

# スキウタ集計結果のデータマイニング

## NHK紅白歌合戦に関する視聴者ニーズの構造化

慶應義塾大学大学院

政策・メディア研究科 後期博士課程

小野田 哲弥

### 1. 背景と目的

一九六三年に視聴率八一・四%という金字塔を打ち立てた『NHK紅白歌合戦』（以下『紅白』）も、近年では「かつての勢いはない」（村上、二〇〇六）などと揶揄されている。だが視聴率の減衰は『紅白』固有の問題ではなく、社会環境の変化に拠るところが大きいと思われる。個人化とインターネットの登場が「マス・マーケティング」から「One-to-One マーケティング」へのパラダイムシフトを促したとされるが（eg. Kotler, 2000）、これは何も消費や流通に限った現象ではない。人々の価値観の多様化、そして接触可能なメディアの爆発的増加ともパラレルである。したがって視聴率の低下傾向を悲観し、その回復ばかりを主眼に据える主張には賛同しかねる。それよりも時代のニーズを的確に捉え、なおかつ『紅白』の持つ歴史的価値を損なわないための建設的議論が成されるべきではあるまいか。

その議論にはたたき台が必要であり、それを構築するための客観的データが不可欠である。だがこれまで十分な資料を得ることは叶わなかった。毎年NHKは『紅白』のための世論調査を行っているが、非公開が常で、公開された場合もその信憑性を疑う議論（文芸春秋b、二〇〇四など）が絶えないからである。しかし二〇〇六年一月、貴重な資料が一般に公開された。それが「スキウ

タ「紅白みんなでアンケート」(以下「スキウタ」)の集計結果である。ただし残念なことに、それを下敷きとした議論が起きることなく半年を経過してしまった。ゆえに本研究により、その手付かずともいえるデータに先鞭をつけた。だがそれは、誰かが選ばれるべきといった選考結果の検証を意図するものではない。本稿の目的は、未来の『紅白』を展望する議論に必要な、まさに「たき台」を提供することである。そして、より具体的には今日の複雑極まる視聴者ニーズを捕捉し、『紅白』の選考に活かすために有効な、一つの方法論を提示することにある。

## 2・使用データと解析技法

### 2 1 珠玉の「スキウタ」データ

「スキウタ」の調査期間は、二〇〇五年八月一日から一〇月三十一日までの三ヶ月間である。四媒体(「投票ハガキ」「携帯電話」「パソコン」「データ放送」)を利用した非サンプリング調査であり、希望する者は誰でも自由に調査に参加できた。文字通り「好きな歌」を投票するのだが、投票できるのは一人四曲以内である。うち三曲はあらかじめ用意された六〇〇曲の中からの選択、残る一曲はそのリストにはない曲を自由記述する方式である。なお厳密には「曲名」ではなく、必ず「歌手名」を併記することが条件とされた。

調査上の問題点については後述するとして、第一に指摘しておかなければならないのは、「日本人の好きな歌」に関して、これほどまでに大規模な調査が実施されたことは前例がないという点である。重複回答があつたにせよ一五〇万人強から、四〇〇万件近い回答が寄せられた。これは実に国民の一〇〇人に一人以上が回答した計算になる。第二に、複数媒体を一つの調査に融合的に活用した点も特筆すべきである。一年を締め括る『紅白』の原初的コンセプトは「その年を代表する歌手が持ち歌を披露する」ことにある(合田、二〇〇四)。

だが今日、その代表性を一義的に規定する指標は存在しない。かつてのCD

至上主義は鳴りを潜め（烏賀陽、二〇〇五）、カラオケ選曲、ライブ動員、音楽配信など様々な享受形態によって複合的な音楽市場が形成されている（電通総研、二〇〇五）。媒体の網羅性に問題があるとはいえ、複数媒体からのデータ取得を許容した調査設計は、その難題にアプローチする門戸を開く英断であった。「スキウタ」の集計結果はNHKのウェブサイトに、票数一件のものまで限なく公開された（図1）。この類稀な集計結果を基に、以下、この活用方法について議論を進める。

## 2.2 データマイニングの適用事由

『紅白』のターゲット層は「マス」か「ニッチ」か。それは紛れもなく前者である。その意味で同調査は飽戸（一九八七）に代表される、確立された社会調査法に則るべきであった。標本抽出を行わず、四つもの媒体を用い、選択肢以外の自由回答をも認めた「スキウタ」は、「日本人が好む上位曲」を決定するマス向けの調査としては、余りに無駄の多い調査だったと言わざるをえない。だが図らずもこの調査設計は今日の世相を色濃く反映するものであった。プライバシー意識の高まりから（白石、二〇〇五など）、綿密なサンプリング調査の実施は困難である。また「標本」の前提となる明確な「母集団」の定義も難しい。国民のほぼすべてが『紅白』視聴者に想定できた時代では、母集団は国勢調査に等しく「日本国民」であったろう。しかし「受信料完納者」「『紅白』視聴希望者」に限定するとすると、それは日本人の部分集合でしかなくなる。このような状況下だったからこそ、NHKは「膨大な量のマイクロな「動き」を「全体」として把握する」（梅田、二〇〇六）手段を選んだに違いない。

固有値計算に代表される多変量解析技法は、その適用が相応しいデータ、すなわち実験計画に基づいて回収された非欠損データに対して、その効力を最大限発揮する。しかし「ニッチ」にチャンスを求める実業界、ことに新興企業では、日々、不確実性の高い疎行列データをハンドリングする必要性に迫られている（豊田、二〇〇一など）。このような要請から、マーケティング分野で目覚

しい発展を遂げてきた新たな解析技法として「データマイニング」(以下、DM)がある。「スキウタ」のデータ的特性を踏まえれば、多変量解析とDMのどちらの適用が理に適っているかは議論を待たない。よって本稿ではDM技術を応用してデータの構造化を図る。

### 3・入力データの準備

#### 3 1 大規模調査におけるノイズの必然性

実験計画に基づかず調査前の制御が不完全な非サンプリング調査においては、事後のデータ加工が最終的なアウトプットの成否を大きく左右する(eg. Bigus, 1997)。次章でのDM適用が円滑に進むように、集計データに入念な加工を行い、安定した入力データを実現するのが本章の目的である。

「スキウタ」集計データにおいて第一に憂慮されるのは、完全なる自動集計調査ではなかった点である。例えば「ハガキ」のようなアナログ媒体では、手作業による集計が不可避である。そして一〇〇万件を超える投稿量を捌くためには、非熟練者の協力も仰ぐしかない。四件もの複雑な「歌手名+曲名」の組み合わせ、しかも自由回答の中には集計者が未知の名称も数多く含まれていたと考えると、相当数の誤集計があつたとしても不思議ではない。実際に集計結果をダウンロードして各媒体票数の総和を求め、それをNHKが概要として公表した総数と比較してみたところ、表1のような誤差が見られた。だが上述の過酷な集計条件を考慮すれば、わずか一%程度の誤差に収まった点はかえって評価されてよく、この問題に拘泥することも賢明ではない。

第二の問題点は参加者サイドの誤植である。前段で述べた集計者側のミスだけでなく、参加者が自由回答入力する際に表記を誤るケースも考えられる。しかし集計公表時の曲数は一二、五三一曲にものぼった。その一つ一つについて「表記揺れ」(笠原ら、二〇〇二など)を緻密にチェックする作業は非現実的であろう。よってこの点も解決法の開発は今後の研究に委ねたい。

### 3 2 歌手単位へのデータ変換

より本質的な問題点は、『紅白』の選考方法と「スキウタ」集計方法との乖離だと思われる。そしてこの問題解決は、前節で述べた二つの問題点を緩和する効果もある。その方法とは「曲」ではなく「歌手」を単位に集計し直すことだ。そもそも『紅白』の選考プロセスは、まず出場歌手が発表され、次に曲が発表される順である。だとしたならば「スキウタ」も、持ち歌の得票数の総和によって、まず歌手のランク付けが成され、その次に出場候補と目される上位歌手の中で、歌ってほしい曲が決められるのが自然な流れではなからうか。

「歌手」を集計単位としてグループ化を行うと「スキウタ」集計データは一二・五三一「曲」から四・一七九「人」(グループを含む)へと大幅に縮減される。これは第一の問題点としての集計ミスが全体に及ぼす影響が相対的に小さくなることを意味する。そして第二の問題点である誤植もそのチェックが容易となる利点がある。NHKが公開前、そのチェックに多大な労力を注いだことは想像に難くないが、「歌手」に限定し筆者が目視した限りでも、なおも九五〇件の表記揺れが確認された。表2はその代表例である。これは補正可能な問題であり、またそうすべきであるから、統一表記に置換してグループ化を実行し直した。その結果「スキウタ」に投票された「歌手」の総人数は三、六五七人となった。曲単位から歌手単位に置き替えて再集計を行うことにより、行数は当初の二九・二%にまで凝縮されたことになる。以上の加工によりデータの安定性は格段に向上したと考えられる。

### 3 3 メディアバイアスの考慮

ここで改めて「スキウタ」順位の決定プロセスについて確認したい。これは次の通りである。まず曲ごとに媒体別得票数を集計する。次にそれを基に媒体内での順位を出す。続いて曲ごとの四媒体順位の平均を算出する。そしてその値が小さいものから最終順位を決定する。表3はこの方法に基づく白組の上位

一〇曲である。この方法からは、複数投票や組織票に苦慮し、かつ視聴者に説明しやすい決定方法を模索したNHKの苦慮も覗かれるが、最良の方法とは評価しがたい。

筆者の対案は次の通りである。第一に、前節で述べたように「曲」ではなく「歌手」を集計単位とする。第二に、「順位」ではなく「正規化した値」（以下、正規化値）を用いる。そして第三に、最終的な順位は回帰式  $Y = a * X_1 + b * X_2 + c * X_3 + d * X_4 + e$  に基づき決定する（ $Y$  は従属変数。  $X_1, X_2, X_3, X_4$  は独立変数であり、それぞれ「投票ハガキ」「携帯電話」「パソコン」「データ放送」の得票数を正規化した値が入る。  $a, b, c, d$  は各媒体に与えられる係数。  $e$  はプラス要素としての外部項）。本稿ではこの従属変数  $Y$  を「支持度」と呼ぶ。

上記提案理由について補足したい。第一の点に関しては、順位レンジの平準化もさることながら、ニーズの多様化に対応する意図が含まれる。実際に表3でもその半数で歌手名の重複が見られた。「歌手」を単位とすることで、このような重複問題が解決され、上位に含まれる歌手名が増える。その結果、様々な傾向を反映した人選リストが得られるのである。第二の点については「得票率」も考えられる。しかし最大四曲までの複数回答方式は「シェア」の概念に反する。したがって「正規化値」の方がより適した指標と考えられた。第三の点については、厳密に補正するためには各媒体内で層化抽出などを行うべきだが、集計データからそこまでの情報を引き出すことはできない。よって「四媒体」がそのまま「日本人の四層」を表象するものと捉え、上掲のシンプルな回帰式を提案した。なお四媒体の係数決定（ウエイト付け）も本稿の主眼ではないため、すべてを等価（ $a=b=c=d=1$ ）として以降の議論を進める。も考慮しない（ $e=0$ ）。

#### 4・視聴者ニーズの可視化プロセス

##### 4 1 階層区分

以上から最終順位を決定するための尺度「支持度」が得られた。この数値が「ロングテール」(Andreson,2004)で注目された Amazon.com の売上データと異なるのは、正規化値の平均は0であるため、その合計である「支持度」には負の値が散見される点である。サブカルチャー研究やニッチ・マーケティングであればこれらにも注視しなければならないが、『紅白』はあくまで「マス」ターゲットなのであるから、四媒体の正規化値を合計して、なおも0を超えない歌手は解析対象外としたい。このラインを超える歌手は表記揺れ補正後人数の1・8%に当たる四三一人である。これは『紅白』の選考に参考とする人数としては過不足ない候補数であろう。

ただし限定した上位歌手の「支持度」もべき乗分布(図2)を成す。このように入力データ成分の総和に大きな格差がある場合、それが相関係数の高さに直結し、成分間の違いが過小評価されることが知られる(大村、一九八五)。これはデータの質的構造を正しく把握する上での大きな障害だ。よって「レイヤー分割」(小野田、二〇〇三)の手法を参考に、図1の左側ほど「メイン」、右側ほど「サブ」の傾向が強いと考え、その中を四つの階層に区分する。本稿では裾野が累乗的に広がった構造を想定し、レイヤー番号を二乗した比率で階層を生成する。この結果、レイヤー1に一四人、レイヤー2に五七人、レイヤー3に一三〇人、レイヤー4に二三〇人が属する四レイヤーが生成される。表4は以上の過程をまとめたデータのうち、上位五〇件について掲載したものである。

#### 4 2 レイヤー別クラスタリング

前節のデータ処理により、第一の「量的」な分類が完了した。次に求められるのは、「質的」な分類である。DMの諸技法の中で、本件の分類(クラスタリング)に最も適していると考えられるのは、情報損失量を抑えながら、多次元情報を二次元平面に視覚的に再現することのできる(徳高ら、二〇〇二)「自己組織化マップ」(SOM: Self-Organizing Map)である。集計結果で質の違い

を表す指標は、四媒体の得票傾向の違いしかない。よって四媒体の正規化値を成分とする入力データを準備し、それをSOMで解析する。

一歌手の持つ情報量はレイヤー1ほど大きく、レイヤー4ほど小さいため、クラスタリングは下位レイヤーほど巨視的に行う。すなわち下位レイヤーほど一つの歌手群（クラスタ）に、より多くの歌手を帰属させるのである。こうすることで「歌手」データを統合して「クラスタ」単位にしたとき、レイヤー間の情報量格差の解消が期待できる。レイヤー1ではクラスタリングを行わないこととし、クラスタ数がレイヤー番号の比（1:2:3:4）に等しくなるよう解を求めると、各レイヤーのクラスタ数は14:28:42:56と定まる。

#### 4 3 レイヤー間クラスタ結合

各レイヤーでクラスタリングが図3のように完了する。全体構造を把握するためにはレイヤー別に散在するクラスタを統合しなければならない。この目的は次のようにして達成できる。まずレイヤー2以降では「歌手」単位のデータを「クラスタ」単位のデータへと変換する。これは同一クラスタに属する歌手の正規化値を媒体ごとに平均すればよい。続いて各媒体の値が四媒体の総和中でどれだけの比率を占めるかを算出する。以上によって成分単位はレイヤーに関係なく、各媒体率の和を1・00とする尺度で統一される。

クラスタ結合は「非類似度」を基に行う。「非類似度」は式 $\frac{f_{1i}}{f_{1i}+f_{2i}}$ によって得られる。 $f_{1i}$ はあるクラスタの媒体1における比率、 $f_{2i}$ はそれは異なるクラスタの同じ媒体1における比率、 $\sum$ は媒体数（本件では4）である。すなわち「非類似度」とは、異なるクラスタ間の、四媒体それぞれのシェアの差の絶対値を総和した尺度である。その値が大きいほど異質、逆にいえば「非類似度」が小さいほど得票傾向が似通っていることになる。

実際の統合過程は次の通りだ。まず下位レイヤーのクラスタとその一つ上位のクラスタの「非類似度」を比較し、その値が最小のクラスタ同士を結合する。この工程を経ると、レイヤー2までのすべてのクラスタは異なるレイヤーの何



かからのクラスタと結合される。そして、最上層のレイヤー1ではレイヤー内で「非類似度」が最小の歌手へベクトルを引く。こうすることでレイヤー2からのリンクが全くなかったレイヤー1の歌手も、浮遊することなく全体構造へと組み込まれる。以上によって『紅白』出場候補歌手四三一人は、解釈可能なツリー図として構造化される。

## 5・結論

### 5 1 成果と活用法

第三章の前処理、第四章のDM適用によって、視聴者ニーズを反映した邦楽構造は、図4～図10のように可視化できる。図上各クラスタの先頭行は「クラスタID」「ハガキ率」「携帯電話率」「パソコン率」「データ放送率」を示し、クラスタ内の各行は当該クラスタに含まれる歌手を表す。歌手列の構成は左から順に「支持度順位」「歌手名」「代表曲」である。コネクタ線上の数値は二つのクラスタ間の「非類似度」を示す。以上が筆者の最良と考える「スキウタ」集計結果の分析プロセスと最終アウトプットである。

重要な点は大きく三点挙げられる。第一は歌手を単位としたこと、第二は階層を分割したこと、そして第三は領域ごとに構造化したこと、である。第一の点が『紅白』の歌手選考と整合的である点は再三述べてきた通りだ。第二の点は選考に使用する以上、優先順位をつけることは必須である。ただし誤差は不可避のため、その順位が數位入れ替わることには大意はない。よってわかりやすく四階層に分けて提示した。

本研究の新奇性ともいえる第三の点については、詳説したい。他のコメディティとは異なり、音楽コンテンツの購入に関しては強い「非代替性」が知られる(吉田就彦、二〇〇五)。つまり歌手AのDVDを購入しに来店した消費者が、それが売り切れだったからといって替わりに歌手BのDVDを買って帰るか、といえどということだ。この例はファン意識や対価とも関連するやや極端な

ものだが、『紅白』の選考においても、そのコンフリクトを最小限に抑える必要がある。すなわち、あるJ POP歌手に出演を断られた場合、代わりに演歌歌手を選んだのではその補填にはならない。このようなケースでは、同一ジャンルの中で次点候補を立てることが当然望ましい。しかし同一世代内における「島宇宙化」(宮台ら、一九九三)が明らかにされてから、さらに十年以上の時を経た今日、ジャンルの細分化、新規ジャンルの勃興、既存ジャンルの融合などが複合的に進行し、もはや演繹的にジャンルを規定することは不可能となっている。かような状況下では、本研究のように、帰納的に領域(ジャンル)を規定するしかあるまい。本件で提示した構造は、ランキングといった一元的なスケールだけでなく、質的な違いを反映している。そのため、選考のバランスを考慮する目的に適合的である。またレイヤー間結合とその線上の「非類似度」は、次候補、次々候補を選定する際の明確な基準として活用できる。

## 5 2 課題と展望

しかしながら上掲の図が、現実の邦楽構造を余すところなく再現したものであるとは考えていない。本研究の成果は「スキウタ」集計結果の活用にも適な「科学的手続き」を、具体例をもって提示した点に限定される。よって本件の課題は無数に存在しよう。主要な課題を二点挙げ、それらと関連する展望を述べて本稿を閉じたい。

第一は、変数不足という点だ。上記解析に用いたのは、わずか四変数(「投票八ガキ」「携帯電話」「パソコン」「データ放送」の得票数)のみである。年齢構成や男女比もまったく考慮していない。むしろ、このような不完全なデータから真相を明らかにできるはずがないのである。しかしながら、上掲の図は一定の信頼性を保持していると思われる。それはなぜか。言うまでもなく、「メディア」が「人」を表した、ということである。調査媒体の違いは母集団の違いを反映し、各媒体の得票数は、その媒体の利用者属性と密接に関連していたと考えられる。図11は図3をベースに図4〜図10の位置関係を図示したものだ。こ

の図には大きく二軸存在する。横軸は「パソコン」の成分強度と相関の高い「年齢軸」であり、左ほど若年、右ほど年配となる。上部左に「携帯電話」、上部右に「ハガキ」の成分が強く反応することはこの証左である。他方縦軸は、下部に働く成分が「データ放送」であることから、対極に位置する「携帯電話」& 「ハガキ」との差異が軸の説明となる。筆者の仮説では、それは「視聴人数」である。「ハガキ」や「携帯電話」を利用した回答者は、個人単位で調査に参加し、自らの意思で投票した可能性が高い。対する「データ放送」は、リビングに集った家族全員が協議の上で投票した可能性が高いからである。この点は解析上、属性の影響を曖昧化する作用をもたらす。ゆえに図11の中央領域では世代の異なる歌手同士がクラスタリングされる奇妙な現象が起きている。その傾向が顕著なのが図6であり、この図に収まり切らないクラスタは図12に表示せざるを得なかった。『紅白』が「家族揃って観る番組」を標榜し続けるなら、「データ放送支持領域」は最重要領域であり、適切な変数を加えた精査が必要である。

第二の課題は、本論が『紅白』の既成概念に沿って展開されているという点である。それはいうなれば「全方位」であり「総花的」だ。冒頭で述べた「マス」から「One-to-One」への時代変容を踏まえるならば、全ジャンルから均等に歌手を選出するべきとする選考基準も見直さざるを得なくなる。また歌唱順も「一家全員」に配慮すべく、異なるジャンルの歌手を交互に織り交ぜるのが常であるが、個人視聴の一般化と、世代差による決定的な嗜好の違いを念頭に置くとき、時間帯によってターゲットを鮮明にする戦略も存在する。

第一の問題点において言及したように、本稿の解析結果は入力情報不足である。大局的に見れば「ハガキ」はオールドメディアだが、「携帯電話」を利用できない幼年層からの投票もそこに含まれる。また媒体が同一だからといって、主婦とサラリーマンの回答を同属性としてまとめてしまう弊害もある。回答者の属性が判明すれば、それらを説明変数として加えることができる。調査を行わずに変数を入手する方法も存在しよう。例えばサーチエンジンでのヒット件

数、ブログの検索ワード数などである。その他、CD・DVDの売上データ、着うた・着メロダウンロード数、カラオケ選曲数など、リアルな購買データの活用も選択肢の一つに数えられる。説明変数が増えれば、構造化の精度は確実に向上する。四媒体以外の説明変数を動員し、より精緻な構造化を行う必要性を痛感する。

そして第二の、ターゲティングの発想を導入する議論だが、本稿で提示した分析結果はあくまでベーシックなものに留まる。「支持度」の算出式において、ターゲットの比重が高い成分の係数を高める、階層数を変更する、歌手をあらかじめ紅白（男女）に分けて構造化を行う、データを回答者の属性に分けた上で構造化を行うなど、様々な応用可能性が考えられる。生データが入手できるならば、別のDM技術を用いた異なる方法論の開発さえ可能である。

このように『紅白』の予備調査である「スキウタ」の効果的活用法に限っただけでも議論が尽きない。本稿はDMを専門とする筆者独自の視点から、その巨像に微かな光を当てたに過ぎない。これを足掛かりとして、『紅白』に関する活発な議論が喚起され、学際的研究が進捗することを期待する。

#### 参考文献

- 飽戸弘（一九九七）『社会調査ハンドブック』日本経済新聞社
- Anderson, Chris (2004), 『The Long Tail』, 『Wired Magazine』  
<http://www.wired.com/wired/archive/12.10/tail.html>
- Berry, Michael J.A. and Linofit Gordon S. (1997), 『Data Mining Techniques: For Marketing, Sales, and Consumer Support』, John Wiley & Sons.
- Bugus, Joseph P. (1996), 『Data Mining with Neural networks』, The McGraw-Hill.
- 文芸春秋（二〇〇四a）『NHK紅白プロデューサーが制作費八〇〇万円を横領していた！ みなさまの受信料で甘い汁』、『週刊文春』第四六卷（三〇号） 二四二頁
- 文芸春秋（二〇〇四b）『紅白にホントに出てほしい歌手・ほしくない歌手 NHK世論調査とは大違い！三六〇〇人小誌大アンケート』、『週刊文春』第四六卷（四四号） 四六二頁
- 文芸春秋（二〇〇五）『NHK紅白四〇％割れ つまらなかつたシーン第一位は「旗揚げゲーム」』、『週刊文春』第四七卷（二〇号） 一四一～一四三頁
- 電通総研（二〇〇五）『情報メディア白書二〇〇六』、『ダイヤモンド社』
- 合田道人（二〇〇四）『怪物番組 紅白歌合戦の真実』、『幻冬舎』
- 引田惣彌（二〇〇四）『全記録 テレビ視聴率五〇年戦争 そのとき一億人が感動した』、『講談社』
- 笠原要・永森千晴・加藤恒昭（二〇〇二）『自由回答アンケートにおける単語の表記揺れとその解消』、『Jとは工学研究会』第七回  
<http://ulti.mavi.arc.net.my/banana/Workshop/Papers/7/kasahara.pdf>

- 岸田功(一九九七)、「視聴率」、『マス・コミュニケーション研究』第五〇号 八〇 八六頁  
 権容爽(二〇〇五)、「紅白に別れを告げる新しい日本を見た!」、『ニューズウィーク』第二〇巻(四九号) 一七頁
- Kotler, Philip (2000), *Marketing Management: Millennial Edition 10th ed.*, Prentice-Hall.
- 宮台真司・石原英樹・大塚明子(一九九三)、『サブカルチャー神話解体』PARCO  
 村上卓史(二〇〇六)、『大晦日・紅白に勝つためのシナリオ』、『GALAXY』三月号 二〇二―二三頁
- NHK(二〇〇六)、『スキウタ最終データ発表(五〇音順)』、『第五六回NHK紅白歌合戦』  
<http://www3.nhk.or.jp/kouhaku/>
- 大村平(一九八五)、『多変量解析のはなし』、『日科技連出版社』  
 小野田哲弥(二〇〇三)、『マンガ事象におけるサブカルチャーの領域規定』、『マンガ研究』第四号 一九〇―二二五頁
- 尾関光司(二〇〇四)、『構造変化・総論』、『デジタルを笑い飛ばそう!』、『月刊民放』七月号 五―一四頁
- 白石孝(二〇〇五)、『国勢調査をめぐるプライバシー問題』、『都市問題』第九六巻(一一二号) 一九―二三頁
- 烏賀陽弘道(二〇〇五)、『Jポップとは何か』、『岩波書店』
- 徳高平蔵・藤村喜久郎・山川烈(二〇〇二)、『自己組織化マップ応用例集』、『海文堂』
- 豊田秀樹(二〇〇一)、『金鉱を掘り当てる統計学』、『講談社』
- 梅田望夫(二〇〇六)、『ウェブ進化論』、『筑摩書房』
- 吉田就彦(二〇〇五)、『ヒット学』、『ダイヤモンド社』
- 吉田理恵・中野佐知子・渡辺洋子(二〇〇六)、『日本人の生活時間・二〇〇五』、『放送研究と調査』四月号 二二―二五頁

注

我が国のテレビ視聴率調査は、現在ビデオリサーチ社によって行われている。通常「視聴率」とは関東地区の「番組平均世帯視聴率」を指す(岸田、一九九七)。

メディア環境の変化と高視聴率番組数の低下との因果関連を論じた研究としては、リモコンの普及に伴うザッピングの一般化(尾関、二〇〇四)、インターネット世代のテレビ離れ(吉田理恵ら、二〇〇六)などが知られる。

二〇〇五年は戦後六〇年の節目の年であったため、各年のヒット曲一〇曲ずつがあらかじめリストとして用意された。

当時、NHK受信料の支払い拒否が社会問題化していた。その発端は「紅白」担当経験もあるチーフプロデューサーの番組制作費不正流用の発覚である(文芸春秋、二〇〇四aなど)。前例のない大プロジェクト「スキウタ」が実施された背景には、NHKが直面していた開局以来の危機意識がある。

この問題点はTBS『輝く!日本レコード大賞』(以下、『レコ大』)がより強く内包する問題点である。『レコ大』も同じく大晦日恒例だが、昭和時代は両者の放送時間が異なるため相乗効果さえ期待できた。しかし平成以降(一九八九年〜)、『紅白』が二部制を採用したことに伴い競合関係へと突入する。二〇〇六年、TBSは『レコ大』を例年より一日早い一二月三〇日に放送することを決定した。『紅白』の将来展望を図る上で、『レコ大』との比較分析も重要である。

「スキウタ」は「曲」単位で集計すべく設計されていたため、歌手欄に複数の歌手名が並列される問題も存在する(例えば曲『卒業写真』に対応する歌手名は「荒井由実/ハイファイセット」)。集計結果データを分離させることは不可能であるため、このようなケースはその表記のままとした。楽曲提供、リメイク、デュエット曲などがこれに該当する。例えば同じ〇票の場合、「データ放送」では二、八五九位のところ、「投票八ガキ」では四、一九八位となるなど、媒体間の順位レンジに無視できない格差が存在するからである。レイヤー1が1の二乗、レイヤー2が2の二乗、レイヤー3が3の二乗、レイヤー4が4の二乗を、それぞれ占めるように分割する。これは分布をあらかじめ30に当分しておき、上位から1:4.9:16の間隔で割り与えることと同義である。

四媒体の比率が〇%〜一〇〇%の範囲に収まるよう、正規化値平均が負の場合は〇に置

換してから比率を求めた。

「クラスタID」はクラスタを一意に識別することが主目的である。千の位だけはレイヤー番号に対応するが、百位未満の値は順位等とは関係ない（レイヤー1のみ対応）。なお各媒体率の単位はいずれも「パーセント」である。

「曲」に関して「歌手」同様、四媒体で正規化を行い、それらを加算して「支持度」を求めた。「当該歌手の中で支持度が最大の曲」、それが本件における「代表曲」の定義である。

奇妙なレイヤー間結合も散見されよう。これらに関しては「非類似度」を真偽の判別に利用されたい。

抜本的な改革を望む声もある（権、二〇〇五など）。「紅白」という男女分け、「紅勝て！白勝て！」という合戦形式が今日これからも支持され続けるかどうかは定かではない。

曲間に挿入されるゲームや小ネタも好き嫌いが分かれるところである（文芸春秋、二〇〇五）。ターゲティングの議論は、時間帯によって「NHKによる管理」と「歌手による自由な表現」とのバランスを調節することにも通じよう。

例えばアソシエーションルールが適用できる。これはAmazon.comに実装されている協調フィルタリングと同原理であり、「この曲（歌手）に投票した人が投票している別の曲（歌手）」を定量的に抽出することが可能である。

図 1 「スキウタ」集計結果の公開ウェブページ

## 第56回 NHK紅白歌合戦

# スキウタ 最終データ発表

(50音順)

第56回 NHK紅白歌合戦  
投票結果発表へ➔

1/13

次のリストへ➔

曲名	歌手名	投票ハガキ	携帯電話	パソコン	データ放送	スキウタ
		順位 / 票数	順位 / 票数	順位 / 票数	順位 / 票数	順位
A BOY- ずっと忘れない-	GLAY	4198 / 0	5182 / 0	9125 / 0	1452 / 1	
A CROSS OF SADNESS	Hawaiian6	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
A DAY GOOD NIGHT	吉川晃司	4198 / 0	3377 / 1	9125 / 0	2859 / 0	
a Day in Our Life	嵐	1298 / 11	1723 / 6	1766 / 7	2859 / 0	
a LOVE story	SEAMO with BENNIE K	4198 / 0	5182 / 0	3365 / 2	2859 / 0	
A MIRACLE FOR YOU	中島美嘉	4198 / 0	3377 / 1	4736 / 1	2859 / 0	
a presio us pride	麻帆良学園中等部2-A師匠と やめるオトメ組	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
a song dedicated to...	PUHIM(プシン)	4198 / 0	3377 / 1	4736 / 1	2859 / 0	
a song for ××	浜崎あゆみ	1608 / 4	1137 / 16	1296 / 14	1452 / 1	
a song for Life	中山譲	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
a song for love	TUBE	4198 / 0	2644 / 2	9125 / 0	2859 / 0	
A SONG FOR MAMA	BOYZ II MEN	2419 / 1	5182 / 0	4736 / 1	2859 / 0	
A SONG FOR YOU	野口五郎	2419 / 1	5182 / 0	9125 / 0	2859 / 0	
A SONG FOR YOUR LOVE	SMAP	4198 / 0	5182 / 0	2764 / 3	2859 / 0	
a song is born	浜崎あゆみ & KEIKO	2419 / 1	2306 / 3	2764 / 3	2859 / 0	
A SONG OF OLD HAWAII	アンディウィリアムス	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
a walk in the park	安室奈美恵	4198 / 0	5182 / 0	3365 / 2	2859 / 0	
a whole new world	ディズニー映画「アラジン」のテーマ曲	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
AIR	愛内里奈	4198 / 0	5182 / 0	4736 / 1	2859 / 0	
A・RA・SHI	嵐	700 / 258	222 / 283	589 / 92	782 / 5	

注) 50音順で20番目の曲までをスクリーンキャプチャーしたものの。

実際は1ページ当たり1,000曲が掲載されたページが13ページ続く。

公開期間は2006年1月1日から2週間であった。

表 1 「スキウタ」の集計誤差

媒体	NHK公表		筆者再集計	票数誤差	
	参加人数	票数	票数	差の絶対値	比率
はがき	1,163,507	2,931,146	2,967,666	36,520	1.25%
パソコン	159,563	456,091	455,662	429	0.09%
携帯電話	146,540	310,812	310,008	804	0.26%
データ放送	25,875	61,947	62,000	53	0.09%
無効投票	15,996	-	-	-	-
合計	1,511,481	3,759,996	3,795,336	37,806	1.01%

表 2 「スキウタ」の公表データにおける歌手名の表記揺れ（主な例）

表記揺れが見られた歌手名	得票数				補正後の統一表記
	投票ハガキ	携帯電話	パソコン	データ放送	
B'z	2,072	2,603	3,803	319	B'z
B`z	1,239	453	1,100	50	
L'Arc~en~Ciel	905	3,471	4,886	94	L'Arc~en~Ciel
ラルク・アン・シエル	488	1,314	2,031	35	
Mr.Children	5,183	5,911	9,346	542	Mr.Children
Mr.Childen	608	271	732	28	
ミスターチルドレン	0	19	48	7	
ミスチル	1	42	56	0	
T.M.Revolution	5,117	6,995	4,941	130	T.M.Revolution
T.M Revolution	3,548	3,275	810	55	
一青窈	232	123	174	18	一青窈
一青 窈	7,055	3,429	6,339	564	
一青窈	0	2	0	0	
冠二郎	4,811	11	14	2	冠二郎
冠二朗	39,305	61	56	2	
橋幸夫	23,211	111	971	6	橋幸夫
橋 幸夫	62,181	423	2,007	74	
橋幸雄	1	1	0	0	

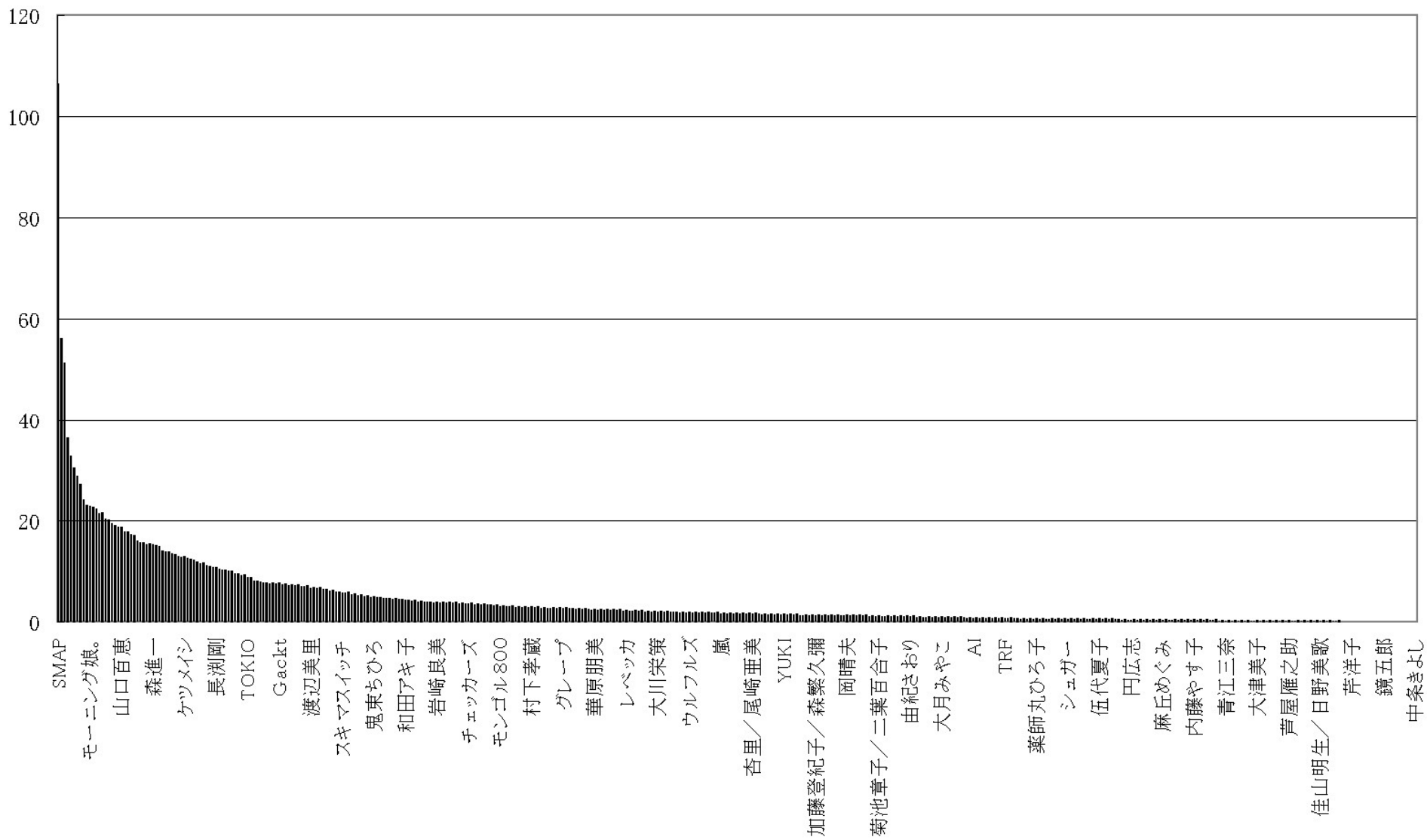


表3 NHKの方法に基づく「スキウタ」の白組上位10曲

最終順位	曲名	歌手名	投票ハガキ		携帯電話		パソコン		データ放送		平均順位
			順位	票数	順位	票数	順位	票数	順位	票数	
1	世界に一つだけの花	SMAP	45	17,853	1	11,595	1	15,313	1	1,715	12.0
2	花	ORANGE RANGE	80	8,437	4	6,230	2	8,228	2	778	22.0
3	きよしのズンドコ節	氷川きよし	4	70,621	29	1,831	56	1,827	11	481	25.0
4	栄光の架橋	ゆず	91	5,900	6	4,070	5	6,036	4	650	26.5
5	箱根八里の半次郎	氷川きよし	3	71,671	36	1,583	60	1,624	14	425	28.3
6	瞳をとじて	平井堅	102	4,968	18	2,787	7	5,299	10	484	34.3
7	夜空ノムコウ	SMAP	125	3,575	7	3,973	3	6,622	20	382	38.8
8	マツケンサンバII	松平健	76	8,954	51	1,225	23	3,168	7	503	39.3
9	桜	リュ・シウォン	36	21,958	39	1,508	32	2,668	76	171	45.8
10	らいおんハート	SMAP	127	3,436	15	2,910	15	4,455	30	279	46.8

注) 4媒体の「平均順位」は公開されなかったため、筆者が求め加筆した。

図 2 支持度正の歌手分布

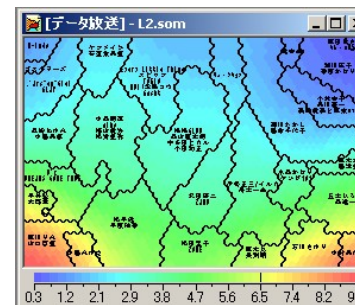
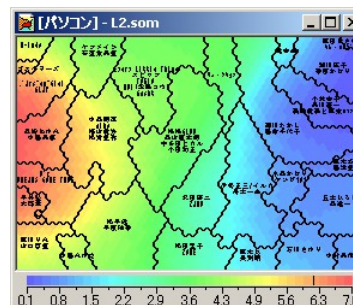
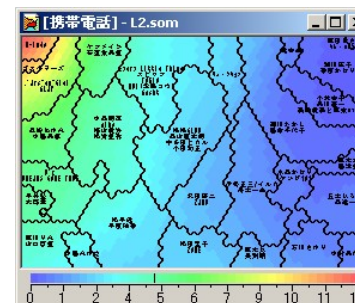
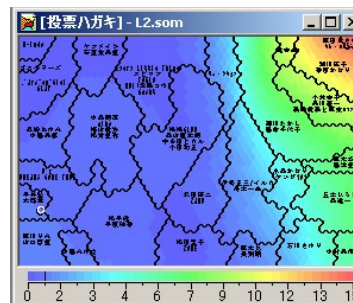
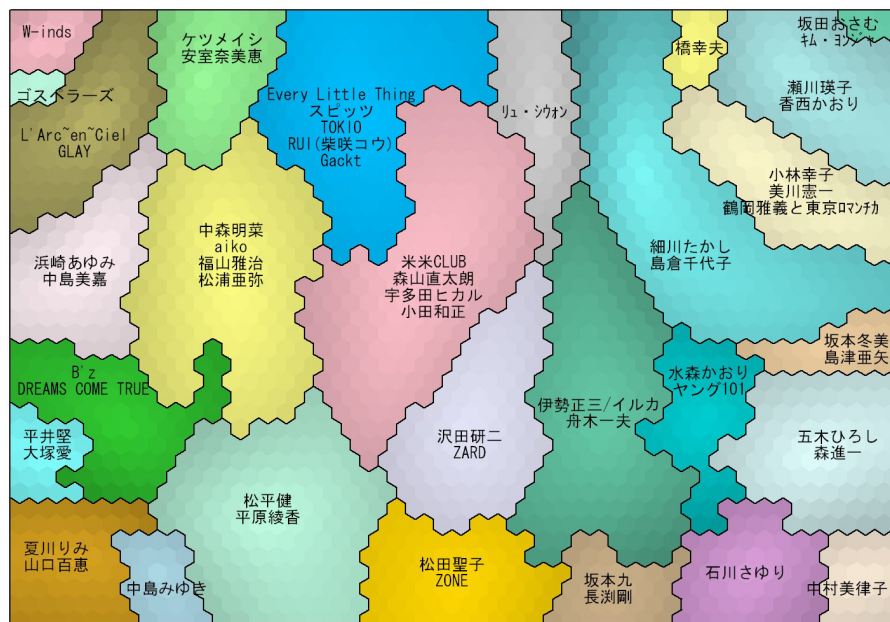


注) ラベルの表示間隔は 10 である。実際には各ラベルの間に 9 人の歌手名が存在する。

表4 歌手の得票数・正規化値・支持度および所属レイヤー（上位50件）

順位	歌手名	投票ハガキ		携帯電話		パソコン		データ放送		正規化合計 (支持度)	レイヤー
		票数	正規化	票数	正規化	票数	正規化	票数	正規化		
1	SMAP	31,505	4.32	21,010	34.67	29,622	37.17	2,550	30.35	106.50460	1
2	氷川きよし	202,776	28.42	4,292	6.97	5,036	6.19	1,229	14.52	56.09851	1
3	ORANGE RANGE	16,772	2.25	11,200	18.42	12,969	16.19	1,225	14.47	51.32026	1
4	ゆず	7,379	0.93	7,344	12.03	10,596	13.20	854	10.03	36.17522	1
5	KinKi Kids	8,179	1.04	8,423	13.81	11,031	13.74	372	4.25	32.84927	1
6	Mr.Children	5,792	0.70	6,243	10.20	10,182	12.67	577	6.71	30.28767	1
7	北島三郎	150,780	21.10	1,600	2.51	732	0.77	383	4.39	28.76367	1
8	T.M.Revolution	8,665	1.11	10,270	16.87	5,751	7.09	185	2.01	27.08400	1
9	サザンオールスターズ	7,481	0.94	3,163	5.10	5,042	6.20	1,006	11.85	24.08479	1
10	美空ひばり	31,857	4.37	1,441	2.25	2,562	3.07	1,121	13.23	22.91426	1
11	モーニング娘。	8,904	1.14	4,428	7.20	6,727	8.32	530	6.15	22.80170	1
12	ポルノグラフィティ	2,825	0.28	5,335	8.70	7,755	9.62	344	3.92	22.51618	1
13	天童よしみ	134,259	18.78	462	0.62	538	0.52	216	2.38	22.30853	1
14	一青窈	7,287	0.91	3,554	5.75	6,513	8.05	582	6.77	21.47944	1
15	平井堅	6,071	0.74	3,466	5.60	6,696	8.28	587	6.83	21.45305	2
16	W-inds	912	0.02	10,242	16.83	2,540	3.04	52	0.42	20.30712	2
17	大塚愛	9,095	1.17	3,406	5.50	5,198	6.39	598	6.96	20.02325	2
18	中村美律子	54,104	7.50	329	0.40	434	0.39	949	11.17	19.46012	2
19	夏川りみ	10,143	1.31	2,187	3.48	4,340	5.31	748	8.76	18.86690	2
20	坂田おさむ	129,080	18.05	112	0.05	422	0.37	42	0.30	18.76925	2
21	山口百恵	7,516	0.94	2,035	3.23	2,678	3.22	952	11.20	18.59502	2
22	L'Arc~en~Ciel	1,393	0.08	4,785	7.79	6,917	8.56	129	1.34	17.77179	2
23	ゴスペラーズ	616	-0.03	6,114	9.99	4,887	6.00	152	1.62	17.58180	2
24	GLAY	1,578	0.11	4,764	7.75	5,638	6.95	223	2.47	17.27744	2
25	橋幸夫	85,393	11.90	535	0.75	2,978	3.60	80	0.76	16.99916	2
26	浜崎あゆみ	4,086	0.46	3,554	5.75	5,204	6.40	308	3.49	16.09703	2
27	キム・ヨンジャ	114,762	16.03	40	-0.07	31	-0.12	8	-0.11	15.73530	2
28	B'z	3,311	0.35	3,056	4.92	4,903	6.02	369	4.22	15.51439	2
29	中島みゆき	4,054	0.46	1,787	2.82	3,400	4.13	684	7.99	15.39620	2
30	五木ひろし	53,794	7.46	522	0.72	780	0.83	545	6.33	15.33236	2
31	森進一	53,665	7.44	749	1.10	1,047	1.16	481	5.56	15.26003	2
32	石川さゆり	23,225	3.16	996	1.51	1,462	1.69	740	8.66	15.01192	2
33	中島美嘉	3,893	0.43	2,651	4.25	5,791	7.14	281	3.16	14.99004	2
34	中森明菜	11,157	1.46	2,736	4.39	3,917	4.78	289	3.26	13.88732	2
35	DREAMS COME TRUE	3,512	0.38	2,385	3.81	4,513	5.53	364	4.16	13.87962	2
36	aiko	5,451	0.65	2,807	4.51	4,550	5.58	258	2.89	13.62836	2
37	松平健	10,480	1.36	1,287	1.99	3,426	4.16	520	6.03	13.54004	2
38	福山雅治	5,916	0.72	2,481	3.97	3,844	4.69	327	3.71	13.09065	2
39	瀬川瑛子	91,660	12.78	52	-0.05	68	-0.07	43	0.31	12.97051	2
40	香西かおり	84,390	11.76	166	0.13	272	0.19	76	0.71	12.78885	2
41	ケツメイシ	4,008	0.45	3,146	5.07	3,107	3.76	306	3.46	12.74365	2
42	松浦亜弥	8,096	1.03	2,165	3.45	3,825	4.66	306	3.46	12.59832	2
43	小林幸子	65,320	9.08	644	0.93	755	0.79	138	1.45	12.24892	2
44	平原綾香	3,127	0.33	1,860	2.94	3,286	3.98	430	4.95	12.20014	2
45	坂本冬美	48,252	6.68	510	0.70	777	0.82	315	3.57	11.77351	2
46	リュ・シウォン	23,327	3.17	1,666	2.62	3,097	3.75	190	2.07	11.60764	2
47	島津亜矢	56,833	7.88	195	0.18	203	0.10	293	3.31	11.47216	2
48	松田聖子	3,479	0.38	1,732	2.73	2,377	2.84	447	5.15	11.09581	2
49	坂本九	6,229	0.76	1,031	1.57	1,460	1.68	586	6.82	10.83103	2
50	安室奈美恵	1,476	0.09	3,491	5.64	2,835	3.42	156	1.67	10.81922	2

図3 SOMによる歌手の分類結果(レイヤー2の例)



注) 使用ソフトは Viscovery SOMine Enterprise Edition Version 3.0J。

右は各成分の影響力をクラスタリング結果上に表示させたもの。

後掲ツリー図の領域区分は、これらの成分表示を参考にしている。

図 4 邦楽構造 1 (「八ガキ支持」領域)

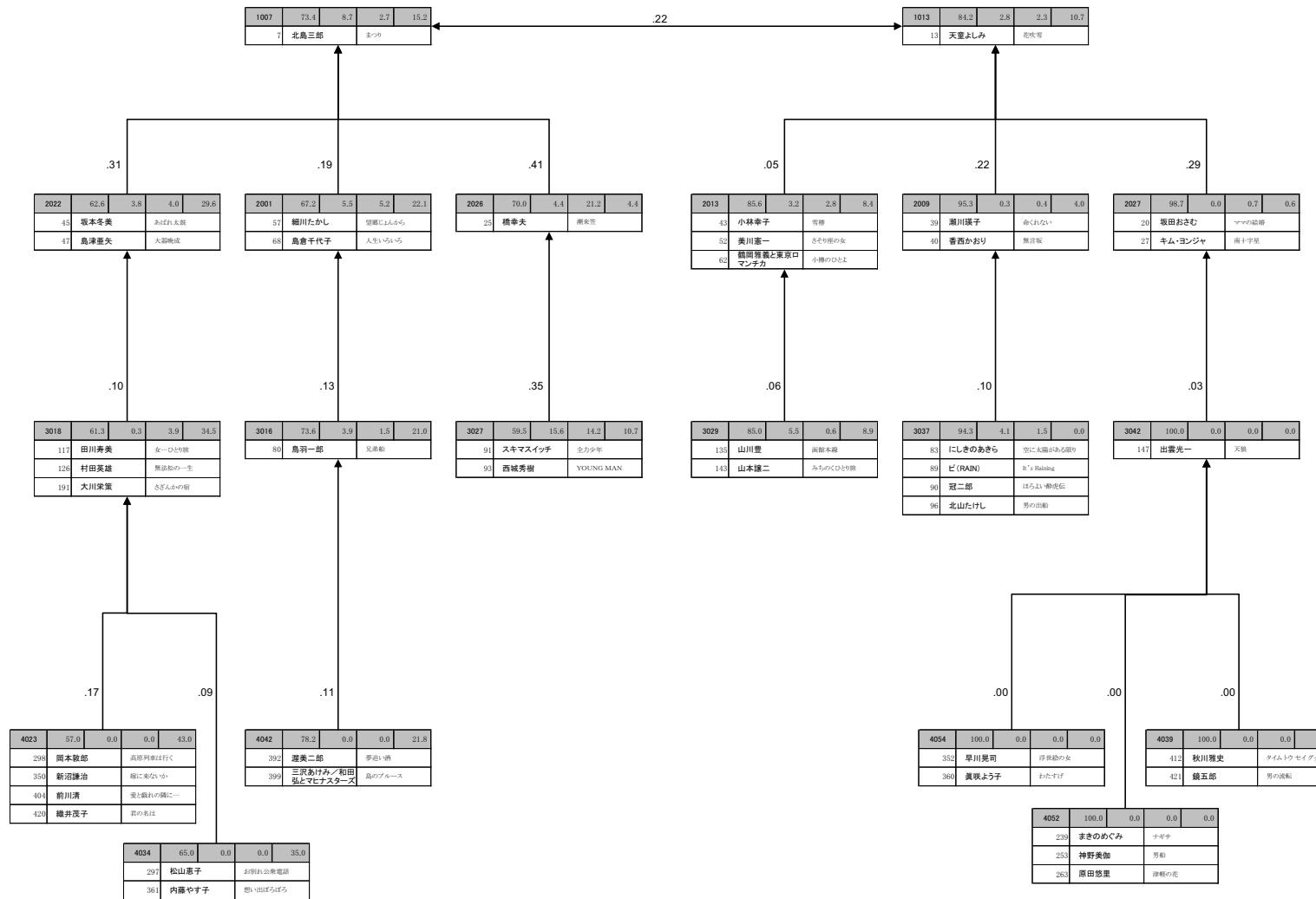




図 6 邦楽構造 3 (「データ放送支持」領域)

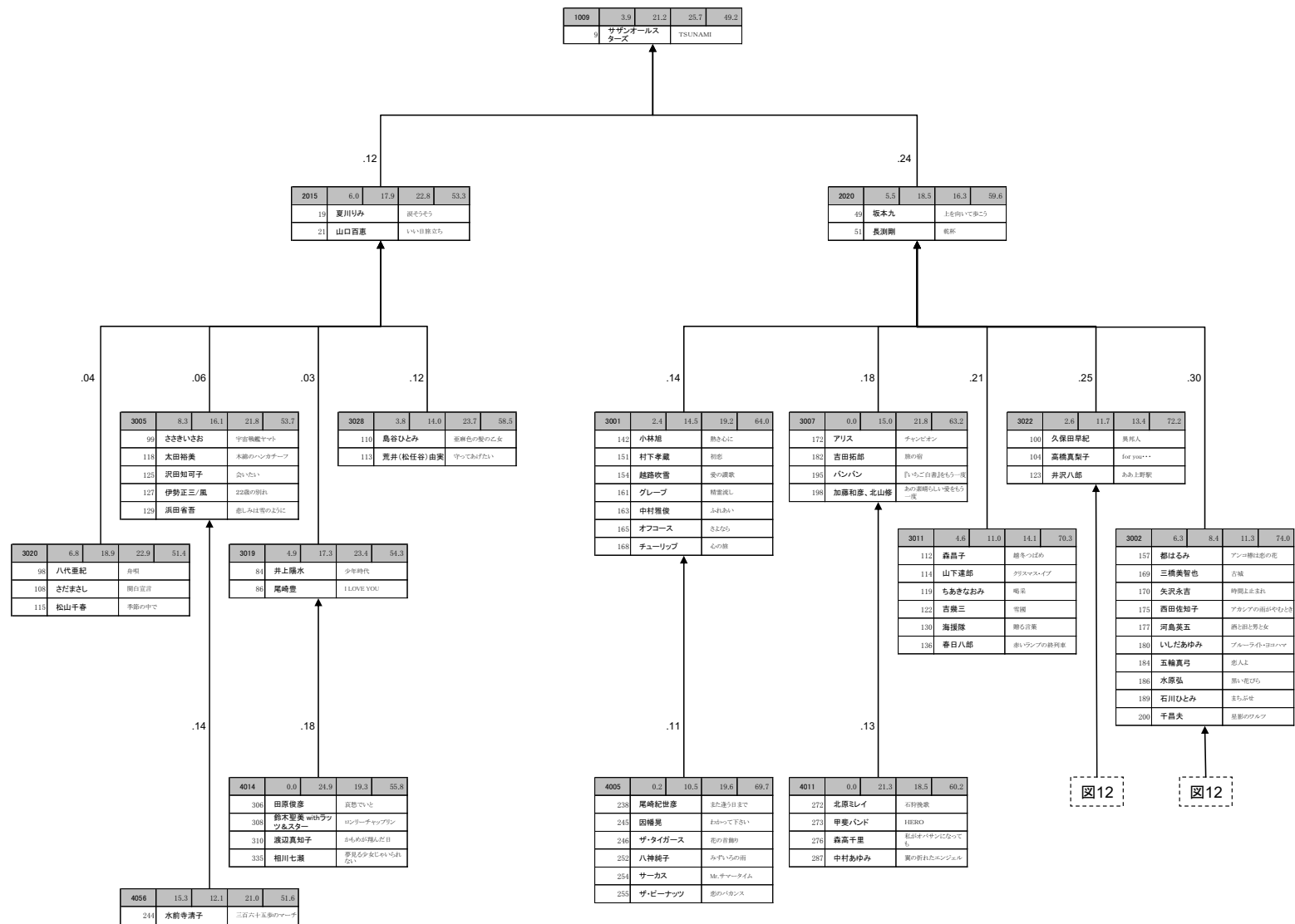


図7 邦楽構造4（「データ放送支持」+「パソコン支持」領域）

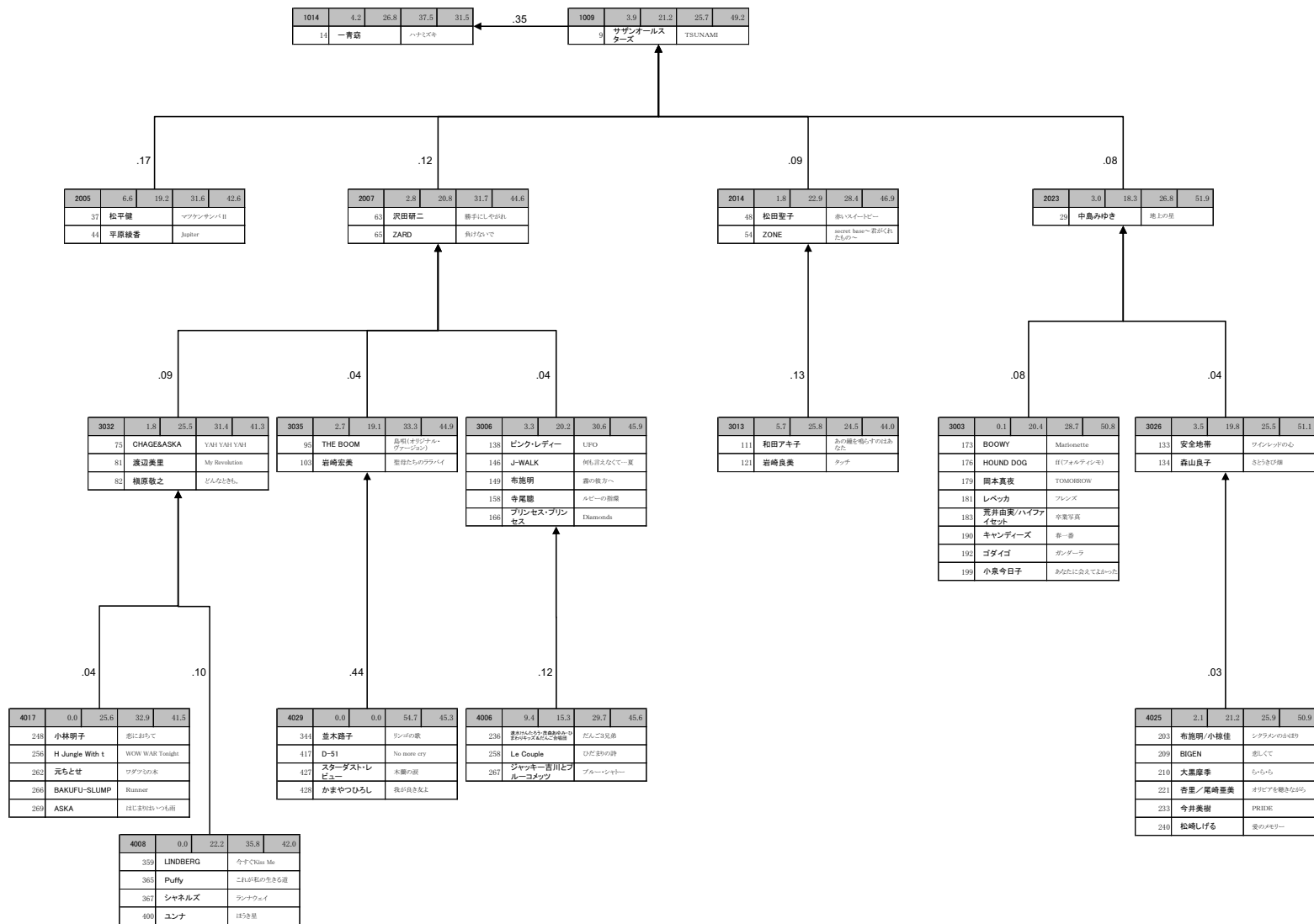




図 8 邦楽構造 5 (「データ放送支持」+「パソコン支持」+「携帯電話支持」領域)

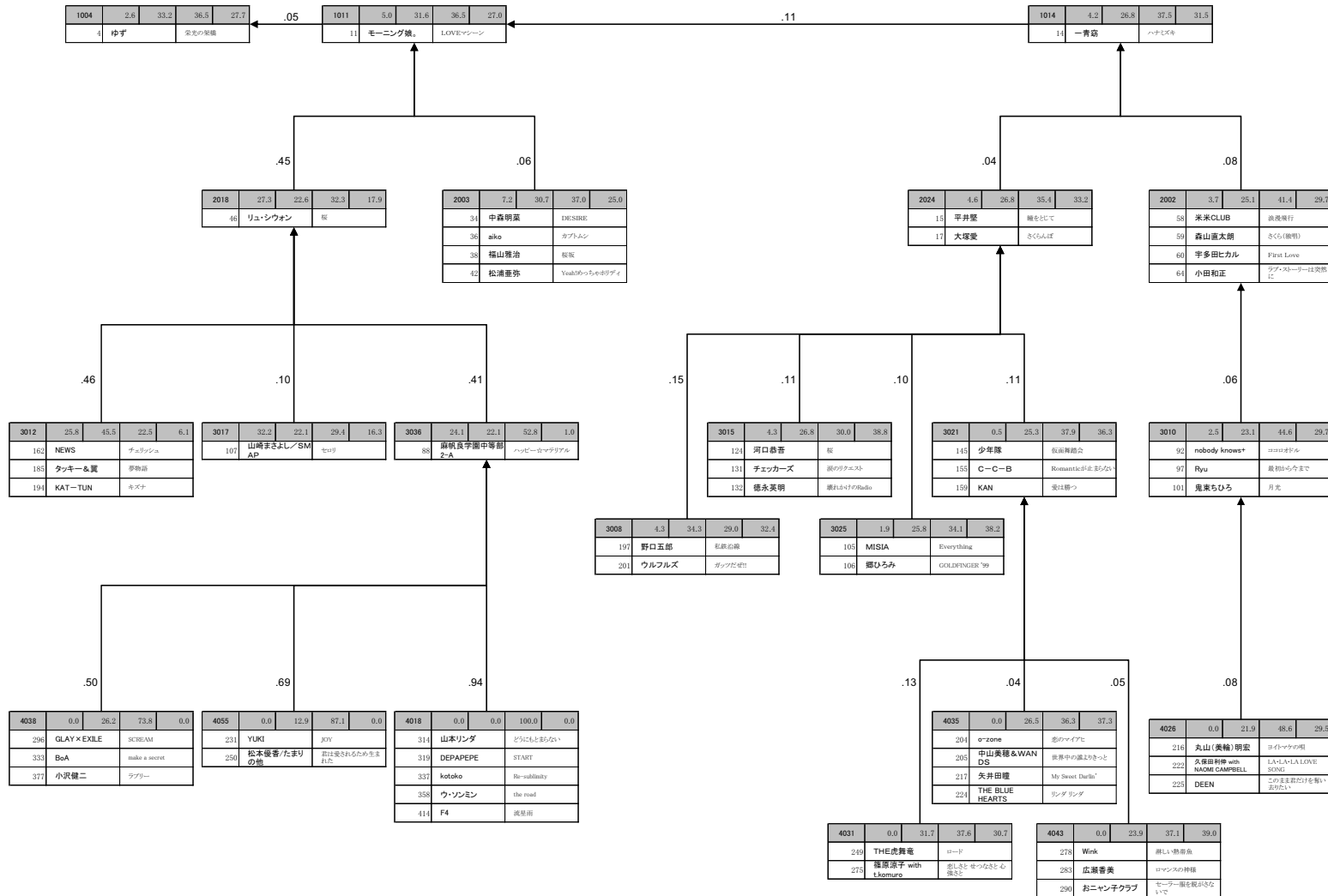


図9 邦楽構造6（「パソコン支持」+「携帯電話支持」領域）

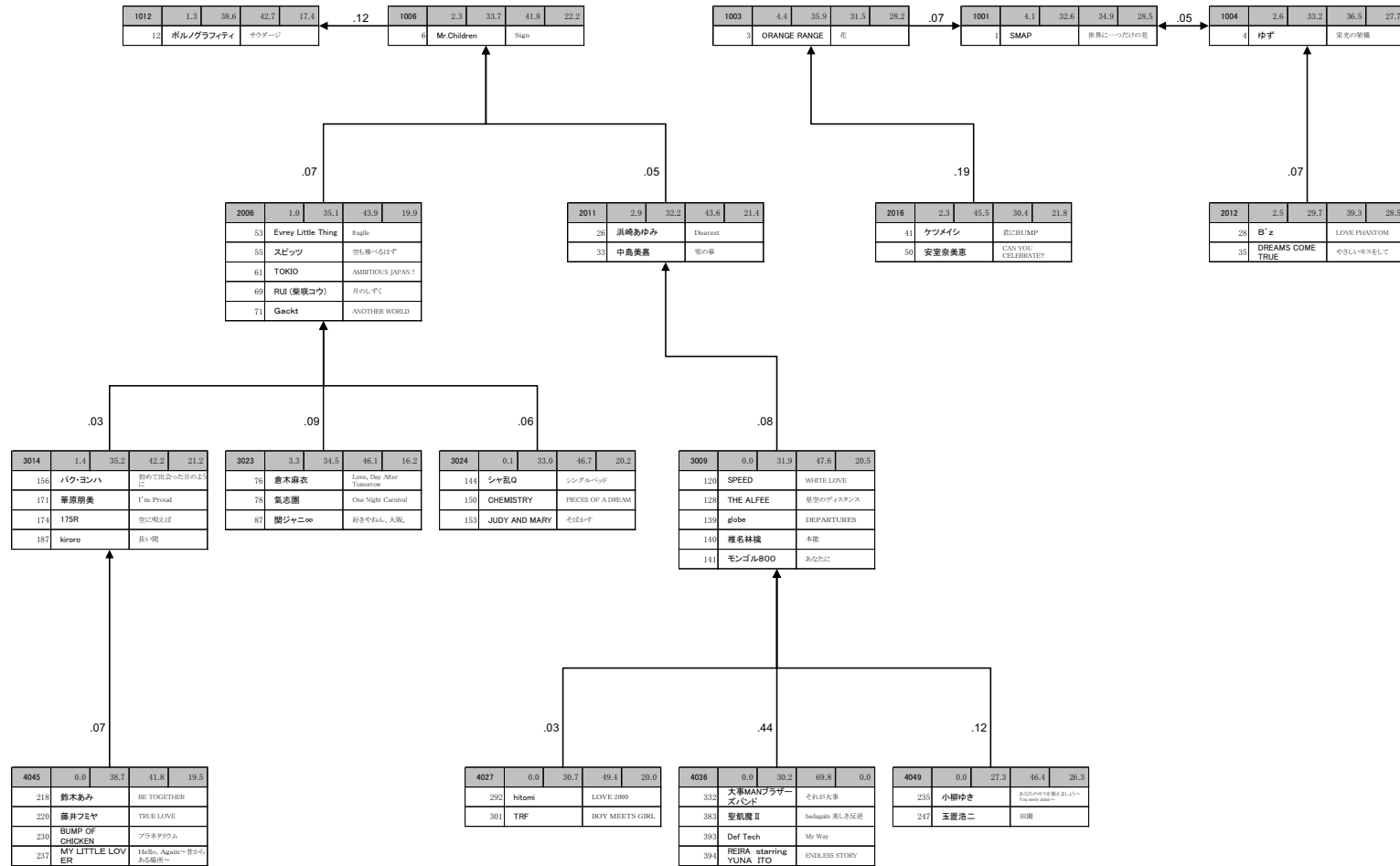


図 10 邦楽構造 7 (「携帯電話支持」領域)

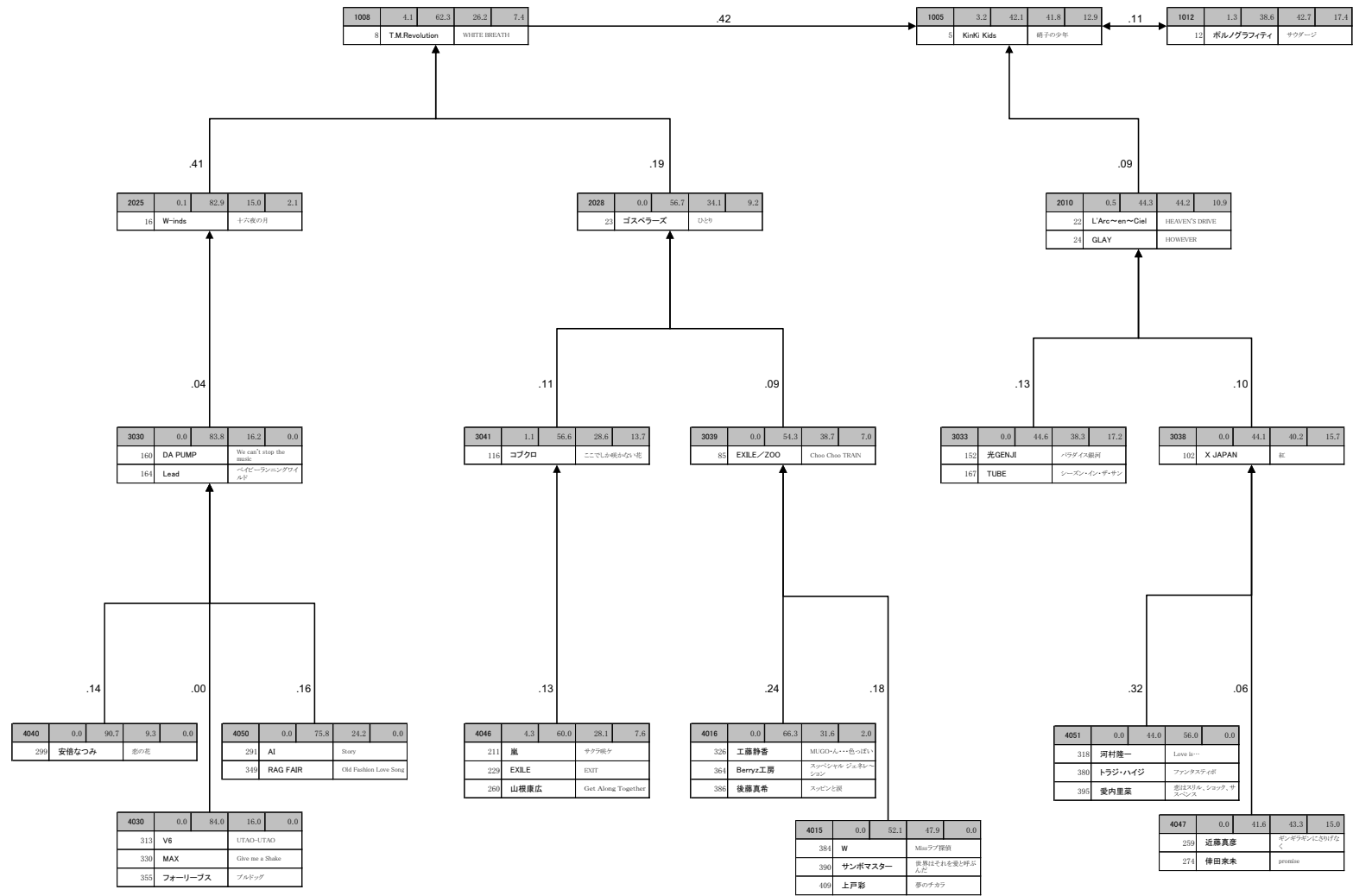


図 11 4 媒体の利用者層と邦楽構造との関連性

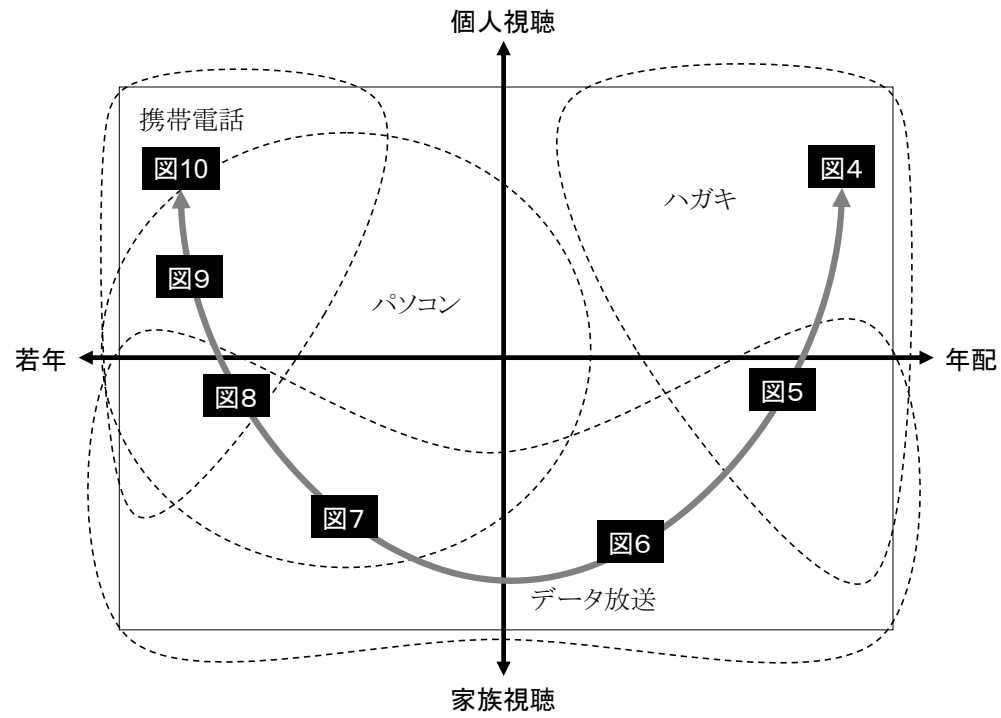


図 12 図 6 の補足

