1K2-4

大規模テレビ視聴データクラスタリングによる視聴パターンの分析

Analysis of Television Viewing Pattern with Clustering Large-Scale Log Data

水岡 良彰[*1] Yoshiaki Mizuoka　陶 亜玲[*1] Yaling Tao　中田 康太[*1] Kouta Nakata　折原 良平[*1] Ryohei Orihara

[*1] 株式会社東芝 研究開発センター 知識メディアラボラトリー
Knowledge Media Laboratory, Corporate Research and Development Center, Toshiba Corporation

This paper presents an analysis on television viewing patterns using large scale television log data. Television log data, collected from televisions with users' permission, contains information of when, how, and which channel are users watched in seconds. In this paper, we extract viewing patterns from television log data with a clustering technique. We focus on when users watch live and recorded contents on television and discovering a user group that regularly watches contents by utilizing time-shift functions. New viewing patterns lead to discovery of audience groups which are valuable to marketing or targeted advertising.

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配信できるテレビメディアは，インターネット広告市場などが伸びている現在においても巨大な広告市場であり，2016 年の総広告費は約 2 兆円（広告全体の約 30%）を占める[電通 2016].

このテレビメディアの価値を測る手段として、(株)ビデオリサーチの提供する視聴率が存在するが，コンテンツや視聴者を掘り下げて調査・分析するには，視聴率だけでは不十分な場合がある．例えば，一般にゴールデンタイムと呼ばれる夜の時間帯にテレビを視聴するユーザが一定数存在することは知られているが，該当するユーザの規模や，他の時間帯のテレビ視聴行動との組み合わせなどを詳細に掘り下げた視聴傾向の分析は難しい．アンケート調査などによる視聴傾向調査は行われているものの，回答の粒度が粗く，また回答が不正確な場合もある．

近年，ネットワーク接続型テレビが普及しており，[菊池 2016]では，ユーザから利用許諾を得て取得した視聴データを使って，番組や CM の視聴分析を行っている．この報告では，ドラマを例として視聴した話数の組み合わせの分布や，子供の有無で分けた視聴傾向などの分析が行われている．このように，事前に分析対象とするコンテンツや視聴者集合，すなわち分析の切り口が明確であれば詳細な分析が可能であるが，そもそも知見が少ない場合は，切り口を事前に決めること自体が難しい．本研究では、分析の切り口を決めることなく視聴分析を行う．さらに分析の切り口をデータから発見する．

本稿では、時間帯ごとの視聴行動の組み合わせを視聴パターンと呼ぶ．特徴的な視聴パターンの抽出や，その視聴パターンを持つ視聴者集合の傾向分析ができれば，視聴者の特性を生かした番組や CM を作成しやすくなり，テレビメディアの価値を向上できるとともに，視聴者も魅力的なコンテンツを視聴できるようになるといえる．さらに，視聴者の視聴パターンを考慮することで番組推薦の精度向上にも繋がる．

ところで，視聴方法にはライブ視聴と再生視聴があるが，視聴方法に違いがあらわれる視聴者集合の切り口を事前に決めることは難しい．そこで本研究では，テレビ視聴した時間帯ごとの，テレビの視聴方法（ライブ視聴・再生視聴）に注目し，各曜日の習慣性のある特徴的な視聴パターン集合の抽出を試みる．以下，2 章で視聴パターン集合の抽出に利用するテレビ視聴データに関して説明し，3 章で視聴パターン集合の抽出を行い，特徴的な視聴パターン集合の分析や，視聴パターン集合に対する名称の自動付与について述べる．4 章でまとめを述べる．

連絡先：水岡良彰，（株）東芝 研究開発センター，
〒212-8582 川崎市幸区小向東芝町 1，
yoshiaki.mizuoka@toshiba.co.jp

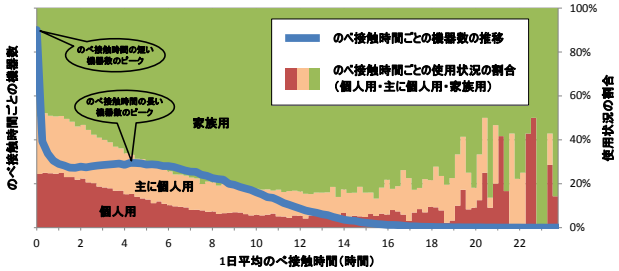

図1 のべ接触時間ごとの機器数の推移と使用状況の割合

2. テレビ視聴データ

一部の東芝製テレビでは，放送中の番組を視聴（ライブ視聴）するほか，本体に接続されたハードディスクに事前に録画しておいた番組を視聴（再生視聴）することができる．本研究では，このような視聴方法ごとの視聴時間に注目し，東芝製テレビのユーザから利用許諾を得て取得したデータを利用して，視聴パターンの抽出を行う．

本章では，視聴パターンの抽出に利用するテレビ視聴データについて説明する．

2.1 視聴パターン抽出対象期間

視聴パターンの抽出対象とする期間は 2016 年 11 月とした．これは，10 月を終えて番組改編が落ち着いている期間であり，また 12 月の冬休みのような長期休暇が無い期間であるため，習慣的な視聴行動の抽出に適切と考えて決定した．

2.2 視聴パターン抽出対象機器

分析対象期間中にユーザがテレビに接触している時間，すなわちテレビがライブ視聴あるいは再生視聴しているのべ時間（のべ接触時間）を，機器ごとに集計した．図1は，1 日平均のべ接触時間を 15 分間ごと区切ったときの，該当するテレビ機器数の推移と，該当するテレビ機器内のアンケートベースのテレビ使用状況（個人用，主に個人用，家族用）の割合を示している．

まずのべ接触時間ごとの機器数のヒストグラムの形状を見ると，1 日平均 2 時間程度での機器数が谷となっており，その左右に山がある．すなわちピークが 2 つ存在している．

ここでテレビ使用状況の推移を見ると，のべ接触時間が短いテレビは個人用テレビが多く，のべ接触時間が増えてくると家族用のテレビが多くなる．個人用テレビの利用がユーザの気まぐれに左右されやすいのに対し，家族の間では各人の行動に暗黙の枠組みがあり，テレビの視聴行動もその枠組みに含まれるであろう．その結果，習慣性が無く気が向いた時だけテレビ視聴するようなテレビは個人用テレビが多く，家族用のテレビは習慣的なテレビ視聴が行われやすいと考えられる．この仮説の通り，習慣性の有無と，テレビの使用状況に相関があるとすれば，視聴時間が長い方のピーク側には，習慣性のある視聴者層が多く含まれていると期待できる．

本研究の目的は，テレビの視聴方法に注目して習慣性のある特徴的な視聴パターンを抽出することである．図 1 のヒストグラムから，習慣性のある視聴パターンは一定以上の視聴がある機器にあらわれると期待できるので，ライブ視聴と再生視聴のそれぞれが一定時間以上の機器から抽出を行う．

本研究では，図 1 を参考に，ライブ視聴と再生視聴がそれぞれ 1 日平均 1 時間以上（このとき，のべ接触時間は 2 時間以上となる）の機器に絞って分析を行うこととした．条件に当てはまる機器数は，116,370 台であった．

2.3 視聴パターン抽出用特徴量

一般に人々の行動は曜日に基づいており，またテレビで放送される内容も曜日によって決まっている場合が多い．よってテレビ視聴の習慣性は曜日単位で現れやすいと考えられる．しかしテレビ視聴は，たまたま見る・見ないといった状況が起こり得るため，実データをそのまま扱うにはノイズが大きい．

そこで本研究では，各機器について，曜日（日曜～土曜）ごと，時間帯は 1 時間ごと，視聴方法（ライブ視聴，再生視聴）ごとに，視聴時間の割合を求める．これにより，数値は月内の平均となるため，ノイズの影響が緩和される．なお本研究では，祝日を除いて特徴量算出を行う．

例えば 2016 年 11 月の場合，水曜日は 5 日間あるが，そのうちの 1 日間は祝日（11 月 23 日，勤労感謝の日）であり，平日は 4 日間である．よって，水曜日の 05:00～06:00 のライブ視聴については，各週の水曜日の 05:00～06:00 にライブ視聴を行ったのべ時間（秒）を，18,000（3,600 秒×4 日間）で割った値が特徴量となる．各条件の特徴量を同様に計算することで，視聴パターンを 336（7 曜日×24 時間帯×2 視聴方法）の値で表現できる．

3. 視聴パターン抽出

視聴パターンの抽出は，2.3 節で示した各機器の特徴量を用いて，k-means 法でクラスタリングを行い，各クラスタの視聴パターンを可視化した上で，特徴的な視聴パターンを目視でピックアップする．ピックアップした視聴パターンに対して，詳細な検証を行う．また，視聴パターンへの自動名称付与を行う．

3.1 k-means 法

k-means 法は，以下の手順に従い，n個のデータをk個のクラスタに割り振るアルゴリズムである．

1. 各データ$\boldsymbol{x}_i\,(i=1,\cdots,\mathrm{n})$をランダムにクラスタに割り振る
2. 各クラスタの中心$\boldsymbol{V}_j\,(j=1,\cdots,\mathrm{k})$を計算する
3. 各データ$\boldsymbol{x}_i$を最も近い$\boldsymbol{V}_j$のクラスタに割り振る
4. 手順 2 と 3 を収束するまで繰り返す

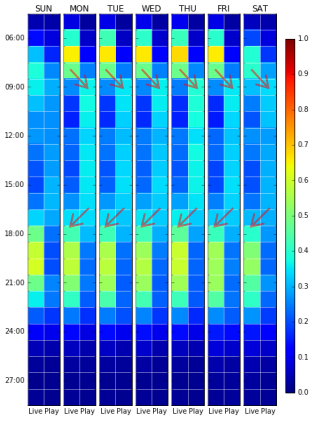

図2 朝夜はライブ視聴
昼間は再生視聴の
視聴パターン

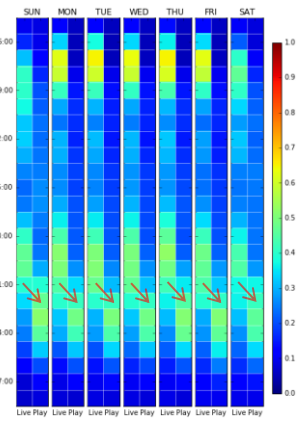

図3 ライブ視聴から
再生視聴への
視聴パターン

表1 図2クラスタの
デモグラフィック属性リフト値

	男性	女性
未婚	0.500	0.720
既婚	1.121	1.775

上記の手順で最終的に割り振られた結果がクラスタとなる．

なお本実験では，i番目の機器における 2.3 節で説明した特徴量を 1 列に並べた 336 次元の特徴ベクトルをデータ$\boldsymbol{x}_i$とし，距離計算にユークリッド距離を用いる．

3.2 視聴パターンの抽出と可視化

2.1 節の期間かつ 2.2 節の機器を対象に，2.3 節の特徴量を用いて，3.1 節の k-means 法（k=40）を適用し，クラスタリングを行った．クラスタリング結果を可視化し，ライブ視聴と再生視聴にまたがっている視聴パターンを確認したところ，図 2 および図 3 を抽出できた．

図 2 および図 3 は，各クラスタに含まれる機器の平均値をヒートマップで可視化したものである．縦軸が 1 時間単位の時間帯，横軸が曜日と視聴方法の組み合わせとなっており，各条件での視聴時間の割合をヒートマップの各マスの色の違いで示している．各図とも，左から順に日曜から土曜に対応する形で，曜日別に 7 列に分けている．さらに各曜日の列ごとに，左がライブ視聴で右が録画視聴に対応する形で，視聴方法別に 2 列としている．なお，図に重畳してある赤色の矢印は，視聴方法の移り変わりを示している．

図 2 は，クラスタサイズ（機器数）は 2,223 台であり，朝と夜はライブ視聴を行っているが昼間は再生視聴を行っているところが特徴的である．図 3 は，クラスタサイズ（機器数）は 1,879 台であり，朝にライブ視聴を行っているところと，22 時頃にライブ視聴から再生視聴に切り替わっているところが特徴的である．

3.3 抽出した視聴パターンの検証

3.2 節で抽出した各視聴パターンについて検証を行う．

(1) 朝と夜はライブ視聴，昼間は再生視聴（図2）

図 2 の視聴パターンは，昼間に再生視聴をしている一方で，平日の朝と夜は規則正しくライブ視聴を行っている．平日に出勤する家族を見送った上で，昼間に再生視聴している視聴者が，このような視聴パターンになっていると推測できる．

本クラスタに属す機器について，アンケートで収集した男性・女性および未婚・既婚のデモグラフィック属性の分布を調べたところ，リフト値は表 1 となった．リフト値とは，あるクラスタにおけるある属性の割合を，全体におけるある属性の割合で割ったもの

表2　図3クラスタの
機器属性リフト値

	有効	無効
タイムシフト 再生視聴機能	1.448	0.692

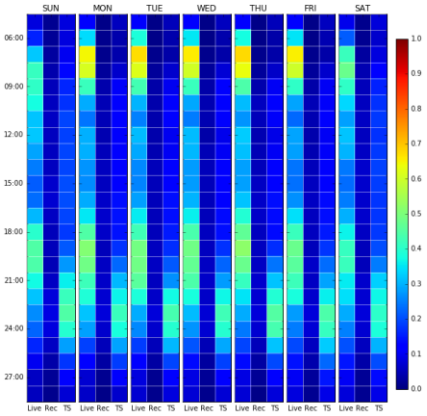

図4　図3クラスタのうち
タイムシフト再生視聴機能が有効な機器の視聴パターン

表3　平日と休日の対応例

曜日		平日・休日
日曜	→	休日
月曜	→	平日
火曜	→	平日
水曜	→	平日
木曜	→	平日
金曜	→	平日
土曜	→	休日

表4　時間帯の名称例

時間帯		時間帯名
05:00–09:00	→	朝
09:00–12:00	→	午前
12:00–13:00	→	ランチ
13:00–19:00	→	午後
19:00–22:00	→	ゴールデン
22:00–29:00	→	夜

表5　視聴方法の名称例

視聴方法		視聴方法名
ライブ視聴	→	ライブ
再生視聴	→	再生

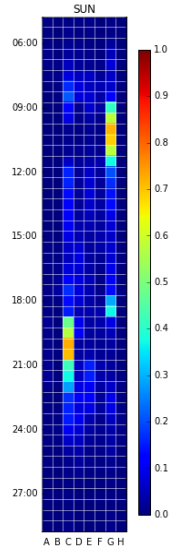

図5　視聴チャンネルに
注目した視聴パターン例

であり，母数の偏りに関わらず，相対的に多いか少ないか判別するのに用いる．

これを見ると，この視聴パターンには既婚女性が多いことが分かる．前述のような行動をとる利用者の例として専業主婦があげられることを考えると，前述の推測を支持するデータと言える．

(2)　22 時頃にライブ視聴から再生視聴へ（図3）

図 3 の視聴パターンは，ゴールデンタイム（19:00～22:00）にライブ視聴をしており，それが終わる頃から再生視聴に移行している．自分の見たいゴールデンタイムの番組を視聴し終わってからもテレビを楽しみたい視聴者が，このような視聴パターンになっていると考えられる．

再生視聴についてさらに考察を進めるために，再生視聴の方法に注目する．一部の東芝製テレビには，過去一定期間の全番組を録画しておくことで，明示的な録画予約をせずとも過去の番組を遡って視聴できる機能が搭載されている．この機能を用いた再生視聴を本稿ではタイムシフト再生視聴と呼び，ユーザが明示的に録画を行った番組を再生視聴する方法（録画再生視聴）と区別する．

本クラスタに属す機器について，タイムシフト再生視聴機能が有効な機器の分布を調べたところ，リフト値は表 2 となった．さらに，タイムシフト再生視聴機能を有効にしている 1,108 台を対象に，録画再生視聴とタイムシフト再生視聴を分けて視聴パターンを確認すると，図 4 となった．以上の結果から，本クラスタの視聴パターンは，タイムシフト再生視聴機能が有効な機器が多く，またそれらの機器の再生視聴のほとんどはタイムシフト再生視聴によるものと分かる．本クラスタの視聴者は，タイムシフト再生視聴機能をゴールデンタイムが終わった後もテレビを楽しむ際に活用されていることが分かる．タイムシフト再生視聴の特性から，本クラスタのユーザには，目的のコンテンツが存在するかに関わらず，ゴールデンタイム後もテレビを楽しみたいと思うユーザが含まれていると考えられる．本分析によって，タイムシフト再生視聴機能の活用事例を発見することができた．

3.4　抽出した視聴パターンへの名称付与

視聴パターンの抽出では，クラスタリングによって自動的に多くのクラスタを生成するため，目視による特徴的な視聴パターンの抽出はコストがかかる．また複雑な視聴パターンの場合，解釈が難しいという問題もある．そこで視聴パターンの数値から，各視聴パターンに名称を自動付与することを考える．

そこで，これまでの図のような視聴パターンの各マスのうち，閾値以上だったマスに基づいて名称を生成する手法を提案する．例えば，休日と平日の対応例を表 3，時間帯の名称を表 4，視聴方法の名称を表5と決めておく．図3の視聴パターンに対して閾値を 0.45 とすると，「平日と土曜の朝はライブ，ゴールデンはライブ，夜は再生」という名称が生成できる．この方法は一見，単純な方法であり情報量も落ちてしまうが，説明上分かりやすく直感的であることから，クラスタの名称を一覧しておき，詳細な可視化結果の観察前のフィルタリングに利用することができる．

この方法は，視聴方法以外に注目した視聴パターンに対しても適用できる．図5は，日曜日における，テレビ視聴した 30 分単位の時間帯ごとに，視聴放送局 A～G に注目して抽出した視聴パターンの一例を，これまでと同様の方法で可視化したものである．図 5 の視聴パターンに対して同様に名称を生成すると，「朝は放送局 G，ゴールデンは放送局 C」という名称が生成できる．

視聴パターンに自動付与した名称は，分析の切り口を自動抽出した結果と捉えることもできる．生成された名称にしたがった視聴パターンで視聴者層を定義してもよいし，該当するクラスタを分析して特徴的だったデモグラフィック属性で視聴者層を定義してもよい．いずれにしても，特徴的な視聴者層を分析の切り口として決めることで，様々な分析が行えるようになる．

4.　おわりに

本研究では，実データを用いて，テレビ視聴データから視聴パターンを自動抽出できることを示した．また抽出した代表的なパターンについて考察し，抽出される視聴パターンが妥当であることや，新しい知見を発見できることを示した．さらに抽出した視聴パターンを分析の切り口とする方法について説明した．

本研究では，曜日と時間帯と視聴方法で分けたのべ視聴時間を特徴量に利用したが，特徴量の切り口の変更や，他の情報と合わせることなどにより，別の新たな未知の視聴パターンや視聴者に対する知見の抽出が期待できる．また抽出できた新たな視聴パターンは，特徴的な視聴者セグメントとして特徴を分析

することで，テレビ視聴データを利用したより詳細な分析が可能になるだろう．

今後は，目視で行っている注目すべき視聴パターン特定の自動化や，自動抽出で多くの分析の切り口が抽出された場合の絞り込み，分析によって発見される知見の具体的な活用方法などについて検討を進めていく．

参考文献

[電通 2017] 株式会社 電通: 2016 年 日本の広告費，http://www.dentsu.co.jp/news/release/2017/0223-009179.html．

[菊池 2016] 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，情報処理学会デジタルプラクティス Vol.7 No.4，情報処理学会，2016．