情報処理学会論文誌　デジタルプラクティス　Vol.4　No.1　34–44 (Jan. 2023)

放送局を横断する大規模テレビ視聴履歴データの統合手法の提案と実践

松田 裕貴[1,2,a)]　榊原 太一[2]　松田 裕貴[1]　水本 旭洋[3]　安本 慶一[1]

受付日 2022年6月20日, 再受付日 2022年8月9日/2022年10月6日,
採録日 2022年10月21日

概要：近年，各テレビ放送局において，個人を特定しない形式で，インターネット接続されたテレビから視聴開始時刻や視聴終了時刻等を含む非特定視聴履歴データを収集し，利活用する取り組みが進められている．しかし，各放送局は自局の非特定視聴履歴データしか利用できないため，膨大なデータを蓄積しているにも関わらず，有用な知見を得るまでに至っていないのが現状である．さらに，非特定視聴履歴データの収集方式やデータ粒度は，各社各様となっており，各局が蓄積したデータを統合し，利用することもできていない．そこで本稿では，各局が独自の方式で取得している非特定視聴履歴データを放送局間でマッチングする手法を提案し，データ統合を行う．提案手法では，視聴履歴データ収集時に集めているIP アドレス・郵便番号・メーカ ID・ブラウザメジャーバージョン・ブラウザマイナーバージョンの 5 項目とチャンネル遷移タイミングが一致するテレビを同一テレビと推定する．在阪放送局 4 社にて，放送局間での非特定視聴履歴データ連携が技術的に可能か検証した「テレビ視聴データ連携に関する共同技術検証実験」において本手法を適用した結果，各放送局で取得された約 376 万台分のデータのうち約 267 万台分（約 71.0%）のテレビをマッチングできることを確認した．

キーワード：テレビ，視聴履歴データ，ビッグデータ，IoT，クロスデバイスマッチング

Proposal and Practice of Integration Method for Large Scale TV viewing log data Across Broadcasters

Hiroki Matsuda[1,2,a)]　Taichi Sakakibara[2]　Yuki Matsuda[1]　Teruhiro Mizumoto[3]　Keiichi Yasumoto[1]

Received: June 20, 2022, Revised: August 9, 2022/October 6, 2022,
Accepted: October 21, 2022

Abstract: Recently, TV broadcasters have been collecting and utilizing non-personal TV viewing log data, including start and end times of viewing, from TVs connected to the Internet in a format that does not identify individual viewers. However, since each broadcaster can only use its own non-personal TV viewing log data, it has not yet been able to obtain useful knowledge despite the vast amount of data it has accumulated. In addition, the collection methods and data granularity of non-personal TV viewing log data vary from station to station, and the data accumulated by each station cannot be integrated and used. In this paper, we propose a method for matching non-personal TV viewing log data collected by each broadcaster using its own method, and integrate the data. In the proposed method, TVs whose channel transition timing matches five items collected at the time of non-personal TV viewing log data collection (IP address, ZIP code, TV receiver manufacturer ID, browser major version, and browser minor version) are presumed to be the same TV. This method was applied to a "joint technology verification experiment on TV viewing log data linkage" conducted by four broadcasters in Osaka to verify the technical feasibility of linking non-personal TV viewing log data among broadcasters. We confirmed that we were able to match approximately 2.67 million TV sets, which corresponds to 71.0% of the 3.76 million unique TV sets acquired by each broadcaster.

Keywords: TV, TV viewing log data, big data, IoT, cross-device matching

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[2] 読売テレビ放送株式会社
Yomiuri Telecasting Corporation, Osaka 540–8510, Japan
[3] 大阪大学大学院情報科学研究科
Osaka University, Suita, Osaka 565–0871, Japan
[a)] hiroki.matsuda@vtv.co.jp

1. はじめに

2007 年頃にインターネット接続可能なスマートテレビが登場したことにより，テレビは単方向に放送コンテンツを受信するだけではなく，ネットワークを介して，コンテンツにアクセスできる双方向の通信が可能となった．2021 年には，41.8% のテレビがインターネットに接続され [1]，視聴者は，データ放送コンテンツや動画配信サービス（VOD：Video On Demand）などが利用可能となっている．これに伴い，インターネット接続されたテレビに関して，各テレビがどの番組を視聴しているか把握可能になっており，放送局は放送サービスを向上させるために，このような情報を**視聴履歴データ**として収集・蓄積している．

視聴履歴データは，本人許諾や個人情報の有無により，オプトイン型特定視聴履歴データ，オプトイン型非特定視聴履歴データ，オプトアウト型非特定視聴履歴データに分けられる．

オプトイン型非特定視聴履歴データの利活用について，様々な研究が行われている．菊池ら [2] は，東芝製テレビに絞った分析を行っており，インターネットに結線された東芝製テレビ視聴者が番組ジャンル別にどのような視聴傾向を持っているのかを明らかにした．また，水岡ら [3] は，テレビを視聴パターンにより分類する手法を提案している．

オプトアウト型非特定視聴履歴データは，個人を特定しない形式で収集されるデータを指し，新たな価値を生み出すビッグデータとして，放送局のみならず，スポンサーや広告代理店からも利活用が期待されている．オプトアウト型非特定視聴履歴データの利活用については，2019 年に一般社団法人放送セキュリティセンターにより公表された「オプトアウト方式で取得する非特定視聴履歴の取扱いに関するプラクティス [4]」を基に，各民間放送局により収集が開始され，利活用に向けた取り組みが進んでいる．具体的には，筆者らの研究グループによって研究されている．松田ら [5] はテレビ CM 視聴がその後のインターネット検索行動に与える影響について分析を行っている．また，吉村ら [6] は CM の完視聴率にどのような地域差が存在するのか分析している．このように利活用に向けた研究は行われているが，各放送局が個別に収集している本データ単体を分析したとしても，1 つの放送局におけるテレビ視聴状況しか把握できないため，新たな価値を生み出すには至っていないのが現状である．

また，電通の調査 [7] によると**図 1** のように 2019 年に地上波テレビ広告費がインターネット広告費に追い抜かれたとの報告結果も出ているが，テレビ広告は未だに数千万人以上にリーチすることができる約 2 兆円規模の巨大市場を有している．現在，テレビ広告の価値を測る指標としては株式会社ビデオリサーチが提供する視聴率が存在しているが，標本調査による統計データとなっているため，テレビ CM や番組コンテンツをより深く分析するには不十分な場合がある．たとえば，株式会社ビデオリサーチのデータを用いた年齢や性別による視聴傾向分析が行われているが，最近の視聴者は趣味趣向が個別化しているため，年齢や性別による単純なカテゴリ分類では議論することが難しい．より詳細な分析を行うためには，同一の番組出演者や同一番組ジャンル等を視聴した視聴者がどのような視聴傾向を持っているかなど，視聴者の行動を起点とした視聴傾向を把握することが重要だと考えられる．これらの視聴者行動は，抽出条件を増やすほど視聴者行動をグループ化することが可能となるが，視聴率データは関西地区の 1,200 世帯を対象としており標本数が少ないため，同様の分析が難しい．

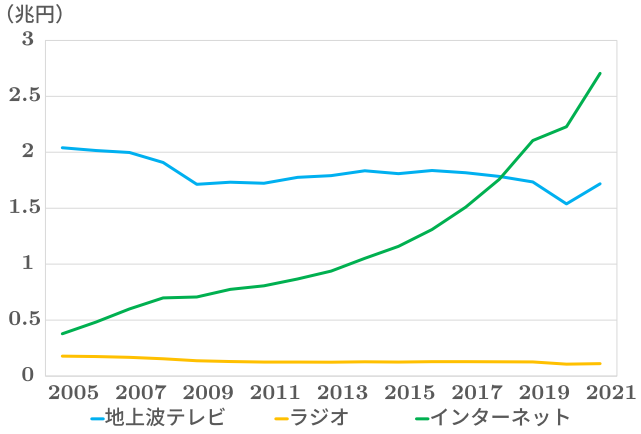

図 1　日本の広告費

Fig. 1　Advertising expenditures in Japan.

この課題に対して，関西地区だけでも 100 万台以上のテレビから各放送局が個別に収集している非特定視聴履歴データを連携・統合することによる解決手法の検討が進んでいる．在阪の 4 つの民間放送局においても，放送局間での視聴履歴データの連携が可能か議論および，共同技術検証実験を実施している．しかし，非特定視聴履歴データは，各放送局が取得にあたりコスト検討や仕様検討し，各放送局の経営判断により取得を開始しているため，放送局ごとに方式やフォーマットが異なっており，放送局間のデータを統合することは難しい．そこで本稿では，各放送局が収集した非特定視聴履歴データについて，視聴者のチャンネル切り替えを把握できるように，放送局間の非特定視聴履歴データをマッチングすることで各放送局が各テレビ端末を特定するために割り振っている ID（以下，テレビ ID[*1]）を統合する手法を提案し，在阪放送局が収集している実データへの適用を実施する．

提案手法では，IP アドレスをキーにして，その他に視

[*1] 放送局ごとに呼び方は異なるが，本稿ではそれらを「テレビ ID」と総称する．

聴履歴データ収集時に取得している郵便番号，メーカ ID，ブラウザメジャーバージョン，放送局が付与した約 376 万台分のブラウザマイナーバージョンの 5 項目すべてが一致するテレビのうち，更にチャンネル遷移タイミングが一致するテレビを同一テレビと推定する．

放送局間の視聴履歴データの連携に関する技術検証を目的として実施された「テレビ視聴データ連携に関する共同技術検証実験 [8]」（以下，在阪視聴データ連携技術実験）において，在阪の 4 つの民間放送局で収集された実際の非特定視聴履歴データに対して本手法を適用した．その結果，各放送局が付与した約 376 万台分のテレビ ID のうち約 71.0% に相当する約 267 万台分のテレビ ID をマッチングできることを確認した．

本稿の構成は次のとおりである．まず，2 章で視聴履歴データについて説明し，3 章で提案方式と検討項目の精査を行い，4 章で実データを用いた検討項目の検証と提案方式の実データへの適用結果を示す．最後に 6 章にて本稿をまとめる．

2. 日本国内における視聴履歴データ

本章では，視聴履歴データの分類および現状の問題についてそれぞれ述べたのち，本稿の位置づけについて述べる．なお，本稿プラクティスは**表 1** のとおり，オプトアウト型非特定視聴履歴データに対して実施したものである．

2.1 視聴履歴データの分類

日本国内において視聴履歴データは，表 1 のように，本人許諾や個人情報の有無により，オプトイン型特定視聴履歴データ，オプトイン型非特定視聴履歴データ，オプトアウト型非特定視聴履歴データの三種類に分類される．各データの特徴を表 1 に示すとともに，以降で詳述する．

2.1.1 オプトイン型特定視聴履歴データ

オプトイン型特定視聴履歴データは，視聴者から許諾を得たうえで収集される，メールアドレス等の本人特定が可能な個人情報と紐づけられた視聴履歴データである．放送局は，視聴者がパソコンやスマートフォンなどからウェブページを通して会員登録を行い，インターネット接続したテレビのデータ放送画面からログインすることで取得可能な状態となる．本人許諾を得る際に，多様な個人情報を取得できるため，視聴履歴データの分析が容易に行える．しかし，視聴者自身の手で，事前にテレビ端末以外から会員登録を行い，さらにテレビ端末からログインする必要があるため，負担が大きく，多くのデータを集めることは難しい．また，会員登録を行う視聴者層にも偏りが大きいと言われており，収集されるデータの多様性が低い．

2.1.2 オプトイン型非特定視聴履歴データ

オプトイン型非特定視聴履歴データは，視聴者から許諾を得たうえで収集されるが，個人情報は含まない視聴履歴データである．

放送局は，視聴者がインターネットに接続されたテレビのデータ放送画面を用いて許諾を行うことで取得可能な状態となる．この方法では，許諾時にアンケート形式で性別，生年月等の視聴者属性に回答して貰うことで，個人情報とはいかないまでも，視聴者属性を持った視聴履歴データを取得可能である．オプトイン型特定視聴履歴データとは違い，テレビ端末のみで許諾が得られるため，視聴者の負担は軽減されるが，この手法においても，視聴者が能動的に取得のための手続きを行う必要があるため，参加者層の多様性が低く，多くのデータを収集することも難しい．

2.1.3 オプトアウト型非特定視聴履歴データ

オプトアウト型非特定視聴履歴データは，本人の許諾無しに収集される，個人情報を含まない視聴履歴データである．放送局は，視聴者がテレビをインターネットに接続するだけで取得可能な状態になる．視聴者がデータを提供したくない場合は，データ放送画面からデータ提供を拒否（オプトアウト）できる．視聴者の能動的な会員登録や許諾を必要とせず，個人情報や視聴者属性を取得することはできないため，視聴開始/終了時刻や，テレビに登録されている郵便番号（住所ではない），テレビ ID などの個人が特定できないデータのみが収集可能である．オプトイン型と異なり，会員登録や許諾を必要としないため，視聴者の負担は無く，多様な視聴者から多くのデータを収集することができる．しかし，個人情報や視聴者属性を用いることができないため，各放送局で蓄積されている自局のデータのみでは，簡単な分析しか行えない．

2.2 視聴履歴データが抱える現状の問題

放送局が収集する視聴履歴データは，視聴者がインターネットに接続されたテレビで，特定の放送局にチャンネル

表 1　視聴履歴データの分類
Table 1　Classification of TV viewing log data.

種類	データの特徴				本稿で扱うデータセット
	本人許諾	個人情報	データの多様性	データの数	
オプトイン型特定視聴履歴データ	有り	有り	低い	少ない	-
オプトイン型非特定視聴履歴データ	有り	無し	低い	少ない	-
オプトアウト型非特定視聴履歴データ	無し	無し	高い	多い	○

を設定したタイミングで各放送局で提供しているデータ放送プログラムを利用して収集される．つまり，各放送局は，自局を視聴中のテレビから，独自の項目および粒度で視聴履歴データを収集している．しかし，視聴履歴データの更なる活用を進めていくためには放送局を横断する視聴履歴データを生成し，今以上に視聴者を理解する必要がある．また，横断的な視聴履歴データ生成は，テレビ広告価値を可視化するための手段として，スポンサーや広告代理店からの期待は大きい．

そして，横断的な視聴履歴データを生成するためには，少なくとも複数の放送局をまたいだデータのマッチングが必要である．データのマッチングについては，2018 年度から継続的に，総務省により視聴履歴データに関わる実証事業が行われており，在京放送局が非特定視聴履歴データの連携・統合に向けて技術実証を実施している．具体的には，テレビ受像機に内蔵されている NVRAM と呼ばれる不揮発性メモリに割り当てられた放送事業者共通保存領域に共通 ID を書き込むことによりデータをマッチングさせる方式である．本方式を使えば，確実に放送局間の ID をマッチングさせることができるため，優れた手法ではあるが，2 点の課題がある．まず，1 点目として，放送事業者共通保存領域は民間地上放送事業者だけでなく，個人情報を大量に保有する NHK や有料放送事業者も読み込むことができる．共通 ID は改正個人情報保護法において，個人関連情報と定義されているため，個人関連情報をあらゆる放送事業者が参照できる場所に保管することに対して，個人情報保護委員会から疑義がでている．2 点目として，法律では問題ないと判断されたとして，放送事業者共通保存領域を民間放送局の利益のために使用するためには，テレビ受像機の規格改定が必要となる．規格改定には，放送事業者だけでなく，テレビ受像機メーカの理解も必要となり，テレビ受像機を利用したテレビ視聴履歴データを用いてサービス提供している各メーカの理解を得ることは難しいと考えられる．

2.3　本稿の位置づけ

本稿では，在阪の 4 つの放送局（読売テレビ・毎日放送・朝日放送テレビ・関西テレビ）がそれぞれ独自に収集したオプトアウト型非特定視聴履歴データを用いて，総務省実証とは違う NVRAM の放送事業者共通保存領域を使用しないデータ統合の手法の提案およびプラクティスを示す．

3.　放送局横断テレビ視聴履歴データ統合手法 Non-NVRAM TIME マッチングアルゴリズム

本章では，NVRAM の放送事業者共通保存領域を使用しないテレビ視聴履歴データ統合手法である，Non-NVRAM TIME マッチングアルゴリズム（以下，NNTM アルゴリズム）を提案する．本手法は，データ収集を行っているデータ放送プログラムは変更せず，各局各様で収集しているオプトアウト型非特定視聴履歴データをそのまま利用することを前提とするが，オプトアウト型非特定視聴履歴データ以外への適用も可能である．

3.1　NNTM アルゴリズム概要

NNTM アルゴリズムを適用するステップを**図 2** に示す．

NNTM アルゴリズムは，クロスデバイスマッチングの一手法である IP アドレスマッチングを基本とする．しかし，IP アドレスはマンション等の集合住宅において 1 つの IP アドレスを複数宅で共用していることも多く，IP アドレスだけではテレビを一意に特定することは難しい．また，IP アドレスは一定期間で変更されていくが，どの程度の期間で変更されるかはプロバイダに依存しており，データを観測すると数分〜数日で変更されるものが多い．

そこで，図 2 の Step1 では，各社が同一のテレビ端末から視聴履歴データを集めているという特徴を利用する．**図 3** に示すとおり，IP アドレスだけでテレビを分離するのではなく，その他にデータ放送プログラムを用いて同時に収集している郵便番号，メーカ ID（各放送局がテレビ受像機メーカを区別するために割り振っている ID），テレビ受像機を起動させているブラウザ[*2] のメジャーバージョン情報，マイナーバージョン情報を使ってテレビの分離を行う．

Step2 として，各放送局のデータを IP アドレスや郵便番号等で分離した後にマッチングさせていくが，マッチング精度を高めるために更に「視聴時刻」をマッチング条件に使うこととする．具体的には，**図 4** で示すとおり，ある 1 台のテレビが A 局・B 局・C 局とチャンネル遷移を

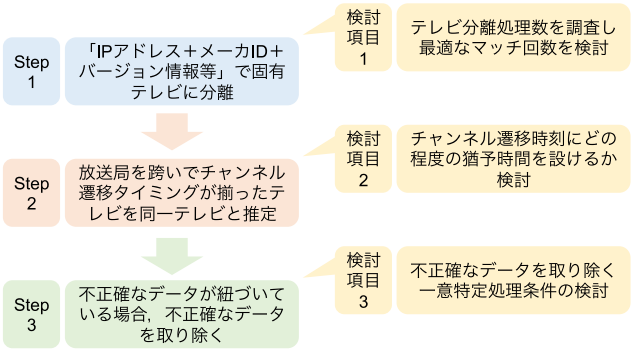

図 2　NNTM アルゴリズム適用ステップと検討項目

Fig. 2　Non-NVRAM time matching algorithm application steps and considerations.

[*2] テレビ端末がデータ放送を起動する場合，専用のブラウザを利用している．メーカやテレビ型番により利用されているブラウザは異なる．

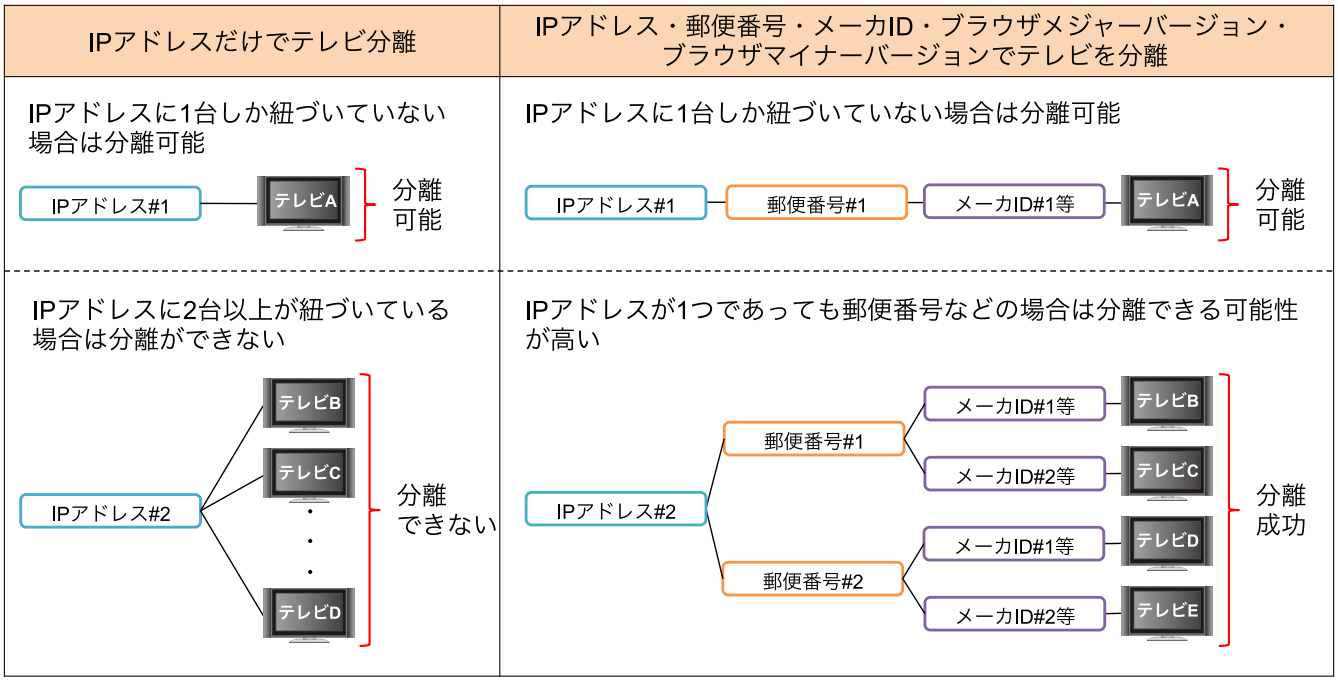

図 3　テレビ分離のイメージ

Fig. 3　Example of TV separation.

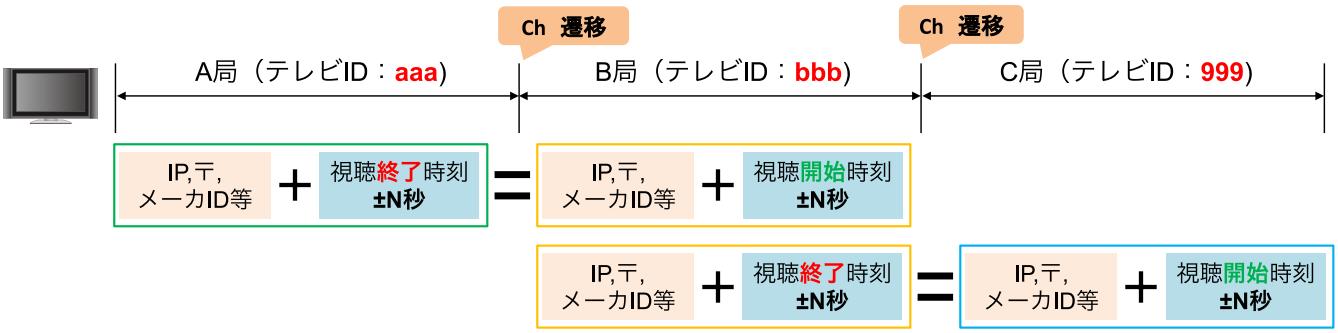

図 4　あるテレビにおけるマッチングイメージ

Fig. 4　An example of matching in one TV.

しながら継続視聴している場合を想定する．A 局・B 局・C 局は同一テレビに対して，それぞれ独自のテレビ ID を付与している．これらのデータを持ち寄り，前述のとおり「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」がすべて一致するテレビのうち，さらに視聴終了時刻と視聴開始時刻に一定の猶予を見た時刻が一致したものを同一テレビと推定する．最後に Step3 として，不正確なデータが同一テレビと推定されている場合の処理を行う．具体的には，A 局の 1 台に対して，B 局等で複数台のテレビが紐づいてしまった場合，経験的に不正確なデータを特定できる場合にはその不正確なデータを取り除くことを目指す．

3.2　NNTM アルゴリズム検討項目

前節のとおり，NNTM アルゴリズムを実装するうえで，図 2 のとおり，各 Step で必要な検討項目（1）マッチング回数，（2）猶予時間，（3）一意特定処理の 3 点を検討する．本節では，これらの検討項目の論点を整理する．

3.2.1　検討項目 1：マッチング回数の検討

NNTM アルゴリズムでマッチングした異なる放送局のテレビは 1 回のマッチングで同一テレビと推定しても良いのか検討する必要がある．そのためには，IP アドレス以外の要素を用いることでテレビそのものをどれだけユニークと識別できるのか調査する．具体的には，IP アドレスが変更されないと想定される短期間における「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」で分離できるテレビの割合を調べることで，インターネットに結線されたテレビと IP アドレスやその他データの関係を調査する．同一集合住宅内で同一メーカ同型式テレビを保有する世帯が多くなるほどテレビを分離できなくなり，チャンネル遷移タイミングが一致するテレビが多数生じる可能性が高いため，

一定期間内における最低マッチング回数などを検討する必要がある．

3.2.2　検討項目 2：猶予時間の検討

タイムマッチングにおける視聴終了時刻と視聴開始時刻の猶予時間を検討する必要である．そのために，全放送局の組み合わせにおいて，複数の猶予時間でのマッチング数を調査する．理論的に，タイムマッチングにおける視聴終了時刻と視聴開始時刻は，まったく同一時刻であることが望ましい．しかし，実際はチャンネル遷移後にデータ放送プログラムが起動してデータ収集を開始するタイムラグやデータの正確性に差が存在する．そこで猶予時間を設定することで，より正確なマッチングを目指す．ただし，この猶予時間を取り過ぎると，**図 5** のように，同じ集合住宅に住んでいる同一メーカ同一機種を所有している世帯間において，偶然チャンネル遷移のタイミングが一致する可能性がある．一致した場合，1 対 1 マッチング数が減少し，1 対多のマッチング数が増加する．

そのため，各社の視聴履歴データ特性から最適な猶予時間を検討する必要がある．

3.2.3　検討項目 3：一意特定処理の検討

放送局間で 1 台のテレビに複数台のテレビが紐づいてしまった場合の，一意特定処理を検討する必要がある．具体的には，明らかに誤ったテレビが紐づいている場合にそのマッチングを除去する処理を検討する．図 5 にあるとおり，偶然チャンネル遷移のタイミングが一致すると，1 台のテレビに複数台のテレビが多数紐づくことが想定される．また，マッチング期間を増やせば増やすほど，たとえば 100 回は同じテレビが 1 対 1 で紐づいているが，1 回だけ偶然別テレビが紐づいているなどのデータが増えていってしまう．このように複数台がマッチングしたデータをすべて捨ててしまうとマッチング期間を増やせば増やすほどマッチング数が減少することが想定される．そこで，経験的にみて，1 台のテレビに複数台のテレビが紐づいているが，明らかに不正確なデータが紐づいているであろう場合において，不正確なマッチングデータを排除し，1 対 1 になるような一意特定処理を組み込むことでマッチングデータの救済を実施する．

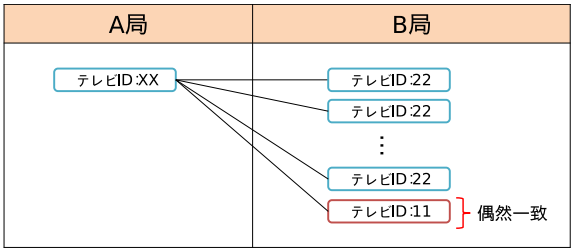

図 5　A 局 1 台に対して B 局 2 台紐づいたパターン

Fig. 5　Pattern with two units of TV station B tied to one unit of TV station A.

4.　実視聴履歴データへの適用事例プラクティス

本章では，在阪視聴データ連携技術実験にて収集した実データと実際に適用したマッチング事例について述べる．

4.1　在阪局のオプトアウト型非特定視聴履歴データ

本節では，在阪の 4 つの放送局が取得している非特定視聴履歴データの特徴について説明する．今回の検証で使用したデータは，在阪 4 局にて実施した在阪視聴データ連携技術実験で取得・交換したデータを用いる．データ期間は，2021 年 10 月 4 日～2022 年 1 月 5 日である．

4.1.1　データ取得方式と取得対象メーカ

データ放送プログラムを利用して収集する方法には，ビーコン方式と From-To 方式の 2 種類の方式があり，**表 2** のように，放送局間で方式や収集条件が異なる．

ビーコン方式は，視聴者が視聴しているときに一定間隔で視聴履歴データをサーバに送信する方法である．リアルタイム性には優れているが，送信間隔次第で視聴終了時刻が正確では無くなってしまうという特徴がある．また，ビーコンの送信間隔は放送局によって違いがあり，表 2 のように，A 局は 60 秒，B 局は 15 秒を採用している．

From-To 方式は，視聴者が視聴し始めた時刻を NVRAM に保存しておき，また視聴終了時にも時刻を NVRAM に保存しておき，視聴者が当該チャンネルに戻ってきたタイミングで前回視聴開始・終了時刻をサーバに送信する方法である．From-To 方式は，データ送信タイミングが視聴者行動に依存しているため，リアルタイム性はないが視聴開始・終了時刻を正確に記録できる特徴がある．

データを取得する対象メーカも放送局によって違いがあり，A 局と B 局はすべてのメーカを対象に取得している．しかし，C 局は一部の主要メーカを取得対象としておらず，D 局は国内シェアの大きいメーカのみを対象としている．

4.1.2　データ比較

在阪視聴データ連携技術実験で交換した 2021 年 10 月 4 日～2022 年 1 月 5 日のデータ比較結果を**表 3** に示す．

表 3 にあるようにデータ取得の対象がすべてのメーカになっている A 局と B 局はテレビ ID 数が多いことが確認できる．また，ビーコン方式で取得している A 局と B 局

表 2　各放送局の視聴履歴データ特徴

Table 2　Viewing history data characteristics for each TV station.

	A 局	B 局	C 局	D 局
方式	ビーコン	ビーコン	From-To	From-To
ビーコン間隔	60 秒	15 秒	-	-
対象メーカ	全て	全て	一部未取得	主要メーカ

表 3　在阪 4 局の視聴履歴データ比較
Table 3　Comparison of viewing history data of 4 TV stations in Osaka.

放送局	テレビ ID 数	データ量	行数
A 局	4,639,071	404GB	11 億行
B 局	4,482,199	13.5TB	645 億行
C 局	3,614,554	64GB	6.7 億行
D 局	3,008,024	56GB	5.8 億行

表 4　データ処理後の在阪 4 局視聴履歴データ比較
Table 4　Comparison of viewing history data of 4 TV stations in Osaka after data processing.

放送局	テレビ ID 数	データ量	行数
A 局	3,760,849	72GB	4.0 億行
B 局	3,727,169	104GB	5.7 億行
C 局	3,214,689	60GB	6.2 億行
D 局	2,604,325	64GB	5.0 億行

は行数も多く，From-To 方式で取得している C 局と D 局は行数が少ないことが確認できる．A 局はビーコン方式でデータ取得を行っているが，データベースへ格納するタイミングで一部のデータを From-To 方式のデータ形式に変更することでデータ量と行数の削減を図っているため，同じビーコン方式の B 局よりも少なくなっている．

4.2　実視聴履歴データへの適用事例プラクティス

本節では，在阪視聴データ連携技術実験の実データを利用したマッチング事例について述べる．

4.2.1　実データの前処理

前提条件を揃えるためにすべてのデータを C 局と D 局が対象としているメーカに絞る．また，NNTM アルゴリズムを適用するにあたり，全局のデータテーブルを統一し，データを From-To 形式に変換することでデータ量の削減に努めた．データ処理後のテレビ ID 数，データ量，レコード数を**表 4** に示す．テレビ ID 数は，全メーカを対象にデータ取得を行っていた A 局と B 局では約 17–19% 減少している．また，全放送局のデータのうち，各放送局が独自で取得しているデータを削除し，共通項目だけを共通テーブルとすることでデータ量・レコード数ともに削減できた．

4.2.2　NNTM アルゴリズムにおける検討事項

3.2 節で記載した検討事項について，実データを参照しながら条件を確定させていく．なお，参照するデータは在阪視聴データ連携技術実験で得られたデータのうち，2021 年 10 月 4 日～10 月 17 日のデータとする．

検討事項 1：マッチング回数

3.2.1 項に前述のとおり，同一集合住宅内で同メーカ同型式テレビを保有している世帯が多い場合は，チャンネル

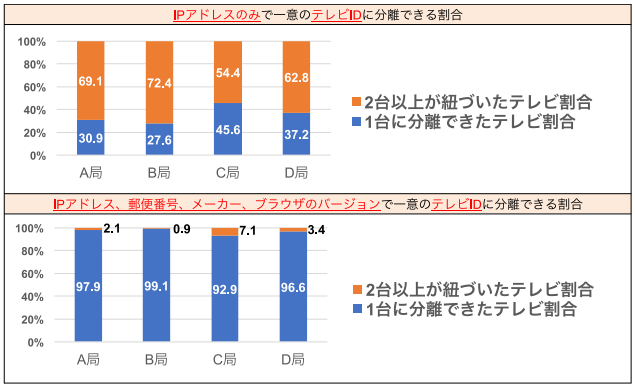

図 6　項目別のテレビ分離割合
Fig. 6　Percentage of TV separation by item.

表 5　各放送局の時刻データ正確性
Table 5　Accuracy of time data for each TV station.

放送局	A 局	B 局	C 局	D 局
開始時刻の正確性	正確	正確	正確	正確
終了時刻の正確性	60 秒以内の誤差	15 秒以内の誤差	正確	正確

遷移タイミングが一致するテレビが多くなってしまうことが想定される．そこで，実データを用いて「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」で分離できるテレビの割合を調査する．なお，本調査は 2021 年 10 月 5 日 19 時～21 時のデータに対して実施した．

調査結果を**図 6** に示す．図 6 の上部は「IP アドレス」のみでテレビを分離した結果を示している．その結果，IP アドレスとテレビ ID が 1 対 1 になるテレビは 54–72% であった．そして，同一のデータに対して「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」で分離した結果を図 6 下部に示している．その結果，93–99% のテレビが一意に分離されることを確認できた．本データより「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」で分離をすることでほとんどのテレビを一意特定することができ，同一集合住宅において 1 つの IP アドレスを複数宅で共用していることにより起こる問題も解決できる．さらに NNTM アルゴリズムでは，これらの分離したデータに対して更にチャンネル遷移タイミングをマッチング条件に加えるため，4 局データにおいてはマッチング数を 1 回に設定すれば良いという知見が得られた．

検討事項 2：猶予時間

3.2.2 に前述のとおり，マッチング条件に利用する猶予時間を調査する必要がある．条件として，視聴終了時刻と視聴開始時刻が同一時刻であることが望ましいが，実データではプログラム起動ラグや，表 2，**表 5** 記載のとおりビーコン方式にはデータ取得間隔があり，データの正確性

に差が存在するため，猶予時間を設ける必要がある．ただし，猶予時間を取りすぎると誤マッチングが増えてしまうため，各社の視聴履歴データ特性から最適な猶予時間を検討し，実データから最適な猶予時間を確認する．

表 5 のとおり，A 局からその他局へチャンネル遷移する組み合わせの相性が悪いため，許容誤差は ±60 秒が最適だと推測される．そこで実データを用いて，各局データのマッチング数を 0 秒，15 秒，30 秒，60 秒，120 秒で分析を行った結果を**表 6** に示す．

この結果から，事前推測していた許容誤差 ±60 秒でマッチング数の増加率が鈍化していることが確認できる．また，**表 7** のとおり，1 対多のマッチング数は許容誤差を拡大すると単調に増加していくことが確認できる．

以上の結果より，4 局データにおいては猶予時間を ±60 秒と設定すると良いという知見が得られた．

検討事項 3：一意特定処理

3.2.3 項に前述のとおり，マッチングするために使うデータ期間を延ばすほど，1 台に対して複数台が紐づいてしまうことが想定される．そこで経験的に見て，明らかに不正確なデータがマッチングしている場合には，その不正確なデータを排除する一意特定処理を組み込むこととする．

具体的には，マッチング数 10 回以下となっているサンプルを抽出し，出現しうるパターンを整理したうえで経験的に条件式 (1) とした．条件式 (1) を用いて，一意特定処理を実施する．なお，合計マッチング数が 3 以下の場合は対象外とし，一意特定処理を実施しないこととした．

表 6　許容誤差別の放送局 1 対 1 マッチング数
Table 6　Number of TV station one-to-one matches by tolerance time.

組合せ	許容誤差 ± 0 秒	± 15 秒	± 30 秒	± 60 秒	± 120 秒
A 局⇔ B 局	3,194	911,032	1,087,265	1,187,329	1,219,404
A 局⇔ C 局	2,847	1,122,730	1,242,122	1,300,492	1,310,827
A 局⇔ D 局	3,523	1,222,884	1,267,231	1,302,679	1,304,329
B 局⇔ C 局	4,880	1,368,043	1,422,740	1,429,611	1,425,659
B 局⇔ D 局	8,047	1,236,956	1,296,801	1,309,869	1,310,638
C 局⇔ D 局	3,732	1,351,503	1,389,753	1,382,557	1,364,098

表 7　許容誤差別の放送局 1 対多マッチング数
Table 7　Number of TV station one-to-many matches by tolerance time.

組合せ	許容誤差 ± 0 秒	± 15 秒	± 30 秒	± 60 秒	± 120 秒
A 局⇔ B 局	475	83,328	136,388	193,984	246,993
A 局⇔ C 局	873	130,394	194,810	266,340	343,817
A 局⇔ D 局	513	130,243	176,600	236,494	301,756
B 局⇔ C 局	1,374	107,988	162,100	225,706	298,432
B 局⇔ D 局	817	82,319	126,824	181,890	246,738
C 局⇔ D 局	1,652	141,589	211,027	303,482	423,896

$$\text{1 位のマッチング数} - \text{2 位のマッチング数} > \text{全マッチング数} \times \frac{1}{3} \quad (1)$$

適用イメージを**図 7** に示す．適用例では，A 局 1 つの ID に対して，B 局 2 つの ID が紐づいた場合に，一意特定処理が成功したパターンを示している．

実データへ一意特定処理を行った結果を**表 8** に示す．表 8 のとおり，どの組み合わせにおいても約 20–30% の一意特定に成功した．以上の結果より，4 局データにおいては条件式 (1) にすると良いという知見が得られた．

なお，IP アドレスは時間とともに変化するが，本方式では個社が独自で割り振っているテレビ ID をキーとして使い，さらに同一時間帯のチャンネル遷移タイミングでマッチングしているため，同じテレビ ID に対して紐づく IP アドレスが途中で変更になったとしてもマッチング数への影響は少ないと想定される．

4.2.3　NNTM アルゴリズム適用結果

在阪視聴データ連携技術実験で交換した 2021 年 10 月 4 日～2022 年 1 月 5 日までのすべての実データを利用して，NNTM アルゴリズムを適用した結果を**表 9**，**図 8** に示す．結果は，すべての局を特定できた「4 局マッチング数」と 1 つの局だけ特定できなかった「3 局マッチング数」と 2 つだけ特定ができた「2 局マッチング数」で示す．また，適用期間別のマッチング数を確認することで，最低限マッチングに必要な期間を調査する．

94 日分のデータを突合することで，約 159 万台の 4 局特定，約 267 万台の 2 局以上のマッチングに成功した．特に 2 局以上の特定に成功した約 267 万台は，近畿 2 府 4 県

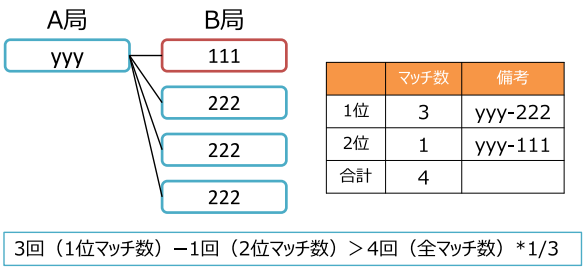

図 7　一意特定処理の適用例
Fig. 7　Example of application of unique identification process.

表 8　一意特定処理数
Table 8　Unique identifier count.

組合せ	特定台数	1 対多台数	特定率
A 局⇔ B 局	37,410	193,984	19.3 %
A 局⇔ C 局	57,939	266,340	21.8 %
A 局⇔ D 局	57,152	236,494	24.2 %
B 局⇔ C 局	44,204	225,706	19.6 %
B 局⇔ D 局	36,792	181,890	20.2 %
C 局⇔ D 局	91,653	303,482	30.2 %

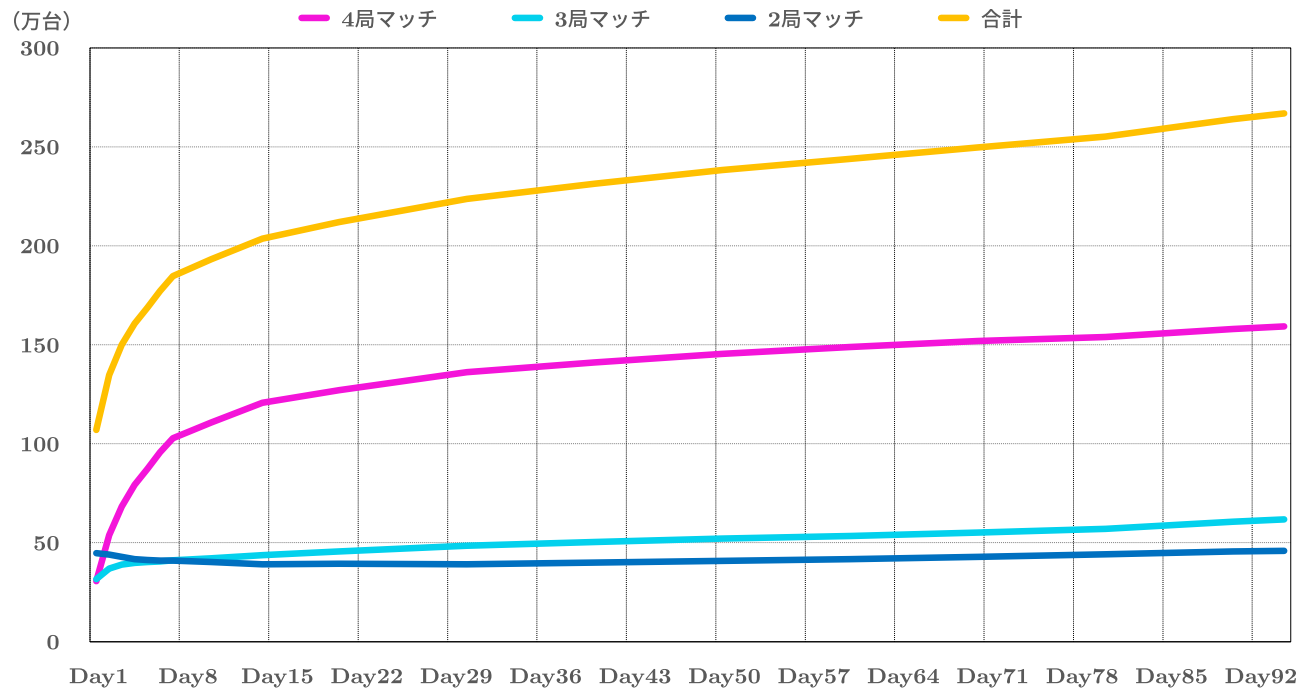

図 8　期間別のマッチング数

Fig. 8　Number of matches by time period.

表 9　期間別のマッチング数

Table 9　Number of matches by time period.

期間	4 局 マッチ数	3 局 マッチ数	2 局 マッチ数	合計
Day1	305,241	316,252	447,669	1,069,162
Day7	1,027,334	410,474	409,591	1,847,399
Day10	1,107,861	421,441	402,491	1,931,793
Day14	1,207,100	437,224	391,053	2,035,377
Day20	1,270,870	455,894	393,352	2,120,116
Day30	1,361,738	484,211	390,990	2,236,939
Day40	1,410,476	503,561	399,475	2,313,512
Day50	1,453,540	520,780	407,639	2,381,959
Day60	1,488,446	534,054	416,394	2,438,894
Day70	1,518,303	550,592	428,481	2,497,376
Day80	1,539,263	570,155	441,695	2,551,113
Day90	1,579,022	605,566	455,274	2,639,862
Day94	1,592,786	617,408	458,872	2,669,066

の約 922 万世帯（R2 国勢調査）と 2 人以上世帯の平均テレビ保有台数 2,076 台から推定される近畿に存在する約 1,915 万台の約 13.9% に該当する．これは放送局が放送サービス向上などのためにマーケティング利用するには充分な台数である．また，図 8 のグラフよりマッチング台数の伸び率が減少し，伸び率が一定に近づくためには 30 日程度のデータ突合が必要であることが分かる．

大阪には 5 つの民間放送局が存在しているが，今回の在阪視聴データ連携技術実験は 4 つの民間放送局で実施されている．また，本取組みの中で，各局が記録している IP アドレスの取得タイミングも視聴開始時，視聴終了時，次回視聴開始時など揃っていないことが判明した．今後は IP アドレス取得タイミングも考慮することでアルゴリズムの精度向上を目指したい．

5.　議論と考察

本方式では「IP アドレス＋郵便番号＋メーカ ID+ ブラウザメジャーバージョン＋ブラウザマイナーバージョン」をいう 5 項目とチャンネル遷移タイミングをマッチング条件としている．今回は在阪放送局が取得している情報のうち，共有している項目をすべて利用しているが，たとえば IP アドレスがあれば郵便番号は必要ないように思える．そこで，各項目の寄与度を調査した結果を**表 10** に示す．表 10 は，5 項目すべてを使った場合と各項目のうち 1 つだけ除いた 4 項目を使った場合のマッチング数を比較している．なお，チャンネル遷移タイミングの許容誤差は 60 秒としている．この結果から，IP アドレスまたは郵便番号を除外した場合に大幅なマッチング数の低下がみられたため，本方式への寄与度が高いといえる．また，その他の項目についても除外時にマッチング数が低下することから，本方式で提案した 5 項目の組み合わせが有効であることが確認された．IP アドレスがあれば不要に思える郵便番号の寄与が高い理由として，実際の居住地とは異なる初期設定値や引越し前の郵便番号などが登録されていることにより，本来 IP アドレスと郵便番号が同一になる集合住宅などにおいて，テレビの分離が行えているためだと考えられる．

また，本方式で使用している IP アドレスは数日もしく

表 10　項目別のマッチングへの寄与
Table 10　Contribution to matching by item.

チャンネル遷移タイミング以外のマッチング条件	マッチング数
5 項目（全項目）	1,027,334
4 項目（IP アドレス以外）	691,894
4 項目（郵便番号以外）	976,809
4 項目（メーカ ID 以外）	1,022,129
4 項目（ブラウザメジャーバージョン以外）	1,021,737
4 項目（ブラウザマイナーバージョン以外）	1,015,200

は数週間という時間経過とともに変更されるが，変更されるタイミングはインターネットプロバイダ会社依存となるため，ユーザ側での把握は難しい．しかし，たとえば 3 ヵ月程度の長期間データの期間中に IP アドレスが変更になったとしても，本方式では個社が独自で割り振っているテレビ ID をキーとして使い，さらに同一時間帯のチャンネル遷移タイミングでマッチングしているため，マッチング数への影響は少ないと想定される．

最後に，今回の参加放送局の中には 1 局だけテレビ ID を半年程度で定期的にリセット処理している局がある．リセットされた場合であっても，残りの 3 局ではリセットされていないため，問題なく同一テレビを追跡することが可能と推測される．全国的にもテレビ ID をリセットする放送局は少ないため，テレビ ID リセットが実施されても問題ないと認識している．ただし，本方式を利用したい放送エリアにおけるすべての放送局がテレビ ID を定期的にリセットする場合，IP アドレス変更時に同一テレビを継続的に認識することが難しくなる可能性は存在する．

6. おわりに

本稿では，放送局が取得する非特定視聴履歴データに対する利活用ニーズと，利活用ニーズに応えていくための課題，その課題に対する従来アプローチとはまったく違う手法による解決手法の提案と実データへの適用結果を述べた．その結果として，放送局が利活用するためには充分なマッチング数を得ることに成功し，最低限必要なデータ期間が 30 日程度であることを示した．今後，本手法を用いて在阪放送局でデータ交換を行い，編成・営業利用を進めていきたい．

謝辞　NNTM アルゴリズムの検証にあたって多大なご協力をいただきました株式会社毎日放送，朝日放送テレビ株式会社，関西テレビ放送株式会社の皆様に，この場を借りて深く感謝申し上げます．

参考文献

[1] 株式会社マクロミル：2021 年年末最新のテレビ利用動向調　査，〈https://www.macromill.com/press/release/20211223.html〉（参照 2022-10-06）．
[2] 菊池匡晃，坪井創吾，中田康太：大規模テレビ視聴データによる番組視聴分析，デジタルプラクティス，Vol.7, No.4, pp.352–360 (2016).
[3] 水岡良彰，中田康太，折原良平：大規模テレビ視聴データによる視聴パターン推移の分析，人工知能学会全国大会論文集，pp.1P203–1P203 (2018).
[4] 一般財団法人放送セキュリティセンター視聴関連情報の取扱いに関する協議会：オプトアウト方式で取得する非特定視聴履歴の取扱いに関するプラクティス（ver2.1），〈https://www.sarc.or.jp/documents/www/NEWS/hogo/2021/optout_practice_ver2.1.pdf〉（参照 2022-10-06）．
[5] 松田裕貴，榊原太一，木俣雄太，鳥羽望海，真弓大輝，松田裕貴，安本慶一：テレビ視聴における非特定視聴履歴データとインターネット検索データの関係性分析，第 14 回データ工学と情報マネジメントに関するフォーラム（DEIM'22），日本データベース学会，pp.1–6 (2022).
[6] 吉村　啓，水本旭洋，榊原太一，松田裕貴：テレビ視聴時の CM 離脱と地域傾向分析，人工知能と知識処理研究会，Vol.121, No.439, pp.43–48 (2022).
[7] 株式会社電通：2020 年日本の広告費，〈https://www.dentsu.co.jp/knowledge/ad_cost/2020/〉（参　照 2022-06-20）．
[8] 読売テレビ放送株式会社：「テレビ視聴データ連携に関する共同技術検証実験」について，〈https://www.ytv.co.jp/privacy/experiments/index.html〉（参　照 2022-10-06）．

松田 裕貴（非会員）

読売テレビ放送株式会社デジタル戦略局主査．2011 年に奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2019 年から読売テレビ放送株式会社デジタル戦略局で，視聴履歴データを含むデータ分析業務に従事．2021 年より奈良先端科学技術大学院大学情報科学研究科博士後期課程．

榊原 太一（非会員）

読売テレビ放送株式会社デジタル戦略局．2020 年に京都大学博士前期課程修了．2020 年から読売テレビ放送株式会社デジタル戦略局で，視聴履歴データを含むデータ分析業務に従事．

松田 裕貴（正会員）

奈良先端科学技術大学院大学先端科学技術研究科助教．2019 年同学情報科学研究科博士後期課程修了．博士（工学）．ユビキタスコンピューティングに関する研究に従事．

水本 旭洋（正会員）

大阪大学大学院情報科学研究科特任講師（常勤）．2014 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．サイバーフィジカルシステムに関する研究に従事．

安本 慶一（正会員）

奈良先端科学技術大学院大学情報科学研究科教授．1991 年大阪大学基礎工学部情報工学科卒業．1995 年同大学大学院博士後期課程退学．博士（工学）．ユビキタスコンピューティングに関する研究に従事．