Vol.2014-ICS-173 No.8
2014/1/23

国際放送コンテンツの潜在視聴層検出における帰納的分析方法

岡崎孝太郎[†1]

ビジネスの現場で活用されるデータマイニング手法では，観測や相関が頻出するほど統計的に有意であり，解決に貢献すると考えられている．しかし国際放送コンテンツの潜在視聴層の検出においては，この前提が通用しにくく，むしろ逆に分析精度向上の妨げにすらなる．今こそ視聴層の観戦リテラシーや生活様式の動的遷移を捉えた，仮説発見から検証そして周到な観測へつながる知識展開の閉ループの実践が求められる．自由文データベースを特徴空間とする帰納的分析プロセスを使った国際放送コンテンツ視聴に関する比較実証実験で，この問題を具体的に提示する．

Inductive Approach for Prospect Discovery

KOTARO OKAZAKI [†1]

1. はじめに

特定の国で制作され世界各国へ供給される国際放送コンテンツのビジネスが，今世紀に入り活況を呈している．アジアや南米地域の経済発展が，国民の生活水準を向上させ，スポーツやドラマといった娯楽コンテンツへの興味や関心を促した．制作から供給，視聴に至る全てのプロセスが，コンテンツをデジタル情報として処理するようになり，事業が高度に洗練されるようになった．多国籍の視聴者に求められている放送コンテンツを，直接視聴者に有償で供給する以外に，放送権や配信権をマーケティング企業にライセンスするビジネスも発達した．そうした企業にとっては国際放送コンテンツへの活用が，文化や生活様式や価値観の異なる国々へ，一斉に自社に有利な戦略を刷り込む環境整備の役割を果たすからである．

様々な国際放送コンテンツと様々な企業同士が，事業目的の達成に貢献する取り組みを追究している．そして国際放送コンテンツが世界中のどのような視聴者から支持されているのかという問題は，コンテンツの投資価値を企業が計る際の重要な尺度の一つである．そのコンテンツの視聴者が各々の国にどれだけいて，どのような属性を有しているかが，ライセンシー企業のマーケティング計画の成否に直結する．そのため放送・配信事業者は視聴者層の正確な把握に努め，さらなる獲得に注力している．

コンテンツを供給するプラットフォームがグローバルな規模で急速に整備され，顧客情報と視聴履歴データが大量に蓄積されている．大量の顧客データを分散処理できる高性能のコンピュータの登場により，データマイニング技術の活用が進んでいる．そしてかねてより潜在視聴層の検出という問題が，放送事業者にとってとりわけ難しい課題であり続けている．国境を越えて放送圏が拡張するにつれ，文化圏や国毎に異なる視聴習慣が散見され，背後にある視聴環境と生活様式をデモグラフィックに捉えにくくなってきているからである．一つの国で視聴実態を説明できる要因が他の国で全く通用しないため，グローバルに各国間を比較できる説明変数が発見しにくくなっている．

本稿はこうした国際放送コンテンツの潜在視聴層検出という問題へのデータマイニングの実践について論じる．まず国際放送コンテンツの顧客データ様式とマーケティング活用における一般的特徴について解説する．次に当分野で活用されている主要なデータマイニング手法の，分析面からみた共通の特質に触れる．そして国際放送コンテンツの潜在視聴層の検出において，何故これらの手法がうまく働きにくいのかを導く．

本研究ではその解決への糸口として，インダクション(induction; 帰納)的解析方法を採り入れたテキストマイニングに焦点をあてる．まず視聴行動をグローバルに分析できる説明変数の減衰につれて，顕在する変数間に擬似相関が生じやすくなる点を指摘する．この問題に対し交絡を可能な限り回避できる特徴空間として，自由回答文によるサンプルデータベースを提示する[a]．自由回答文データを効果的に活用するマイニングプロセスと，インダクション型の解析アプローチについて述べ，これらの手法について従来型のマイニング手法との比較実証実験を試みる．最後に，実験結果への考察と論証を通じて本研究の成果と課題について論じる． 潜在視聴層の検出問題への取り組みは，放送コンテンツのみならず，グローバルに事業展開するあらゆる分野の企業が抱えるマーケティング上の挑戦に大きく寄与するものと信じる．

†1 (株)ソナー
SONAR Inc.

a) 交絡とは， 統計モデルにおいて説明変数と従属変数の両方に相関する外部の変数が存在している第一種過誤の状態をさす．

2. 国際放送ビジネスとデータマイニング

国際放送コンテンツの顧客データは，世帯毎の視聴履歴と契約者情報をセットにしてデータベースとして管理する．これを番組送出・編成情報と照合させて，世帯がいつどんな番組を視聴したかを読み取ることができる．しかし多くの場合こうしたデータベースは，当初より課金管理を用途として構築されている.そのため個人情報保護の観点から，契約解除と同時に記録を破棄しなければならず，契約と解約のモデル化に使える顧客情報が調わないことが多い．

放送・配信事業者の意思決定の根拠は，概ね二種類の相関関係を参照しているにすぎない． 第一は，マーケティングコストの投入時機と加入解約件数との相関関係である．第二は，契約者のデモグラフィック属性と世帯が視聴した番組内容との相関関係である．これらの統計量から顧客の視聴行動モデルを着想するのは難しい．因果則を導く前提としての反事実的状態を仮定できないからである[b].

その一方でまた，各国毎の視聴行動を説明できる特徴変数，すなわち視聴習慣や背後に横たわる生活様式や社会環境の特質については，データとして収集も解析もできていない．ここではその顕著な例として，インドネシアとブラジルの視聴層を比較する．

インドネシアの経済を牽引する新興中間層は約 4,000 万人おり，全人口の約 17%を占めている．彼らの月間世帯支出額は 250 万インドネシアルピア=約 20,000 円である．ブラジルのスラム街に暮らしている最貧困層は約 1,000 万人で，全人口の約 5%を占めている．彼らの月額所得は 500 レアル=約 20,000 円である．有料テレビの普及率は，インドネシアでは国内の全中間取得層の 4%である．ブラジルでは国内の全スラム地域の 28%である．そして好きなコンテンツの第一位は，両国ともにサッカーとなっている．

放送・配信事業者がこれら二つの視聴層を分析する際に利用できる顧客データは，「世帯年収」「好きなコンテンツ分野」「パラボラアンテナの有無」程度に限られる．

この制約の下で分析すれば結果は次のようになるだろう．円換算した世帯年収金額で両視聴層の間に差は見られない．そして両層ともにサッカー好きである．それぞれの国内では成長著しい新興富裕層と依然停滞している最貧困層との比較という格差は仮定できる．にもかかわらず嗜好品に相当する有料テレビの利用実態は，所得に反比例している．パラボラアンテナの有無と有料テレビ利用率とは強い相関にあるという程度になる．これでは潜在視聴者層の検出はおろか，現状把握すら難しい事態に陥ってしまう．

実際に両視聴層をデモグラフィックに分析比較する場合には，両国の為替レートの決定要因や，可処分所得別の生活必需品，国内のスポーツ振興実績や文化的素地，政府の社会政策の方針と実績等の情報を加味するべきと考えられるが，勿論現行の顧客データベースからこのような統計情報を観測することはできない．社会事情の異なる多くの国を跨るほど，各国に共通して説明力を発揮するデモグラフィック属性の収集はさらに困難さを増すのである．

それでも放送・配信事業者はこれらの本質課題に留意する暇もなく,課金管理情報と二種類の相関に依存し続ける．本来なら研究対象として取り組むべき顧客データベースが,同時に火急のマネジメントすべき対象でもあり，中長期を視野に入れた研究環境を整備するモードより，経営意思決定というモードをどうしても優先させてしまうからである．

3. 経営慣習とマイニング手法の複合作用

データマイニングは，大量のデータを分析するために機械学習やパターン認識，人口知能，データベースシステム等の各分野からの解析技術を統合した広範な技術群である．異なる特性を備えたデータに対しても性能を果たすように,これらの幅広い分野からの技術を展開して様々なマイニング手法が生成されてきた．一方で，問題解決のため取捨選択されるマイニング手法は，常に限られた時間やコストといった現実的な事業局面でマネジされなければならないという制約の下にある．その際，取り扱うデータの性質に適切にフィットする方法の選択が，良好な解析結果を出力する最も重要な条件となる．しかし，事業者によるマネジメント慣習が解析アルゴリズムの背景にある概念特徴と結びつくと結果の精度に大きく影響することがある．業種を問わず顕著なマネジメント慣習とこれに結びついて誤謬や解析精度の低下を招くマイニング手法の概念特徴について，その関連を列挙していく．

(1) 解決目的と解決手段の錯誤

現実の経営決定では，問題解決の精度よりも速度が優先されることが多く，解析に十分な観測の時間や規模，サンプル収集の正しい手続きが無視されやすい．特徴数が疎であるほど高速かつ高精度となる単純ベイズ等の線形分類器を利用している現場では，特徴空間の把握に必要なサンプルデータの収集を怠ってしまう傾向がある．

(2) 例外なくアプリオリな事前データ

所与として訓練例を設定しなければならないサポートベクターマシン等の機械学習全般の特徴として，十分な観測やサンプル収集がなされていない事前データであっても，そのまま使用してしまい，結果として精度が下がる傾向をさす．特徴空間を所与とする，最尤法の概念を含む決定木なども同様で，統計分類全般について言える懸念である．

b) Judea Pearl によれば，科学的実験と統計的手法は，世界の反事実的状態を可能な限り近似することを主な目標の 1 つとしている．

(3) **所与のルールを設定する困難さ**

ロジットや重回帰等の回帰分析全般における前提としての仮説設定が，特徴空間である事前データの性質に影響を受け困難になる傾向がある．ニューラルネットワークにおける誤差逆伝播法や，ベイジアンネットワークにおける独立変数の定義等, 教師あり学習全般にも同様のことが言える．

(4) **教師なし学習とデータベースシステム**

データマイニングは本来，調査データから学んだ既知の特徴の予測ではなく，未知の特徴を発見しデータの背後にある本質的な構造を抽出する研究分野である．出力すべきものを予め定めない教師なし学習，特に未知の学習領域を開拓する強化学習を含む手法が脚光を浴びている．例えば自己組織化写像等のデータクラスタリングや，アソシエーション分析等の頻出パターン抽出がこれに相当する．これらの手法も実際には，解析対象となるデータベースが十分に蓄積され，問題を表す特徴空間と呼べる状態へ遷移するまで解析精度を発揮しにくい．

事業者のマネジメント慣習と複合して，解析精度の低下を引き起こす危惧を抱えたマイニング手法は，背景となる科学分野の境界を越えて多岐にわたり，様々な現実の事業局面に幅広く浸透している．2006 年に IEEE が発表した，データマイニング分野で大きな影響を及ぼしている上位 10 位までのアルゴリズム(1)においては, 十種類全てが解析プロセスや概念特徴の面で同様の特質に該当している．データマイニングとマネジメント慣習，特に国際放送コンテンツ事業の顧客データベースシステムが抱える本質的な課題は，あらゆるグローバル企業の意思決定プロセスの現状にとっても，潜在的な脅威なのである．

では潜在顧客の発見に最も貢献すると考えられるマイニング手法は何か？ Ngai, Eric WT, Li Xiu, and Dorothy CK Chau.による，CRM におけるデータマイニング技術に関するメタアナリシス(2)によれば，分析対象となった論文全体の約 62%が，線型分類器や頻出パターン抽出を使った顧客維持を研究主題としている一方，潜在顧客獲得に関連する論文は, 主に教師なし学習である自己組織化写像や決定木，あるいはクラスタリングを使用しており，論文数は分析対象全体の約 15%のみであった．現行の主要なマイニング手法，とりわけ教師あり学習や統計分類や回帰といった，事前データや仮説を所与とする解析手法は，こと潜在顧客の発見という問題に限り，適合しにくい可能性がある．

国際放送コンテンツの潜在視聴層検出問題とは，事前の訓練例を与えずに未知の特徴を発見する，データマイニングに特有の解析手法の開発と実践への適用である．そして同時に，不完全なサンプルデータベースをもっていかに交絡変数を見極め，特徴空間における疑似相関を回避するか？という挑戦でもある．

事業マネジメント慣習が及ぼすデータ品質への悪影響は，急速に進むデータマイニング手法の実用局面故に生じる新種の副作用である．本研究ではこの問題に，疫学分野等で活発に進むランダム化比較と因果則研究の角度からアプローチする．より精度の高い因果則を抽出する解析環境のアイディアとして，自由文解析に新しい価値と研究の可能性を付け加えようと目論む．そしてまたインダクションという仮説発見のための推論形式を通じて，知識発展の実践的な方法論を拡張する．以上が本研究の新規性である．

4. 研究の方法

本研究の狙いは，特徴空間からの交絡変数の抽出における仮説発見推論コンセプトの活用である．主にインダクション型の推論形式を使って，国際放送コンテンツ潜在視聴層を検出するデータ解析プロセスの実現を目指す．

1990 年代以降，帰納論理プログラミング(inductive logic programming; ILP)研究の発展によって注目されはじめた，インダクション型の仮説発見プロセス(3)を活用して，国際放送コンテンツの視聴行動との因果関係を説明する仮説を発見する．さらに実際の潜在顧客データの解析を通じ，背景知識の抽出と仮説の更新を行う．インダクションから導かれる背景知識や検証済み仮説は，それらが発見される前の時点では，不完全な特徴空間における交絡変数群に含まれていると考える．これらを抽出することによって疑似相関を制御する．さらに仮説を検証して実験計画を決定し，新たな観測を得て仮説を更新していくプロセスを CRM において実践する(2)．そして Peirce によって提示された知識発展の閉ループ手法の実現へ発展させていく(4).

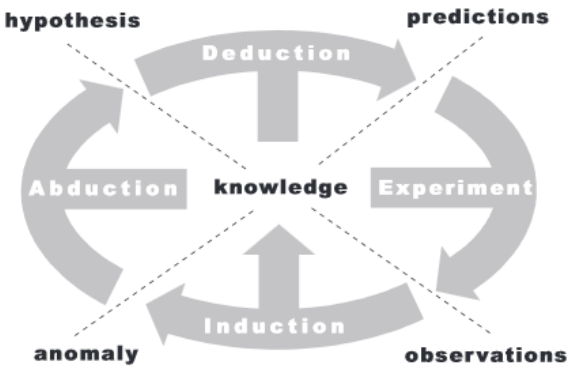


図 1　推論形式の閉ループ構造

インダクションの定式化(3)は次の通りである．

背景知識 B，仮説言語 $\mathcal{L}$，および観測式 O が与えられた

とき，論理式の集合 H が O の$(B,\mathcal{L})$に関する仮説であるとは，

(1)O は $B \cup H$ へ論理的に帰結する，
(2)$B \cup H$ が無矛盾である，
(3)H は$\mathcal{L}$から構成される，が成り立つことをいう．

例として以下の一階論理式を考える(s はソクラテスを表す定数，x は変数とする)

$F_1 = man(s)$,
$F_2 = mortal(s)$,
$F_3 = \forall x(man(x) \rightarrow mortal(x))$.

ここで仮説言語$\mathcal{L}$を言語全体とし，観測を $O=F_2$ とする．今，背景知識を $B_1=\{F_1\}$としたとき，O を説明する仮説として $H_1=\{F_3\}$を求めることがインダクションである．F_2 は $\{F_1\} \cup \{F_3\}$へ論理的に帰結するため，O も $B_1 \cup H_1$ へ論理的に帰結する．

これを前提として潜在視聴層検出問題を改めて次のように定義する．

国際放送コンテンツにおける視聴誘引における背景知識を B，仮説言語$\mathcal{L}$を顧客データ全体とし，データ解析結果を観測 O とする．このように定義される問題空間においてインダクションを用い，仮説 H を発見する．

本研究におけるもう一つの狙いは，マイニング手法の初段階で収集する解析データの改善である．具体的には，潜在視聴者や顧客の自発的な自由文を特徴変数と定義してデータベースをつくる．これにより仮に解析データに残余交絡（residual confounding）が存在し，この交絡が視聴行動の因果則を示す背景知識集合や仮説群であるとしても，解析データは当問題の特徴空間としてランダム化されていると見做すのである(5)．この仮定が妥当と考えられる理由を以下に列挙する．

(1) 通常解析対象となるデモグラフィックな独立変数に比べ，その未然已然を問わずコンテンツの視聴態度に関する自由文のほうが，視聴行動への因果則を反映している可能性が高い．仮にデモグラフィックな環境改善が視聴の直接誘因となるとしても，前提として視聴態度が既に形成されていなければならず，従って態度を反映した自由文からなる特徴空間の方が，より優れた独立変数を含むと考える．
(2) 視聴態度を反映する自由文全体を仮説言語$\mathcal{L}$と定義することで，当問題の特徴空間を，可能なあらゆる背景知識集合 Bと，あらゆる仮説集合 Hの両方を含む空間として設定できる．
(3) 仮に解析データが当問題における残余交絡を含んでいる状態だとしても，視聴態度を反映した自由文 $\mathcal{L}'$ の集合 Lがサンプルの収集段階でランダム化されていれば，観測 Oが帰結するべき $B \cup H$もランダム化されているはずである．
(4) 収集された複数のデモグラフィック変数を関連づけて生成する顧客個人別のデータベースに比べ，自由文という変数同士を関連づけて生成する視聴態度別のデータベースの方が，反事実的条件法に則して因果則を導きやすい．
(5) twitter や LINE 等のソーシャルネットワークサービスと，iPhone を始めとするスマートフォンの急速な普及が進んでいる現状，リアルタイムに収集できる行動履歴データとしての自由文に着目する解析プロセスは，現実の問題への適用性が広い．
(6) 経験的に，質問回答に見られる自由回答文に含まれる語彙の種類とその出現数は冪乗則にならう．従って，当件で取り扱う自由文集合はパラメトリックに取り扱うことができる．

5. 実験計画

国際放送コンテンツの潜在視聴層検出手法に関する解析比較の実証実験を試みる．まず実験上の質問に対する協力者の回答結果群を収集し，ここからリレーショナル型データベースをつくる．これを仮の顧客データベースと見做し分析対象とする．この同一の仮定の顧客データベースに対して異なる 2 つの分析手法を試みる(6)(7). 一方は自由回答文のデータ群に対して出現する，語彙同士のアソシエーション分析を行う．これは従来型のデータマイニング手法を代表する．他方はインダクション型の推論形式を援用した仮説の発見と，仮説に基づく自由回答文データ群への，特別集計結果の解析を通じて仮説の更新を行う．それぞれの分析から出力される知見を，潜在視聴層の検出という目的から比較対照する．分析結果から潜在視聴層のみを判別する手続きが導かれるか，そして，分析を経てデータベースから交絡を減衰させることに寄与したかを優劣判定の基準とする．この実験は，国際放送コンテンツの放映権者であるテレビ局と全国規模のモニターを運営する調査会社との協力によって行われる．

(1) 実験対象である国際放送コンテンツ

モータースポーツ分野であるフォーミュラ 1 世界選手権をとりあげる[c]．近年，開催国が急速に全世界規模に拡大し

c)フォーミュラ 1(通称 F1)は，ヨーロッパを中心に世界各国を転戦する最高峰自動車レースである．2014 年は暫定で，22 か国での全 22 戦開催となっ

ているため，文化圏によって競技人気にばらつきが生じている．競技観戦に関連する語彙に専門用語を多く含み，他競技からの知識流用が困難なため，自由回答文データの質を高く保持できる．

(2) 実験対象である調査協力者

協力者の招集地域を日本全国とする．調査会社が保有するインターネット調査モニターから，各歳別総人口いわゆる人口ピラミッドを全射した約 50000 人の母集団に対して，「フォーミュラ 1 世界選手権というスポーツを聞いたことがありますか」という質問を提示し，肯定回答を得た協力者を F1 認知層と定義する．この F1 認知層からランダムに抽出された 15 歳～69 歳の男女協力者 5000 名(指令ベース)を実験対象とする．実際のコンテンツ視聴経験の有無は問わないため，この集団にのべ潜在視聴層が含まれていると考える．非認知層から得られる F1 に関する自由回答文は実験上無効と考え，実験対象から除外する．

(3) 実験を行う時機

1 月上旬とする．F1 世界選手権の開催は近年，その年の 3 月中旬から 11 月中下旬までの期間に定まりつつある．シーズンオフに当たる 12 月から 2 月のうち，当年開催結果への振り返り報道と次期開催への予習報道が行われる時機を除外し，実験時点での協力者の自然想起内容に偏向が生じないように配慮する．

(4) サンプルデータの収集方法

「フォーミュラ 1 世界選手権について思い浮かぶワードをご自由にいくつでも書いてください(最大 10 個まで)」 という質問提示に対して回答される自由文を収集する．回答数の上限数を定めない方がより多くサンプルを収集できる利点がある一方，協力者の回答負荷が高まり結果的に精度を損なう可能性がある．またデータベースとしての解析容量の規模から上限を設けることもできる．今回はこれらの点を勘案して最大 10 個とする．インターネット調査による．自由回答文をサンプルデータとする場合，筆記式アンケートに比べ回答結果のデジタル化処理過程をスキップできる．収集データは CSV ファイル形式で保持する．

(5) サンプルデータの収集手順

サンプル収集に使用する質問画面は，最大 20 文字入力できる回答欄を，最大で 10 回画面更新する形で提示し，ここに記入させる．記入可能な文字数の上限は，そのコンテンツ領域の関連語彙の最大文字数がどの程度かを勘案して定める．回答中においては質問提示画面以外のいかなるページへも遷移させない．また回答時間にも制限を加え，全体で 10 分として一回答欄当たり 1 分程度とする．一旦回答を確定したら次の回答欄へ自動に不可逆に遷移させる．実験協力者がそれ以上想起できないと感じた時点で，いつでも質問回答を終了できるようにする．これらの措置は，回答途中で知識を他から流用して回答する事態を可能な限り防ぐためにとられる．

(6) 付加的な収集内容

現実の顧客データサンプルの収集と同程度のデモグラフィック属性について，自由文回答終了後に入力を要請する．性別・年齢・居住地・可処分所得・未既婚・家族構成・子供有無等についてである．

(7) データベースの生成

回収したサンプルデータを形態素解析し，語彙単位のデータへと整備する．これにより回答欄に文章が入力されていても，そこから有効サンプルを抽出できる．

次に専門家(本実験ではテレビ局)と検討して作成した関連語辞書とサンプルデータを照合させ，サンプルデータをクレンジングする．明らかなタイプミス，欠損値等以外は最大限協力者の自由文をなるべくそのまま新規に定義する方針をとり，忘却や誤認を含めて自然想起される知識体系の実相を保持できるように最大限配慮する．

こうして準備された語彙群に，その語彙を回答した協力者のデモグラフィック属性と所要回答時間などの回答履歴とを連結し，語彙毎のリレーショナルデータベースを生成する．これを実験上の顧客データベースと見做す．

デモグラフィック属性については，本実験上必須の属性ではないので，協力者(今回は調査会社)のプライバシー管理基準に照らして妥当な水準の内容を保持しておく．少なくともデータベースから具体的な個人を特定できないようにする必要がある．

(8) 分析結果の優劣判定手順

事前に専門家(本調査ではテレビ局)と検討し，注目ワードを複数定義しておく．この注目ワードは，F1 中継を視聴し続けることで初めて興味・関心をもって認識されるワードで，必ずしも競技自体に関する語彙だけで構成するものではない．これらの注目ワードは実験協力者，分析担当者に対して全く開示されない．注目ワードが分析結果にどの程度反映されているか，そして分析結果に視聴態度に関する知見が含まれているかどうかを優劣判定基準とする．

ている．オリンピック，FIFA ワールドカップと共に世界的な人気も高いが，近年では景気後退によるスポンサーの撤退や開催費用の負担などから，最盛期に比べ縮小傾向にある(8).

6. 実験結果

6.1 結果全体

実験協力依頼者全員（指令数）5,000 に対して，有効回答となった協力者数（回収数）は 4,641 であった．この数が本実験において，F1 世界選手権での関連語彙同士の組み合わせ出現数の総合計である．4,641 パターンの組み合わせの中で，データクレンジング後に関連語彙として定義したワードの総合計は 21,173 語であった．

6.2 アソシエーション分析の結果

組み合わせ総数 4,641 パターンを対象とした，アソシエーション分析の結果のうち，評価指標である 支持度・確信度・リフトの上位を示す(9).

NO	lhs	⇒	rhs	support	confidence	lift
1	{シューマッハ}	⇒	{セナ}	0.069	0.645	9.346
2	{セナ}	⇒	{シューマッハ}	0.069	0.445	6.452
3	{ホンダ}	⇒	{セナ}	0.042	0.472	11.236
4	{セナ}	⇒	{ホンダ}	0.042	0.271	6.452
5	{フェラーリ}	⇒	{セナ}	0.032	0.386	12.048
6	{セナ}	⇒	{フェラーリ}	0.032	0.206	6.452
7	{ホンダ}	⇒	{シューマッハ}	0.030	0.337	11.236
8	{シューマッハ}	⇒	{ホンダ}	0.030	0.280	9.346
9	{フェラーリ}	⇒	{シューマッハ}	0.027	0.325	12.048
10	{シューマッハ}	⇒	{フェラーリ}	0.027	0.252	9.346
11	{車}	⇒	{セナ}	0.023	0.767	33.333
12	{フェラーリ}	⇒	{ホンダ}	0.023	0.277	12.048
13	{ホンダ}	⇒	{フェラーリ}	0.023	0.258	11.236
14	{セナ}	⇒	{車}	0.023	0.148	6.452
15	{トヨタ}	⇒	{セナ}	0.020	0.377	18.868
16	{セナ}	⇒	{トヨタ}	0.020	0.129	6.452

図 2　F1 自由回答文アソシエーションルール評価

評価指標 [d]上位のアソシエーションルールの中には，注目ワードは殆ど含まれていなかった．次に注目ワードを含むルールのうち，評価数値が高いルールを幾つか列記する．

NO	lhs	⇒	rhs	support	confidence	lift
1	{モナコGP ☆}	⇒	{佐藤琢磨 ☆}	0.009	0.220	24.390
2	{佐藤琢磨 ☆}	⇒	{モナコGP ☆}	0.009	0.161	17.857
3	{鈴鹿GP ☆}	⇒	{モナコGP ☆}	0.006	0.207	34.483
4	{モナコGP ☆}	⇒	{鈴鹿GP ☆}	0.006	0.146	24.390
5	{鈴鹿GP ☆}	⇒	{佐藤琢磨 ☆}	0.005	0.172	34.483
6	{佐藤琢磨 ☆}	⇒	{鈴鹿GP ☆}	0.005	0.089	17.857
7	{鈴木亜久里 ☆}	⇒	{佐藤琢磨 ☆}	0.002	0.077	38.462
8	{佐藤琢磨 ☆}	⇒	{鈴木亜久里 ☆}	0.002	0.036	17.857
9	{鈴木亜久里 ☆}	⇒	{鈴鹿GP ☆}	0.001	0.038	38.462
10	{鈴鹿GP ☆}	⇒	{鈴木亜久里 ☆}	0.001	0.034	34.483
11	{鈴木亜久里 ☆}	⇒	{モナコGP ☆}	0.000	0.000	–
12	{モナコGP ☆}	⇒	{鈴木亜久里 ☆}	0.000	0.000	–

図 3　注目ルールを含むアソシエーションルール例

上位ルールの評価数値に比べて，改めて統計検定しなければ有意であるか判定できなかった．また，協力者のデモグラフィック属性別に分類した後のパターン群における分析結果も全体と同様の傾向だった．

d) 支持度 $supp(X⇒Y)$は，$X⇒Y$ にしたがう組み合わせの数 $\sigma(X \cup Y)$の出現数が組み合わせの総出現数に占める比率をさす．確信度 $conf(X⇒Y)$は，$\sigma(X \cup Y) / \sigma(X)$ をさす．リフト $lift(X⇒Y)$は，$conf(X⇒Y) / supp(Y)$ をさす．

6.3 インダクション型推論形式を援用した仮説発見～更新による分析の結果

6.3.1 初期仮説の発見（第一インダクション）

実験協力者のうち，注目ワードの設定にも携わった専門家(テレビ局の F1 国際中継制作部門のプロデューサー5名)に，抽出されたデータベース全体 4,641 パターンを列記した表，21,173 ワードを出現率の降順に列挙したもの(図 4)を提示し，自由に討論してもらいその発言内容から，F1 国際中継番組と視聴者の番組認知との関係における知見を抽出した．

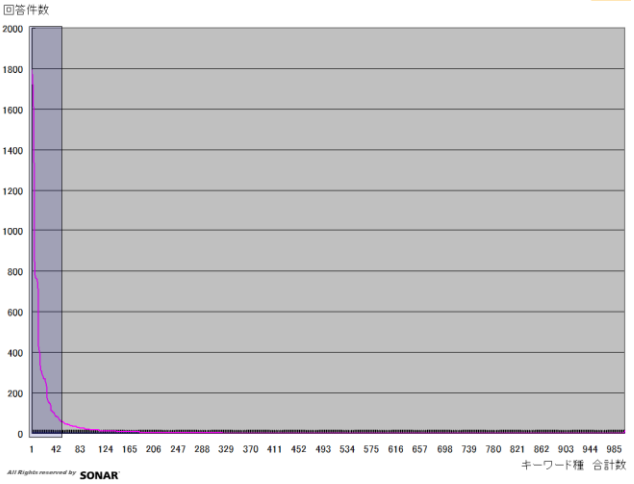


図 4　自由回答ワード別出現率

列記された知見群を専門家グループに提示して，知見群の背後で共通に関連づけられるワード群を改めて各自列記してもらった．このワード群の中から発見した初期仮説は以下である．

F_4 ：スポーツリテラシーを備えているか否かと，国際放送コンテンツの満足度の間には関係がある[e].

6.3.2 初期仮説を利用したデータ解析

発見した初期仮説 F_4 を，背景知識 B_1 とした．

B_1 に関連づけて自由回答文データを観測するためにデータベース内の独立変数を検討した結果，実験協力者の回答履歴についての変数である

「その人は何個のワードを回答したか(最大回答数)」，

「そのワードが回答時間経過の中で何番目に回答記入されたか(回答順位)」，

「そのワードはどれだけ語られたか(語彙頻出度)」

の 3 つに着目することとした．

e)スポーツリテラシー(Sports Literacy)とは，観戦コンテンツに込められた環境情報，隠れた意味やメッセージを見出，読み解き，堪能する能力のこと．

（最大回答数）×（語彙頻出度）の相関(図 5)と，（回答順位）×（語彙頻出度）の相関(図 6)について図示する．

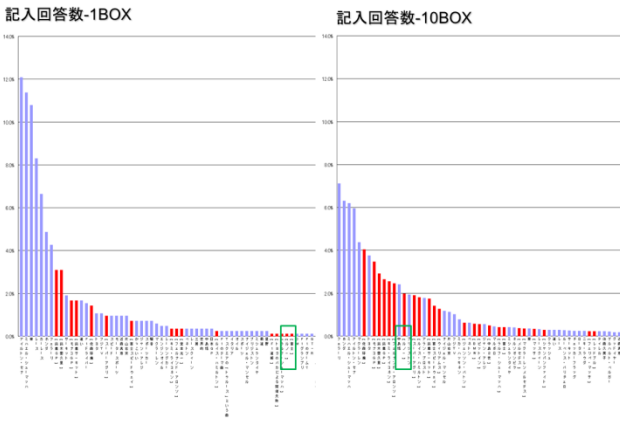


図 5 （最大回答数）×（語彙頻出度）の回答数 1 対 10 比較

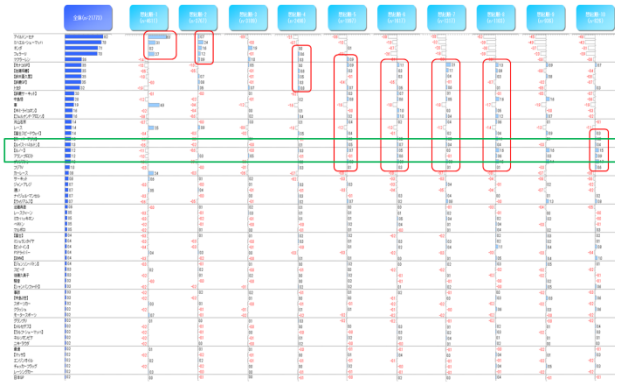
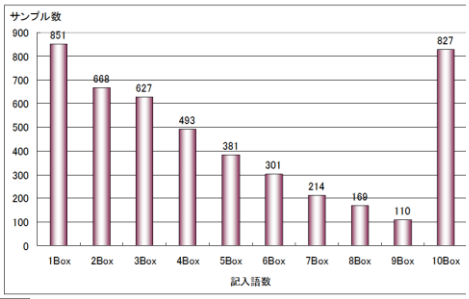


図 6 （回答順位）×（語彙頻出度）の相関

F1 に関連していて頻出度の低い語彙ほど最大回答数が高くなる傾向がみられた．また，F1 に関連していて頻出度の低い語彙ほど，回答順が遅くなる傾向もみられた．逆に，頻出度の高い語彙ほど最大回答数は少なく，かつ回答順位が早い傾向がみられた．また最大回答数別に，回答者の出現率をみると最大回答数と反比例している．



記入語数	サンプル数	%
1Box	851	18.3%
2Box	668	14.4%
3Box	627	13.5%
4Box	493	10.6%
5Box	381	8.2%
6Box	301	6.5%
7Box	214	4.6%
8Box	169	3.6%
9Box	110	2.4%
10Box	827	17.8%
総計	4641	100.0%

F1(Formula 1)と聞いて思い浮かぶキーワードを何個まで書けるのか

図 7 最大回答数別の回答者出現率

最大回答可能数である 10 個回答したサンプルは全体の 17%いるが，回答数の上限をさらに引き上げればそれだけ，この分布は線形に近づくと考えられる．

語彙毎に見て，そのワードがどの最大回答数に多く含まれていたかを示す．

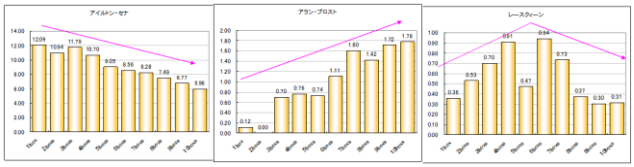


図 8 語彙別最大回答数の比較

語彙の種類とは概ね，最大回答数と正比例するもの，反比例するもの，中間的な回答数で多く出現するもの，の 3 種類であった．

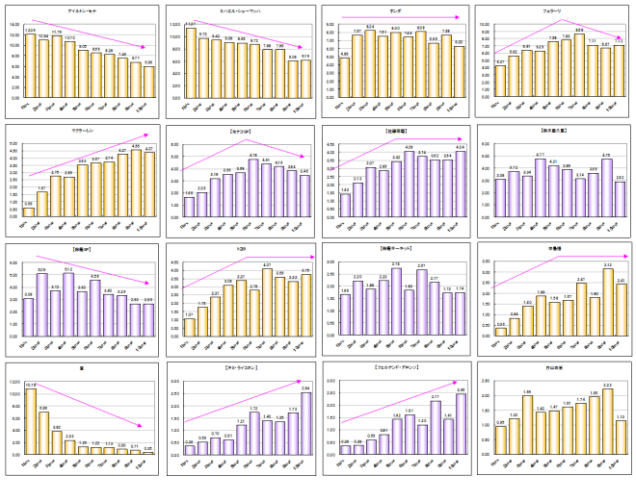


図 9 語彙別最大回答数(青は注目ワード)

6.3.3 注目ワードとの照合

最大回答数と正比例する語彙群と，山型に分布する語彙群に注目ワードが多く含まれていた，また注目ワードは回答される順位が低く遅い順番で出現しやすい．

6.3.4 仮説の更新（第二インダクション）

以上から得られた観測 O_i $(i = 1, \ldots, 4)$ を列記する．

O_1: リテラシーの程度に応じた関心領域がある．
O_2: リテラシーが増すと観戦が楽しくなる．
O_3: リテラシーが増すと誰かにそれを伝えたくなる．
O_4: 知っていることはなるべく多く話したくなる．

第一インダクションと同じ手続きで，背景知識 B_1 にお

いてこれらの観測群から，新たに仮説を導き出す．
更新した仮説は以下である．

H_2. スポーツリテラシー多寡と視聴態度は比例する．
H_3.関心を喚起するコンテンツ事象には視聴態度の成長に応じて遷移していく．

7. 考察と結論

実証実験の結果，潜在視聴層の検出問題においては，インダクション型推論形式を援用した自由文テキストマイニングの方が，より優れているといえる．

解析過程からの出力データ中の注目ワードも，インダクション側により多く含まれており，この事象を視聴層検出への具体的なプロセスにつなげることに成功した．また，スポーツリテラシーという説明変数の発見は，日本のみならずグローバルでの視聴層検出，ひいては CRM への応用が可能である．更新仮説 H_2，H_3 をもとに，スポーツリテラシーに準じる特定の語彙群をマーカーとして，潜在視聴層を検出できる可能性と，さらに閉ループ手法の活用により仮説を更新し続けることで確実に交絡の回避を前進させられることも判明した．そしてその際に，自由回答文データベースを保持していれば残余交絡問題を懸念せずに済む．

自由文テキストマイニングにおいても，語彙を解釈する技術のみならず，語彙の発生履歴について定量解析する重要性が示唆された．より日常的なライフログとしての自由文データベースの重要性は高まっていくことだろう．

アソシエーション分析結果からは，残念ながらワードの連関という知見のみでは視聴層検出に近づけないと示唆された．析出したアソシエーションルールを専門家に提示しても結果は同様で，直接の問題解決には貢献しなかった．

残された課題は，今回の実験での課題解決への決め手が，依然として専門家の属人的技能の活用に依存していた点である．いかにメタレベルで，初期仮説の更新へ向けた解析の切り口を最適化できるか．そうした解析に適うサンプルデータを，限られた資源でいかに整えるか．そして今回の実験結果はそのまま，アソシエーション分析を始めとする主要なデータマイニング手法の有用性に疑問を投げかけるというものではない．問題の特質に応じた手法選択の重要性を改めて示唆している．

本研究において何よりも重要な発見は，因果則研究と自動推論とテキストマイニングとの邂逅による学術と事業両面における機会の拡がりである．また，現実社会における実践的技能や蓄積を，学術的に再評価する大切さである．そうした取り組みこそが学術分野とビジネス分野の双方に有益なシナジーを生んでいくと確信するものである．

参考文献

1) Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and Information Systems 14.1 (2008): 1-37.
2) Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert Systems with Applications 36.2 (2009): 2592-2602.
3) 井上克巳. "アブダクションとインダクション (< 特集> 論理に基づく推論研究の動向)." 人工知能学会誌 25.3 (2010): 389-399.
4) Ray, Oliver. Hybrid abductive inductive learning. Diss. University of London, 2005. Peirce, Charles Sanders. Collected papers of charles sanders peirce. Vol. 3. Harvard University Press, 1974.
5) Greenland, Sander, James M. Robins, and Judea Pearl. "Confounding and collapsibility in causal inference." Statistical Science (1999): 29-46..
6) Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." ACM SIGMOD Record. Vol. 22. No. 2. ACM, 1993. .
7) Brin, Sergey, Rajeev Motwani, and Craig Silverstein. "Beyond market baskets: generalizing association rules to correlations." ACM SIGMOD Record. Vol. 26. No. 2. ACM, 1997.
8) 「フォーミュラ 1」『フリー百科事典 ウィキペディア日本語版』(http://ja.wikipedia.org/)。2013 年 11 月 24 日 8 時（日本時間）現在での最新版を取得。.
9) Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." ACM SIGMOD Record. Vol. 22. No. 2. ACM, 1993..

謝辞 本研究にご協力頂いた皆様に，謹んで感謝の意を表する．