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

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

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

[*1] 株式会社東芝　研究開発本部　研究開発センター　アナリティクスAIラボラトリー
Analytics AI Laboratory, Corporate Research and Development Center, Research & Development Division, Toshiba Corporation

This paper presents an analysis of television viewing patterns using large-scale television log data. Television log data, collected from televisions upon users' permission, contains information of when, how, and which channels are watched in seconds. In this paper, we extract user groups based on long-term transitions of viewing patterns from television log data with a clustering technique, and we analyze the user groups from viewpoints of a fixed habit and change in a life style. As a result, new long-term transitions of viewing patterns have been found. They are expected to lead to the 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，日本における 2017 年の総広告費は約 2 兆円（広告費全体の約 30%）を占める[電通 2018].

このテレビメディアの魅力を向上させるため，近年普及したネットワーク接続型テレビの視聴データを使用することで，従来のアンケート調査では難しかった長期の視聴分析がしやすくなった．これにより，ある番組を見始めた，あるいは見なくなった原因の分析において，ライフステージの変化などのマクロな観点を加えることで，番組内容によるものか，その時間帯の生活が変化したためかを区別して考察することができる．また，この種の長期的な視聴習慣の種類や割合を把握することで，番組内容の企画や CM の妥当性の検討にもつながる．ただし，長期間の視聴データ分析は膨大なデータを使用するため容易ではなく，長期的な視聴習慣をとらえる手法が必要である．

[Xu 2012]では，ウェブベースのテレビサービスにおける番組推薦を目的に，10 週間の視聴ログを分析しているが，1 クール以上の視聴習慣までは対象にしていない．[菊池 2016]では，ネットワーク接続型テレビの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，特定の番組の視聴の有無や家族構成など，分析対象を事前に決めて分析を行っており，さらに[水岡 2017][陶 2017]では，事前知識が少なく分析対象を決められない場合であっても，抽出対象期間から特徴的な視聴習慣を抽出し，抽出した視聴習慣を持つユーザに絞った分析を可能としている．しかしこの方法では，抽出対象期間が限られるため長期的な視聴習慣の分析は難しい．

本研究では特定の期間における視聴習慣を，時間帯ごとの視聴時間の組み合わせによる視聴パターンで表現し，複数の期間で視聴パターンがどう推移するか考えることで，長期的な視聴習慣を分析する．まず 1 ヶ月間の視聴データから視聴パターンを抽出し，次に抽出した視聴パターンの長期的な推移について，1 年 1 ヶ月間の長期間の視聴データを用いて分析を行う．

連絡先：水岡良彰，（株）東芝　研究開発本部　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

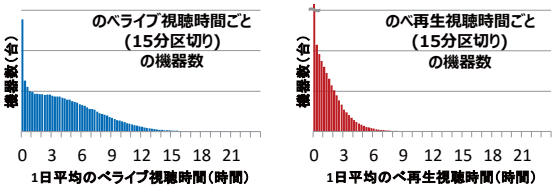


図 1　ライブ視聴または再生視聴の
のべ視聴時間ごとの機器数

以下，2 章では本研究で利用するテレビ視聴データについて説明する．3 章で注目する視聴パターンを抽出し，4 章にて注目する視聴パターンの長期的な推移を抽出し，5 章で抽出した長期的な推移について検証する．6 章でまとめる．

2. テレビ視聴データ

一部の東芝製テレビはネットワーク接続に対応し，視聴データを用いた番組推薦や自動録画等の機能をもつクラウドサービスが利用可能である．本研究では，上記サービスを提供するにあたって，テレビ機器のユーザから視聴データの利用許諾を得て取得したデータを利用し，分析を行う．

また，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．このうち，ライブ視聴はテレビ番組の放送時間に合わせて視聴することから，時間的な視聴習慣が表れやすいと考えられる．また，ライブ視聴と再生視聴それぞれの視聴時間ごとの機器数を図 1 に示す．のべ視聴時間の長い機器が多いと，分析する機器数が多いことが期待できる．図 1 を見ると，再生視聴に比べてライブ視聴を長時間視聴するユーザの方が多いことが分かる．以上から，本研究ではライブ視聴を分析対象とする．

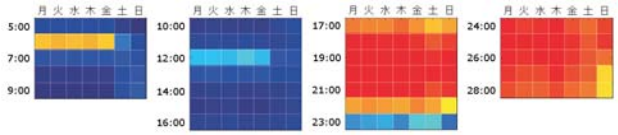
本稿において，視聴パターンの抽出，および，その長期的な推移の抽出で用いるデータは次に述べる通りとする．

(1) 視聴パターンの抽出に用いるデータ

分析対象とする期間は，2016 年 11 月の 1 ヶ月間の期間とした．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．

表 1　視聴パターン抽出時の時間帯区分

朝帯	5~10時
昼帯	10~17時
ゴールデン帯	17~24時
深夜帯	24~29時



(a) 朝帯 5～10 時　(b) 昼帯 10~17 時　(c) ゴールデン帯 17~24 時　(d) 深夜帯 24~29 時

図 2　時間帯区分別の視聴パターンクラスタリング結果（抜粋）

分析対象とする機器は，この期間中にテレビ視聴を行ったデータのある全機器とした．このとき，対象となる機器は 327,721 台であった．

(2)　視聴パターンの推移の抽出に用いるデータ

分析対象とする期間は，1 週間の開始を月曜日として，2016 年 10 月 3 日の週から 2017 年 10 月 23 日の週までの 52 週間，すなわち 1 年 1 ヶ月間とした．

分析対象とする機器については，本研究では習慣的な視聴パターンの分析が目的であるため，この期間中に毎週ライブ視聴が 1 回以上ある機器を対象とした．すなわち，2016 年 10 月 3 日より後に購入された機器や，1 週間以上ライブ視聴しない期間があった機器は対象外とした．この条件に当てはまる対象機器は，187,891 台であった．

3.　視聴パターン抽出

本研究では，各機器の視聴パターンをクラスタリングして注目する視聴パターンを抽出する[水岡 2017]の方法を用いる．ただし，抽出に用いた具体的な特徴量については本研究に合わせて変更する．本章では，視聴パターンの抽出に用いる特徴量について説明し，抽出結果を可視化して示す．

3.1　抽出に用いる特徴量

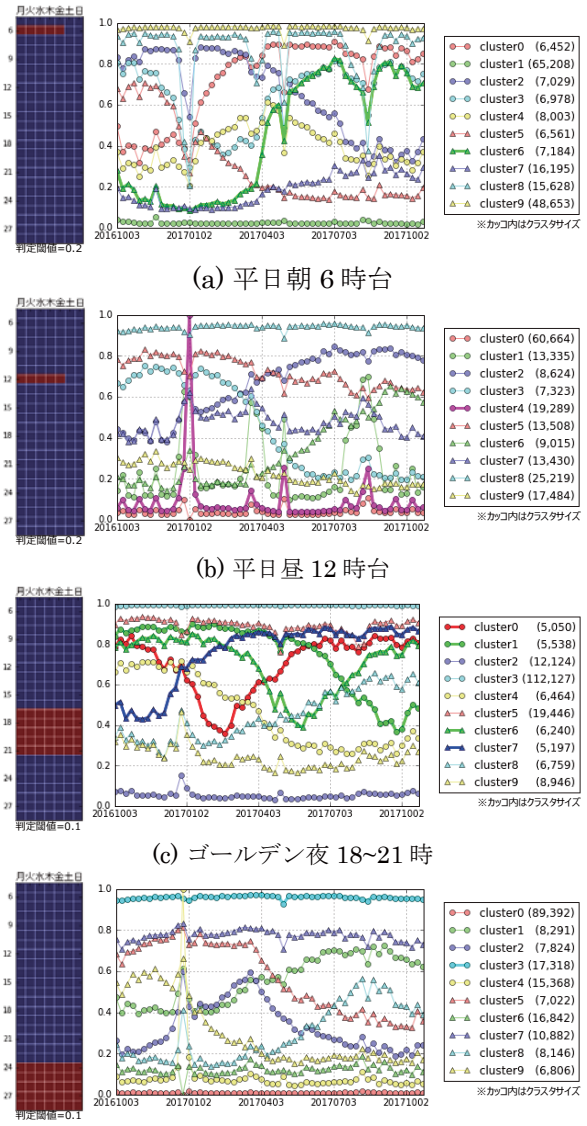
人々の行動は曜日に基づいていることが多く，またテレビで放送される内容も曜日によって決まっている場合が多い．よってテレビ視聴の習慣性は曜日単位で現れやすいと考えられる．しかしテレビ視聴は，たまたま見る・見ないといった状況が起こり得るため，実データをそのまま扱うにはノイズが大きい．

そこで特徴量として，各機器について，曜日（月曜～日曜）と時間帯（1 時間単位）ごとに，ライブ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．この数値を連結したベクトルをクラスタリングの特徴量として用いる．

なお本研究では，祝日を除いて特徴量算出を行う．また，細かく特徴を抽出できるよう，4 つの時間帯区分を定義し，特徴量をそれぞれの時間帯区分に分割してクラスタリングを適用する．各時間帯区分は表 1 の通りである．

3.2　抽出結果

2 章の(1)の機器を対象に，3.1 節の特徴量を用いて，k-means 法(k=40)を適用し，クラスタリングを行った．視聴パターンのクラスタリング結果（k-means で得られた各クラスタの中心点）



(a) 平日朝 6 時台

(b) 平日昼 12 時台

(c) ゴールデン夜 18~21 時

(d) 深夜 24 時以降

図 3　視聴パターンテンプレート（各図左）と視聴パターン判定の推移のクラスタリング結果（各図右）

から，視聴する時間帯が特徴的な各時間帯区分から 1 クラスタずつ抜粋して可視化し，図 2 に示す．

各クラスタの視聴パターンはヒートマップで示されている．各時間帯のクラスタは，横軸が曜日で縦軸が 1 時間区切りの時間帯を表しており，マスの色が青→黄→赤になるに従って対応する曜日の時間帯にテレビを視聴している時間が多いことを示している．

本研究では得られたクラスタのうち，図 2 で示したクラスタを各時間帯区分の代表例として注目し，抽出する視聴パターンとして選択する．

4.　視聴パターンの長期的な推移の抽出

3 章のクラスタリングによって，抽出対象期間に注目した視聴パターンを持つ機器を特定できるようになったが，抽出対象期間外については不明である．本章では 3 章で注目した視聴パターンをテンプレートとして定義し，抽出対象期間外であっても各機器が視聴パターンを持つか否か判定できるようにした上で，

判定結果をクラスタリングすることで視聴パターンの長期的な推移を抽出する.

4.1 視聴パターンのテンプレート化

1 週間を 1 時間単位の時間帯に分割し, テンプレート化したい視聴パターンにおいて視聴している時間帯を“1”、そうではない時間帯を“0”としたベクトルでテンプレートを表現する. 次元数や各次元が意味する時間帯は, 視聴パターンに等しい.

3.2 節で注目した各視聴パターンをテンプレート化して可視化すると, 図 3 の各図の左のヒートマップとなる. 本ヒートマップは, 対応する各時間帯について, “1”を赤, “0”を青で表している. 例えば(a)の平日朝 6 時台に視聴するテンプレートであれば, 月火水木金の朝 6 時台を赤, それ以外を青で示している. 本テンプレートは, “1”となっている曜日の時間帯に視聴しており, “0”となっている曜日の時間帯は視聴していてもいなくても構わないことを意味している. テレビ視聴においては, 習慣的に見る訳ではないが, たまたまテレビを点ける場合もあり, 習慣の判定としてはノイズとなり得るが, このようにすることでノイズ除去の役割を果たすことができる.

最後に, 視聴するとした時間帯(“1”とした時間帯)を, どの程度以上視聴したときにテンプレートが表す視聴パターンを持つと判定するかの割合を, 判定時のパラメータとして設定する.

4.2 テンプレートによる視聴パターンの判定

ある 1 週間のテレビ視聴が, テンプレートの表す視聴パターンに合致するか判定する.

まず判定したい 1 週間を視聴パターンの形式で表現する. すなわち, 1 週間を 1 時間区切りとして各時間帯の視聴割合を特徴量としたベクトルを生成する. ここで各特徴量は視聴割合なので 0 から 1 の実数を取り得る. 次に, テンプレートのベクトルの“1”を, ベクトルに含まれる“1”の個数で割った上で, 前述の 1 週間の視聴パターンのベクトルとの内積を計算する. この結果が閾値以上の場合に, テンプレートが表す視聴パターンを持つと判定する.

なお本稿では, 視聴している“1”と視聴しているか問わない“0”の 2 値を用いることとしたが, 明示的に視聴していない場合を考慮したい場合も考えられる. この場合は, 視聴していない時間帯を“-1”として計算するか, あるいは視聴していない時間帯を“1”とするパターンを別途作成し, このパターンに合致すると判定された場合は判定から外す操作を行えばよい.

4.3 視聴パターン判定のクラスタリング

4.2 節の方法によって, 注目する視聴パターンを持つか否かの判定を, 期間を問わずに, すなわち視聴データが存在すれば長期に渡って可能となった. しかし分析対象とする機器が多い場合, 個別に判定結果を観察しても大きな傾向を把握することは困難である.

そこで本研究では, 4.2 節の方法による判定結果をクラスタリングすることで特徴的な推移を抽出する. 本稿での視聴パターン判定の推移の抽出対象期間は 52 週間であるため, 1 機器に付与されている判定数は 52 個となる. この 52 個の 0 または 1 の判定結果を連結し, 1 次元のベクトルとしてクラスタリングで用いる特徴量とする.

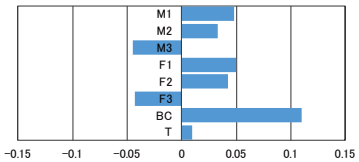
4.4 抽出結果

2 章(2)の機器を対象に, 4.3 節の方法を用いて, クラスタリング手法として k-means(k=10)を適用し, クラスタリングを行った. 視聴パターンの推移のクラスタリング結果(k-means で得られた各クラスタの中心点)を可視化した結果を図 3 の各図の右側に示す. 各クラスタの視聴パターンの推移が折れ線グラフで示されており, 横軸は対象期間内の週単位の時期, 縦軸はクラスタに含まれるテンプレートが表す視聴パターンを持つ機器の割合を表している.

以上のようにクラスタリングおよび可視化によって, 代表的な視聴パターン判定の推移を観察できるようになった. この可視化結果を観察することで, 視聴パターン判定の特徴的な推移を抽出することができる.

性年齢区分	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

(a) 世帯属性の区分



(b) 世帯属性の分布

図 4　平日朝 6 時台に 4 月から視聴する層の世帯属性の分布

表 2　平日朝 6 時台に 4 月から視聴する層の番組視聴割合リフト値上位の番組（上位 5%から抜粋）

番組名	開始時刻	チャンネル
すくすく子育て「きょうだいの子育て」	2016/12/02 13:05	ＮＨＫＥテレ
プチプチ・アニメ　ポンユタウン「ポンユタウン」	2016/12/02 08:50	ＮＨＫＥテレ
まいにちスクスク　子育てママの体のケア（１）「腰痛」	2016/11/28 10:55	ＮＨＫＥテレ
いないいないばあっ！	2016/12/02 16:15	ＮＨＫＥテレ
おかあさんといっしょ	2016/12/02 16:30	ＮＨＫＥテレ

5. 抽出した視聴パターン判定の推移の検証

本章では, 4 章の結果に基づいて, 視聴パターン判定の特徴的な推移を持つ機器のクラスタを検証する. 以下, 図 3 の各視聴パターン判定の推移から, クラスタをピックアップして検証する.

5.1 検証に用いるデータ

各クラスタの分析を行うために, 複数のデータを用いる. まず, テレビ機器利用者の世帯属性データを用いる. ここでの世帯属性は, 図 4 の(a)に示す 8 つに区分される性年齢区分であり, 時間帯ごとの平均視聴時間からランダムフォレストで推定する[菊池 2016]の方法によるものである. また属性の分布はリフト値の対数(底は 10)で示す. リフト値は, クラスタ内の対象区分の比率を, クラスタリング対象とした全機器内の対象区分の比率で割った値であり, 対数を取ることで数値間の比較をしやすくしている. 本稿では 2017 年 9 月末の推定結果を用いる.

また, テレビ機器のエリア設定データを用いる. ここで用いるエリアはテレビの視聴設定から判別できるデータであり, 世帯属性と同様にリフト値の対数で検証する. 本稿では 2017 年 9 月末時点の判別結果を用いる.

さらに, 番組視聴割合データを用いる. これは, ある期間に放送された各番組の, 放送時間に対する視聴時間の割合のランキングである. ただし単純に人気の番組が上位にならないよう, 世帯属性と同様にリフト値のランキングを用いる. 本稿では, 検証のしやすさのため関東圏の機器かつ NHK および地上波キー局のチャンネルに限定し, 2016 年 11 月 28 日からの 1 週間に放送された 1,973 番組をランキング対象とする.

表 3　各クラスタ検証結果概要

テンプレート	クラスタ	サイズ	クラスタの概要	世帯属性	エリア設定	番組視聴割合ランキング
a	6	7,184	平日朝6時台に4月から視聴している層	BC層が多い	特徴なし	Eテレが上位を占める. 子供向けあるいは親向けの番組がメイン.
b	4	19,289	平日昼12時に正月だけ視聴している層	偏りなし	東北や関東が比較的多く 近畿や九州は比較的少なめ	Eテレが上位を占める. 語学番組や短歌・手話などの教養番組がメイン.
c	7/0/6/1	5,197/5,050/6,240/5,538	17時から22時に4半期単位で視聴していない層	M1層とF1層とBC層が多い M3層とF3層とT層が少ない	特徴なし	未明や早朝の番組と, 土日の日中のEテレの教養番組が混在.
d	3	17,318	深夜にいつも視聴している層	特にT層が多く M2層とF2層も多い M1層とF1層とBC層が少ない	近畿や四国比較的多く 九州の一部も比較的多い 北海道や東北や中部は少ない	複数のチャンネルが視聴されており, ジャンルもドラマやバラエティなど 幅広く視聴されている.

5.2 検証結果

代表例として, 本節では図 3 の(a)の cluster6 に注目する. このクラスタは, 平日朝 6 時台に 4 月から視聴するようになった, 7,184 機器からなるクラスタである. 4 月は進学や就職といった生活スタイルが変化する時期であるため, 朝早く起きて活動するよう変化した層を表していると考えられる.

世帯属性の分布は図 4 の(b)の通り, 特に BC 層が多いことが分かる. このことから, 幼稚園や小学校などに通うようになった子供本人や, 朝に子供を送り出す親が早く起きて活動するようになったと推測できる.

番組視聴割合のリフト値でのランキング上位の番組を確認すると, 上位 5%の 97 番組全てが NHK の E テレであった. 具体的な番組を見てみると, 表 2 に示す番組が上位となっており, 子供や子供を持つ親向けと思われる番組が目立っていた. この結果からも子供がいる世帯であると推測でき, 前述の推測とも矛盾しない.

一方で, テレビ機器のエリア設定の分布については, 大きな傾向は見られなかった. エリアに関わらず子供がいる家庭が早く起きて活動する生活スタイルは考えられるため, これまでの推測に矛盾していないといえる.

以上の検証から, 注目する視聴パターン判定の推移を持つ機器集合に対して, 別の観点のデータから分析し矛盾がないことを確認した. この例の場合, 抽出したクラスタのユーザを分析していくことで, 幼稚園や小学校などへの入園や入学に合わせた番組や CM の制作に活用できるだろう. また, ある番組の視聴者における抽出したクラスタのユーザの割合を調査することで, 幼稚園や小学校などに入園・入学した視聴者の分布といった観点で, 番組の視聴者層を分析することもできるだろう.

その他の各クラスタについても検証し, 結果の概要を表 3 にまとめた. 表 3 のテンプレートやクラスタは, それぞれ図 3 の各図のアルファベットや各クラスタの番号と対応している.

図 3 の(b)の cluster4 は, 平日昼 12 時に普段は視聴していないが正月だけ視聴した層である. 教養番組の視聴が多いことから, 平日は仕事に外出し, 休日の日中の時間を活用してテレビで情報収集や勉強を行っていると考えられる.

図 3 の(c)の cluster7,0,6,1 は, 四半期単位で 17～22 時に視聴しない時期がある層である. 17～22 時はテレビ視聴率が高いと知られている時間帯を含んでおり, 主にドラマなど, 四半期の区切りで改変される番組が放送される時間帯である. これらクラスタは四半期の区切りに合わせて視聴しなくなっているため, それぞれの期間で見たい番組が無かった層と推測できる. また世帯属性の結果から, M1 層・F1 層・BC 層は改変期ごとに選択的に番組を視聴する傾向が高いと推測できる.

図 3 の(d)の cluster3 は, 深夜にいつも視聴している層である. エリアによる偏りは, 地域ごとの就寝時刻に起因している可能性がある. 平成 28 年社会生活基本調査の結果をもとに公開されている『夜更かし！？ランキング』[総務省 2017]を参照すると, 近畿・四国・九州の一部(福岡や沖縄など)に夜更かし上位の, 北海道・東北・中部に夜更かし下位の都道府県が多くなっており, 複数の都道府県で傾向が一致する様子が観察できる.

以上の(b)から(d)についても, それぞれのクラスタから分かる視聴者層に向けた, 番組や CM の制作に活用することができるだろう. あるいは, ある番組の視聴者層の分析にも活用できるだろう.

本章では, 提案するアプローチによって視聴パターン判定の推移から抽出した機器のクラスタを複数分析した. 機械的に得られた機器集合であるクラスタに対して, 世帯属性・エリア設定・番組視聴割合ランキングといった別の観点から検証して矛盾がない仮説を立てることができた. また, 番組や CM の制作への活用の可能性を示すことができた.

6. おわりに

本研究では, 注目する視聴パターンに該当するかどうかを判定する手法, および視聴パターンの推移を自動抽出する手法を提案した. また, テレビ視聴の実データから得られた視聴パターンの推移や, その推移を持つ機器集合を検証することで, 視聴パターン判定の推移を自動抽出する手法の妥当性を示した. これまでは難しかった, 長期的な観点からの実データの分析ができるようになったため, 新しい知見の発見や, それに基づくマーケティングなどが期待できる.

有用な視聴パターン判定の推移を抽出したら, 視聴パターンと同様にテンプレート化することで, 抽出の対象期間や対象機器以外についても判定が可能となる. ただし, 期間が異なると祝日が同じ曜日とは限らない点に考慮して計算する必要がある.

今後は, 視聴パターン判定の推移のさらなる分析や, 再生視聴を含めた分析, 分析によって発見される知見の具体的な活用方法について, 検討を進めていく.

参考文献

[電通 2018] 株式会社 電通: 2017 年 日本の広告費, http://www.dentsu.co.jp/news/release/2018/0222-009476.html .

[Xu 2012] Mengxi Xu, Shlomo Berkovsky, Irena Koprinska, Sebastien Ardon, Kalina Yacef: Time Dependency in TV Viewer Clustering, UMAP Workshops, vol. 872, 2012.

[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.

[水岡 2017] 水岡 良彰, 陶 亜玲, 中田 康太, 折原 良平: 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, 人工知能学会全国大会(第 31 回), 人工知能学会, 2017.

[陶 2017] 陶 亜玲, 水岡 良彰, 中田 康太, 折原 良平: 1K2-3 大規模テレビ視聴データを用いたクラスタリング技術の開発, 人工知能学会全国大会(第 31 回), 人工知能学会, 2017.

[総務省 2017] 総務省統計局: 統計局ホームページ/平成 28 年社会生活基本調査 47 都道府県ランキング http://www.stat.go.jp/data/shakai/2016/rank/index.htm.