をグラフで表示することで調査結果を把握している．

4 Orion Annotator のインタフェース

Orion Annotator はアノテーションを付与するための Web インタフェースを提供する．本節ではアノテーションを効率的に付与するためのインタフェースと操作について説明する．

4.1 アノテーション操作

図 2.2.2 は Orion Annotator で提供しているアノテーション画面の一例である．アノテーション画面の左側にはエントリが表示され，右側にはアノテーションを選択するコントロールが表示される．アノテーション作業者は以下の作業を繰り返す．

1. 左側のエントリを閲覧する
2. それに該当するラベルを右側のコントロールで選ぶ
3. 確定ボタンを押下する

アノテーションを確定すると自動的に次のエントリが表示される．確定ボタン以外にも保留ボタンによってアノテーションを後回しにしたり，データ欠損ボタンでブログが削除されていて閲覧できないことを報告できる．場合によってはテキストメモでアノテーション管理者へのメッセージを記述する．

このように 1 つの画面でアノテーション作業が完結するインタフェースを提供することで，余計な操作を排した効率的な作業が可能となっている．

[*3] https://zeppelin.apache.org

図 2.2.2　アノテーション画面のスクリーンショット

他の効率的なアノテーションのための工夫として，アノテーションの入力や確定・保留・データ欠損ボタンにショートカットキーを割り当てることで各種操作を省力化している点や，シングルページアプリケーション（SPA）を採用したことによる高速な画面遷移を実現している点がある．

4.2　インタフェースのカスタマイズ

多くのデータ形式やアノテーション需要に対応するため，Orion Annotator のインタフェースは柔軟にカスタマイズできる．管理画面から設定することで対応できるアノテーションの種類と，インタフェースの拡張について説明する．

エントリの設定

Orion Annotator は URL 形式，テキスト形式，画像形式のエントリをサポートしている．図 2.2.2 で例示したアノテーションは Ameba ブログの記事 URL をエントリとして扱い，URL の示すページをインラインフレームで表示する例である．

コントロールの設定

ラベルを入力するためのコントロールの種類をタスクに合わせて指定できる．現在，具体的なコントロールとしてボタンとドロップダウンリストを実装している．ボタンでのアノテーションの場合はラベルに確信度を付与でき，「発リンクスパムである可能性 50%」といった段階的な確信度を伴ったラベルの付与が可能である．ドロップダウンリストの場合は，ラベル数が多い場合に対応するための入力補完を備えている．いずれのコントロールを利用した場合でも，1 つのエントリに「発リンクスパムかつアフィリエイトスパム」のように複数のラベルを与えられるかを設定できる．

アノテーション形式の設定

ここまでで紹介したアノテーションはエントリ全体に対してラベルを付与するものである．しかし，エントリの一部に対してラベルを付与するといった特殊なアノテーションのニーズもある．画像ではセグメンテーション，テキストでは固有表現抽出などのタスクが該当する．Orion Annotator では現在テキストの範囲に対してアノテーションする機能をサポートしている（図 2.2.3）．テキス

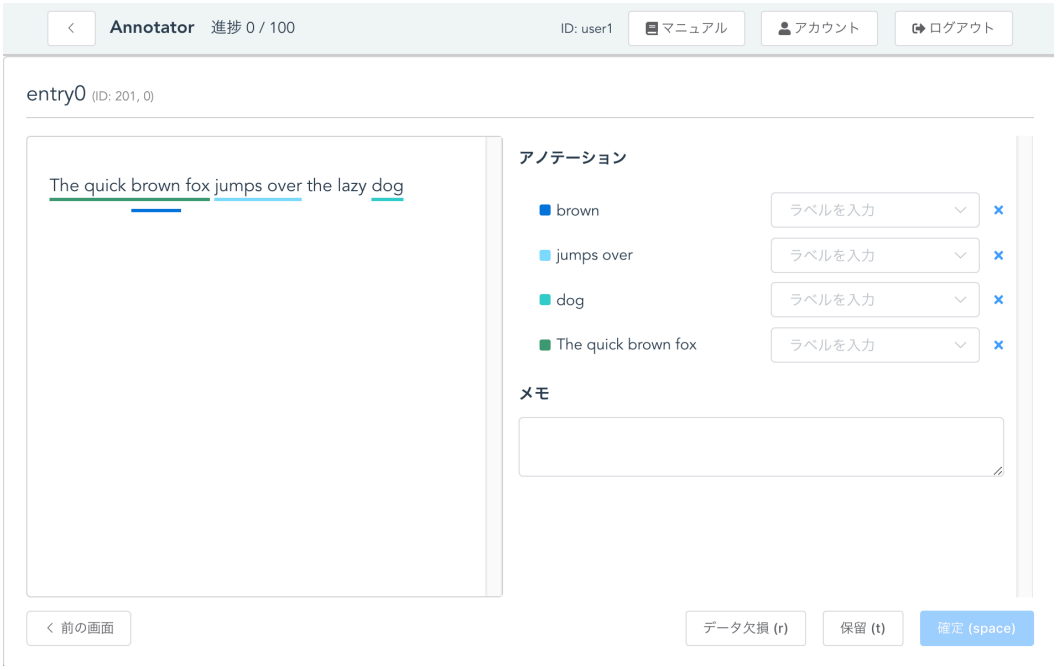

図 2.2.3 テキスト範囲に対するアノテーション

トの一部をマウス操作で選択し，選択範囲に対してラベルを付与できる．

インタフェースの拡張

Orion Annotator の Web インタフェースは Vue.js[*4]で記述されており，新たなエントリ形式やコントロールを簡単に追加できるようプラガブルな設計を行っている．特別なインタフェースを要するアノテーションも多いため，新たなインタフェースを実装しやすいことは重要である．

たとえば，エンティティリンキング[*5]というタスクでは，テキストの範囲に対し Wikipedia の見出し語をラベルとして付与するアノテーションが必要である．このアノテーションではラベル選択のドロップダウンリストに Wikipedia の見出し語が並ぶので，見出し語に該当する Wikipedia ページを確認するためのリンクがあると便利である．こういった需要に対応するため，Wikipedia ページへのリンク付きドロップダウンリストを新たなコントロールとして実装した（図 2.2.4）．このような拡張性を備えることで，Orion Annotator は多くのアノテーションに対応できる．

5 まとめ

本稿ではアノテーション環境を改善するためのシステムである Orion Annotator を紹介した．Orion Annotator はアノテーション作業の効率化とデータフローの自動化を実現し，Ameba ブログを中心に活用されている．現在の課題としては，アノテーション作業者ごとにアノテーションの判断基準が異なっているなどの理由によるアノテーションの正確性の問題がある．今後は，誤った基準でのアノテーションを未然に防ぐしくみの実装や，アノテーションの評価，傾向が異なるユーザの可視化などに取り組む予定である．

参考文献

[1] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J.: “brat: a Web-based Tool for NLP-Assisted Text Annotation,” Proceedings of the European

[*4] https://jp.vuejs.org/
[*5] 文書中の単語に対して Wikipedia などの辞書の見出し語を関連付ける自然言語処理のタスク

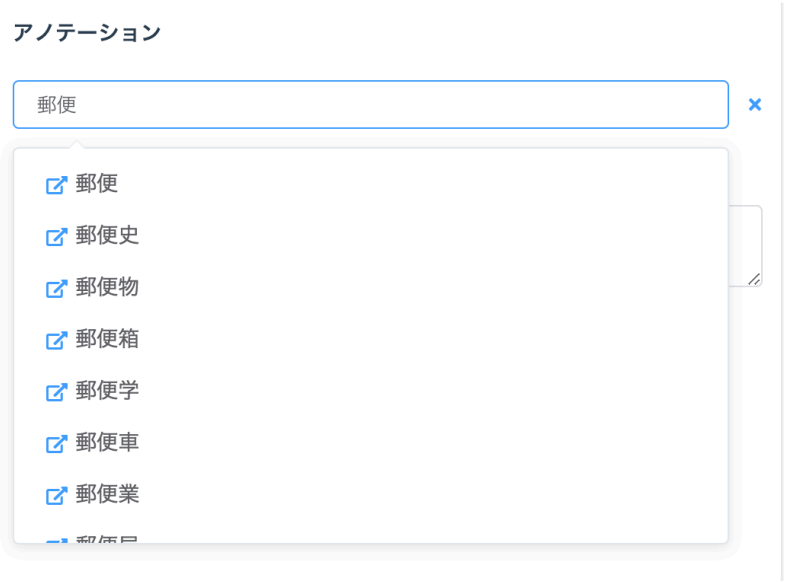

図 2.2.4　Wikipedia リンク付きドロップダウンリスト

Chapter of the Association for Computational Linguistics (EACL 2012), 2012.

[2] Zenmyo, T., Iijima, S., and Fukuda, I.: “Managing a Complicated Workflow based on Dataflow-based Workflow Scheduler,” 2016 IEEE International Conference on Big Data (IEEE BigData 2016), short paper, 2016.

[3] 善明晃由, 津田均, 田中克季: “可視化のための非構造データの表化手法”, 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM2017), 2017.

Akihabara Lab

3 音楽ストリーミングサービス「AWA」

技術報告 vol.2

AWA[a]は当社とエイベックス・デジタル株式会社の共同出資による音楽ストリーミングサービスである．世界最大規模の 5000 万曲超の楽曲を配信しており，高いデザイン性に加えて再生レスポンスの早さやインタラクションの滑らかさなどの操作性を重視したユーザインタフェースを特徴とする．

利用にあたっては，PC やスマートフォン，カーナビ，テレビデバイス，ウェアラブルデバイス，スマートスピーカーなどに用意された専用アプリケーションを用いる．月額課金のサブスクリプション型の音楽配信サービスに分類できるが，月あたりの再生時間や一部の機能が制限された無料プランも用意されている．

ユーザが作成したプレイリストを軸としたユーザ体験が得られることを特徴としており，2015 年 5 月にサービスを開始して以来，継続的に機能の追加が行われてきたが，2019 年 1 月にアプリケーション内の構造の煩雑化を解消する目的でメジャーバージョンアップが実施され，同時に「パーソナライズプレイリスト」「タグ」などの新機能が発表された．

本技術報告は，秋葉原ラボで実装したシステムや実施した分析について，以下の 2 つを取り上げる．3.1 章では，AWA に実装されている類似プレイリストの推薦機能の背後にある理論と実践について述べている．3.2 章では，AWA のユーザの行動ログに基づいて人間の行動に普遍的に現れるとされる現象のモデル構築を試みている．

前者は，AWA で配信している数千万もの楽曲の不定サイズの集合であるプレイリストどうしの類似度の計算という，素朴なアルゴリズムでは膨大な計算時間を要する問題について，時々刻々と生成されるプレイリストと，比較的変化が少ない楽曲と，いうなれば時定数が異なる 2 つの要素の特徴量の作り方に工夫を施すことで高いリアルタイム性を実現している点にご注目いただきたい．

後者は，音楽の鑑賞行動というさまざまな要因が考えられて行動原理も明らかでないものが，べき乗則というシンプルな法則にしたがっていることと，一方で従来の研究で明らかになっているモデルではこの現象を説明しがたいこととを発見している．さらにこれまで説明できなかった現象を部分的にではあるが説明するモデルを提唱しており，これは AWA のユーザのみならず，普遍的な人間の行動原理の解明につながる試みである．

[a] https://awa.fm

Akihabara Lab

3.1 AWA における類似プレイリスト探索システムの構築事例について

水上 裕貴，福田 鉄也，山下 剛史，Juhani Connolly，武内 慎，數見 拓朗，福田 一郎

概　要　本稿ではストリーミング型音楽配信サービス AWA におけるコンテンツ推薦システムの構築，およびシステム稼働に至るまでの実践的なプロセスについて紹介する．AWA ではユーザに対する適切なコンテンツの推薦，とりわけ類似コンテンツの推薦が，体験向上の大きな要因となる．本サービスにおいてユーザに推薦すべきコンテンツとしては，単一の「楽曲」の他に，楽曲の集まりである「プレイリスト」などがあるが，後者のプレイリストは多くの場合ユーザが高頻度で内容を更新することが多い．このため従来の大規模集計による手法では推薦内容のリアルタイム性と網羅率に課題があった．そこで我々は適切な特徴量加工と近似最近傍探索を用いた類似プレイリスト探索システムを構築しリアルタイム性と網羅率を大幅に改善した．

Keywords　推薦, AWA, feature embedding, product quantization

1　はじめに

近年では，レコード産業においても情報処理技術の利活用が活発になってきている．レコード産業とは音楽産業のうち，とりわけ録音・録画された楽曲や映像を流通させる産業を指すが，流通の形態は常に変化を続けている．レコード産業の誕生以降，音楽はレコードやカセットテープそして CD などの物理メディアを通じて流通していた．やがて情報通信技術が発展しスマートフォンなどが普及すると，インターネットを通じて楽曲や映像が流通するようになった．これらは総称してデジタルメディアと呼ばれている．また，レコード産業では物理メディアからデジタルメディアへの急速な代替が起きている [1, 2]．2001 年時点でレコード産業の市場のほぼ全体を占めていた物理メディアの市場規模も 2015 年にはデジタルメディアの市場規模が逆転した．また，同時にレコード産業全体の規模もこの年以降増加傾向に転じている [1]．デジタルメディアを通じた音楽体験の向上は，レコード産業において今後ますます重要な役割を果たすと考えられる．消費者の立場からみれば，デジタルメディアの普及により音楽の体験は大きく変化した．物理メディアにおいて発生していた生産や流通コストの大部分はデジタルメディアによって効率化され，インターネットと端末があればどこでも楽曲が入手できるようになった．そして楽曲の流通の効率化に伴い，楽曲を見つけるコストが顕在化し課題になった．この課題に対する技術的回答として，ユーザが指定したクエリに合致する情報を探索する「情報検索」などがあるが，近年では行動ログなどのデータを用いて必要な情報を予測して，高い精度で提示する「情報推薦」も応用

本稿は情報処理学会デジタルプラクティス論文誌 38 号（Vol.10 No.2）に寄稿した「《特集号招待論文》AWA における類似プレイリスト探索システムの構築（水上 裕貴，福田 鉄也，山下 剛史，Juhani Connolly，武内 慎，數見 拓朗，福田 一郎）」を改稿したものである．

範囲を広げている．この情報推薦の技術は，利用者に対して必ずしも情報要求の言語化・クエリ化を課さず，音楽体験においても未知の楽曲との接触機会を増加させると考えられており，レコード産業からも注目を集めている．

AWA は 5000 万曲超の楽曲を再生時にインターネット経由でスマートフォンなどの端末に配信するストリーミング型音楽配信サービスである．筆者らは，利用者の音楽体験の向上に向けて日々ユーザインタフェースの洗練や配信コンテンツの拡充，データを活用した楽曲推薦の品質向上に取り組んでいる．本稿では，AWA の特徴的な機能であるプレイリストに関して，自然言語処理分野で実績のある skip-gram モデルを応用して構築した推薦システムの事例について紹介する．

1.1 推薦システム

推薦システムとは，価値の高い商品やコンテンツを特定するのを助ける道具である [3]．マーケティングの言葉に言い換えれば顧客が購入する商品・アイテムを予測し，訴求する技術といえる．

アイテムの推薦は基本的には利用者の嗜好に沿うように行われる．例としては「近しい嗜好の利用者どうしは近しいアイテムを好む」という仮説に基づく協調フィルタリング（Collaborative Filtering）がある．この協調フィルタリングは多くの場合，利用者の行動履歴を蓄積・比較することで実現される．また似たような方針として，アイテム自体の情報に基づく，内容ベースフィルタリング（Content-Based Filtering）がある．この内容ベースフィルタリングは，「興味を示したアイテムの類似アイテムにも興味を示す」という仮説のもとで，何らかの意味でアイテム間の類似度を定義することにより実現される．

また実際の推薦システムの推薦結果には往々にして，推薦者側の都合や意向が反映される．サービスの継続的な発展のために収益性を担保することや，知見獲得のために情報を収集することも推薦システムの重要な役割であり，その他にも総合的なユーザ体験の設計において情報供給者の思想を強く反映することもある．たとえば推薦されるアイテムとして，酒類などの成人向けの商品やコンテンツを取り扱う際，当然推薦者は未成年に対する影響を考慮しなければならないと考えるだろう．本稿では，このような推薦者側の意向に基づく推薦方式を総称して意向ベースフィルタリング（Intention-Based Filtering）と呼ぶこととする．本稿では，上記で紹介した複数の根拠を加味したアイテム選定基準，そしてそれを利用者に届けるインタフェースおよびそのバックエンドを含むシステム全体を推薦システムと呼ぶこととする．

1.2 AWA とプレイリスト

多くの音楽再生ソフトウェアでは，あらかじめ順番が定められた楽曲の列は「プレイリスト」と呼ばれている．AWA におけるプレイリストは，主に利用者によって作成され，公開される時点では最大で 8 曲から構成される．また大部分のプレイリストには，作成者によりタイトルと説明文が付与されており，利用者がプレイリストを選択する際の補助的な情報として機能している（図 3.1.1）．以降，推薦対象となる AWA におけるプレイリストを混同のない範囲で単にプレイリストと呼ぶこととする．プレイリストは利用者による作成や公開が活発で，作成されたプレイリストの数は累計で 1100 万を超える．また多くの利用者は，プレイリストやその作成者に対してフィードバックを送り合いながら新しい音楽と出会っていくが，このことは AWA における音楽体験の大きな特徴である．実際に，楽曲再生利用者数に対する再生動線毎の利用者数の比率をみれば，物理メディアとしての販売単位であるシングル・アルバムに匹敵する割合でユーザ作成プレイリストが利用されていることが確認できる（図 3.1.2）．

そこで，AWA ではプレイリストに注目した推薦機能を実現することにした．利用者から見ると，プレイリストの再生画面の最下部には「Related Playlist」と呼ばれる類似プレイリスト推薦枠が表示されるが，この部分にシステムが推薦するプレイリストを表示するようにしている．この類似プ

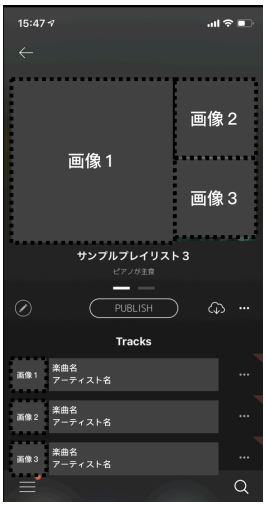

図 3.1.1　AWA におけるプレイリストのレイアウト

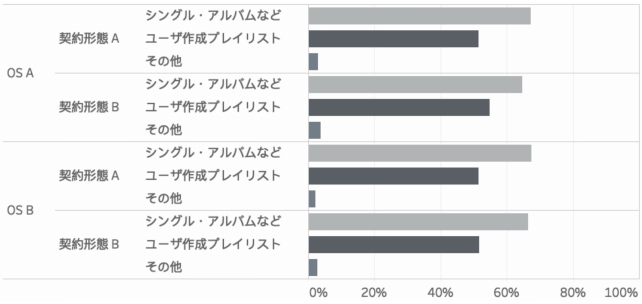

図 3.1.2　再生経路の集計結果

図 3.1.3　類似プレイリストのレイアウト

レイリスト推薦によって，AWA の利用者に対して連続的な音楽体験を提供する機能を実装した（図 3.1.3）．

2　類似プレイリスト探索における課題

AWA における類似プレイリスト推薦システムは，従来，プレイリストの類似度計算を，含まれる楽曲の楽曲情報の重複数を総当たりで計算する大規模なバッチ処理で行っていた（これを旧仕様と呼ぶ）．この旧仕様は構成楽曲の楽曲情報の一致を根拠にした内容ベース推薦の一種といえる．この章では旧仕様が抱える 2 点の課題について述べる．

2.1　網羅率の低さ

1 点目は網羅率が低いことである．ここでは網羅率を，再生されるプレイリストのうち，類似プレイリストが 1 つ以上表示できるプレイリストの比率と定義する．網羅率が低い原因の一つは，計算資源の制約から，上述の大規模なバッチ処理において集計対象を作成日時や再生回数などいくつかの条件で限定していることである．もう一つの原因は，厳密な楽曲情報の一致を条件とするため，プレイリストに採用実績がない楽曲のみからなるプレイリストは類似プレイリストが計算できないことである．

2.2　プレイリスト作成・更新に対する結果反映の遅延

2 点目はプレイリストの作成・更新に対して，類似プレイリストへの結果反映が遅延することである．楽曲の内容情報は一度登録されると更新されることはとても希であるが，プレイリストでは利用者によって内容情報の作成や更新処理が頻繁に行われる．このため従来の手法ではプレイリストの作成や更新からバッチ処理までの時間に，類似プレイリストが表示されない，もしくは古い情報に基づいた類似プレイリストが表示されていた．一

方でアプリケーションの設計上，プレイリストは作成・更新の直後が最もよく再生されるため，類似プレイリストの欠損は利用者の体験を低下させる．

なお，旧仕様で適切な類似プレイリストが表示されるまでの待ち時間の内訳は以下の通りである．

作成待ち時間

プレイリストが作成・更新されてから次の処理が始まるまでの時間である．バッチ処理には計算機を効率的に稼働させられるなどの利点がある一方で，最大で実行間隔分の待ち時間が発生する．

特徴量の計算時間

次にプレイリストの類似度計算のもとになる特徴量を取得するための時間が必要になる．今回，内容ベースの推薦ならば，多くの場合，各楽曲に関連するアーティスト ID などの，参照が容易な情報が特徴量となるので，このコストは無視できるほど小さい．

一方，たとえば協調フィルタリングを行う場合は，再生ログなどのソフトクラスタリングによって得られたベクトルが特徴量となる [4]．しかしソフトクラスタリング自体に大きな計算を要する他，ソフトクラスタリングに必要なログの収集にも時間が必要になる．高頻度で内容の作成・更新処理が発生する今回のケースに，そういった大規模な学習や集計を要する手法は適さない．

最近傍探索の時間

プレイリストに対して特徴量が得られたあとは，その特徴量を用いてプレイリスト間の類似度や距離を定義する．そのうえで最も類似すると思われる複数のプレイリスト ID を得たい．最も単純なアイデアで n 個のすべてのプレイリスト間で最近傍を求める場合，計算量オーダー $O(n^2)$ もの距離の評価が必要になる．また使用する類似度や距離にもよるが，特徴量の次元の大きさも近傍探索の計算時間に大きな影響を与える．プレイリストに対して含まれる楽曲 ID を特徴量とするとき，この特徴量の次元は楽曲 ID の 8 倍ほどの大きい次元になり，この場合，空間計算量が楽曲 ID の数に比例する．

プレイリストのように高頻度で作成や内容の更新が行われるケースではより効率的な探索アルゴリズムや距離計算，そして低次元な特徴量が求められる．

ネットワークの通信時間

上記の他にもクライアント-サーバ間の通信やサーバどうしの通信にも時間を要する．上記の他の要素に比べれば基本的には無視できるほど小さいが，プレイリストのように高頻度で作成・更新などのリクエストを受け取るシステムでは，各構成要素において，効率的なキャッシュの設計などの工夫がパフォーマンス向上においても重要である．

3 要素技術

これまでよりリアルタイムな類似プレイリストの計算のために，より低次元な特徴量と，効率的な近傍探索が重要であることを述べた．この章では本稿で紹介する推薦システムで活用される先行研究を紹介する．

3.1 分散表現を用いた低次元特徴量の獲得

自然言語処理分野において文章を単語の系列とみなして学習する，単語のソフトクラスタリング手法の一つに skip-gram がある [5]．Skip-gram は系列データのある要素から，その周辺の要素を推測する 3 層のニューラルネットワークで，中間層が入出力層と比べて小さい次元になっている．このモデルは「任意の単語の意味はその周辺語によって特徴づけられる」という仮説に基づき周辺語からその単語の低次元ベクトル表現を獲得する．実際に得られた単語のベクトルについて「Madrid」−「Spain」+「France」=「Paris」のような意味上の演算が構成できるなど，単語の意味上の関係を低次元の空間上で幾何学的・代数的に再現するケースが確認されている．この変換は特徴埋め込み（Feature embedding），変換後のベクトルは分散表現（Distributed representation）と呼

ばれている．

3.2　直積量子化に基づく距離の近似計算

ユークリッド距離の近似計算には直積量子化（Product quantization）を用いることができる．いま，検索対象となる有限個の特徴量ベクトルの集合 $Y \subset \mathbb{R}^d$ から，クエリベクトル $x \in \mathbb{R}^d$ とのユークリッド距離を最小にする特徴量ベクトル $y \in Y$ を探索したい．先に述べたようにこの計算には膨大な計算が必要になるが，直積量子化により距離の計算を簡略化できる．

直積量子化の要点は 2 つある．1 点目は特徴量の空間をあらかじめ k-平均法によりクラスタリングしておき，クエリベクトルとの距離をクラスタの重心との距離で近似し，計算量を削減できる点である．2 点目は 2 つのベクトル間の 2 乗距離が複数の区分に分割して計算できる点である．$\mathbb{R}^d$ 上のベクトル y に対して $y = (y_1, y_2, \cdots, y_k) \in Y_1 \times Y_2 \times \cdots \times Y_k \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_k}$ のように k 個の区分への分割を考える．このとき 2 つのベクトル $x, y \in \mathbb{R}^d$ の距離の 2 乗 $\|x - y\|^2$ は

$$\|x - y\|^2 = \sum_{j=1}^{k} \|x_j - y_j\|^2$$

のように区分毎の距離の 2 乗の和と一致する．この 2 つの事実を組み合わせて，分割した空間でクラスタリングした結果を元に，近似的に距離を計算するのが直積量子化による距離計算である．本システムでは文献 [6] で紹介されている非対称型の距離計算（Asymetric Distance Computation; ADC）を用いる．

4　類似プレイリスト探索システムの構築

本章では類似プレイリスト推薦システムで利用する推薦モデルと，その実装例を紹介する．大まかな方針としてプレイリストの低次元特徴量の幾何学的な探索により網羅率の向上を図り，また分散表現を用いたリアルタイムな特徴量計算と近似最近傍探索アルゴリズムを用いて，結果反映までの所要時間の短縮を図るものである．

4.1　プレイリストの特徴量モデル

プレイリストの特徴量の計算にはプレイリストに含まれる楽曲情報を用いる．AWA では利用者による再生楽曲の系列データを保有している．この再生楽曲の系列データについても単語の系列データと同様の「楽曲の個性は前後に再生される楽曲で特徴づけられる」という仮説をとる．この仮説のもとで，各楽曲に対して，Skip-gram による分散表現を獲得しておく．そして，プレイリストの特徴量としてはそのプレイリストに含まれる楽曲の分散表現の重心を用いる．前述の通り楽曲の内容情報の更新は希であるので，楽曲の分散表現については事前に学習できる．したがってプレイリストの作成や更新イベントの度に簡単なベクトルの演算でプレイリストの低次元特徴量を得ることができる．

4.2　意向ベースフィルタリング

AWA では 1100 万を超えるユーザ作成プレイリストのデータを保有している．距離の近似計算について直積量子化により計算量を削減しているが，距離の計算対象の絞り込みによっても計算量を削減することが期待される．そのため，推薦対象となるプレイリストについて事前に何らかの意味で絞り込みを行いたい．また，上記で作成したデモを通じていくつか，アプリケーション設計上の意向を正しく反映していない例が見つかった．そういった課題に対しては，単純なルールベースの処理により推薦対象を選定した．意向ベースのフィルタにより推薦対象を約数十万にまで絞り込み，計算量の削減と品質の向上を実現している．そのような意向ベースのフィルタの構築やその動機および実験手順やなどの実践的な内容の詳細については文献 [7] を参照されたい．

4.3　システム構成

この章ではこれまで紹介した要素技術を組み合わせた類似プレイリスト探索システムの概要について説明する．本システムはユーザによる楽曲再

生ログ，プレイリスト作成・更新ログの 2 種類のストリームデータを処理する Parser, Learner, Filter, Registerer, Finder の 5 つのマイクロサービスからなる．それぞれの概要は以下の通りである．

Parser: 学習データの前処理

Parser では利用者の楽曲再生ログを学習フレームワークが学習しやすい形式に処理してストレージに格納する役割を果たす．各データに施す処理は大まかに 2 つある．1 点目は必要な情報の選択で，利用者の楽曲再生ログにはクライアントの状態を表すさまざまな情報が格納されている．このログから利用者 ID，楽曲 ID，時刻のみを取り出す処理を行う．2 点目は ID を整数値に変換することである．ID は一般にランダムな文字列だが，学習させる際には 1-of-k 表現のための整数値への対応が必要になる．この ID と整数値の一意な対応を作成しながら変換することが 2 つ目の処理である（図 3.1.4）．

Learner: skip-gram による分散表現の学習

Learner では格納された再生系列データ（図 3.1.5）を用いて skip-gram の学習を行う．学習が終了するとこれを再び楽曲 ID を Key，分散表現を Value とする Key-Value Store（以下 KVS）のテーブルに格納する．

Filter: 意向ベースのフィルタ処理

Filter ではプレイリスト作成・更新ログを受け取りそのログの内容を評価して，そのプレイリストが推薦対象となるかを判定してそのフラグを付与する．

Registerer: プレイリスト特徴量の計算

Registerer では Filter が処理した推薦対象フラグ付きのプレイリスト作成更新ログから，プレイリスト ID とそれを構成する楽曲 ID の列と推薦対象フラグを取り出す．次に楽曲 ID の列からそれぞれの分散表現を取得しその重心を計算する．この重心をプレイリストの特徴量としてプレイリスト ID を Key，この特徴量と推薦対象フラグを Value とする KVS のテーブルに格納する．

Finder: 特徴量の近似最近傍探索

Finder は API からプレイリスト ID を受け取り類似プレイリストの列を返すサービスである．起動時に直近一定期間に作成・更新されたプレイリストのうち検索対象となるもののみの特徴量を前述のテーブルから読み込み，直積量子化に用いるためのクラスタリング処理を行い各クラスタの重心（＝辞書）をメモリ上に格納する．その後数十秒程度の一定間隔で直近の更新分をも量子化したうえでメモリ上に格納する．プレイリスト ID を受け取るとその特徴量を前述のプレイリスト特徴量テーブルから取り出し，あらかじめ量子化した推薦対象リストから近似的な近傍プレイリストを，要求された数だけ返却する．

5 評価

提案法を 5% の利用者に適用し 1 ヵ月の A/B テストを行った（現在は全利用者に適用済み）．その結果，網羅率と反映までの時間差の 2 項目について大幅な改善がみられた．また実際の利用者への適用を通じて，類似プレイリストの利用率についても改善が見られた．以下に各項目の詳細を述べる．

5.1 網羅率

検索元のプレイリストに対して類似プレイリストが存在する比率，すなわち網羅率を比較する．従来では楽曲情報の厳密な紐付けを元に類似プレイリストを探索していたため，一部のプレイリストには類似プレイリストが割り当てられなかった．本システムでは分散表現を用いた幾何学的な近傍探索を行うことで，楽曲の類似性から類似プレイリスト探索が行え，そのため一定数の再生実績がある楽曲が 1 つでも含まれていれば類似プレイリストの探索が行える．このことにより従来およそ 64% であった網羅率が現在ではおよそ 98% に改善した．これは再生実績のない楽曲のみからなるなどの，例外的なプレイリストを除くおよそすべ

図 3.1.4　ID-整数値 Mapping の例

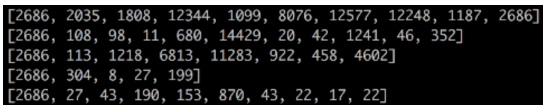

図 3.1.5　処理された再生系列データの例

てのプレイリストを網羅していることに匹敵する．

5.2　反映待ち時間

ここでは，類似プレイリストが計算されるまでの最大の時間差を比較する．従来の反映待ち時間はバッチ間隔と集計の実行時間の和で，最大で 24 時間程度要していた．一方本システムでは Registerer による処理時間と Finder による近似最近傍探索の処理時間の和となり，負荷ピーク時においても 1000ms 程度のほぼリアルタイムな推薦を実現している．

5.3　類似プレイリスト利用率

AWA では各機能を通じた新しい楽曲との出会いを重要視している．このことを測る尺度としてここでは対象機能を通じて楽曲を再生開始した比率を比較した．契約形態に応じて利用動向が大きく異なることがわかっているため契約形態毎にその比率を比較したところ，各料金プランで 2 倍程度の比率のユーザが対象機能経由で楽曲を再生したことが確認された．

6　まとめと今後の課題

本システムの実装と実験を通じて自然言語処理分野で実績のある skip-gram モデルが，楽曲の再生系列データに対しても有効な特徴量を獲得できること，そしてプレイリスト，つまり楽曲の集合においてその分散表現の演算結果がプレイリストの特徴量として機能することを確認した．また従来法との比較を通じて，いずれの指標においてもユーザ体験を向上させていることを確認した．またシステム設計の観点で見れば，特定の学習モデルに依存しない汎用的な学習データの前処理を行う Parser，特定のアプリケーションに依存しない学習モデルである Learner および Registerer，また汎用的なベクトルの近似最近傍探索を行う Finder と複数のサービスからなる再利用性を意識したアプリケーションの構築事例を紹介した．また，今後の課題としては，モデルの継続的な更新やライフサイクル管理がある．総合的なアプリケーションとしての魅力に対して有力なオフライン評価がない今，学習データやモデルの更新には，利用者や管理者のフィードバックを取り入れる必要があると考えられる．本システムはその意味でモデル更新の完全なる自動化には至っていない．文献 [8] や [9] などで人間のフィードバックを取り入れた継続的デリバリのしくみが提案・実装され始めており，また学術的な観点では，AI 応用システムにおいて旧来のソフトウェア工学を適用する際の課題に対する提言として文献 [10] がある．今後，本システムを含む多くの機械学習を応用したアプリケーションにおいても開発プロセスの改善やモデルの更新自動化・高性能化および品質向上を通した，利用者の音楽体験の向上に取り組んでいきたい．

参考文献

[1] IPFI: Global Music Report 2018 — ANNUAL STATE OF THE INDUSTRY, 2018. https://www.ifpi.org/downloads/GMR2018.pdf

[2] 一般社団法人日本レコード協会: Statistics Trends — the Recording Industry in Japan, 2018. https://www.riaj.or.jp/f/pdf/issue/industry/RIAJ2018.pdf

[3] Dan Cosley, et al.: “Is seeing believing?: how recommender system interfaces affect users’ opinions,” Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, pp.585–592,

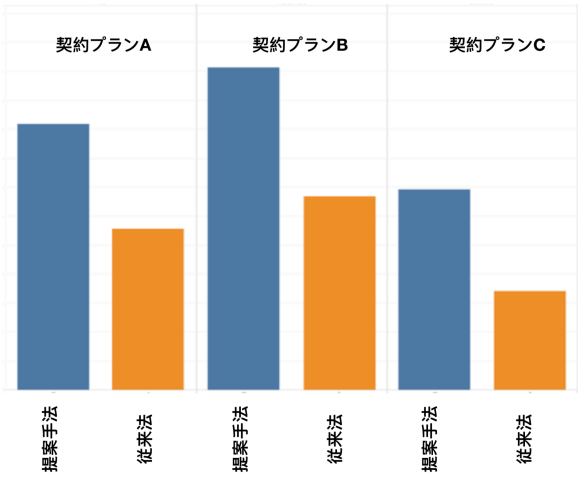

図 3.1.6 類似プレイリスト利用者集計

2003.

[4] Deepak K. Agarwal and Bee-Chung Chen: Statical Methods for Recommender Systems, Cambridge University Press, 2016.

[5] Tomas Mikolov, et al.: “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, pp.3111–3119, 2013.

[6] Hervé Jégou, Matthijs Douze, and Cordelia Schmid: “Product quantization for nearest neighbor search,” IEEE transactions on pattern analysis and machine intelligence, vol.33, no.1, pp.117–128, 2011.

[7] 水上 他: “AWA における類似プレイリスト探索システムの構築,” デジタルプラクティス論文誌 第 38 号, 情報処理学会, vol.10, no.2, 2019. (in press)

[8] Denis Baylor, et al.: “Tfx: A tensorflow-based production-scale machine learning platform,” Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017.

[9] Spinnaker.io: “Spinnaker Continuous Delivery for Enterprise, Fast, safe, repeatable deployments” https://www.spinnaker.io/

[10] JST 研究開発戦略センター システム・情報科学技術ユニット: AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立, 2018. https://www.jst.go.jp/crds/pdf/2018/SP/CRDS-FY2018-SP-03.pdf

Akihabara Lab

3.2 AWA に見られるバースト行動に対する現象論的モデルの検証

武内 慎

概　要　人間のさまざまな行動において，短期間で急に行動頻度が増加するバースト行動が確認されており，その発生メカニズムを解明することは人間の行動を理解するために重要である．本研究では，個人の音楽鑑賞行動においてもバースト行動が存在することを確認し，それがランダムな外的要因と記憶効果のみから生じるという仮説を，Hawkes 過程を用いて検証した．その結果，Hawkes 過程で行動をモデル化できたユーザについてはそのパラメータの解釈についても妥当であった．一方で，ユーザの一部の行動は Hawkes 過程に基づくモデルには従わなかった．このことから，興味駆動の行動を説明するためには，記憶効果以外のメカニズムも考慮する必要がある可能性が示された．

Keywords　Human Dynamics，バースト，記憶効果，Hawkes 過程

1　はじめに

人間の行動の背後にあるルールを探求する Human Dynamics 分野において，集団から個人までのさまざまなレベルで普遍的に確認されているバースト現象のメカニズムの解明がその重要な課題の一つとなっている [1]．バースト現象は，短期間に高頻度でイベントが発生するバーストと呼ばれる状態と長期間に低頻度で行動が発生する状態が交互に繰り返されるようなパターンで特徴付けられる．このような特徴は，イベントが一定確率でランダムに発生するような定常ポアソン過程では説明できないため，バースト現象はその背後にランダム以外のメカニズムが存在することを示唆している．本稿では，人の行動にみられるバースト現象のことをバースト行動と呼び，特に個人のバースト行動に着目する．

個人のバースト行動に関しては，e-mail 送信等の社会的な相互作用を前提とした行動についての先行研究が多く存在する [2–4, 8]．一方で，音楽鑑賞のような個人の興味駆動と考えられる行動については先行研究は少なく，そのメカニズムも十分に解明されていない．

以上の背景の下，本稿では個人の興味駆動と考えられる音楽鑑賞行動の理解のために，それを説明できるモデルを提示し，さらに実データに基づいてモデルの妥当性を検証する．具体的には，音楽ストリーミングサービス AWA ユーザの再生行動データから，音楽鑑賞行動においてもバースト行動が存在することを確認し，そのバースト行動が記憶効果とランダムな外的要因から生じるという仮説を，自己励起点過程の一つである Hawkes 過程を用いて検証する．記憶効果は，あるイベントが発生した場合にそのイベントがその後発生しやすくなるという効果であり，相互作用を仮定しないバースト現象を説明するメカニズムとしては一般なものの一つである．

2 先行研究

前述のように人のバースト行動についての先行研究は，e-mail 送信 [2]，携帯電話 [3, 8]，対面コミュニケーション [4] といった社会的相互作用に基づくものが比較的多く存在する．これらについては，個人レベルの行動を生成できる代表的なモデルがいくつか提案されている．一方で，動画閲覧 [5] や，音楽鑑賞 [6] 等の興味駆動に基づくものについては，集団行動に焦点を当ててマクロなモデルを議論しているか，もしくは個人の行動の実験的な研究にとどまり，それが生成されるミクロな視点のメカニズムは提案および検証がされていない．すなわち，興味駆動に基づく個人レベルの行動の生成メカニズムについては，ほとんど解明されていない．

なお，先行研究における代表的なバースト行動モデルとしては，個人がタスクリストをもつことを仮定し優先順位に従ったタスク処理によりバースト行動を説明する Queuing モデル [2]，集団行動を対象とした自己組織化臨界モデル [7]，記憶効果をもつ自己励起点過程 [8] 等が提案されている．このうち，Queuing モデルはタスク処理としての行動を対象としているので，興味駆動に基づく行動を説明するモデルとしては不適切である．また自己組織化臨界モデルは集団行動を対象としているので，個人レベルの行動を説明するモデルとしては不適切である．一方，自己励起点過程は，過去のイベントがその後のイベントに影響を与えるという自己励起のメカニズムを仮定することでバースト現象を説明する比較的汎用性の高いモデルである．本研究では自己励起点過程モデルの一つである Hawkes 過程を用いる．

3 モデルと評価方法

本研究では，各ユーザの音楽鑑賞行動をイベント時系列データとして用いて，Hawkes 過程に基づくモデルに対して最尤法で Fitting を行い，残差を見ることでモデルの適合度を評価する．Hawkes 過程は，単位時間あたりのイベントの平均発生回数を表す強度関数が過去の系列に依存して決まることで特徴付けられる．Hawkes 過程の時刻 t における強度関数 $\lambda(t)$ は

$$\lambda(t) = \lambda_0 + \sum_{t_i<t} \nu \exp\left(-\frac{t-t_i}{\tau}\right) \qquad (3.2.1)$$

と書ける．第 1 項の λ_0 はベースのイベント発生確率を表し，これはランダムな外的要因によるイベント発生に対応する．第 2 項は過去のイベントに起因する現在のイベント発生確率の励起の総和であり，これは自己励起のメカニズムに対応する．ν はその励起のジャンプ幅を，τ はその励起の時間経過による指数減衰の速さを，それぞれ表すパラメータである．また t_i はイベント発生時刻であり，その間隔を $T_i = t_i - t_{i-1}$ とする．ここで想定している自己励起のメカニズムは人個人の記憶効果によるものなので，その自己励起の時間経過による減衰は人の記憶の減衰を模倣すべきである．人の記憶の減衰曲線に関しては，初期の Ebbinghaus [9] の研究以降さまざまな研究がなされており，記憶は線形に減衰しないことはほぼ確実であるが，指数関数かべき関数かについてはあまり明確になっていない．ここでは減衰関数に指数関数を用いる．

Hawkes 過程における重要な指標として branching ratio があり，次式で定義される．

$$n = \int_0^\infty \nu \exp\left(-\frac{t}{\tau}\right) dt = \nu\tau \qquad (3.2.2)$$

これは全体のイベント数のうち自己励起項によって生成されるイベントの割合を示している．この branching ratio が $n > 1 - 1/\sqrt{2} \approx 0.2929$ の場合，その過程で生成されるイベント時系列は非定常的であり，バーストが発生しやすくなることが解析的に示されている [10]．また，モデルの適合度は，既存の評価手法に倣い，ランダム時間変更

$$\xi_i = \Lambda(t_i, t_{i-1}) = \int_{t_{i-1}}^{t_i} \lambda(s) ds \qquad (3.2.3)$$

による点過程の残差が，理論的には定常ポアソン過程となることを利用する．定常ポアソン過程の重要な性質として，それによって発生したイベン

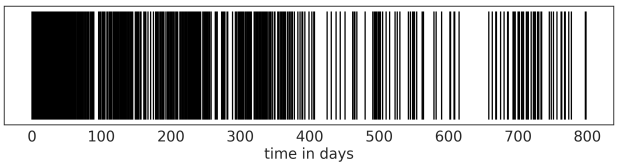

図 3.2.1　ユーザ個人のイベント時系列データのバーコードプロット

トの時間間隔の存在確率分布が指数分布となること，および一定時間内に発生するイベント数の分布がポアソン分布となることが挙げられる．これらの性質を用いて点過程の残差（3.2.3）が理論値と合っているかどうかを評価する．

4　解析結果

4.1　データ

本研究では，AWA の各ユーザの音楽再生ログを分析対象とした．参照期間は 2015 年 6 月から 2018 年 7 月の約 3 年間である．これは音楽鑑賞行動の先行研究 [6] の参照期間の 4 ヵ月間弱と比較すると長期間であり，興味駆動の行動の長期的な時間スケールにおける新たな特徴をとらえられる可能性がある．データの粒度は，サービスの機能に大きく依存すると考えられる連続再生行動の影響と，概日リズムの影響とを除外するように決定した．具体的には，日次の再生有無をバイナリ変数（その日に再生したかどうかを 0, 1 で表現したもの）のイベント時系列データとした．図 3.2.1 は，あるユーザのイベント時系列データを，横軸を日数としてバーコードプロットしたものの例である．

また，分析対象のユーザについては，2 年間以上の期間 AWA を利用し，合計再生日数が 30 日間以上のユーザとした．AWA には有料プランと無料プランが存在するため，参照期間において，「常に有料プラン」「常に無料プラン「有料プランと無料プランが混在」の 3 つの分類毎に 100 人，合計 300 人分をランダムサンプリングした．

4.2　イベント間隔の分布

ここでは，AWA ユーザのバースト行動の有無について確認する．バースト現象の最も一般的な特徴は，イベント間隔 T_i の存在確率分布が，指数分布 $f(x) \propto \mathrm{e}^{-x}$ ではなく，べき分布等のファットテールな分布となる点である．そこで，対象データのイベント間隔，つまりある AWA ユーザが音楽を聴いてから次に聴くまでの日数の分布を調査する．

べき分布とは次式で表現できる分布のことである．

$$f(x) \propto x^{-\alpha} \tag{3.2.4}$$

この関係式において x を定数倍しても関数形は不

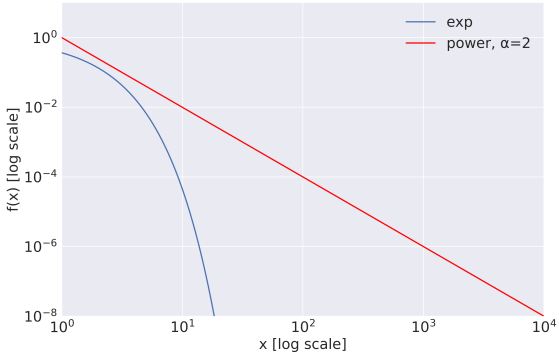

図 3.2.2　指数分布 $f^{\mathrm{exp}}(x) = \exp(-x)$ とべき分布 $f^{\mathrm{power}}(x) = x^{-2}$

変であり，これはべき分布の重要な特徴である．べき分布と指数分布との違いの一つに，$x \gg 0$ における減少のしかた，つまり 0 への漸近がある．指数分布 $f^{\exp}(x)$ とべき分布 $f^{\text{power}}(x)$ それぞれの対数をとると

$$\log f^{\exp}(x) \propto -x \ (x \geq 0)$$
$$\log f^{\text{power}}(x) \propto -\alpha \log x \ (x \geq 1)$$

となり，べき分布においては $\log x$ との比例関係が示せるのでこれらを両対数グラフ（図 3.2.2）で比較するとわかりやすい．このグラフから，べき分布の 0 への漸近は指数分布よりも遅い様子が分かる．このような 0 への漸近が指数関数よりも遅い分布のことを一般にファットテールな分布という．

対象データを確認すると，ユーザ毎のイベント間隔 T_i を対象ユーザ全体で合算した分布は，ファットテールな分布を示している（図 3.2.3）． 図中の赤線は，データがべき分布に従うことを仮定した場合の Clauset ら [11] の手法を用いた Fitting 結果であり，α の値はその際のべき指数である．図 3.2.4 は，個人のイベント間隔 T_i の分布例であり，おおよそべき分布を示しているように見える．

ただ，対象ユーザ毎に確認すると個人差があり，指数分布からファットテールな分布まで多様な分布が存在した．つまり，対象ユーザの中には，バーストしているユーザとバーストしていないユーザが混在していた．本研究で用いる Hawkes 過程は，その強度関数 (3.2.1) の第 2 項が第 1 項と比較して無視できるほど小さい場合に定常ポアソン過程とみなせるため，イベント間隔が指数分布となるバーストしていないユーザも含めてモデリングでき，個人差はパラメータの値として表現される．

4.3 Hawkes 過程のモデリング結果

ユーザ毎のイベント時系列データに対して，それぞれ最尤法を用いて Hawkes 過程のパラメータ推定を行い，その残差（3.2.3）を Q-Q プロットで評価した．行動をモデル化できていればこの残差は理論分布である指数分布となり，Q-Q プロットは傾き 1 の直線上にプロットされる．図 3.2.5，図 3.2.6 は実際の Q-Q プロットでの評価例であり，図中の R^2 値は残差の理論値と実際の値についての決定係数である．これらの結果から Hawkes 過程で説明できるユーザとできないユーザが存在することが分かる．

分析対象とした 300 人を Hawkes 過程でモデリングした結果，決定係数が 0.8 以上となったのは約 2/3 の 202 人であった．恣意的ではあるが，決定係数が 0.8 以上のユーザを Hawkes 過程でその行動を説明できるユーザとし，そのパラメータ分布を確認する．

図 3.2.7 は，ユーザを 3 つのプラン分類に分け，推定したパラメータの分布を箱ひげ図で比較したものである．まず，プラン毎の傾向の違いとして，常に有料のユーザは，ベースの確率 λ_0 が高く，記憶効果によるジャンプ幅 ν が低い．これは，有料

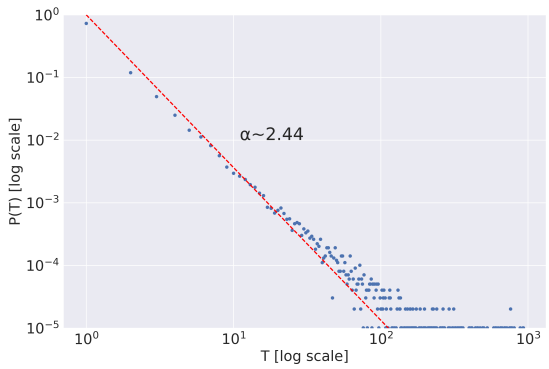

図 3.2.3 イベント間隔分布 $P(T_i)$（全体）．

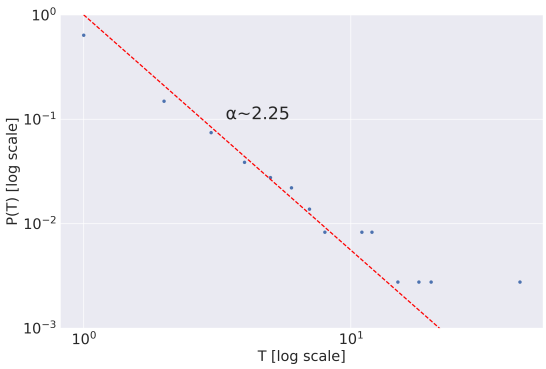

図 3.2.4 イベント間隔分布 $P(T_i)$（個人）．

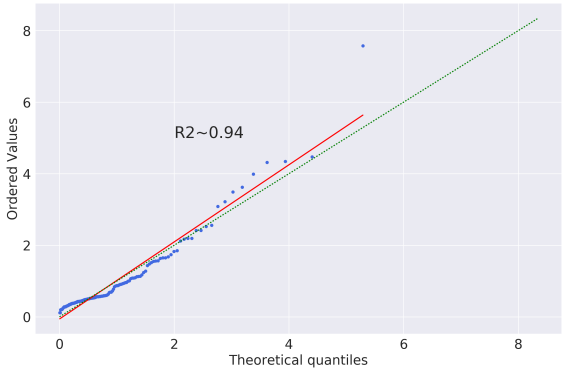

図 3.2.5　行動をモデル化できている場合，傾き 1 の直線上にプロットされる．

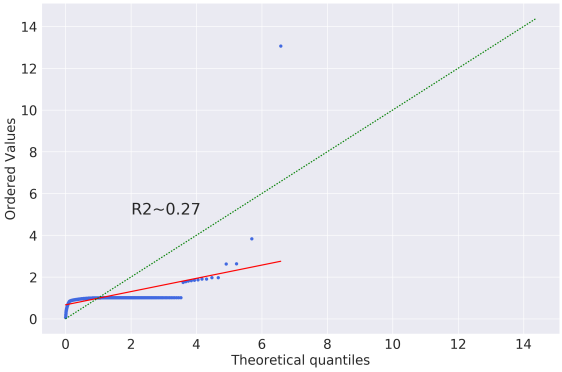

図 3.2.6　行動がモデルに従わない場合，傾き 1 の直線上からプロットが乖離する．

の場合は音楽鑑賞に対するモチベーションが高く定期的に再生するのに対し，無料の場合は普段から再生しているわけではなくときどき再生しやすい状態（バースト）になりやすいことを示している．これは branching ratio の傾向とも矛盾しない．つまり，常に無料のユーザおよび有料と無料の混在ユーザの各グループにおいては，そのほとんどが branching ratio の閾値 0.2929 を超えた非定常状態でありバーストを生じやすいのに対し，常に有料のユーザグループは，比較的多くの定常状態ユーザを含んでいる．

これらの定常状態ユーザは，音楽鑑賞行動の傾向が長期間において安定しているユーザだと解釈できる．また，τ は記憶効果の平均寿命と解釈でき，中央値は 10 付近の値となっているが，これは人の週周期（7 日間周期）のリズムを反映している可能性がある．全体としては，τ の分布は 0 から 20 付近まで幅広く，記憶効果の寿命には大きな個人差があることが分かる．

一方で，残りの約 1/3 のユーザはその行動を Hawkes 過程で説明できなかった．これらのユーザの Fitting 結果を確認したところ，主に以下の 2 つのケースが存在した．図 3.2.8, 3.2.9 はその代表例を示しており，それぞれ図の左上が Q-Q プロット，右上が残差のイベント数の移動合計，下がイベントの時系列プロットを示している．残差のイベント数の移動合計は理論的にはポアソン分布に従うはずである．グラフにおける水色破線は，これを正規分布で近似した場合の $1\sigma, 2\sigma, 3\sigma$ の標準誤差である．一つは，図 3.2.8 に示すような休止期間とバースト期間が極端なケースであり，長期の全くイベントのない休止期間を Hawkes 過程で表現できていない．この極端な休止期間は，ユーザが AWA サービスから離脱状態にあると考えられる．また，もう一つは，図 3.2.9 に示すような途中でベース確率 λ_0 か，もしくは記憶効果のパラメー

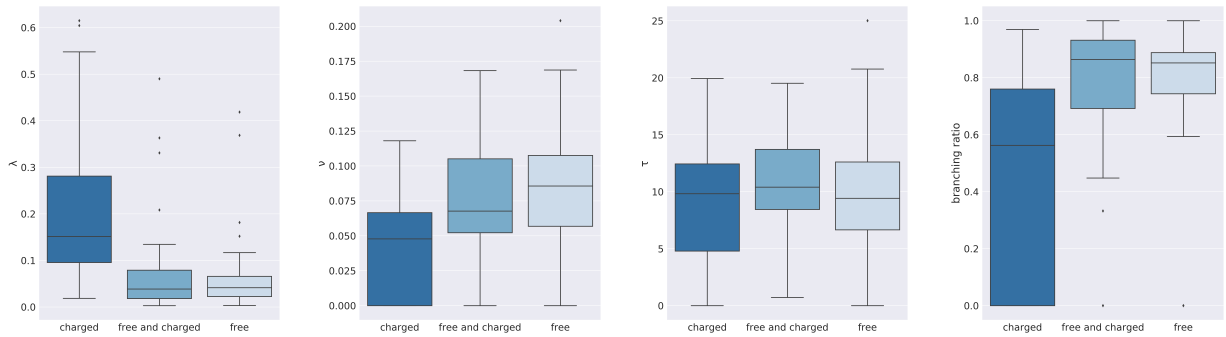

図 3.2.7　プラン分類毎の各モデルパラメータの分布．左から λ_0, ν, τ, branching ratio.

タ τ，ν が変化しているように見えるケースであり，このユーザの音楽に対する関心の変化などに伴い行動の傾向が変化していると考えられる．これらの Hawkes 過程で説明できないユーザは，その音楽鑑賞行動をランダムな外的要因と記憶効果で説明できるという仮説に対する反例であり，興味駆動の行動を説明するためには，記憶効果以外のメカニズムも考慮する必要がある可能性を示している．

5 まとめ

本研究では，AWA ユーザに見られる音楽鑑賞のバースト行動が，ランダムな外的要因と記憶効果のみから生じるという仮説を，Hawkes 過程を用いて検証した．その結果，行動をモデル化できたユーザについては，そのユーザの記憶効果や定常か非定常かといった音楽鑑賞行動における特徴をパラメータから定量的に抽出できた．一方で，Hawkes 過程に基づくモデルには従わなかったユーザの分析から，音楽鑑賞のバースト行動はランダムな外的要因と記憶効果のみではすべて説明できない可能性が示唆された．これらのケースでは，ユーザの比較的長期的な状態遷移を考慮するなどの工夫が必要かもしれない．

個人の，音楽鑑賞行動を含む興味駆動の行動の背後にあるメカニズムを理解することは，その集団的な振る舞いや，コンテンツの流行メカニズムの

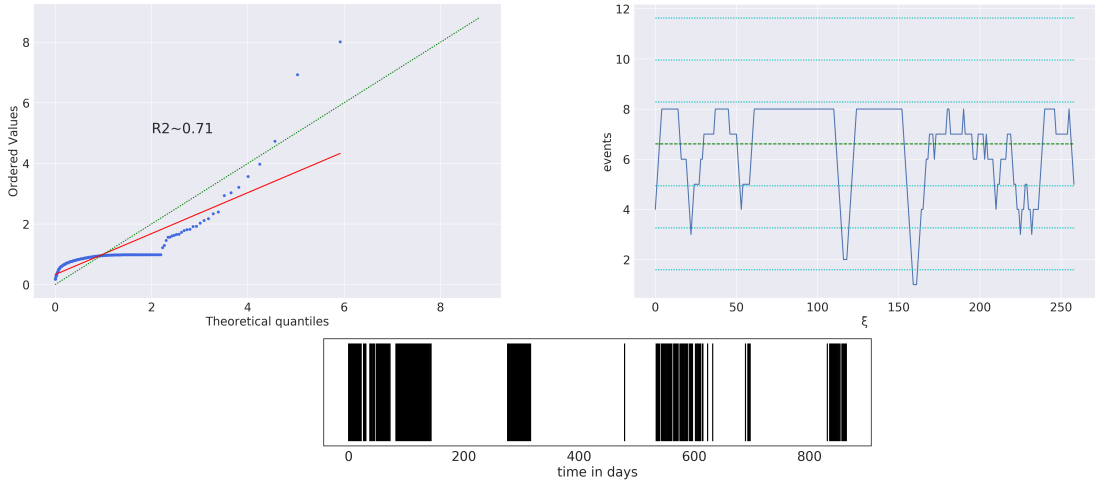

図 3.2.8　休止期間とバースト期間が極端なケース．右上図の残差のイベント数の移動合計において，3σ を超える値が生じており，Hawkes 過程による表現に無理があることが分かる．

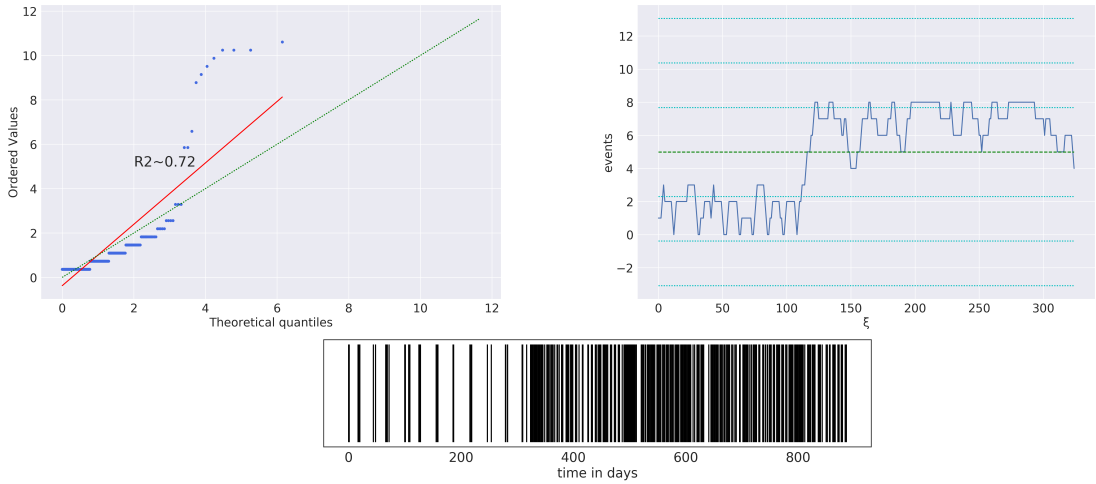

図 3.2.9　途中で記憶効果が変化しているように見えるケース．

より深い理解につながる可能性がある．また，コンテンツ推薦技術への応用に役立つ可能性もある．今後は，今回モデリングできなかったユーザの行動メカニズムの解明や，これらの応用を視野に入れた研究も行っていきたい．

参考文献

[1] M. Karsai, H. Jo and K. Kaski, Bursty Human Dynamics, Springer (2018).

[2] A. L. Barabasi, The origin of bursts and heavy tails in human dynamics, Nature 435, 207-211 (2005).

[3] M. Karsai, K. Kaski, A. L. Barab á si, J. Kert é sz, Universal features of correlated bursty behaviour, Sci. Rep. 2, 397 (2012).

[4] J. Fournet and A. Barrat, Contact patterns among high school students, PLoS ONE 9(9), e107878 (2014).

[5] R. Crane and D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system. Proc. Natl. Acad. Sci. 105(41), 15649-15653 (2008).

[6] H. Hu and D. Han, Empirical analysis of individual popularity and activity on an online music service system, Phys. A Stat. Mech. Appl. 387(23), 5916-5921 (2008).

[7] A. Vazquez, Self-organization in populations of competing agents, Phys. Rev. E 62, R4497 (2000).

[8] H. Jo, J. I. Perotti, K. Kaski and J. Kertesz, Correlated bursts and the role of memory range, Phys. Rev. E92(2), (2015).

[9] H. Ebbinghaus, Memory: A contribution to experimental psychology, Number 3. Teachers college, Columbia university, (1913).

[10] T. Onaga and S. Shinomoto, Bursting transition in a linear self-exciting point process. Phys. Rev. E89(3), (2014).

[11] A. Clauset, C. R. Shalizi and M. E. J. Newman, Power-law distributions in empirical data, arXiv: 0706. 1062, (2007).

[12] 武内 慎，人の音楽鑑賞行動に見られる冪分布に対する現象論的モデルの検証，日本物理学会 2018 年秋季大会，9aM203-3，(2018).

書籍・解説記事

- クリストファー・G・ブリントン（原著）, ムン・チャン（原著）, 臼井翔平（翻訳）, 鬼頭朋見（翻訳）, 浅谷公威（翻訳）, 坂本陽平（翻訳）, 高野雅典（翻訳）, 伏見卓恭（翻訳）, 池田圭佑（翻訳）, 鳥海不二夫（解説）: “パワー・オブ・ネットワーク: 人々をつなぎ社会を動かす 6 つの原則”, 森北出版, 2018.

論文誌・国際会議（査読付き）

- H. Hayano, M. Takano, S. Morishita, M. Yoshida, and K. Umemura: “Analysis of the Influence of Internet TV Station on Wikipedia Page Views,” The 3rd International Workshop on Application of Big Data for Computational Social Science (workshop at IEEE BigData2018), 2018.
- K. Kakiuchi, M. Nishiguchi, F. Toriumi, M. Takano, K. Wada, and I. Fukuda: “What influences people to broaden their horizons?,” IEEE/WIC/ACM International Conference on Web Intelligence (WI2018), 2018.
- K. Kakiuchi, F. Toriumi, M. Takano, K. Wada, I. Fukuda: “Influence of Selective Exposure to Viewing Contents Diversity,” The 10th International Conference on Social Informatics (SocInfo), 2018. [Best Paper]
- M. Takano: “Two types of social grooming methods depending on the trade-off between the number and strength of social relationships,” Royal Society Open Science, Vol. 5, Issue 8, 2018.
- M. Takano: “Two Types of Social Grooming Methods Depending on the Trade-Off Between the Number and Strength of Social Relationships,” 4th Annual International Conference on Computational Social Science (IC2S2), Extended Abstract, 2018.
- M. Takano and T. Tsunoda: “Bullied-Experience Talks in Online Communications: Support for Bullied Children,” 4th Annual International Conference on Computational Social Science (IC2S2), Extended Abstract, 2018.
- V. Trivittayasil: “MovingBubbles : Animated d3 bubble chart,” useR!, Poster Session, 2018.
- M. Takano, G. Ichinose: “Evolution of Human-like Social Grooming Strategies regarding Richness and Group Size,” Frontiers in Ecology and Evolution, Vol. 6, Article. 6, pp. 1–8, 2018.

国内学会/セミナー

- 岩井二郎：“画像の特徴量を利用した CTR 予測”, 第 11 回 Web とデータベースに関するフォーラム（WebDB Forum 2018）, 2018.

- 福田鉄也：“AbemaTV における推薦システム”, 第 11 回 Web とデータベースに関するフォーラム（WebDB Forum 2018）, 2018.
- 松田和己, 和田計也, 福田一郎：“インターネットテレビ局 AbemaTV における番組視聴データを用いたユーザー行動分析”, 統計関連学会連合大会, 2018.
- 高野雅典, 森下壮一郎, 小川祐樹：“インターネットテレビ局 AbemaTV におけるニュースの「チラ見」がニュース視聴の動機に与える影響”, 日本行動計量学会 第 46 回大会, 2018.
- 高野雅典：“仮想空間における自己開示とオンラインソーシャルサポート：アバターチャットサービスにおけるいじめ経験の告白と相談”, 電気通信大学人工知能先端研究センター シンポジウム「心ふるわせ、社会うごかすエージェントデザイン」, 2018.
- 武内慎：“人の音楽鑑賞行動に見られる冪分布に対する Hawkes 過程を用いた分析”, ネットワーク科学セミナー, 2018.
- 高野雅典：“コミュニケーション方法を二分するしきい値 —社会関係の数と強さのトレードオフパラメータ—”, ネットワーク科学セミナー, 2018.
- 西口真央, 鳥海不二夫, 高野雅典：“複数交流系ソーシャルメディアデータを利用したオンラインリスク未然予測モデル”, ネットワークが創発する知能研究会 + ネットワーク生態学グループ 合同研究会 (JWEIN+NetEco2018), 2018.
- 武内慎：“人の音楽鑑賞行動に見られる冪分布に対する現象論的モデルの検証”, 日本物理学会 2018 年秋季大会, 9aM203-3, 2018.
- Y. Fujisaka: “Orion: An Integrated Multimedia Content Moderation System for Web Services,” ACM International Conference on Multimedia Retrieval (ICMR), Industrial Session, 2018.
- 森下壮一郎：“残差に基づいて匿名性と有用性を両立させる匿名加工に関する考察”, 第 32 回人工知能学会全国大会, 3H2-OS-25b-01, 2018.
- 高野雅典, 角田孝昭：“オンラインコミュニケーションにおける「いじめ経験の告白」”, 第 32 回人工知能学会全国大会, 3O1-OS-1a-02, 2018.
- 平野雄一, 鳥海不二夫, 高野雅典, 和田計也, 福田一郎：“未成年女性のネットリスク分析”, 第 32 回人工知能学会全国大会, 2Z3-02, 2018.
- 高野雅典, 水野寛：“仮想社会における社会関係とソーシャルサポートに関する一考察”, 第 32 回人工知能学会全国大会, 1B3-04, 2018.
- 斎藤貴文, 善明晃由：“ストリーム処理におけるステートフルデータの管理手法”, The 2nd. cross-disciplinary Workshop on Computing Systems, Infrastructures, and Programming (xSIG2018), 2018.
- 高野雅典, 角田孝昭：“オンラインコミュニケーションにおける「いじめ経験の相談」：いじめ被害者に対するソーシャルサポート”, 第二回計算社会科学ワークショップ, 2018.
- 平野雄一, 鳥海不二夫, 高野雅典, 和田計也, 福田一郎：“未成年女性のネットリスク分析”, 第二回計算社会科学ワークショップ, 2018.
- 斎藤貴文, 善明晃由：“コンフリクトフリーなデータ型に基づくステートフルストリーム処理基盤”, 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), 2018.

CyberAgent 秋葉原ラボ 技術報告 Volume.2

発 行 日　2019年3月1日 初版
発 行 所　CyberAgent 秋葉原ラボ 技術報告編集委員会
〒101-0021
東京都千代田区外神田1丁目18番13号
秋葉原ダイビル13階

編　集　數見 拓朗
上辻 慶典
松井 美帆
善明 晃由
森下 壮一郎
デザイン　Design Factory　後谷 莉子
印 刷 所　昭栄印刷株式会社

CyberAgent