

Special Edition

定性情報のマイニング —自由回答データの解析—

大隅 昇 | Ohsumi Noboru

統計数理研究所調査実験解析研究系
パターン解析研究部門教授



■1972年、文部省統計数理研究所、第4研究部研究員。79年理学博士取得。85年同調査実験解析研究系助教授。91年より現職。

1. はじめに—問題の背景—

各種の調査環境の急速な変化が指摘されるようになって久しい。とくに最近は質的調査や定性調査への関心が高まっている。従来の調査、とくに社会調査（例えば面接調査、留め置き調査）では、適度にサンプル数が大きく、また伝統的な標本調査法に従ったサンプリング操作を経て行われる量的な調査が主体であった。しかし、経済的にも労力確保の面からも負担が大きく、また質の高い調査内容が得られにくくなってきたことから、質的調査や定性調査に関心が移行する傾向が見られる。

こうした現れの一つとして、グループ・インタビューやモチベーション・リサーチ等が改めて注目されている。また、少数のサンプルや条件を絞り込んだ回答者を相手に、インターネット調査等を用いて、自由回答や自由記述の設問を設け、これで取得したデータの質的解析を試みる等の例が多くなってきた。

とくにインターネット環境を利用することで、自由回答に代表されるテキスト型データ、文章化データの取得が、内容の質の良否に関わりなく、容易に、しかも大量に取得可能となり、これを多用することが多くなった。

また、従来からの定性調査手法として、面接法（深層、集団）、投影法（言語連想、文章完成、略画完成、絵画解釈）等があったが、こうした方法で取得したデータも、電子テキスト化されることが多くなり、文字情報としての利用が容易になってきた。この他、POSデータや移動体通信（携帯電話、データロガー等）によるデータ収集、さらにはコールセンターにおける消費者・顧客管理のデータ取得技術やデータベース化技術の進歩により、様々な場面でテキスト型データが見られる。関連研究分野としては、自然言語処理・言語情報処理、全文検索他、計量言語学的な研究、計量文献学のアプローチ、人工知能的なツール（発想支援ツール等）の開発と多様である

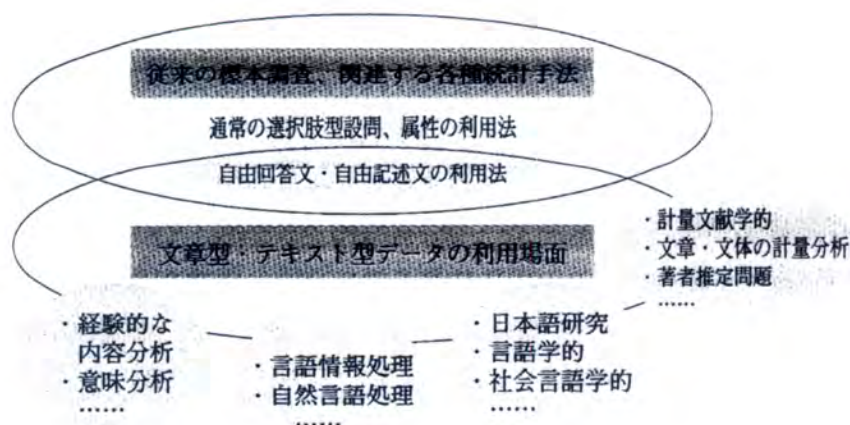


図1 自由回答の解析の位置づけ

が、言うまでもなくその内容にはかなりの違いがある。

調査における自由回答の利用方法や分析方法は、こうした関連研究との接点はあっても、以下に述べるように、必ずしも同じアプローチが可能とは限らない。適切な表現を直ちに思いつかないが、社会調査における統計的な方法論としての自由回答の利用法と上記の各種分野との関係を、模式的に表すと、図1のようになるであろう。筆者の考える自由回答の研究とは、この図に示すように、従来の統計科学やデータ科学を基礎とした科学的な調査法との接点において考察されるものと考えている。

今後は、電子的データ取得環境の進歩だけでなく調査環境も多様化し、それに伴って文章型・文字型データの取得や解析の機会が益々増えるものと考えられる。ここでは、主として自由回答設問で得られるテキスト型データの解析方法について取り上げよう。

2. 自由回答と定量的アプローチとの関連

調査で用いる自由回答とは何かを知るには、

通常の利用型設問と対比することが容易である。自由回答を open-ended questions、free answer 等というように、設問への回答が自由記述的に書かれるデータのことである。これに対して、選択肢型設問 (close-ended questions) とは、あらかじめ選択すべき選択肢を設問中に用意し、回答者がこれから選択する方式

である。また通常の利用型設問の中に、自由回答欄や「その他」欄を設けて意見・回答の記入を求める方式もある。しかし、自由回答としてどのような設問方法が適切であるかの検証が十分になされないまま、選択肢として明示的に表すことが難しいとき、あるいはそれを思いつかないとき等に、回答者が何か他のことを考えたり予想しないアイデアを持っていることを想定して、安易に自由回答設問を設けることが多いであろう。

しかし、このような自由回答取得の考え方に少し誤解がある。自由回答とは回答者が意見を自由勝手に書き入れることではない。選択肢型設問だけでは引き出すことが難しいとき、しかも調査者側が取得したい内容、意見を回答者からの確に引き出し書き入れてもらうことが目的である。従って、白紙を示して「何でも自由に…」といった方式はあまり適切とは言えない。自由回答取得のための設問構成・調査票のあり方や設問方法の研究があってしかるべきと考えるが、現実には、この種の研究は、さほど進んでいるとは言えない。

例えば、通常の利用型設問として「コン

ピュータはこれからの生活になくてはならないものと思うか」との設問に「そう思う」「そうは思わない」と選択肢を設けると、8割近くの回答者は「そう思う」を選択する。これに加えて、「ではそう考える理由は」と自由回答で問うと、「思う」と「思わない」という意見の中で、さらにきめ細かい意見聴取が可能となる。実はこれは筆者等が実際に用いた設問例であるが、「思う」の選択者の中でも前向き意見と消極的意見がどう混在するかを自由回答が教えてくれる[3]。この分析の結果から、自由回答の内容を類型化し、新たな選択肢型設問の作成時の情報とすることも可能である。また、自由回答方式で取得したデータの統計解析法の研究もそれほど行われているとは思われない。とくに自由回答方式の設問では、以下のような特徴があるとされてきたが、いずれもが尤もな指摘であろう。

- ①考えたことがないことには答えにくい、いきなり設問を受けても答えることが難しい。
- ②一方、予想しなかった回答や知見がえられるという期待がある。
- ③無記入が多くなる傾向があるとされてきた。
- ④調査法や標本抽出法との関連性が明らかにできないと言われている（妥当性の問題）。
- ⑤通常の他の設問（選択肢型）の選択肢の影響を受けるのではないかと（回答誘導の懸念）。
- ⑥客観的かつ適切なデータ解析法がないと言われてきた。
- ⑦回答に均一性を欠くと言われている、再現性がない、あるいは信頼性に欠ける（調査の反復可能性）。
- ⑧集計や解析の手間がかかると考えられている。
- ⑨計量的に表現されないため、解釈が恣意的となる（定量化の困難性）。
- ⑩設問文の意図がどう反映されたか、客観的には読みとり難い（解釈が主観的となる）。

ところで、調査における自由回答データ取得で重要とされることは、選択肢設問型調査

と異なり、収集データから数量として定量的かつ客観的な評価が困難であるとされてきたことにある。例えば、回答比率で比較したり、標本誤差を推定するなどの操作に相当する方法がない。標本調査であれば、その標本抽出との関連や調査票や設問の設計をどう考えるか、さらに、調査の反復可能性や再現性の問題をどう捉えるか（通常の回答変動とは異なる事象が見られるであろう）等々、様々の検討事項がある。このような数理的要件に照らすと、いずれも満たされているとは考えられず、正当な評価が得られにくいということがある。一方では、前述のように調査環境の多様化や変化を意識して、従来型の設問法による定量的アプローチよりは、定性的情報として回答者個々人の意見をきめ細かく取得し分析することが望ましいといった意見も多く見られる。

要は、定性調査を定量的調査との優劣や利便性だけで比較するのではなく、またそれぞれを別のことと考えるのではなく、調査における“自由回答データの取得方法”の検討や、従来の選択肢設問形式による調査との“併用”が可能であるかの観点から議論し、適切な方法論を検討することである。従来の選択肢設問形式を定量的評価に用い、一方、自由回答は、その定性情報の計量化を図る方法論を考え、それら両者を併用することで、より客観的な情報要約を探索的データ解析の観点から進めることが肝要である。統計システムの開発に際しても、こうした発想が反映された設計指針が必要とされる。分析事例で用いるInfoMinerは、こうした視点から筆者等が開発を進めてきたソフトウェアである。

3. 自由回答文の処理上の問題

日本語が欧米の言語と大きく異なることの一つに、文章・テキスト型データが「べた書き」（膠着言語）であって「分かち書き」されていないということがある。さらに、日常用いる日本語は漢字、カタカナ、ひらがな、それに外来語（それに充てられた漢字や仮名等）が混在している。しかも、現実の調査で得られる自由回答文は書き方が多種多様であって、これに日本語としての文法規則や記法の概念をそのまま適用して内容分析を行うことはほとんど不可能にちかい。とくに自由回答は整った文章として取得可能とは限らない（回答者は勝手に書く）。

一方、欧米語は「単語」という単位で区切られた言語であることから、個々の単語の抽出や処理が単語を基本単位として扱うことができる。日本語はこれが困難であるだけでなく、複数の語が連結されて複合語を形成することも多い。

こうした状況は、同じく日本語を扱う計量文献学や自然言語処理・言語情報処理等にも共通したことである。しかしこれらと比較してかなりの差異がある。例えば、計量文献学では、初めに素材（分析対象）が与えられていること、しかし素材が限定され特定の著者の少ない著作物であるとか、文芸作品であれば一点しかないとかの制約がある。場合によっては、日本語の歴史、文法、用法・語法等の体系的な知識の援用を必要とする。

また、自然言語処理では、表記の構造を形態素解析¹⁾、構文解析により確認し、続いて意味解析、意味理解といった操作が行われる。

いずれにしてもコンピュータ処理の支援は避けられない。しかも一般には相当量の計算処理時間を要する。

一方、調査の世界では、昔から自由回答で得た情報整理を、アフターコーディングあるいはポスト・コーディング（post-coding）などの手作業で行ってきた。例えば、グループ・インタビューの発言内容をコーダーが解釈・分類して、これをコーディングにより通常のカテゴリー型設問に置き換える等である。しかし、

- ・コーダーの偏りが入る
- ・表現の内容、意味などが別の形に変質するおそれがある
- ・元来の曖昧さ、複雑さ（を重視するなら、その特徴）が損なわれる
- ・解釈が複数のカテゴリーにまたがり判断が難しく、適切な分析方法もない

等の理由から慎重な処理と経験が必要とされてきた。また分析結果に主観的要素が入るという懸念もある。

4. 別の観点からの接近法

ところで、自由回答データの解析を統計的データ解析（とくに記述的・探索的多変量解析）のパラダイムの中で考えてみる。調査で得られる自由回答データには多量のゴミ（ノイズ）が含まれるから精密なモデリングや方法論を展開してもうまい結果は得られないであろう。データ・マイニング風に言うならば宝の山を掘り当てるには、まずは適切なデータ取得方法の研究があってしかるべきと言うのが筆者の主張の骨子である。

¹⁾ 形態素解析の方法論やその周辺技術として多数の解析方式が提案されているが、ここでは触れない。

電子的にテキスト・ファイル化した自由回答文を分かち書き処理により「構成要素」(fragments)に分解する。ここで構成要素とは言語学的な単語・語としての解釈が難しいことを想定した曖昧な意味である。このとき形態素解析に関連した処理技法を用いるが、これの確定的な方式は未だない。そこでテキスト化された日本語文章を(どのような方式であれ)一定の方式で「分かち書き」した各単位を「構成要素」と名付け、これを適当に分解した文(分かち書き文)という程度の緩やかな約束と考えておく。緩やかな決まりとする理由は、元来取得データ自体が持つ曖昧さや多様な表現を許容して、厳密な定義や拘束を避ける方向で分析を進めることにある。要は自由回答だけではなく、通常の選択肢型設問や属性、その他の付帯情報を総合的に併用・突合して解析するための、自由回答データの取得・解析法のあり方の総合的な研究という枠組みの中で考えようというのである。こう考える理由や根拠が多々あるが、ここでは多数の自由回答に関する実験調査を通じてこのような考えに至ったとだけ述べておく。

事例の紹介

InfoMinerを用いた解析事例は既にいくつもあり、ある企業ではこれをエンジンとした定性情報統合解析システムの構築を試みている[4]。ここでは、筆者が複数の調査会社の協力を得て実施したWeb調査から取得した自由回答データの分析例を取り上げる[2]。用いた設問は、他の調査で用いられてきた良く知られた設問を利用している。こうする理由は、既存調査の結果とWeb調査のそれとを比較した

いという意図があるからである。実際に両者の調査で取得したデータを比較分析すると、なかなか興味ある差異や類似が見られるのであるが、その紹介は別に譲り、ここではごく簡単な例を示そう。

[用いる設問]

問1-1: あなたが好きで、よく食べる料理は何ですか?

問1-2: 他に、あなたが好きでよく食べる料理は何ですか?

問1-3: あなたにとって「理想的な食事」とはどんなものでしょうか。

問1-4: 他に、あなたにとって「理想的な食事」がありますか。

この設問は飽戸弘氏他による研究調査で用いられたものに同じである[1]。Web上にアドオンされた調査票にこれらの設問を示し、各設問の後ろに記入欄を設ける。回答者はキーボードから自由回答を記入し、調査者に返信する(調査票はhtmlで記述しCGIを利用してバックグラウンドでサーバ上にデータを取得する)。Web調査票の設計に際して、通常の調査票(質問紙)の作成時に必要となる留意事項の他に、Web調査特有の配慮も必要となる。自由回答では、記入欄の大きさ、デザイン、記入欄のスクロールバーの有無等々である。こうした調査票設計が回答にどう影響するかの研究は十分にはなされていない。

ここでは、InfoMinerによる解析結果の一部を紹介するが、自由回答の分析処理でもっとも手間取る作業は、取得データのランドリや再編集にある。通常の数値型データであってこの種の作業は必須であるが、自由回答では次ページ囲みのように分かち書き処理の前後で様々の手当を必要とする。

- ・元の自由回答文（原文）の確認：誤記、日本語変換のミス等、Web調査であれば回答の入力中断エラー等の点検が必要である。
- ・確認済みの原文の分かち書き処理：分かち書き処理自体は自動化されているが、分かち書きで得た結果の再吟味が必要となる。分かち書きのミス（辞書照合の不一致等）、辞書にない言葉の扱い、人の判断を必要とする判定、類語の括り、同義語や多義語の扱い等である。つまり自由回答の記述内容が曖昧であることから生じるやむをえない手当である。
- ・分かち書きファイルの他、キーワード・ファイル、分かち書き・キーワードの回答者別出現頻度数ファイル等を必要に応じて解析対象として利用する。
- ・これら解析対象ファイルについての統計解析の処理。例えば、総単語数、異なり単語数、異なり率等の統計値の算出や比較を行う。

この他にも様々な細かい手当が必要とされる。このような過程を経てから、初めて多変量解析的手法による分析、それらの結果の視覚化等を行う。コンピュータ利用環境が非常に良くなったことから、PC程度でも十分に対応ができる。また、実用的なソフト構築にはPCでの処理を前提とした設計指針が必要である。

(1)事前処理、初動探査

①回答者の特性

まず、調査対象とした回答者の構成について簡単に述べる。調査法としてはWeb調査を用い、「食生活に関わる意識調査」の多数の設問の中に、前述の設問をおいた。回答者の総数は1,111名、うち男性は659名、女性は452名である。年齢区分構成は後の分析で示す。

②分かち書き処理の結果

分析対象としては上記設問の問1-1、1-2の二つの自由回答を併せて用いる。この原文の分かち書き処理を行うと様々の情報が得られるが、ここでは一部を示す。図2は分かち書きで得た総単語数、異なり単語数と異なり率の情報である。ここで言う「単語」とは前述の意味での構成要素を言う。これは句読点、助詞その他自由回答に現れた記号等々のすべてを含む。異なり単語とは同じ表記の単語を1語とした場合の頻度である。また、閾値とはその指定した値（閾値）以上の単語数を計数したものである。従って閾値が1の場合は全単語ということになる。図表に見るように総単語の場合の異なり率が15.7%であり、閾値の変化に対して異なり単語数（率）は急速に低減することが分かる。多くの例で、類似した傾向を示す。

③単語の編集

次に、単語の頻度集計や長さの分布等を観察し、不要単語の削除、類語・同義語の併合、誤記、分かち書き結果の再編集などを行う。

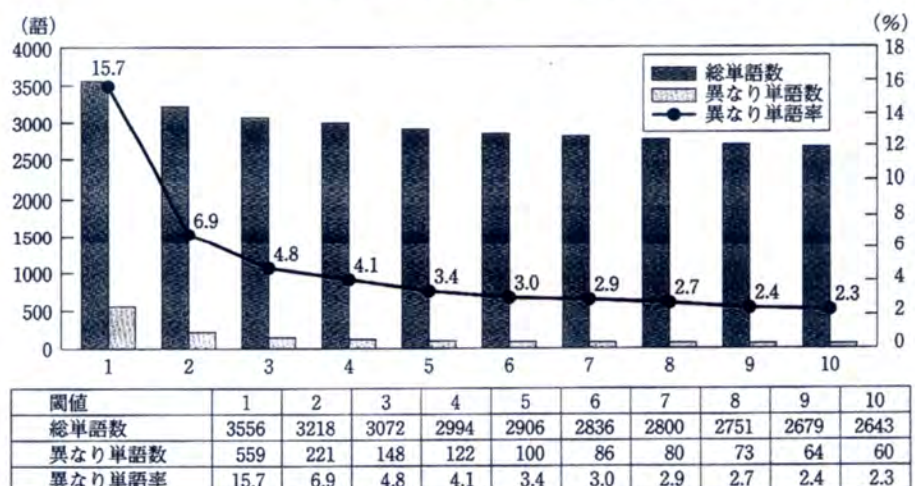


図2 単語の出現頻度の分布（「好きな料理は」）

こうした操作を一切行わないで直ちに多変量解析的な処理を行いデータ構造の大まかな特徴を知ることも初動探査としては重要であるし、InfoMinerではこれに対応する機能がある。ここで削除と置換の簡単な例だけを見る。なお、削除、置換ともに無数にあるので、そのごく一部を示した（いずれも原文のまま）。

[削除の例]、。・：；？！^_ーちゃん～…（
）「」＋－＝＆１１００％のがにをとでな
こともはやかです特なし

これらの文字、記号を削除した。特殊記号や変わった表記（フェースマーク等）、あるいは「特になし」などが対象となっている。

[置換の例]

鶏肉＝鶏、チキン、若鳥、鳥肉、鳥肉類 鳥
漬物＝お漬物、おつけもの、漬け物
お好み焼き＝好み焼 お好み焼
ごはん＝ご飯、ご飯類
スパゲッティ＝スパゲッティー、スパゲッテイ、ス
パゲティ、スパゲティー、スパゲティイ、スパゲティ
丼物＝丼もの、丼モノ、丼料理
刺し身＝刺身、おさしみ、お刺し身、さしみ
何でも＝なんでも
焼き肉＝焼き肉系、焼肉
惣菜＝総菜
天ぷら＝天麩羅、てんぷら
とんかつ＝カツレツ、豚カツ、トンカツ
餃子＝ギョウザ、ぎょうざ、ぎょーざ
メキシカン＝メキシコ、Mexican
エスニック＝エスニックフード、エスニック系
イタリア＝イタリアン、イタリア系、itariann、イ
タリア、イタリア
寿司＝握り寿司、すし、おすし、スシ、鮓、お寿司
安い＝安く、安くて、安ければ、安価、安上がり
定食＝定職
家族団欒＝家族そろって、家族みな、家族揃って、
家族団らん
規則正しい＝規則正しく
日本食＝純日本食、純和食、純和風、和食、和食系、
和風

置換の例に見るように（等号の右側の複数の語を左の１語で置換）、同じ用語の表記が整っ

ていなかったり、誤記（誤変換）やかな・カナ、英文字などの混在もある。こうした例は特別なものではなく、自由回答では日常的に起こる。また類語を括り置換することも必要に応じて必要となる。こうした操作を終えたあとに単語数の再カウントを行って利用単語を確定する。ちなみにここでは、２語以上で１５６語、３語以上で１００語、４語以上が８６語、…となったので、３語以上の閾値指定で残った１００語を以下の分析対象とした。この頻度閾値の制御等、こまごまとした手当が必要であるが、これを省略し、この編集済みデータセット（１,１１１サンプル、１００語）による解析例を見よう。

(2) 解析結果の解釈

ここで用いる手法は多変量解析的な数量化法やクラスター化法による分析が基本となるが、むしろその周辺の種々の機能が用いられる。InfoMinerには様々な機能があり、すべてを述べることはできないので、ごく一部の機能を用いた結果を示す。

① 回答者と単語の特徴を類型化すること

まず知りたいことは「抽出単語と回答者の回答パターンの関係」であろう。回答者によって用いる単語にどんな類似や差異があるのか、あるならばその特徴は何か、といったことである。このためには単語、回答者の自由回答の記述内容を類型化し対比することが必要である。まず回答者をクラスター化する必要があるが、同時に単語のクラスター化や、それら相互の関係を知り、また個々の回答者の回答の内容が分類で得た類型とどう関係す

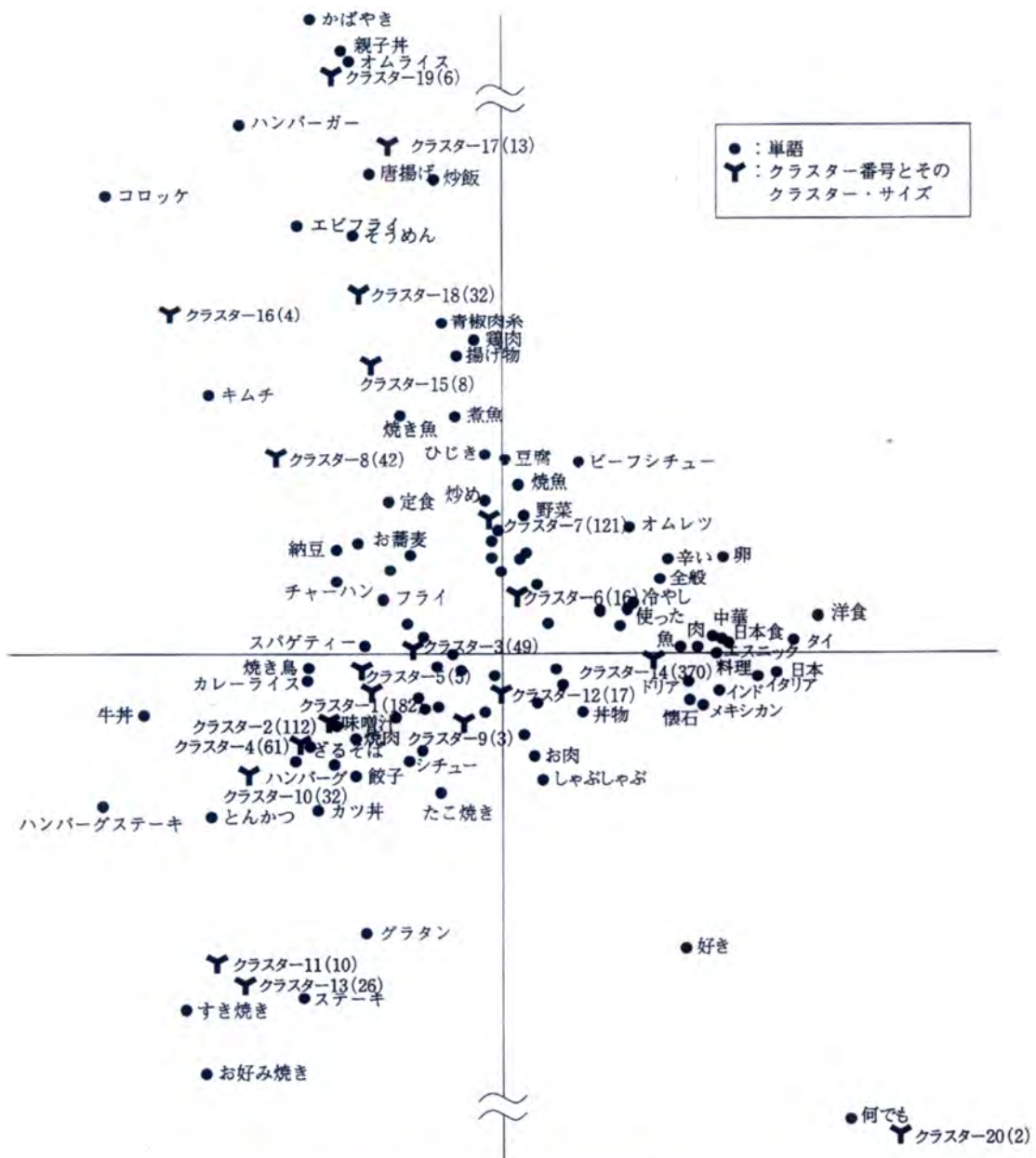


図3 回答者クラスター情報と単語の同時布置図

るかを知りたい。

この解析結果の一つが図3である。ここには、100語の単語と1,111名の回答パターンを類型化したクラスター情報とが同時布置図となっている（クラスター数は20群）²⁾。もち

ろん、1,111名すべてのスコア（数量化成分スコア）と単語のそれとを同時布置図として観察することも可能であるが、あまりに煩雑となるのでここではクラスター化情報としてある。なお図中のクラスター番号の括弧内数字

²⁾ クラスター化処理ではクラスター数をいくつと指定するかが常に問題とされる。InfoMinerでは、クラスター数を変えながらそのクラスター化の履歴を追って、ある種の分類基準を目安に決めることができる。

表1 クラスタを特徴づける単語の一覧

クラスター 1 (182人)	クラスター 2 (112人)	クラスター 3 (49人)	クラスター 4 (56人)	クラスター 6 (121人)
1 カレー	1 焼き肉	1 ごはん	1 ハンバーグ	1 刺し身
2 スパゲッティ	2 カレーライス	2 納豆	2 牛井	2 サラダ
3 餃子	3 ラーメン	3 味噌汁	3 グラタン	3 煮物
4 お蕎麦	4 寿司	4 ピザ	4 パスタ	4 など
5 うどん	5 おにぎり	5 漬物	5 カレー	5 肉じゃが
6 チャーハン	6 焼きそば	6 冷やっこ	6 カレーライス	6 焼き魚
7 ラーメン	7 丼物	7 肉類	7 トマト	7 使った
8 ステーキ	8 生	8 スパゲッティ	8 煮込み	8 もの
9 しゃぶしゃぶ	9 フライ	9 焼き魚	9 チャーハン	9 魚
10 ざるそば	10 懷石	10 おにぎり	10 カツ丼	10 トマト
10 肉	10 魚	10 とんかつ	10 麺類	10 料理
9 お好み焼き	9 ハンバーグ	9 麺類	9 煮物	9 鶏肉
8 サラダ	8 イタリア	8 日本食	8 野菜	8 お好み焼き
7 煮物	7 肉	7 焼き肉	7 寿司	7 お蕎麦
6 野菜	6 中華	6 カレーライス	6 魚	6 寿司
5 ハンバーグ	5 カレー	5 刺し身	5 イタリア	5 ラーメン
4 イタリア	4 日本食	4 中華	4 中華	4 中華
3 日本食	3 パスタ	3 肉	3 肉	3 とんかつ
2 中華	2 スパゲッティ	2 イタリア	2 日本食	2 ハンバーグ
1 料理	1 料理	1 料理	1 料理	1 イタリア

クラスター 7 (42人)	クラスター10(32人)	クラスター14(370人)	クラスター18(32人)
1 野菜	1 とんかつ	1 料理	1 唐揚げ
2 炒め	2 ハンバーグステーキ	2 イタリア	2 ハンバーガー
3 煮魚	3 焼き肉	3 中華	3 鶏肉
4 煮物	4 味噌	4 日本食	4 揚げ物
5 物	5 スパゲッティ	5 肉	5 青椒肉糸
6 ビーフシチュー	6 定食	6 日本	6 焼き魚
7 定食	7 等	7 パスタ	7 オムレツ
8 サラダ	8 肉じゃが	8 タイ	8 お蕎麦
9 ひじき	9 ハンバーグ	9 洋食	9 牛井
10 焼き魚	10 カレー	10 卵	10 刺し身
10 魚	10 パスタ	10 野菜	10 野菜
9 日本食	9 野菜	9 とんかつ	9 魚
8 スパゲッティ	8 刺し身	8 サラダ	8 肉
7 カレーライス	7 寿司	7 煮物	7 イタリア
6 イタリア	6 魚	6 焼き肉	6 中華
5 中華	5 肉	5 カレーライス	5 日本食
4 寿司	4 イタリア	4 スパゲッティ	4 焼き肉
3 焼き肉	3 中華	3 ラーメン	3 パスタ
2 ラーメン	2 日本食	2 ハンバーグ	2 ラーメン
1 料理	1 料理	1 カレー	1 料理

注)クラスター・サイズの大きい主なクラスターのみ示した。

は、クラスター・サイズを表している。こうして、クラスターと回答者が用いた単語との関係がおおよそ分かる。個々のクラスターの意味を特徴付ける単語がその近傍にあるようには見えるが、これでは十分とは言えない。つまりクラスター内の個々の回答の特徴が単語とどう関連するかは、これだけでは曖昧である。そこで次に行う操作として、個々のク

ラスタ内の回答者が用いた単語の傾向を頻度情報やクラスター内の等質性を測る指標を用いて検証する。いまここでは、頻度に注目した場合に得られた結果を表としてみた。表1は各クラスターを特徴付ける単語と、それとは逆にそのクラスターの説明に寄与しない単語とを、ある種の頻度検定を行って一覧とした情報である。実際には指標数値や検定の

有意確率などが表示されるが、また抽出個数も任意に制御できるが、ここではそれぞれ10語だけを表とした。また、クラスター・サイズの小さい群は省略した。

例えば、クラスター1は、そこに含まれる回答者数は182人である。このクラスターを特徴付ける単語は「カレー」「スパゲッティ」「餃子」…と続く。反対にこのクラスターに対して説明力のない単語が「料理」「中華」「日本食」「イタリア」…などと分かる³⁾。回答者数のもっとも大きいクラスター14には370人含まれ、ここでは中華、日本食、パスタ、イタリア…となって和洋混在の傾向が見える。これに対してクラスター6は和食系でしかも家庭料理志向であるように見える。

実はここにはページの制約から出力のすべてを呈示できないのであるが、各クラスター内の個々の回答者を有意性の検証機能を用いて一覧とし、表1の内容と照合する機能もある。つまり、恣意的ではなく、ある指標化操作によって、回答者の個々の回答を客観的に吟味することが可能である。

②属性と単語の関係

クラスター化による類型化を行うことは、分類基準が事前に与えられていない場合に回答パターンと単語の出現情報をもとに自動的に類似したグループを生成することである(古典的な言い方では「教師なし分類」を行ったことになる)。従って、各クラスターを改めて他の項目(選択肢型設問、属性情報等)と

突合分析(二次分析)することが必要となるかもしれない。実際にInfoMinerでは、生成クラスターと他の設問や属性とのクロスチェックや再分析が可能である。こうした操作で得られる情報の解釈も重要である。

一方、抽出した単語が既存情報、例えば選択肢型設問や属性あるいはそれに代わる何らかの情報とどう関連するかを知りたい(「教師ありの分類」ということ)。例えば「性別」「年齢区分」はどう関係するのか、あるいは何らかの設問選択肢との関係などである。

まず、性別に注目して好みの料理を比べると、男性では、「とんかつ、ラーメン、焼き肉、カレーライス、刺身、チャーハン、牛丼、…」と続く。一方女性は、「イタリア、パスタ、スパゲッティ、日本食、サラダ、煮物、…」となる。これをさらに「性年齢区分」で属性データをブレイクダウンし、これと単語との関連性を調べるとどうなるであろう。結果が次ページ図4である。ここでも同時布置図を用いたので「性年齢区分と単語の関係」がポジションニングされている。図がやや見にくいかもしれないが、ここには明らかに性差に加え加齢に伴って単語(好きな料理)の間の関係が変わるということが見える。なおこのデータがWeb調査で取得したものであり、年齢がかなり若年層に偏っていることを勘案する必要がある。実際に性年齢区分別に、どのような料理が好まれるのかあるいは歓迎されないかを、前と同じ要領で一覧としてみる。これが25ページ表2である。男女とも40代前半の

³⁾ 「料理」という語は様々な使い方がなされる。中華料理、日本料理、タイ料理…といった表現だけでない。そこで、ここではあえて併合せずに「料理」という1語で独立させた。従って回答全体に占める利用頻度が高いので、どのクラスターにとっても説明力が弱くなる。

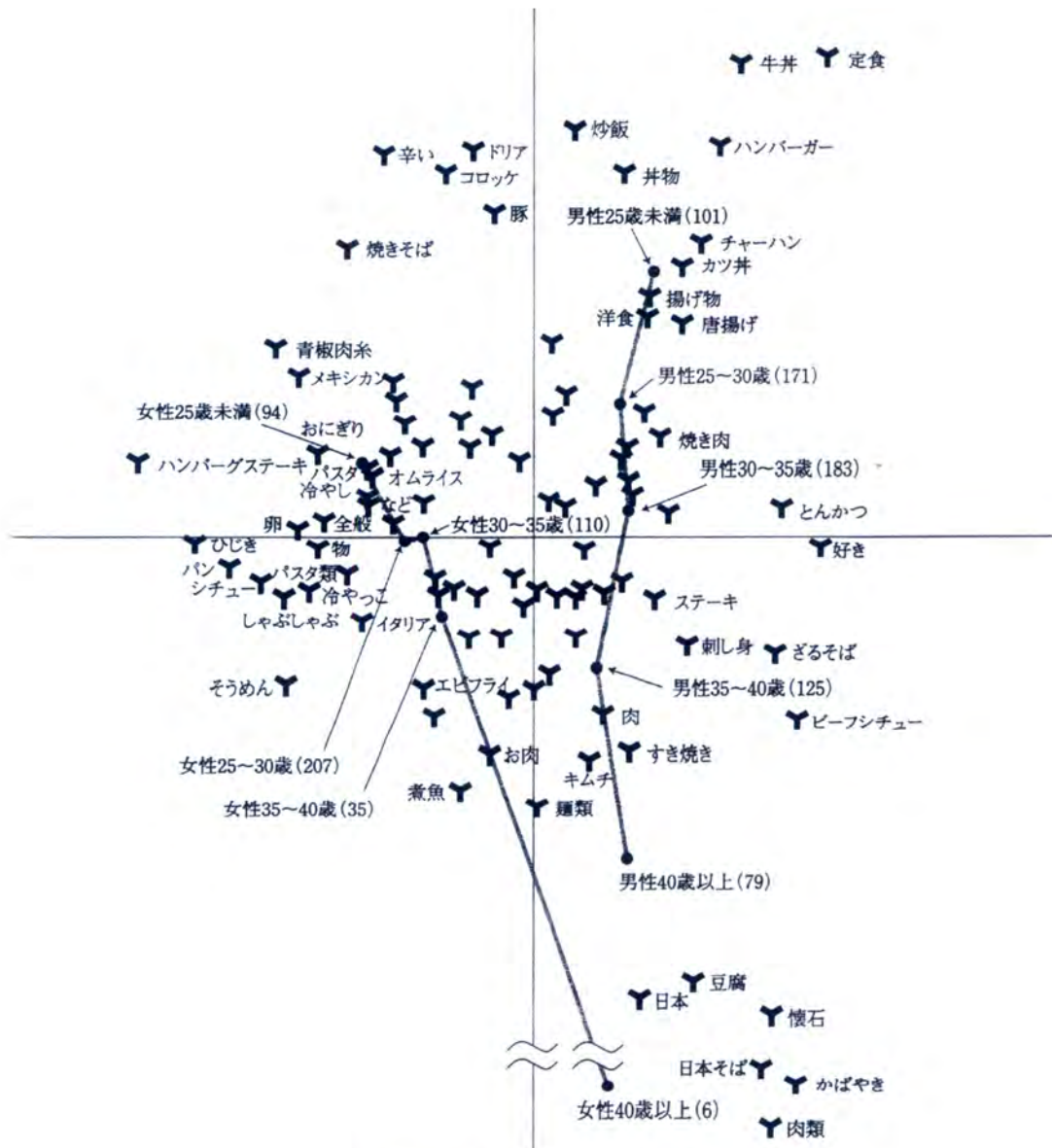


図4 性年齢区分と単語の関係

回答者までしかいないこと、とくに女性は大半が40歳未満であることに注意しよう。これを念頭に図4と表2を見る。

例えば、男性25歳未満では好みの料理として「牛丼」「ラーメン」「丼物」…と続く。また反対にこの属性群では「イタリア（料理）」「日本食」「サラダ」…などは人気がないとなる。また「ハンバーガー」のように、加齢に

伴って「好き」からそうではない方向に移動することが分かる（つまり同じ料理への好意度が年齢層により変化し移動することが分かることが重要である）。一方、加齢に伴って日本食、和食系等に好みシフトする傾向が見えるであろう。

女性については、男性とは異なる特徴が見られる。若年層ではイタリア系やエスニック

表2 性年齢区分を特徴づける単語の一覧

男性25歳未満 (101)	男性25～30歳 (171)	男性30～35歳 (183)	男性35～40歳 (125)	男性40歳以上 (79)
1 牛井	1 カレーライス	1 とんかつ	1 麺類	1 肉
2 ラーメン	2 焼き肉	2 ラーメン	2 刺し身	2 豆腐
3 丼物	3 ラーメン	3 餃子	3 焼き肉	3 料理
4 カツ丼	4 唐揚げ	4 魚	4 お肉	4 麺類
5 定食	5 チャーハン	5 焼き魚	5 料理	5 日本そば
6 ハンバーガー	6 牛井	6 チャーハン	6 タイ	6 日本
7 ピザ	7 ハンバーグ	7 カレー	7 日本そば	7 カレーライス
8 とんかつ	8 刺し身	8 味噌	8 寿司	8 お蕎麦
9 焼き肉	9 ハンバーガー	9 かばやき	9 等	9 すき焼き
10 唐揚げ	10 揚げ物	10 好き	10 肉	10 かばやき
10 全般	10 使った	10 パスタ	10 ハンバーガー	10 うどん
9 パスタ	9 煮物	9 など	9 サラダ	9 煮物
8 日本	8 イタリア	8 ステーキ	8 納豆	8 焼き肉
7 肉	7 スパゲッティ	7 料理	7 うどん	7 チャーハン
6 魚	6 魚	6 サラダ	6 牛井	6 ラーメン
5 餃子	5 炒め	5 中華	5 全般	5 イタリア
4 サラダ	4 等	4 日本食	4 ごはん	4 ハンバーグ
3 日本食	3 料理	3 麺類	3 日本食	3 スパゲッティ
2 イタリア	2 麺類	2 煮物	2 パスタ	2 カレー
1 料理	1 野菜	1 イタリア	1 スパゲッティ	1 パスタ
女性25歳未満 (94)	女性25～30歳 (207)	女性30～35歳 (110)	女性35～40歳 (35)	女性40歳以上 (6)
1 スパゲッティ	1 パスタ	1 など	1 煮込み	1 かばやき
2 サラダ	2 イタリア	2 煮物	2 サラダ	2 肉類
3 肉じゃが	3 パン	3 生	3 日本食	3 日本
4 コロッケ	4 日本食	4 スパゲッティ	4 イタリア	4 ごはん
5 そうめん	5 等	5 焼きそば	5 焼魚	5 豆腐
6 全般	6 物	6 使った	6 フライ	6 イタリア
7 うどん	7 シチュー	7 全般	7 豚	7 炒め
8 エスニック	8 ハンバーグステーキ	8 パスタ類	8 トマト	8 鶏肉
9 中華	9 ひじき	9 エビフライ	9 麺類	9 うどん
10 ごはん	10 煮物	10 日本食	10 煮物	10 野菜
10 唐揚げ	10 唐揚げ	10 唐揚げ	10 など	
9 チャーハン	9 日本	9 ステーキ	9 チャーハン	
8 刺し身	9 豆腐	8 お好み焼き	8 ハンバーグ	
7 ラーメン	7 肉	7 牛井	7 カレー	
6 魚	6 ハンバーガー	6 肉じゃが	6 とんかつ	
5 焼き魚	5 ラーメン	5 焼き肉	5 焼き肉	5 中華
4 日本食	4 焼き肉	4 肉	4 カレーライス	4 焼き肉
3 とんかつ	3 とんかつ	3 炒め	3 お蕎麦	3 カレーライス
2 寿司	2 カレーライス	2 ラーメン	2 ラーメン	2 パスタ
1 焼き肉	1 刺し身	1 とんかつ	1 パスタ	1 ラーメン

注1) 年齢区分の括弧内は回答者数を示す。

2) 10に満たない例は有意となる単語がこれ以上なかったことを示す。

ク・メキシカンが人気がある一方で、肉じゃが、うどん、煮物…と和洋混合が男性よりも顕著である。また明らかに男性とは好みの傾向が異なる。

③用語検索と回答パターンの意味

ところで、単純に料理名の出現頻度数のみで検証すると「料理」「ラーメン」「カレーラ

イス」「和食」「焼き肉」「イタリア」「スパゲッティ」「ハンバーグ」…等が上位に現れる。これは通常言われている好きな料理の多くを含むので（多くの類似調査の結果に見るように）、それなりに納得できる情報を提供してはくれる。しかし、個々の回答者のパターン、つまり料理の選択傾向パターンを知ることまではできない。ここに頻度に依拠する接近法

の限界がある。ラーメンを選ぶと他には何を選びやすいのかといったパターンまでを知るには、やはり解析手法の工夫が必要である。詳細には記述しないがInfoMinerではこうしたことへの手当がなされている。次のコンコードダンス（用語検索）はその探索的機能の一つである。

これは「ラーメン」の使われ方（自由回答）を元の分かち書き文章から検索してこれをどう用いたかの出力である。また、多変量解析的な分析から、これらのパターンの様相を視覚的に知ることができる。

[コンコードダンスの例]

	カレー	ラーメン	
		ラーメン	スパゲッティ
カレーライス	ラーメン		
	ラーメン	カレーライス	
	ラーメン	餃子	
	ラーメン	そば	
	ラーメン	カツ丼	
焼肉	ラーメン		
カレー	ラーメン		
カレー	ラーメン		
寿司	ラーメン		
	ラーメン	ハンバーグ	
カレーライス	ラーメン		
	ラーメン	カレーライス	
味噌	ラーメン	カツカツ	
	ラーメン	サラダ	
(パスタなら何でも)	ラーメン	(しょうゆ味に限る)	
焼肉	ラーメン	うどん	

.....
.....
(以下続く)

5. まとめ

自由回答データの簡単な解析事例を示したが、要点は、単に自由回答だけを用いるのではないことにある。分かち書きを初めとする

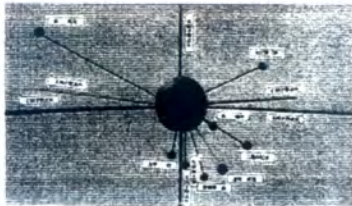
一連の単語処理に始まり、個々の自由回答パターンの意味づけや回答者の類型化、単語の分類情報の分析だけでなく、取得した他の情報（選択肢型設問、属性情報他）を併用して、それらの情報の何が有意に働くか、あるいは自由回答データの解釈にどのように寄与するか等々、調査データの解析法として本来必要とされる情報の提供をいかに行うかの解析手順が重要である。

インターネットの普及に連動して、自由回答等の文章型・テキスト型データの適切なデータ解析手法の登場が期待されている。データ・マイニングに関連してテキスト・マイニング手法が多用されることが良い例である。しかし、調査における自由回答の取得方法や、そのデータ解析技法の方法論の研究はそう進んでいるとは言えないだろう。現実の現象解析にあたって重要なことは、問題とする対象を説明するためにもっとも適したデータ取得法はどうあるべきかを、データ科学の視点から実証的に検証することであると考え、その事例の一つをここに紹介した。

*参考文献

- [1] 鮑戸弘編著 (1994) : 「食文化の国際比較」 : 日本経済新聞社.
- [2] 大隅昇 (2000) : 「調査環境の変化に対応した新たな調査法の研究」 報告書 : 文部省科学研究費特定領域研究, ミクロ統計データ, 公募研究 (研究課題番号 : 09206117) .
- [3] 大隅昇, 丸岡吉人他 (1997) : 自由回答データの解析法についての提案—実験調査におけるいくつかの試み— : 第25回日本行動計量学会大会.
- [4] 博報堂インタラクティブカンパニー (2000) : インターネットマーケティング : 日本能率協会マネジメントセンター.

CONTENTS 目次



特集 ■ 文章の統計分析 ————— 2

因子分析による現代作家の統計的分類／

安本美典（産能大学教授）

古典和歌における表現分析／

竹田正幸（九州大学大学院システム情報科学研究科助教授）、福田智子（福岡女学院大学非常勤講師）、南里一郎（純真女子短期大学助教授）

定性情報のマイニング—自由回答データの解析—／

大隅昇（統計数理研究所調査実験解析研究系パターン解析研究部門教授）

探訪 ————— 27

広がりを見せるパソコンソフト②⑥／高橋三雄（麗澤大学国際経済学部教授）

コラム ————— 32

歴史の中に出生率を読み解く／小林亜子（埼玉大学教養学部助教授）

寄稿 ————— 36

第二の母国インド／米田稔（㈱ガネット代表取締役）

紹介 ————— 40

地域を生かした文化の発進

オリジナルステージ・オリジナルアートを育てる創造空間「浄るリシアター」

特別寄稿 ————— 45

座談会／放送新時代のサッカー

