特集号
招待論文

大規模テレビ視聴データによる番組視聴分析

菊池 匡晃[†1]　坪井 創吾[†2]　中田 康太[†1]

[†1]（株）東芝　[†2]東芝映像ソリューション（株）

本稿では，東芝製ネットワーク対応テレビの大規模視聴データ全国 24 万台分を用い，従来方式のテレビ視聴傾向調査では主にコストの観点から困難だった，リアルタイム視聴とタイムシフト視聴（録画再生）両方を含めた視聴実態や，長期間の視聴行動の変化を調査する集計システムおよび視聴分析事例を報告する．本視聴データは 2012 年から約 4 年間蓄積されており，広告主を含むコンテンツ制作者等に対してドラマ全話を通した視聴実態といった視聴傾向の集計・分析を提供し，ビジネス判断の材料にしていただいている．また，こういった分析サービスの 1 つとして，視聴ユーザの性別・年代といったデモグラフィック情報に基づく集計・分析結果を提供している．すべてのユーザからアンケート回答を得ることはできないため，一部ユーザのアンケート回答を基に，機械学習手法によりテレビの保有世帯構成員の性年代属性を視聴傾向から推定する．実データを用いた実験により，一部属性を除き正例の F 値が 0.7 以上と，実用に耐える精度が得られている．

1. はじめに

数千万人以上の視聴者にニュース，ドラマ，CMなどのコンテンツを同時発信できるテレビメディアは，インターネット広告市場が伸びている現在においても巨大な広告市場であり，2015年の総広告費は約2兆円（全体の約30%）である[1]．テレビメディアの価値を測る数値として，（株）ビデオリサーチの提供する視聴率[2]が存在するが，広告主やコンテンツ制作者が，対象コンテンツをより掘り下げて調査・分析するには不十分な場合がある．たとえば，連続ドラマの何話目から見始めた・やめたといった長期的な視聴傾向を，リアルタイム視聴とタイムシフト視聴（録画再生）を合わせて調べられれば，視聴者のフィードバックを活かした番組・CMを作りやすくなり，視聴者も魅力的なコンテンツを視聴できるようになる．

この課題に対し，近年，ネットワーク接続型テレビの普及により，ユーザから利用許諾を得て取得した視聴データを使った番組やCMの視聴分析が可能となった．当社のテレビ視聴データ集計システム“TimeOn Analytics”では，これまでアンケート調査などで行われてきた視聴傾向を，詳細・高速に把握することができ，さまざまな分野のテレビ広告実施企業やコンテンツ制作企業で活用されている[☆1]．また，ヒット現象の分析への適用も検討されている[3]．

本稿では，これまでのテレビ視聴調査について概観した後，我々のテレビ視聴データ集計システムについて解説する．次に，これまで実際に行われた番組視聴分析事例や「どんな人が」を知るための世帯属性推定アルゴリズムと実験結果を報告し，家庭用ネットワーク対応機器のデータ集計システムとデータ分析時のプラクティスを示す．

☆1 お問合せはanalytics@timeon.jpにて受付．

2. テレビ視聴傾向調査

テレビの視聴傾向を調査する場合，従来は，パネルとして視聴データを提供してくれる希望者を募り，その家にテレビ視聴の計測機器を設置してもらう**パネル方式**を採用していたが，店頭で売られているテレビやレコーダなどの映像機器がネットワーク対応となり，その機器上で利用許諾を取って視聴データを集計する**リアル方式**が台頭してきた．前者は詳細な視聴者プロフィールの分析が可能であり，後者は長期かつ大規模な視聴傾向分析に適する（図1）．両方を兼ね備えた方式は仕組み上難しく，調査目的に適した方式を選ぶ必要がある．

2.1 パネル方式

パネル方式は，主にリサーチ会社に採用されている．対象者（サンプル）に謝礼を支払って調査を依頼するため，その期間を通じて家族構成，年齢，年収，嗜好などの詳細なプロフィールや多くのアンケート回答を確実に取得でき，目的に応じて家庭・個人を対象に視

2016 ©Information Processing Society of Japan

	パネル方式	リアル方式
総論	視聴しているユーザ個人の属性まで詳細に把握可能 一方で，コスト上サンプル数を増やせない	大規模・高精度・長期の視聴分析が可能 一方で，ユーザ属性は限定的（世帯情報までが限界）
ユーザ	調査パネルに応募したユーザ → 定期的に入れ替わる	一般的な映像機器ユーザ → 年単位で継続
維持コスト	ユーザ数に比例して増加	ユーザ数に対してほぼ一定
収集できるデータの特徴	視聴者の情報に優れる	サンプル数と視聴データの種類に優れる

図1　パネル方式とリアル方式の違い

聴傾向を集計することができる．また，リモコンのチャンネルボタンを押す際に家族の誰が視聴しているかを追加で選ばせるなど，視聴時の情報も取得できる．一方で，サンプル数に応じて謝礼額が上昇すること，作業負荷が高いことから同じサンプルを長期間維持することが難しく，継続的な募集活動や計測機器を設置する費用がかかるなど維持費が高額となる．ビジネスが成立するためにはサンプル数を一定数以下に押さえなければならず，視聴傾向の分析内容もサンプル数が少なくて済むものに限定される．代表例としては，前述のビデオリサーチ社のほかに，インテージ社は日用品の購買行動，テレビ番組の接触状況（視聴の有無），Webサイトの接触状況（閲覧の有無）を横断的に調査するi-SSPサービスを提供している．テレビ番組への接触状況は音声収集デバイスを持った6,900のパネルサンプルを用いて調査を行う[4]．またスイッチ・メディア・ラボ社は，関東で約2,000世帯のテレビ視聴調査パネルを維持し，テレビ番組のリアルタイム視聴状況の速報分析サービスを提供している[5]．視聴データに加えて視聴態度の分析を行う例もあり，テレビジョンインサイト社は，マイクロソフト社XboxのKinectをパネルサンプルに配布し，Viewability Index（テレビの視聴可能範囲に視聴者がいるかどうか）と，Attention Index（テレビ画面を注視しているかどうか）を毎秒測定することで，番組やCMの価値データを提供している[6]．この方式はチャンネル切り替え時の追加操作が不要であり，視聴時の状況を正確に取得できる．

2.2　リアル方式

リアル方式は，テレビなどの映像機器を購入したユーザから，許諾を得て視聴データを取得する方式である．映像機器やそれに付随する機能およびサービスを目的に購入した100万人規模のユーザを対象にできるため，大規模なデータを集めやすく，多数のサンプルを必要とする視聴傾向分析が可能となる．視聴データの内容はさまざまだが，機器によってはリアルタイム視聴だけではなく，タイムシフト視聴（録画再生）も計測できる．また，維持のためのシステム運用費はユーザ数の増加の影響を受けにくく，安価に維持できる．一方，ユーザにとっては視聴データを使用してこそ実現できるサービス・機能を利用することが第一目的であり，そこで満足を得られなければ，映像機器のネットワークへの接続や視聴データの利用許諾を得ることが難しい．また，目的外の操作は不満の種になるため，パネル方式のような利用者の情報は取得が難しく，特に，誰が視聴したかといった情報は取得できない．当社では，2005年より，番組録画予約データを収集・集計し番組予約ランキングを生成，それを機器利用者やネットメディアに提供してきた[7]．2016年現在，主要な映像機器メーカの多くが，番組の視聴傾向に基づいておすすめ番組を推薦あるいは自動録画するサービス・機能を無償で提供し，その代わりに視聴データをマーケティング調査向けに使用する許諾をユーザから得ている[8],[9],[10]．北米では，多くのメーカの映像機器にACR（Automatic Contents Recognition）技術を組み込み，ユーザの視聴番組をクラウド側で識別，視聴データとして販売するビジネスを行う会社もある[11]．

2.3　“TimeOn”と視聴分析サービス

当社では，2012年秋以降の当社製液晶テレビ「レグザ（REGZA）」のネットワーク対応機種に対し，クラウドサービスTimeOnを提供するとともに，そこで取得した視聴データを用いた視聴傾向分析サービスを企業向けに提供している．TimeOnでは，視聴データ取得を含むサービス利用許諾に同意していただいたユーザ向けに，日々放送される500以上の番組（関東エリアの場合）や，それを録画した番組，YouTube・TSUTAYA TVなどのVOD（Video On Demand）のコンテンツから，気になるタレントの出演番組や先週見た番組の今週放送分を提示または自動的に録画する機能等を提供している（図2）．特に一部の機種では複数チャンネルを長時間録画し続けるタイムシフトマシン機能を搭載し，約500の録画番組が日々追加されるため，コンテンツ発見の支援は欠かせない．このサービスの実現のため，どんな番組を視聴し，録画したか（興味を持った番組の判定），何を見るのを途中でやめ，再生前に削除したか（興味がなかった番組の判定）などをクラウド側で計算

図 2　TimeOn 画面例：視聴データを分析し，気になる番組やタレント出演番組等を紹介

し，任意回答のアンケートによる世帯属性を使って調整した後，関心がありそうな番組を番組リストとして出力する仕組みを構築している．視聴傾向分析サービスは，この仕組みを応用したものであり，リアルタイム視聴・タイムシフト視聴（繰り返し・早送り再生含む）の実績を計測・集計することができる．

3.　視聴分析に対するビジネスニーズ

当社に寄せられる視聴傾向分析へのニーズは，これまで難しかったタイムシフト視聴を含めた視聴行動の実態把握と，数カ月以上に渡る視聴行動の変化の調査といったものが多い．以下，各ニーズの特徴と集計システム設計時の要件を述べる．

3.1　タイムシフト視聴を含めた視聴実態

これまで，放送されたコンテンツに対する視聴者からのフィードバックは，視聴率の増減，SNSやブログでの評判，アンケート結果などに限られてきた．特に，タイムシフト視聴の実態を把握しコンテンツ制作やCM出稿に活かしたいというニーズは多く，ドラマでは実際に視聴した人数の把握，タイムシフト視聴の際でも視聴されるCM枠の把握[12]等がある．また，タイムシフト視聴のスキップ操作からコンテンツの視聴のされ方を調べ，次回放送時にチャンネルを切り替えられにくい内容に改善する応用もある．

3.2　数カ月以上にわたる視聴行動の変化

連続ドラマやバラエティ番組制作者にとって，先週の放送を見た人が今週の放送を見てくれたかどうかは関心事である．リアルタイム視聴での人気は視聴率で知ることができるが，先週見て今週見なかった人，先週見なかったが今週見た人が混在するため，視聴率が増減した原因を知るためには別の情報源が必要となる．また，CM出稿の観点でも，視聴者の変化に合わせて内容を調整すれば広告効果を高めることができる．この種の分析は調査対象期間が1クール（3カ月）から1年単位になるため，パネル方式では条件の掛合せが多すぎて該当サンプル数がすぐにゼロになってしまう．

3.3　集計システムへの要件

これらのニーズにリーズナブルなコスト・時間で応えるためには，視聴データの取得だけではなく，それを柔軟に扱える集計システムが必要である．2012年のサービス開始当初にこれらのニーズすべてが分かっていたわけではないが，番組のリアルタイム視聴とタイムシフト視聴を個別・合算して扱えること，長期間（数年）の視聴データを秒単位で管理し，技術者でなくても手軽に集計できること，特定条件を満たす機器集合のみで集計を行えること，テレビ販売台数の増加に対してスカラブルであることを要件とした．なお，当社に対する依頼は数カ月前～1日前の視聴傾向分析が多く，リアルタイム視聴の速報集計（放送数分後レベル）など，速報性が要求されるニーズに対しては，優先度を落とす判断をしている．

4.　TimeOn Analytics

TimeOn Analyticsは，以上のビジネスニーズを解決するために開発・運用している視聴データ集計システムである．最初に本システムにおける用語定義と扱う視聴データの特徴について述べ，次に現在のシステム構成を説明する．

4.1　用語定義

前述の視聴率[2]は，日本の人口分布を反映するようサンプル抽出されている．本システムでは当社テレビ利用世帯がサンプルであり，性質が異なるため，**視聴接触率**（以下，**接触率**）と呼び区別する．

4.2　データの特徴

本システムで扱う視聴データは，大きく4つの特徴がある．

- 大規模：全国の当社テレビ利用世帯のデータ
- 秒単位で記録：リアルタイム視聴，タイムシフト視聴ともに特定シーンやCMが判別できる分解能を持つ
- シングルソース：地デジ／BS／110度CSの3波の

リアルタイム視聴／タイムシフト視聴／録画予約データを各機器から取得

- 部分的な世帯属性：アンケート（任意回答）に基づく世帯デモグラフィック属性が付属

データ規模は，対応テレビ販売台数の増加に伴って拡大を続けており，2016年4月現在，約24万台である．地域別の分布は，ほぼ人口分布に比例し，最も少ない放送エリアである高知エリアでも800以上のサンプル数となっている．複合条件での視聴傾向分析を行うためにはデータ規模が重要で，たとえば“接触率10%の新ドラマ第1話を3つすべて視聴した”といった機器を調査しようとした場合，3番組が独立に視聴されると仮定すると出現率は0.1%と大きく減少し，数千台オーダのサンプル数でも数台しか出現しないといったことが起こり得る．また，秒単位で記録されていることで，主に3～30秒の長さで放送される番組宣伝やCMの視聴状況を知ることができ，さらにシングルソースであることで，ドラマにおいてタイムシフト視聴で見ていた人が最終回はリアルタイム視聴で見た，などの把握ができる．なお，任意回答アンケートにより世帯属性を取得しているが，回答は一部の世帯に限られ，また，世帯内の個人全員分の回答が得られるわけではない．一方で視聴傾向と世帯属性には相関があると考えられ，一部のアンケート回答から世帯の属性を推定する試みについて第6章にて詳細を述べる．

ほかに，テレビメディアならではの特徴として，自然災害の発生など，皆がテレビで情報収集したくなる状況では平時に比べてデータ量が突発的に10倍以上になる点に注意が必要である．例として2016年4月14日に発生した熊本地震の際の各地域におけるニュース番組の接触率の秒単位推移を図3に示す．地震発生直後に熊本および福岡では急激に上昇し，関東や北海道では数分遅れてから上昇した．これは九州の多くのユーザが，地震を知覚した途端にすぐにニュース番組に切り替えたり，テレビの電源を入れたりし，九州以外のユーザは，ニュース速報やインターネット経由で地震の深刻さを知り，数分後に同様の行動を行ったと考えられる．このようなケースではオートスケール等は間に合わず事前情報もないが，運用コストがかかりすぎないシステムが求められる．

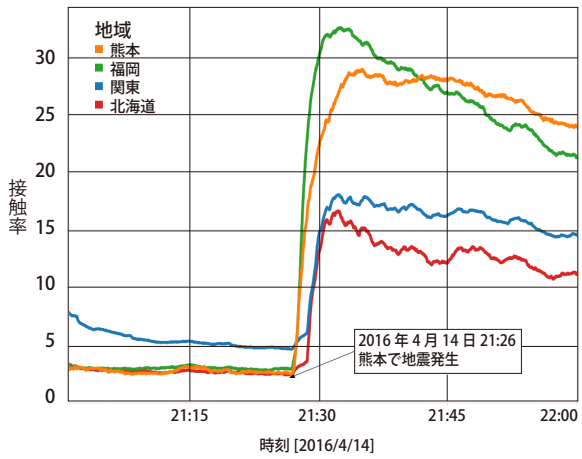

図3　熊本地震発生時の各地域のニュース番組秒単位接触率

4.3　システム構成

TimeOn Analyticsのシステム構成概要を図4に示す．本システムは想定用途を，放送の翌朝までに全国の一般的な視聴状況が把握できること，アナリストが特定番組の長期間の視聴傾向調査等の毎回異なる条件でのアドホックな分析をリーズナブルな時間で行えること，複雑な条件の集計結果でも容易に理解できるようビジュアライズしインタラクティブな操作ができること，の3点とし，それぞれに対応するサブシステムをAWS（Amazon Web Services）上に構築した．以下各サブシステムについて説明する．

4.3.1　定型バッチ集計処理

図4（A）ではEMR（Amazon Elastic MapReduce）を利用した定型バッチ処理で全国・全番組の視聴統計データを集計し，Web APIを介して統計値を提供している．対象テレビ台数が数十万台規模へ拡大してもApache Hadoopを利用し，計算機クラスタをスケールさせることで一定時間内に処理することが可能で，前日分の集計を翌朝までに完了させている．

4.3.2　アドホック集計処理

図4（B）ではAmazon Redshiftを利用し，集計条件に応じたSQLを生成するデータ抽出ライブラリを介してアドホックに集計することができる．地域や世帯属性などで注目する視聴者を絞り込んだ上で，長期間のデータを分析することが主なユースケースとなる．そこで分散列指向DBであるRedshiftを採用し，分散キー，ソートキー，圧縮エンコードなどのパラメータチューニングにより，

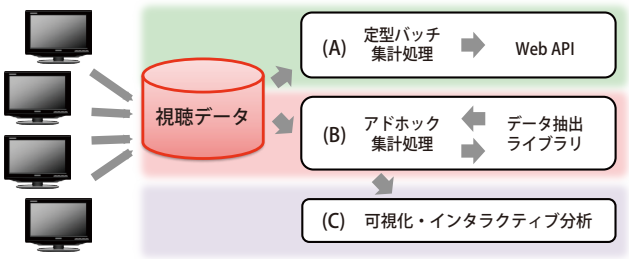

図4　TimeOn Analytics システム構成概要

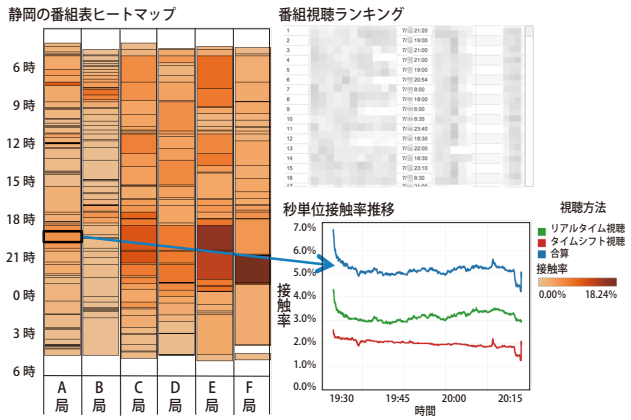

図 5　静岡県における 1 日の番組視聴傾向可視化例

主要なユースケースにおける高速な応答性能を実現している．

4.3.3　可視化・インタラクティブ分析

図4（C）では集計結果の可視化，インタラクティブ分析を行う．可視化については利用者ごとにさまざまなユースケースがあるため，汎用BIツール（Business Intelligence tools）“Tableau”を採用しRedshiftと接続している．静岡県における1日の番組視聴傾向可視化例を図5に示す．ヒートマップは番組表形式になっており，色が濃いほど番組の接触率が高かったことを示している．1つの番組を選択すると右下に選択番組の秒単位での接触率推移が表示され，リアルタイム視聴とタイムシフト視聴を比較するインタラクティブな操作ができる．

5.　番組視聴分析プラクティス 1 ：「番組をどう視聴しているか」

本章では前述のTimeOn Analyticsを用いた，タイムシフト視聴を含めた3つの番組視聴分析事例について述べる．

5.1　番組ジャンルによる視聴傾向の違い

視聴者が1つの番組をどの程度視聴し続けているのかを調査した．図6，図7は番組を視聴した機器の，視聴時間の分布をドラマと音楽で比較したものである．横軸が番組の合計視聴時間，縦軸が機器数を示したヒストグラムである．図6からドラマ番組では視聴者の大多数が番組全体を視聴していることが分かる．本編以外をスキップする操作によりタイムシフト視聴のピークは4,200秒前後にあるが，そこから右側にも広く分布しており，タイムシフト視聴でも多数の機器がCMも視聴している実態が分かる．一方，図7の音楽番組では300秒以下が特に多く，そこからなだらかに分布している．好きなアーティストの出演場面だけを視聴する多数の視聴者が，番組中に激しく入れ替わっており，同一の視聴者が継続的に視聴しているわけではない実態が分かる．つまり全体の平均値としての接触率が同程度であったとしても，その内訳の視聴実態は番組ジャンルによって大きく異なるといえる．

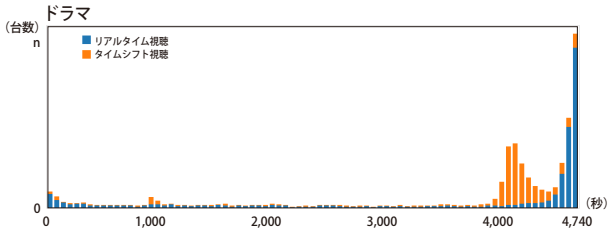

図 6　ドラマ番組の視聴時間分布

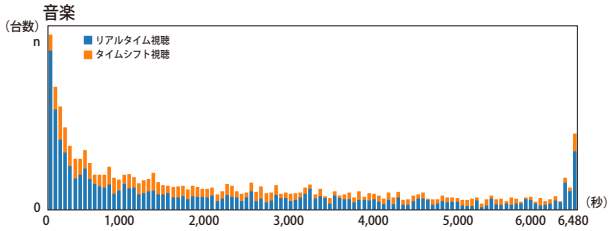

図 7　音楽番組の視聴時間分布

5.2　連続ドラマ 1 話～最終話の視聴傾向

次の事例は，視聴者は連続ドラマを1話から最終話までどのように視聴しているのかの調査である．最終話まで多くの視聴者をひきつけたドラマAと，視聴者の離脱が多かったドラマBの視聴パターン上位20件を図8，図9に示す．なお，リアルタイム視聴，タイムシフト視聴のいずれかで視聴すれば“視聴した”と判定した．ドラマAでは全話視聴したパターンが最多で，途中から最終話まで継続視聴するパターンが上位に目立つ一方，ドラマBでは1話目で脱落してしまうパターンが最多で，途中離脱が多発したことが分かる．

5.3　番組の視聴者ロイヤリティ性

では，バラエティ番組は繰り返し視聴されているのだろうか．始めに，毎週放送される番組の定期的な視聴行動の評価指標として，番組の放送回数をN，視聴者iの番組視聴回数をW_iとし，視聴者iの番組に対するロイヤリティR_iを定義する．

$$R_i = \frac{W_i}{N} \times 100$$

3カ月間で放送されたあるドラマ番組とバラエティ番組について，期間内の視聴者数の推移と番組ロイヤリ

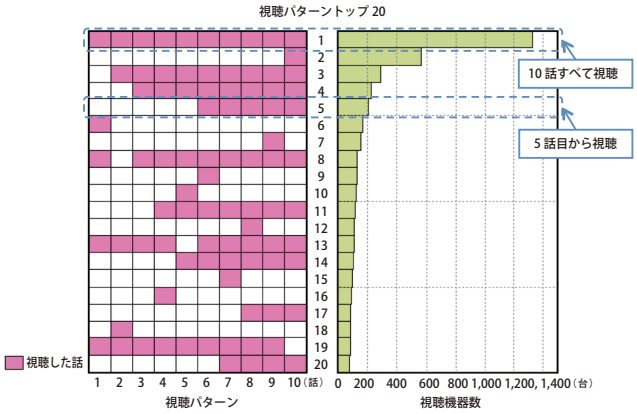

図 8　ドラマ A の視聴パターン

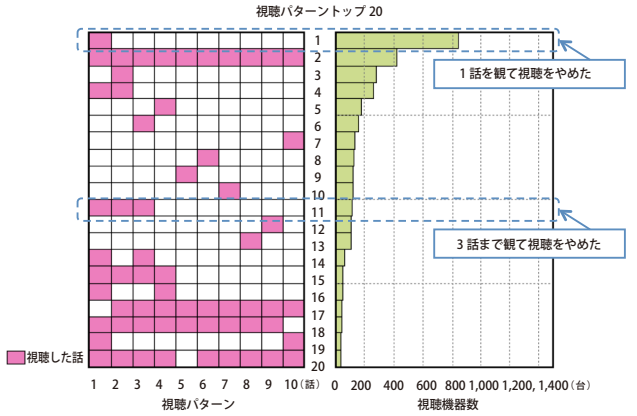

図 9　ドラマ B の視聴パターン

ティの分布をそれぞれ図 10，図 11 に示す．

図 10 の左図から，ドラマ番組では初回から最終回まで視聴機器数は比較的安定して推移しており，リアルタイム視聴よりもタイムシフト視聴している機器が多いことが分かる．なお合算ユニーク視聴は，視聴方法を問わず該当放送回を視聴したかどうかを示している．また図 10 右図は各視聴方法別に，縦軸が 1 放送回以上視聴した機器数を，色が機器のロイヤリティ区分を示している．リアルタイム視聴だけを見ると R が 80 ～ 100 といった高ロイヤリティ層は 5% 程度だが，合算ユニーク視聴では高ロイヤリティ層が大幅に増えている．このことからほぼ毎週観る視聴者もタイムシフト視聴を併用している実態が分かる．毎週リアルタイム視聴することは，ライフスタイルが多様化した昨今では難しいことを示している．

一方のバラエティ番組では図 11 左図から，リアルタイム視聴が中心で，放送回ごとに視聴機器数の変動が大きく，各回ドラマの半数以下と分かる．図 11 右図より低ロイヤリティ層の比率がかなり高いが，ドラマ番組と比較すると 1 回以上視聴機器数がほぼ同水準である．各回の視聴機器数は半数以下でも，期間内でドラマと同程度の異なる視聴者へコンテンツがリーチしている実態が分かる．

5.4　実案件の流れ

これらの分析事例は，TimeOn Analytics が得意なことを端的に表しており，新規顧客の初回商談時に紹介している．顧客はパネル方式には馴れている一方でリアル方式に不慣れで，具体的に何ができるかイメージしにくいことが多い．これら分析事例を説明すると，自らの解決したい課題を聞き出すことができる．たとえば，テレビ広告実施企業の場合，ジャンルによる視聴傾向の違いを可視化して見せると，それならば特定番組群の視聴傾向を可視化することで，それを基に広告出稿番組を選定したい，という要望が得られる．コンテンツ制作企業の場合，連続ドラマの視聴傾向をリアルタイム視聴・タイムシフト視聴横断で把握できることを示すことで，制作番組の人気要因などをより正確に把握し，視聴者の関心を上げる施策を検討したい，という要望が得られる．多くの案件では，これらの要望を集計条件として整理し，いくつかの集計値を取得して分析することが主な流れである．

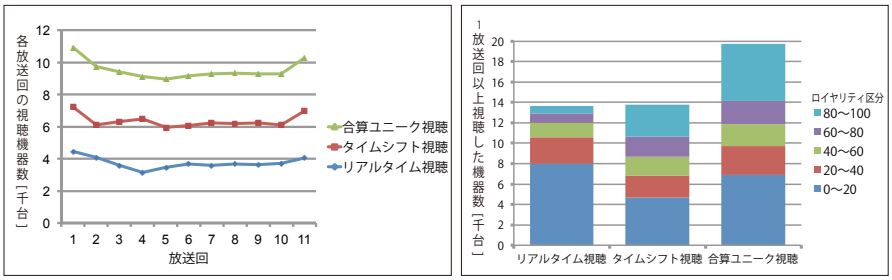

図 10　ドラマ番組の視聴者推移とロイヤリティ分布

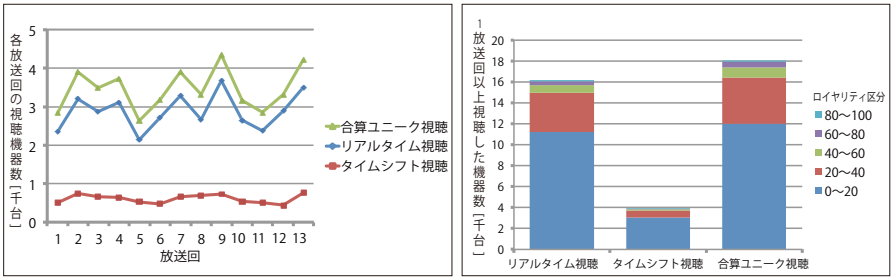

図 11　バラエティ番組の視聴者推移とロイヤリティ分布

6. 番組視聴分析プラクティス２：「どんな人々が視聴しているか」

次に知りたくなるのは，想定していたターゲットにコンテンツが届いているのか，どういった人々に評判が良かったのかという点であり，マーケティングの観点では重要な情報である．本章では厳密な視聴者属性が得にくいシステムにおける属性別分析へのアプローチを述べる．

6.1　属性情報取得の課題と限界

本システムでは大規模データを収集できる反面，ユーザに煩雑な操作を強制できずアンケート回答率が低くなってしまう．またテレビの初期登録作業は世帯の中でも男性が行うことが多いため，女性や高齢層の回答が集まりにくく，アンケート回答者の偏りが大きいといった特徴がある．そのため得られた回答をそのまま利用すると規模が縮小し，属性も偏ってしまうといった課題がある．また世帯の誰が視聴しているのか個人単位での特定はできないといった限界がある．

6.2　機械学習手法による視聴世帯属性推定

低いアンケート回答率と，偏った属性を補うことができれば大規模データの強みを活かしたまま世帯属性による分析が可能となる．そこで回答が得られた機器の視聴傾向から世帯属性を推定するモデルを機械学習手法により構築し，未回答機器の世帯属性を推定する取り組みを行った．

6.2.1　問題設定

表1に示す性年代区分の構成員が世帯に含まれているか否かを視聴傾向から推定する．たとえば30歳男性の一人暮らし世帯には，M1層の構成員が含まれ，それ以外は含まれないと扱う．このような区分はマーケティング分野でよく用いられるものである．通常C層（4 ～ 12歳男女）の区分があり，主体的な行動ができない3歳以下の子供は区分の対象外とされるが，視聴世帯の属性としては0 ～ 3歳も含めて子供の有無を推定できる方が望ましいため，本稿ではBC層（0 ～ 12歳男女）を定義し，C層の代わりに用いる．

6.2.2　正解ラベル

TimeOnではサービスの利用登録時にアンケートを表示し，表2に示す項目の回答を得ている．しかしながら任意回答であるゆえに，たとえばM2層がいないことを示す負例の正解データが直接は得られない．

そこで今回の実験では家族構成についての回答と組み合わせることで，間接的に各層の正例，負例を作成した．たとえばF1層かつ一人暮らし世帯であれば，F1以外の層については構成員がいない負例と扱う．このように直接または間接的に各層の構成員の有無が分かったテレビには正解ラベルを付与する．

6.2.3　視聴特徴量

世帯属性を推定するためには，世帯の視聴傾向が現れやすい特徴量を選択することが重要である．幼児番組などターゲットが明確な番組も多数存在するため，視聴番組のジャンルは1つの候補となるが番組改編期を境に放送内容が大きく変わるため，モデルのロバスト性が問題となる可能性がある．

そこで番組と直接紐付く情報を除いた特徴量として，各曜日を1時間単位に分割し，7曜日，24時間，地上波主要7系列局ごとの3カ月間の平均視聴時間，1,176次元のデータを用いることとする．

なお子供がいる世帯での夏休み期間など，テレビ利用傾向が大きく変わる時期があるため，視聴傾向が日常と異なる春季，夏季，冬季の長期休暇，および祝日を含む週のデータを除外して特徴量を算出した．

6.2.4　モデル構築

上記の正解ラベルと視聴特徴量から，各層の構成員有無を推定するモデルを構築する．なお地上波は地域ごとに放送内容が異なるため，異なる地域のデータを混ぜないよう留意し，モデルも地域ごとに構築した．

表1　推定する性年代区分

性年代区分	定義
M1	男性　20 ～ 34 歳
M2	男性　35 ～ 49 歳
M3	男性　50 歳以上
F1	女性　20 ～ 34 歳
F2	女性　35 ～ 49 歳
F3	女性　50 歳以上
BC	男女　0 ～ 12 歳
T	男女　13 ～ 19 歳

表2　アンケート項目

質問	選択肢	回答
性別	男性／女性／どちらでもない	必須
生年月	年月入力	必須
婚姻状態	未婚／既婚	任意
子供有無	有／無	任意
末子の生年月	年月入力	任意
家族構成	お一人／夫婦お二人／親子二世代（12 歳以下あり）／親子二世代（12 歳以下なし）／三世代／その他	任意

事前実験にてランダムフォレスト，決定木，SVM（Support Vector Machine），ロジスティック回帰の4つの推定手法を比較検証し，最もパフォーマンスの良かったランダムフォレストを採用した．推定精度が低かったT層のパフォーマンスを比較した事前実験の結果を表3に示す．なおモデル学習にはApache Spark上で動作する機械学習ライブラリMLlib[13]を利用し，各手法ともグリッドサーチによりハイパーパラメタを決定した．

6.2.5　実験結果

関東地方における各層の推定精度を表4に示す．精度は10分割交差検証により測定した．BC層は正例のF値が0.8を超えており，子供向け番組の時間帯や子供が好む放送局などの特徴を反映している結果と考えられる．同様にM2層の正例のF値も0.8を超えているが再現率が1.0に近く，多くの世帯でM2層がいると推定していた．これは正解データが偏っておりM2層は正例に対して負例が極端に少ないことが原因と考えられる．

一方で特にT層やF1層の精度が低いが，これらの層はM2層とは逆に正例に対して負例の方が多いことが原因として考えられる．特にT層は正解データが少ない上，テレビの視聴時間が短く，同居する親と比べて視聴傾向に現れにくい，などの可能性がある．精度向上には正解データをサンプリングするなど正例と負例の不均衡を調整する方策が有効と考えられる．

6.3　視聴傾向分析への適用

T層やF1層など課題はあるが，ほかの層では正例のF値が0.7から0.9と実用域といえる．図12に関東地方における世帯のBC層（12歳以下の子供）有無によるリアルタイム視聴傾向比較結果を示す．（1）はBC層あり，（2）はBC層なしの世帯における，ある1日の番組表形式でリアルタイム接触率を示したヒートマップである．両者を比較すると，BC層がいる世帯ではA局，G局が相対的によく視聴されており，逆にB局，D局は相対的に視聴されていないことが分かる．また（1）では夜～深夜の時間帯が全体的に薄く，子供がいる世帯では深夜帯の視聴が少ないという直感に合う結果となった．この種のデータからターゲット層が相対的に多く視聴する時間帯や，逆にテレビの前にいない時間帯の傾向が分かり，マーケティング施策の立案や効果測定に活用できる．一方でT層やF1層には精度に課題が残る．技術的な精度向上アプローチに加え，不足している層のアンケート回答率を高める工夫が必要となってくる．

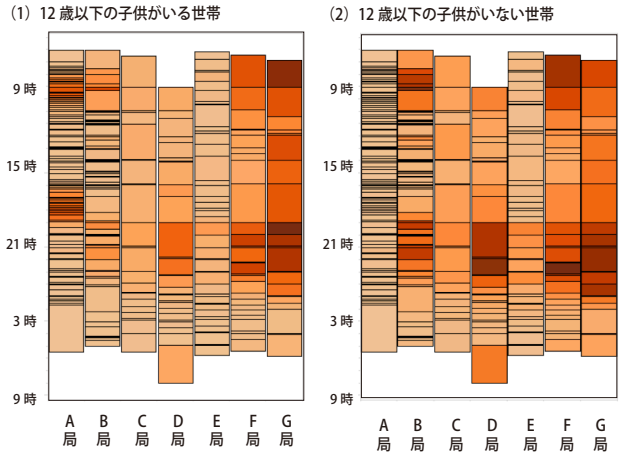

図12　関東地方におけるBC層（12歳以下の子供）の有無による視聴傾向比較

6.4　実案件での活用

顧客の要望の中に，番組やCMにおける各世代の視聴傾向差を把握したいというものがある．前述の通り，アンケート回答のみでは集計対象台数が少なく回答内容の偏りもあるため，従来はテレビの初期設定を行った者の年代・性別であると断って使っていた．近頃の案件では，顧客の要望に応じ，現在の精度を説明した上で，本手法を適用した推定属性に基づく集計・分析を行っている．完全性を求める顧客は最初からパネル方式を選ぶため，リアル方式でこの種の要求が出る場合は，リアル方式の特徴を使いたいが視聴世帯属性もラフでよいので知りたい，という場合であり，現状の精度でも期待に応えることができる．

表3　T層推定精度の4手法比較結果

手法	正例のF値
ランダムフォレスト	0.57
決定木	0.43
SVM	0.41
ロジスティック回帰	0.39

表4　関東地方における世帯属性推定精度

性年代区分	適合率	再現率	正例のF値
M1	0.61	0.79	0.69
M2	0.77	0.99	0.87
M3	0.77	0.68	0.72
F1	0.81	0.46	0.58
F2	0.83	0.76	0.79
F3	0.81	0.68	0.74
BC	0.92	0.74	0.82
T	0.8	0.44	0.57

7. おわりに

本稿では視聴傾向分析に対するビジネスニーズと，それに応えるシステムとして設計されたTimeOn Analyticsの概要，大規模テレビ視聴データの分析事例を述べた．大規模シングルソースデータでのタイムシフト視聴も含めた視聴分析により，番組ジャンルによっては視聴者が番組放送中にも大きく入れ替わっていること，リアルタイム視聴だけを見ても分からなかった番組の高ロイヤリティ層の視聴実態を示した．

また視聴傾向から世帯構成員の性年代区分を推定し，各区分でF値0.57 ～ 0.87の精度を確認した．民生機器での課題であるアンケート回答率の低さを補い，世帯属性による分析のビジネス利用への実用性を示した．本システムによりさまざまな視聴分析が可能となったが，ドメイン知識を持つアナリストが仮説検証を繰り返す必要があり，顕著な特徴を容易に発見できるようにしていくことが今後の展望である．

参考文献

1）（株）電通：2015年日本の広告費，http://www.dentsu.co.jp/news/release/2016/0223-008678.html（2016年7月11日現在）
2）（株）ビデオリサーチ：視聴率データ，http://www.videor.co.jp/rating/wh/index.htm（2016年7月11日現在）
3）石井　晃，川畑泰子：ヒット現象の数理モデル，人工知能学会誌，Vol.30, No.1, pp.97-103 (2015).
4）（株）インテージ：i-SSP（インテージシングルソースパネル），http://www.intage.co.jp/solution/issp（2016年7月11日現在）
5）（株）スイッチ・メディア・ラボ：テレビ視聴分析サービスSMART，https://www.switch-m.com/service（2016年7月11日現在）
6）TVISION INSIGHTS（株）：3Dモーションセンサーによる視聴情報解析アルゴリズム，http://tvisioninsights.co.jp/solution/（2016年7月11日現在）
7）（株）東芝：“おまかせ”と“おすすめ”を追加した「W録」，http://www.itmedia.co.jp/lifestyle/articles/0505/18/news059.html（2016年7月11日現在）
8）（株）ソニー・コンピュータエンタテインメントジャパン：torne，http://www.sie.com/corporate/release/2010/100114b.html（2016年7月11日現在）
9）パナソニック（株）：ミモーラ（MeMORA），http://news.panasonic.com/jp/press/data/jn110214-1/jn110214-1.html（2016年7月11日現在）
10）シャープ（株）：AQUOS視聴ランキング，https://aquos.jp/intro/chranking/（2016年7月11日現在）
11）Samba TV：Interactive TV，https://www.samba.tv（2016年7月11日現在）
12）横山隆治：CMを科学する，宣伝会議（2016）．
13）Apache Spark MLlib，http://spark.apache.org//mllib（2016年7月11日現在）

菊池 匡晃（非会員）masaaki11.kikuchi@toshiba.co.jp
（株）東芝 技術統括部ソリューション開発センター クラウド技術開発部主務．2006年に大阪大学大学院工学研究科知能・機能創成工学専攻修士課程修了．同年（株）東芝入社．テキストマイニングの研究，テレビファームウェアの開発を経て，テレビ視聴データ分析基盤の開発および視聴データの研究に従事．

坪井 創吾（正会員）sougo.tsuboi@toshiba.co.jp
東芝映像ソリューション（株）VSクラウド＆サービス推進室室長附．2000年東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了．同年（株）東芝入社．企業内コラボレーション研究およびシステム開発を経て，2012年より現職．視聴データ分析サービス企画統括．人工知能学会員．

中田 康太（正会員）kouta.nakata@toshiba.co.jp
（株）東芝 技術統括部 研究開発センター知識メディアラボラトリー研究主務．2006年に東京大学理学系研究科地球惑星科学専攻から（株）東芝に入社．データマイニング，テキストマイニング，機械学習技術の研究開発に従事．博士（理学）．人工知能学会会員．

採録決定：2016年7月11日
編集担当：住田一男（(一社）人工知能学会）