筑波大学大学院博士課程

システム情報工学研究科修士論文

カテゴリデータ分析のための視覚的表現及びツールの開発

白石 宏亮

(コンピュータサイエンス専攻)

指導教員　三末 和男

2010年3月

概要

カテゴリデータはビジネスや科学分野など様々な分野で現れ，データ分析の目的で利用されることが多々ある．一般的に，カテゴリデータは多くの属性から成る多次元データであり，データの傾向を把握するには多角的に分析する必要がある．しかし，従来のカテゴリデータの分析方法は表に基づいた方法がほとんどであり，事前に専門的な知識を必要とすることが多い．

本研究では，カテゴリデータの分析を目的とした視覚的表現とその表現を用いた分析ツールを開発した．開発した視覚的表現である「つぶつぶ表現」はカテゴリデータにおけるエンティティを視覚的要素として表示する．そして，視覚的要素の色・配置によってカテゴリを表現する．個々のエンティティを視覚的に表現することで，分析者はカテゴリデータの集合的性質をイメージしやすくなる．

つぶつぶ表現においてインタラクティブな分析が行える手法として，データベースにおけるクエリを模したラベルによる操作を開発した．着目したいラベルを操作すると，視覚的要素がアニメーションによって移動し，視覚的にドリルダウンすることができる．開発した分析ツールは既存のグラフ表現をツールに統合することで，つぶつぶ表現とグラフ表現が互いに利点・欠点を補い合い，属性数の多いデータにおいても柔軟な分析が可能である．

開発した分析ツールを用いて実際のデータの分析を行った例をケーススタディとして示した．ケーススタディでは実際のデータからその傾向を分析できることを示し，つぶつぶ表現及び分析ツールが有効となることが分かった．

目 次

図目次

第1章　序論

1.1　データにおける変数の種類

データの変数はその特徴によって，名義変数 (Nominal variables)，順序変数 (Ordinal variables)，量的変数 (Quantitative variables) の 3 つに分類することができる [1]．名義変数とは名詞的な値をとり，値が同一か否かを評価することだけに意味を持つ変数である．名義変数の例として性別や血液型が挙げられる．例えば，性別は男性と女性という 2 つの値からなり，男性であるか女性であるかといういずれかの値をとる．順序変数とは順序を付けて比較することが可能な変数である．例えば，順序変数として考えられる学年は 1 学年，2 学年というように順序を付けて値同士を比較することに意味を持つ．量的変数とは身長や気温などの数値で表される変数である．名義変数に対して，量的変数は値の大小を比較することや，値の平均を算出することが可能な変数である．

1.2　カテゴリデータとは

カテゴリデータ (Categorical data) とは質的データ (Qualitative data) とも呼ばれ，データ中の変数が名義変数または順序変数によって構成されるデータである．一方，データ中の変数が量的変数によって構成されるデータはカテゴリデータの対として量的データ (Quantitative data) や数値データ (Numerical data) と呼ばれる．カテゴリデータにおける変数は属性 (Attribute) と呼ばれ，変数に含まれる値はその項目によってカテゴリ (Category) と呼ばれる．例えば，性別という属性は男性と女性という 2 つのカテゴリから構成される．

1.3　カテゴリデータの分析

カテゴリデータはビジネスやマーケティングリサーチ，科学分野など様々な領域で現れる．例えば，教育機関において教員が実施する授業アンケートや，研究室に所属する学生が行う主観的な評価実験データなどはその多くがカテゴリデータである．また，ビジネスでは調査を行うためにインターネット上や紙面によるアンケートを行うことが多々あるが，このようなデータもカテゴリデータであることが多い．

次に，カテゴリデータがどのような目的で使用されるかについて代表的なカテゴリデータの例であるアンケートデータを挙げて説明する．ここでは一般消費者向けに製品を販売するメーカーが行う製品購入に関するアンケートデータを考える．アンケートデータは属性とし

て性別や年齢，職業などの基本的な項目に加えて,「製品を購入したことがありますか？」,「製品をどこで知りましたか？」といった各質問が属性として考えられる．このアンケートデータからマーケティングリサーチ部門の担当者は,「製品はどのような年齢層に売れているか？」といったような製品購入に関する分析を行う．そして，分析によって得られたデータの傾向を元に製品のイメージや販売向上に繋げる．

1.4 カテゴリデータ分析のプロセス

ここでは従来の一般的なカテゴリデータ分析についてその方法を述べる．カテゴリデータの分析には SPSS や Microsoft Excel などの表形式をベースとした表計算ソフトウェアがよく利用されている．これらのソフトウェアによる分析のプロセスを図 1.1 に示す．分析のプロセスは大きく 3 つに分けることができる．まず，初めは何の加工もされていない状態の生データである．この生データから属性におけるカテゴリの集計をとることで，集計表としてその分布を表すことができる．集計表を人間が直感的に理解しやすい表現に可視化し，データの傾向を視覚的に分析する．

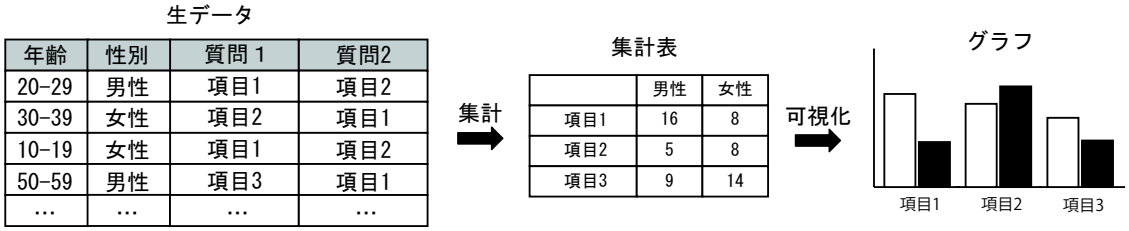

図 1.1: 従来のカテゴリデータ分析のプロセス

1.4.1 生データ

生データとは何も加工されていない状態のデータのことを呼ぶ．カテゴリデータの生データはいくつかの表現形式が考えられるが，一般的には図 1.2 のようなリスト形式の表で表される．表中の列にはカテゴリデータにおける属性 (フィールド) を，行にはエンティティ(レコード) が対応する．エンティティとはデータにおいて，一つの単位としてまとめられる対象のことを呼ぶ．例えば，アンケートデータにおける「人」や，商品データにおける「商品」がエンティティとして考えられる．1 行目には属性が記述され，2 行目以降には 1 行に 1 エンティティが対応する形で全てのエンティティがリスト形式で列挙される．

1.4.2 単純集計とクロス集計

生データにおける各属性のカテゴリを集計することによって，属性の関係を表として表すことができる．一つの属性に着目してそのカテゴリを集計することを単純集計といい，分布

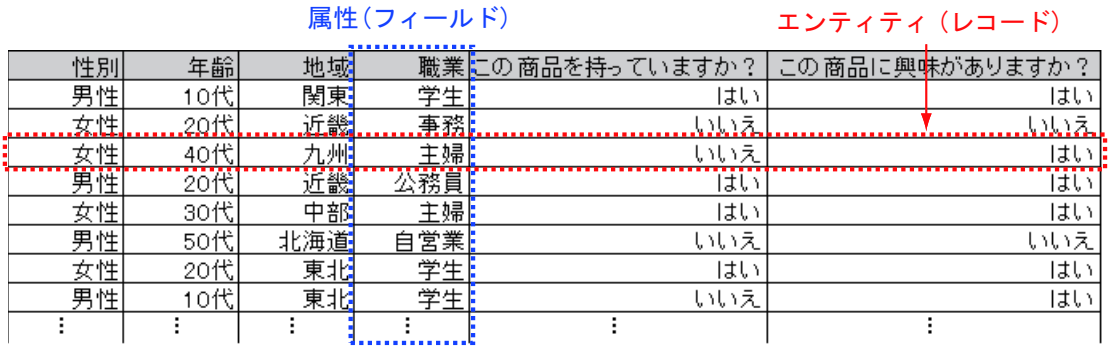

図 1.2: 生データの例

を表として表したものを単純集計表と呼ぶ．例えば，図 1.2 の一つの質問に着目して,「はい」・「いいえ」の項目で集計表を作成すると表 1.1 のようになる．

データ中の 2 つ以上の属性に着目して，そのカテゴリを集計することをクロス集計という．例えば表 1.1 に年齢の属性を加えると，表 1.2 のような表を作成することができる．この表は行と列におけるカテゴリが交差する部分にその集計値が該当することからクロス集計表 (Contingency Table) や分割表と呼ばれる．

表 1.1: 単純集計表

質問 1	はい	いいえ
全体	344	156

表 1.2: クロス集計表

質問 1	はい	いいえ
10 代	85	92
20 代	83	89
30 代	88	87
40 代	144	93
50 代	144	93

1.4.3 可視化

集計表の値からデータの傾向を分析することは不可能ではないが，数値で埋め尽くされた表からそのデータの傾向を発見することは困難である．これを人間が直感的に分かりやすい形にするために，データを可視化して視覚的に表現する．一般的に，誰もが知っているカテゴリデータの可視化表現として幅広く利用されているのが，棒グラフや円グラフである．これらの表現は図形の大きさや比率によって分布を表現するというシンプルな表現であるが，基本的に 1〜3 つの属性の関係を表現するのが限界である．それ以上の属性の関係を可視化すると，グラフにおける要素が多くなるため視覚的に煩雑になってしまう．このため，属性数の多いデータを一度に可視化することは困難である．

1.4.4 多次元データ

世の中の様々な分野に現れるデータは通常 3 つ以上の属性を含むことがほとんどである．データ中の一つの属性は一つの次元 (Dimension) と考えることができ，このような多くの属性を含むデータは多次元データ (Multidimensional data) または多変量データ (Multivariate data) と呼ばれる．カテゴリデータの多くも通常この多次元データである．例えば，アンケートデータは性別，年代や職業などの基本的な属性だけでも 3 つ以上の属性となり，さらには質問の属性を含めると 10 以上の属性を含むデータとなることは珍しくない．

1.5 カテゴリデータ分析における問題点

ほとんどのカテゴリデータは複数の属性からなる多次元データであるが，棒グラフや円グラフなどの一般的なカテゴリデータの可視化手法によって一度に表現できる属性の数は 1〜3 つが限界である．このため，カテゴリデータ分析のプロセスにおいて，データ中の着目したい属性の組み合わせを選び，データの一部を可視化する．または，複数のグラフで可視化し，見比べるというようなアプローチが必要となる．

データを詳細に分析していくような場合は，3 つ以上の属性に関するクロス集計表を作成しなければならず，このような操作は専門的な知識が必要になることや，慣れていなければ時間と労力がかかってしまう．

1.6 本研究の目的

本研究ではカテゴリデータの分析を目的として，カテゴリデータを視覚的に表現する手法を開発する．また，本手法を適用することで，従来では表をベースとして行っていた多次元カテゴリデータ分析に代わり，視覚的な分析が可能なツールの開発を行う．

1.7 本研究の貢献

本研究における貢献は以下の 3 点であると考えている．

一つ目めはカテゴリデータを視覚的に表現する手法を開発したことである．本手法により，カテゴリデータの性質を直感的に理解することが可能になる．

二つ目は本研究で開発した視覚的表現をインタラクティブに操作する上で必要ないくつかの操作を開発したことである．データベースにおけるクエリ操作を模した視覚的ドリルダウンや，視覚的要素を仮想的な力学モデルでレイアウトする方法などが挙げられる．

三つ目は視覚的表現と従来のグラフ表現を統合し，カテゴリデータの分析ツールを開発したことである．本ツールにより，統計分析の知識を持った人はもちろん，専門的な知識を持たない人でも直感的な分析が可能となると考えられる．

第2章　関連研究

本研究の関連研究について述べる．関連研究の分類の仕方として，データの特徴によって，量的データ，カテゴリデータ，量的データとカテゴリデータを含む混合データ，特殊なデータの 4 つに分けそれらのデータを対象とした研究について議論する．なお，いずれの研究も多次元データ・多変量データを対象したものである．

2.1　量的データを対象とした手法

2.1.1　一覧表示する可視化手法

一覧表示する可視化手法とは，多次元データにおける複数の属性の関係を一枚の図として表現する手法のことを示す．Parallel coordinates[2] では各次元に対して座標軸を用意し，座標軸を並列に並べる．そして，各座標軸における点を全て結んでいくことで，多次元データを表現する．

2.1.2　インタラクティブな操作を利用した手法

SCATTERDICE[3] は多変量データの散布図における各次元を切り替える操作を，サイコロをころがすようにインタラクティブに操作する．通常は 2 次元の散布図であるが，次元を切り替える際にその奥行きにあるもう 1 次元の散布図が 3 次元のアニメーションによって変遷することで切り替わる．このようにすることで，一般的な 2 次元の散布図のシンプルの長所を生かしつつ，次元の切り替えを直感的に行うことができる．Dust & Magnet[4] は多変量データを磁石のメタファを利用することで直感的に分析可能な手法である．散布図における各点を磁石によって引き寄せられるダストのように表現し，各属性を磁石のようにマウスで操作をする．各属性において値が大きいほどダストにみたてた各点が強く引き寄せられる．これにより，複数の属性の磁石で操作することによって，多変量データの分布の傾向を概観することができる．

SGVIEWR[5] では Parallel coordinates[2] と Parallel Sets[6] を統合することで，多次元データを一覧表示する．横に積み上げ棒グラフを並べるように表示し，対応する属性を下に重ねていくことで詳細を表示する．

2.2 カテゴリデータを対象とした手法

2.2.1 一覧表示する可視化手法

Cobweb diagram[7] はグラフ理論の網図表現のようにカテゴリ間の関係をエッジで結び，エッジの太さで集計値を表現している．Mosaic Display[8] では矩形の大きさで属性の集計値を表し，縦横に再帰的に並べていくことで 2 つ以上の属性を表現している．全体を概観することで，大きさの目立つ矩形などによりデータの大まかな傾向は概観することが可能であるが，3 つ以上の属性を表現すると，矩形の数が多くなってしまう．Cattrees[9] では Treemap[10] を利用し，属性の値を Treemap の一つのノードの大きさに対応させることでカテゴリデータを表現している．Cattrees では Treemap における各領域をインタラクティブに変更することができる．

Hammock Plots[11] は Parallel coordinates における，線の幅をカテゴリデータにおける集計値に割り当てることで，量的データに加えてカテゴリデータを表現することができる．

2.2.2 インタラクティブな操作を利用した手法

Parallel sets[6] は Parallel coordinates[2] と Mosaic Display[8] を組み合わせることで多次元カテゴリデータを表現している．Parallel coordinates における各点を大きさを持った矩形として表現し，さらに矩形の幅を持った線としてつなぐ．これにより，3 つの属性の関係を一覧して表示している．さらに，Parallel sets ではインタラクティブに次元を切り替える操作を提供している．

SQiRL[12] では円グラフを拡張した手法を提案している．通常の円グラフでは一つの円グラフで一つの属性を表現するが，SQiRL では円グラフの領域をさらに細かく分割することで，複数の属性を表現する．さらに，円グラフの中心をデータベースにおけるクエリ操作のように利用できる．属性のカテゴリを円グラフの中心にドラッグ&ドロップで移動すると，カテゴリの条件によって周りの円グラフがアニメーションによって変化する．

2.3 量的データとカテゴリデータを含む混合データを対象とした手法

Trellis Display[13] は量的データを散布図などで表現し，行と列にカテゴリデータを対応させ，カテゴリによって同じ表現を行列上に並べて行くことで多次元データを表現している．

Pixel Bar Chart[14] は棒グラフにおける各ピクセルを量的データを表現するのに利用する．各ピクセルを量的データにおける値によって色を割り当てることで，従来の棒グラフに加えて量的データを同時に表現することが可能である．Hierarchical Pixcel Bar Chart[15] では Pixel Bar Chart を拡張し，棒グラフを縦に分割していき，高さを変えることで階層情報を表現している．Table Lens[16] は大規模データを表形式で可視化する手法である．大規模表データの一部をフォーカスとズームによって分析することができる．

2.4 特殊なデータを対象とした手法

SellTrend[17] では時系列を含むカテゴリデータを対象とした分析ツールである．航空便のチケット予約・購入の際のデータを対象として，Treemap とヒストグラムを統合することで，時系列におけるカテゴリデータの時間的な傾向の変化を分析することができる．

Set'o'gram[18] は集合型というタイプの多次元データを対象とした可視化表現である．棒グラフの幅を異なるカテゴリとして割当て，一つの棒グラフの上に幅の異なる棒グラフを重ねることで多次元データを表現している．

FanLens[19] は階層型のカテゴリデータを円グラフを拡張した方法で表現する．通常の円グラフにおける各扇の領域のさらに外側に扇を追加していくことで，複数階層の分布を表現している．

2.5 多次元・多変量データ分析ツール

Polaris[20] は多次元データベース向けの可視化分析ツールである．ピボットテーブルを応用したインタフェースにより，クロス集計の操作なしに属性をマウスのドラッグ&ドロップで選択することによって分析が可能である．XmdvTool[21]，GGobi[22] は散布図や Parallel coodinates などを使用したインタラクティブな可視化ツールである．Brushing や Zoom などの様々なインタラクションをサポートしている．

第3章　カテゴリデータの分析要求

3.1　カテゴリデータ分析における要求事項

カテゴリデータの分析では各属性におけるカテゴリの分布を分析することが基本的事項である．カテゴリの分布とは，例えばアンケートデータにおける質問に関して，男女比がどのように分布しているか，年齢層はどうかといったようなカテゴリの集計値の分布である．

1 つの属性に関する分布はデータ全体としてみた時に現れる傾向であり，単純集計された分布からその傾向を分析できる．しかし，2 つ以上の属性に関してはデータの属性数が増えればとりうる属性の組み合わせの数も膨大となり，データの一部にのみ現れる傾向となる．例えば,「ある製品に興味がありますか？」といった質問に対して，データ全体では「いいえ」と答えている人の割合が「はい」と答えている人より多いとする．しかし，年代や性別，職業などの属性を加えて分析すると，ある年代では興味のある人の割合が多く，また一方の年代では興味のない人が大部分を占める，というように，着目する属性の数が増えるとその傾向も局所的なものとなる．

ここでは，このような分析における属性の傾向を区別するために，データ全体を通して現れる傾向を大局的傾向，データの一部のみに現れる傾向を局所的傾向と定義する．

3.1.1　大局的傾向の分析

大局的傾向とはデータ全体としてみた時に現れる傾向である．大局的傾向と局所的傾向の境を明確に区別することはできないが，ここでは 2 つ以下の属性の傾向を大局的傾向と呼ぶ．2 つ以下の属性の傾向とは，1 つの属性に関して単純集計した分布，または 2 つの属性に関してクロス集計した分布における傾向である．

大局的傾向の分析には単純集計表またはクロス集計表を作成し，表中の全体または一部分を可視化する．

3.1.2　局所的傾向の分析

局所的傾向とはデータ全体ではなく，データ中のある一部分に現れる傾向である．ここでは 3 つ以上の属性の傾向を局所的傾向と呼ぶ．3 つ以上の属性の傾向とは，3 つ以上の属性に関してクロス集計した分布における傾向である．

局的的傾向の分析には大局的傾向に比べてより詳細な分析が必要となる．例えば,「10 代の一人暮らし男性」や「商品に興味がある関東に住んでいる人」のように各属性におけるカテゴリの AND 関係をとった分析が必要となる．

3.2 カテゴリデータ分析の流れ

Shneiderman によって提唱された Mantra[23] によると，可視化におけるプロセスとはまず全体を概観し，ズーミングやフィルタリングを行い，さらに必要に応じて詳細に分析するとされている．カテゴリデータの分析においてもまず全体を概観し，得られた知見を元に大局的傾向から局所的傾向へと詳細に分析していくプロセスが必要である．

3.2.1 ドリルダウン

データ分析におけるドリルダウンとは概観から詳細へと掘り下げて分析していく過程のことを呼ぶ．すなわち，大局的傾向を分析し，得られた知見を元に局所的傾向の分析をしていくようなプロセスである．

例えば，初めは性別という 1 つの属性にだけ着目し，その傾向を分析する．得られた知見を元に，さらに着目したい異なる属性を加えると，10 代の男性，20 代の女性…，のようにより詳細へとドリルダウンしていく．

このドリルダウンを実現するためにデータ分析はもちろん，データベースにおけるクエリ操作等でも多くの手法が研究されている．Microsoft Excel などの表計算ソフトウェアではピボットテーブル機能を提供している．これはリスト形式のデータから，属性を指定するだけでクロス集計表を自動に作成する機能である．着目したい属性をドラッグ&ドロップで指定すると集計表が作成され，直感的にドリルダウンすることができる．

第4章　視覚的表現

4.1　カテゴリデータの集合的性質

カテゴリデータはその性質から集合的な見方をすることができる．例えば，アンケートデータにおける性別に着目すると，男性と女性の 2 つの集合から構成されると考えられる．さらに，性別の属性に年代の属性が加わると,「10 代の男性」,「20 代の女性」といったようにカテゴリの AND 関係によってさらに細かい集合へと分けることができる．

本研究で開発する視覚的表現の基本アイデアは，カテゴリデータにおける個々のエンティティを視覚的に表現することで，この集合的性質を分析者に直感的にイメージしやすくすることである．

4.2　集合的性質の表現方法

4.2.1　つぶつぶ表現

本研究で開発したカテゴリデータの視覚的表現について述べる．開発した視覚的表現は，カテゴリデータの集合的性質を分析者にイメージしやすくするために，データにおける個々のエンティティを視覚的要素として表現する．この視覚的要素は大きさを持った円である．円として表現する理由は，図形における最も基本的な形状の一つであるということと，散布図などにおけるデータのプロットは点として表現されるため，個々のエンティティとしての理解が容易であると考えた．視覚的要素を円として粒のように表現する特徴から，この視覚的表現を「つぶつぶ表現」と名付ける．つぶつぶ表現における個々の円を要素と呼ぶ．図 4.1 に本研究で開発した視覚的表現の例を示す．図における要素一つ一つがカテゴリデータにおけるエンティティである．例えば，アンケートデータにおいては「人」が要素に対応する．

開発した視覚的表現との比較として，図 4.2 に図 4.1 における視覚的表現を棒グラフで表したものを示す．棒グラフが棒の長さによって何らかの値の相対量を表現するのに対して，つぶつぶ表現は個々のエンティティを視覚的に表示し，その数によって絶対量を表現する．

4.3　視覚的要素の関係付け

円として表現された視覚的要素はその色や配置に意味を持たせない場合，各要素の違いを区別することができない．

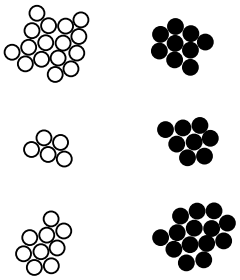

図 4.1: 視覚的表現の例

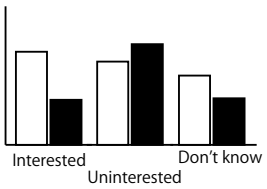

図 4.2: 棒グラフの例

各要素のカテゴリの違い，すなわちカテゴリデータにおける属性のカテゴリを表現するために，ゲシュタルト要因における近接・類同の要因を利用する [24]．図 4.3 は要素に何の関係付けもされていない状態である．すなわち，要素はただ等間隔に配置されているだけであって，各要素の違いを認識することはできない．

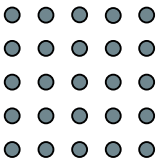

図 4.3: 何の関係付けもない状態

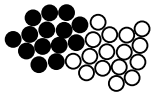

図 4.4: 色による関係付け

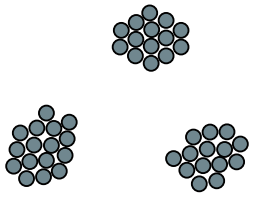

図 4.5: 配置による関係付け

4.3.1 色による関係付け

人間は，色が異なる要素を別々の関係，同色である要素同士を同一の関係と知覚する．この知覚はゲシュタルト要因の類同の要因によるものである [24]．例えば，図 4.4 では要素は白と黒の 2 つの関係に分けられていると知覚する．

要素の色と属性のカテゴリを対応付けることによって，要素のカテゴリを表現することができる．

4.3.2 配置による関係付け

人間は，位置的に近接して配置されている要素群を同一の関係と知覚する．この知覚はゲシュタルト要因の近接の要因によるものである [24]．例えば，図 4.5 では要素はその配置によって 3 つの関係に分けられていると知覚する．

配置によって近接した要素群と属性のカテゴリを対応付けることによって，要素のカテゴリを表現することができる．

4.4 視覚的要素の操作

3.2.1 で述べたように，多次元データを分析する際にはドリルダウンの操作が必要である．ここでは，つぶつぶ表現のおいてドリルダウンを実現する方法について述べる．

4.4.1 クエリを模した視覚的ドリルダウン

配置による関係付けでは，近接している要素群と，何らかのカテゴリを対応付けることによって要素のカテゴリを表現する．これを利用し，分析者の着目したいカテゴリと要素の配置を関係付けることで，ドリルダウンの操作を実現する．この操作を分析者が視覚的に行えるように，ラベルという概念を導入する．ラベルとは属性のカテゴリを表したものであり，分析者はこのラベルを操作することで自身の着目したいカテゴリと要素の配置を関係付ける．初めは何の関係付けもされていない状態である (図 4.6 左)．ここで，10 代のカテゴリに着目したい場合に，10 代と書かれたラベルを適当な位置に配置する．すると，10 代のラベルの近くに，全要素の中から 10 代の要素のみの配置が変わり，要素同士が近接し合い，10 代のラベルの近くに配置される (図 4.6 中央)．同様に，20 代のカテゴリに着目したい場合は，20 代のラベルを別の位置に配置する．すると，20 代の要素がラベルの近くに配置される (図 4.6 右)．

このように，分析者が着目したいラベルを操作することによって，カテゴリと要素の配置が関係付けられることで視覚的なドリルダウンを実現する．

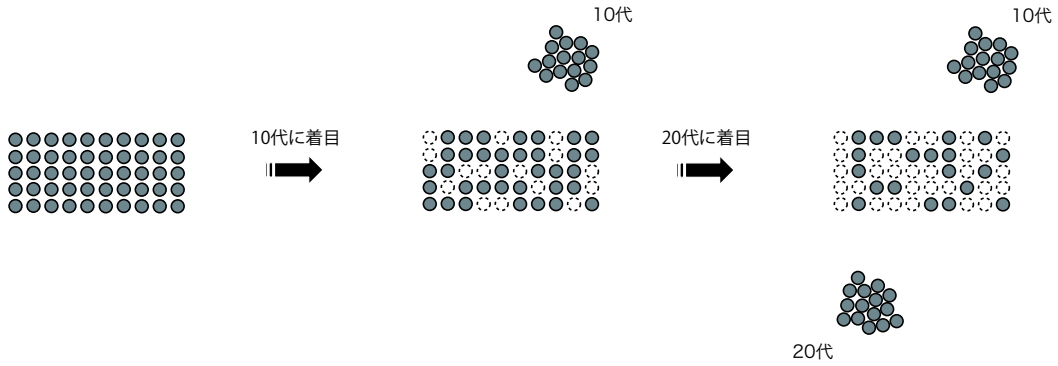

図 4.6: 視覚的ドリルダウン

4.4.2 アニメーション

何らかのオブジェクトがその位置や形を変える際に，その変遷をアニメーションによって表現することは人間がその変遷の前後を知覚するのに有効である [25]．視覚的ドリルダウンにおいても，分析者が常に自分の操作している対象を把握しやすいように，要素の配置が変遷する前後をアニメーションによって表現する．要素の配置が変わる前と，変わった後の間

の変遷を表現するために，要素の配置が変わる際には要素がアニメーションによって移動する表現を用いる．

第5章　ツールの開発

5.1　ツールの設計

5.1.1　既存の可視化手法の統合

カテゴリデータはそのほとんどが多次元データであるため，複数の属性の傾向を一覧して表示できると分析が行い易い．開発する分析ツールでは，データをつぶつぶ表現によって表示するビューと既存の可視化手法によって表示するビューの両方を設け，一つの分析ツール上に統合する．カテゴリデータ分析におけるドリルダウンはつぶつぶ表現のビュー上で行い，既存の可視化手法のビューでは複数の属性の傾向を一覧して表示する．本ツールでは既存の可視化手法として，カテゴリデータを可視化する最も基本的な表現の一つである棒グラフを選択した．

異なる可視化手法を一つのツール上に統合するために，Linking&Brushing[26] の手法を分析ツールに適用する．Linking[27] とは異なる可視化表現間において，同一の要素の色などを統一することである．開発する分析ツールでは要素の色と棒グラフの色を統一することで Linking を実現する．

Brushing[28] とは全体から着目したい部分を指定することによって，何らかの方法で指定された部分をハイライトして表示する手法である．開発する分析ツールでは，つぶつぶ表現でデータが表示されているビュー上で一部の要素を選択すると，その要素だけに関するグラフを作成することで Brushing を実現する (図 5.1).

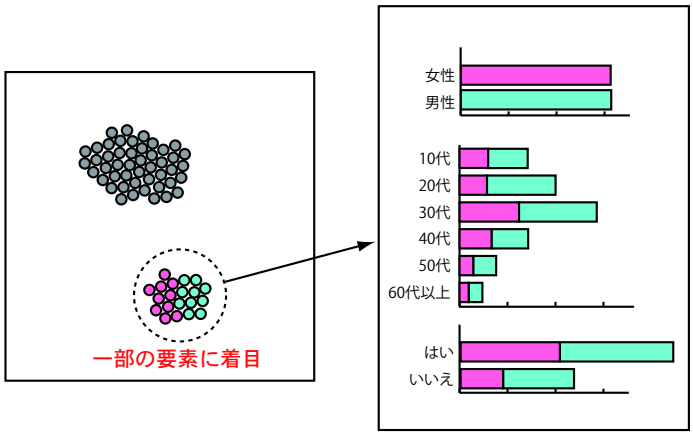

図 5.1: つぶつぶ表現とグラフ表現の統合

5.2　ツールのインタフェース

図 5.2 は開発した分析ツールのスクリーンショットである．本ツールは画面右側のメインパネルと，画面左側の上下に分割されているサブパネルから構成されている．メインパネルはデータがつぶつぶ表現で表示される画面であり，ユーザは主にこの画面で操作を行う．サブパネルは分析していく上で必要な機能や設定がタブ形式で表示される画面である．サブパネルはグラフパネル，属性パネル，詳細パネル，設定パネルで構成されている．

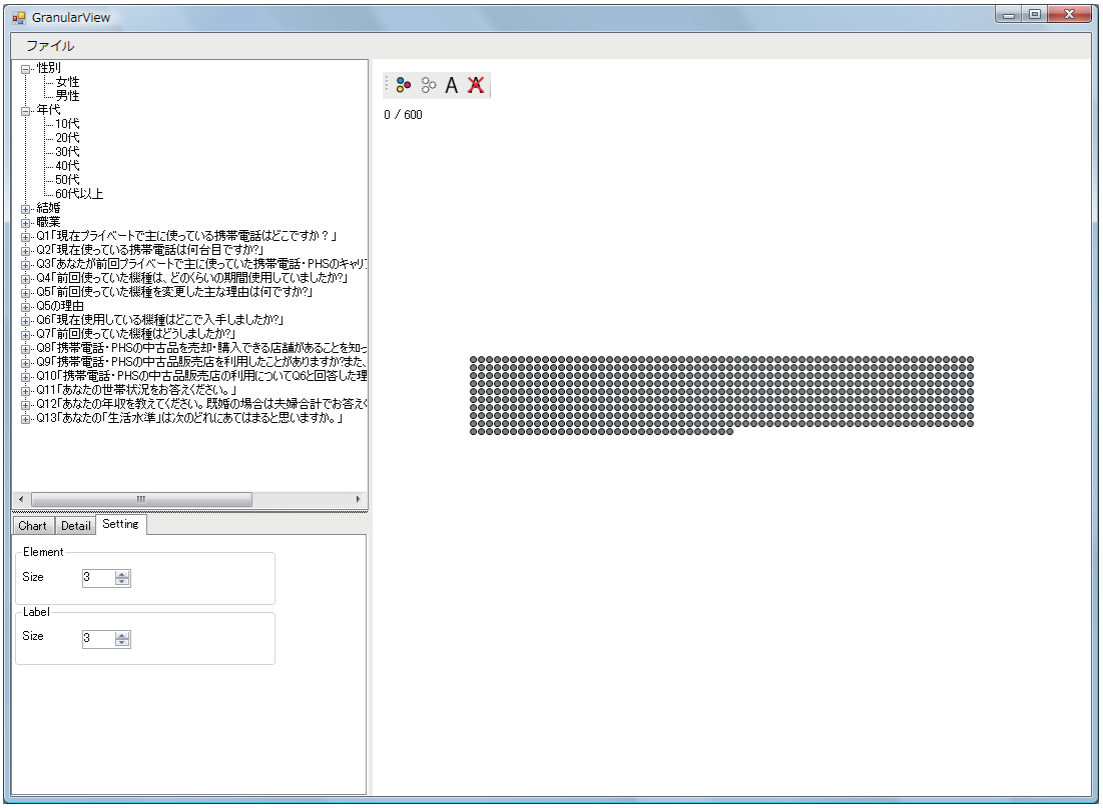

図 5.2: ツールの概観 (初期画面)

5.2.1　メインパネル

メインパネルはデータがつぶつぶ表現で表示される画面である (図 5.3)．画面の左上には操作ボタンが並んでいる．いくつかの機能はこの操作ボタンによって実行する．操作ボタンの下には現在選択している要素の数と，全要素の数が表示されている．

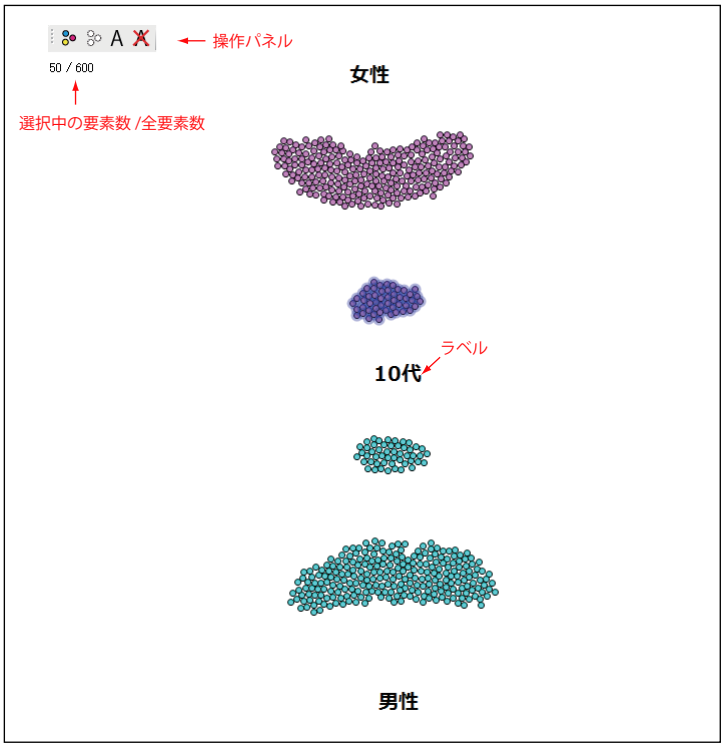

図 5.3: メインパネル

5.2.2　属性パネル

属性パネルは読み込んだデータの属性とそのカテゴリがツリー形式で表示されるパネルである (図 5.4).

5.2.3　詳細パネル

詳細パネルは要素の詳細情報が表示されるパネルである．メインパネル上で要素にマウスホバーをすると，その要素のカテゴリが表示される (図 5.5).

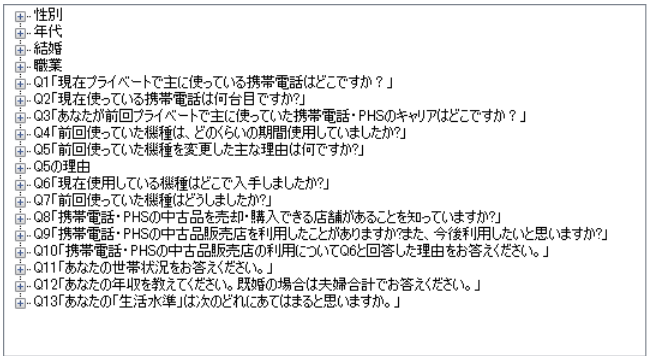

図 5.4: 属性パネル

Detail

性別	男性
年代	10代
結婚	未婚
職業	アルバイト、フリーター
Q1「現在プライベートで主に使って	A社（現在）
Q2「現在使っている携帯電話は何	2台目
Q3「あなたが前回プライベートで主	A社（前回）
Q4「前回使っていた機種は、どのくら	1年以上2年未満
Q5「前回使っていた機種を変更した	新しい機種が発売された
Q5の理由	
Q6「現在使用している機種はどこで	キャリアのショップで購入した(ex.ドコ
Q7「前回使っていた機種はどうしま	販売店に引き取ってもらった
Q8「携帯電話・PHSの中古品を売	知っているが、利用できるところにな
Q9「携帯電話・PHSの中古品販売	利用したことはなく、今後も利用した
Q10「携帯電話・PHSの中古品販	すぐに壊れそうだから
Q11「あなたの世帯状況をお答えく	親と同居（本人未婚）
Q12「あなたの年収を教えてくださ	150万～300万未満
Q13「あなたの「生活水準」は次のど	中の中

図 5.5: 詳細パネル

5.2.4　グラフパネル

グラフパネルはデータが棒グラフで表示される画面である (図 5.6)．ユーザはメインパネルで要素を選択し，選択した要素に関する棒グラフを作成することができる．Bar ボタンでは通常の棒グラフが作成される．Stacked ボタンでは積み上げ棒グラフが作成され，Stacked100 ボタンでは 100%の積み上げ棒グラフが作成される．

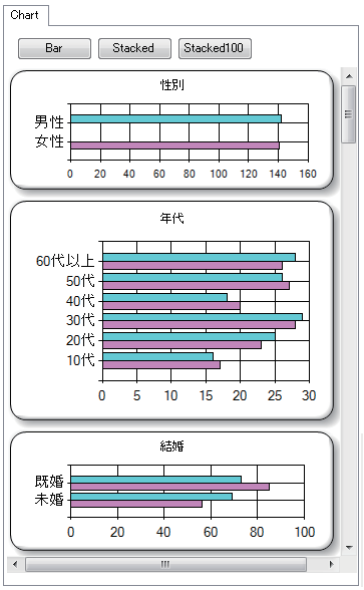

図 5.6: グラフパネル

5.2.5　設定パネル

設定パネルは要素やラベルの大きさ変更など，各種設定を行うパネルである．

5.3　実装

5.3.1　実装言語・使用 API 及びデータ形式

ツールの実装は C#(Microsoft .NET Framework 3.5) を使用した．棒グラフの作成及び表示部分にはグラフ作成 API である Microsoft Chart Controls を利用した．データの読み込みは CSV 形式で行っている．

5.4　ツールの機能

5.4.1　ラベルによる操作

本ツールにおける基本的な操作は 4.4 節で述べたように，ユーザが着目したいカテゴリによって，ドリルダウンする操作である．

データが読み込まれた初期状態では要素は矩形状に並ぶように整列されて配置されている．この状態から，性別における男性に着目したいとする．属性パネルから男性の項目をドラッグ&ドロップでメインパネル上に移動すると，メインパネル上には男性のラベルが表示される．メインパネル上に表示された男性のラベルをマウスで移動させると，メインパネル上の全要素の中から，男性のカテゴリに該当する要素がラベルの方向にアニメーションで移動する (図 5.7).

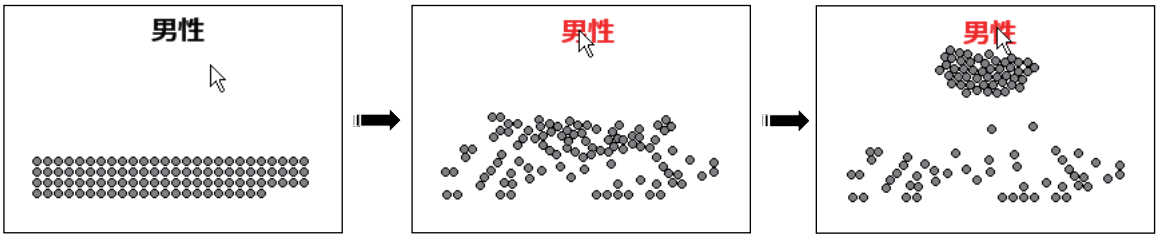

図 5.7: ラベルによる操作

複数のラベルを選択して移動させることで，そのカテゴリに該当する要素を同時に移動させることができる．例えば，年代に着目して，30 代以下の要素と 40 代以上の要素の 2 つに分けたい場合を考える，まず，10 代，20 代，30 代のラベルをメインパネル上に配置し，矩形選択によって全てのラベルを選択する．そして，いずれかのラベルを移動させることで，選択されている他のラベルも同時に移動し，30 代以下の要素が全てラベルの方向に移動する．同様の操作を 40 代，50 代，60 代のラベルに関しても行うことで，40 代以上の要素を分けることができる (図 5.8).

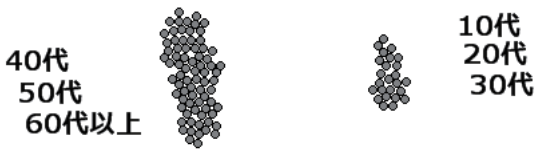

図 5.8: 複数ラベルによる操作

5.4.2　要素の選択

メインパネル上の何もないところでマウスを押し，押した状態で移動をすると矩形選択が開始する．ユーザがマウスのボタンを離すと，矩形の中に含まれる要素は選択状態となる．選択状態となった要素は (図 5.9 右) のように要素の周りが青くハイライトされる．

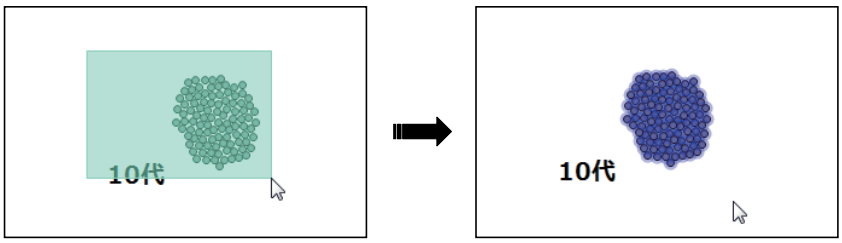

図 5.9: 要素の選択

5.4.3　要素の非アクティブ化

任意の要素を選択した状態で，操作ボタンの非アクティブ化ボタンを押すと，選択した要素を非アクティブ状態にすることができる．非アクティブ状態の要素は図 5.10 のように円の枠が点線になり，ラベルによる操作の影響を受けなくなる．ラベルによる操作はメインパネル上に表示されている要素全てを引き寄せる対象とするため，ユーザの意図しない要素を引き寄せてしまう場合がある．このような時に，着目していない要素を非アクティブ状態にすることで，その要素は一時的に分析から除外することができる．

5.4.4　任意のラベル作成

任意の要素を選択した状態で，操作ボタンのラベル作成ボタンを押すと，選択した要素を引き寄せるラベルを作成することができる．ラベルの名前はテキストボックス上で任意につけることができる．例えば，10 代かつ女性の要素を選択した状態で「10 代の女性」というラベルを作成する．このラベルを移動した際には 10 代かつ女性の要素が引き寄せられる (図 5.11).

任意のラベル作成は次のような場合に利用することができる．一つはユーザが自身のカテゴリを定義して作成したい場合である．年代の属性を含むアンケートデータでは 10 代・20 代・30 代…のように 10 歳ごとにカテゴリ分けされたデータとなっているのが一般的であるが，若い年代とそれ以上の年代との 2 つに分けたい場合などは，任意のラベル作成によって自身で定義したカテゴリを作成することができる．また，全ての要素を選択してラベルを作成することで，全ての要素を引き寄せるラベルを作成することや，ユーザが着目していない要素を選択して,「その他」のようなラベルをつけてメインパネル上の端に寄せるような使い方が考えられる．

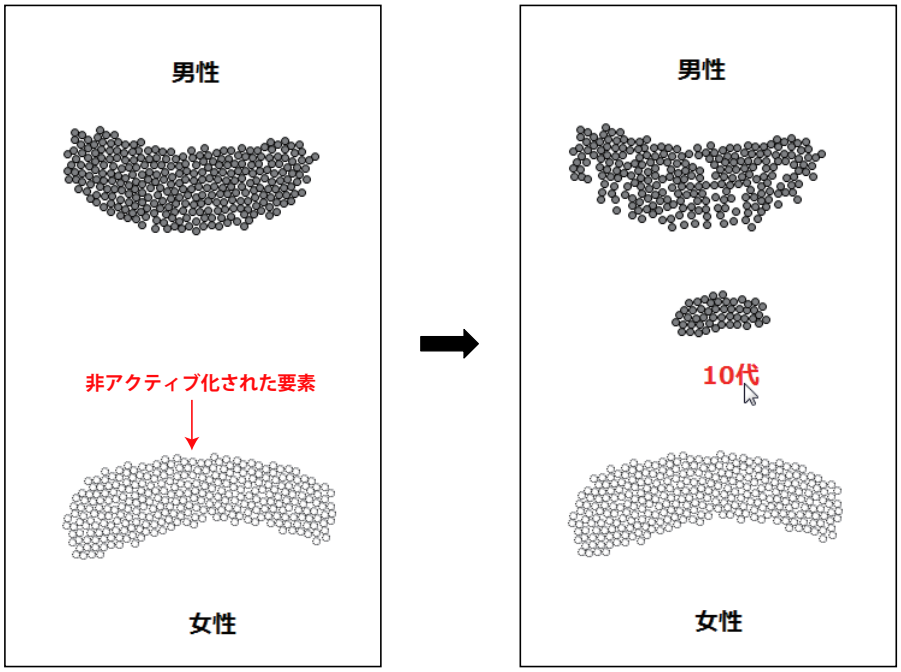

図 5.10: 非アクティブ化状態でのラベル操作

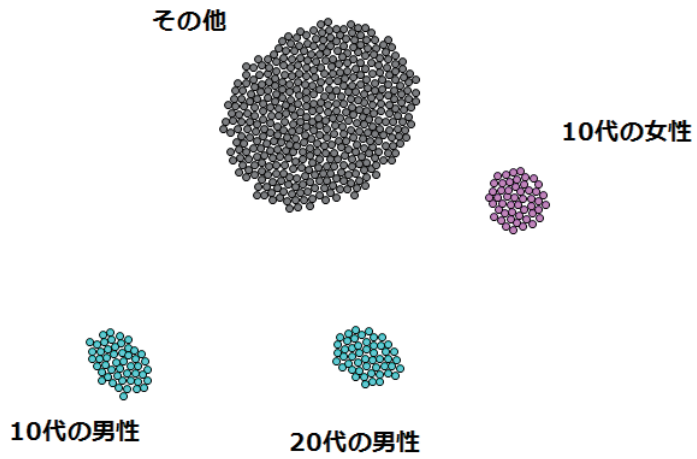

図 5.11: 任意のラベル作成

5.4.5　グラフの作成

要素を選択した状態で，グラフパネル上のグラフ作成ボタンを押すことで選択した要素に関するグラフを作成することができる．グラフは属性ごとに一つの棒グラフが作成され，全ての属性に関するグラフが縦に並んだ状態でグラフパネルに表示される (図 5.6).

5.5 要素のレイアウト

メインパネル上でラベルによって操作をしていく上での要素のレイアウト方法について述べる．要素のレイアウトにはグラフレイアウトなどに利用される力指向アルゴリズム (Force-based algorithms)[29][30] を参考にした．全ての要素間に働く斥力と，ラベルによって要素を引き寄せる引力の 2 つの力を計算することで要素のレイアウトを決定している．

5.5.1 要素間に働く斥力

メインパネル上に表示されている各要素間には互いに反発しあう斥力を計算している．図 5.12 は 3 つの要素が初めは互いに重なっていて，斥力を計算することで反発し合う様子を表したものである．赤い矢印は斥力を表す．各要素間に斥力を計算することで，ラベルによる操作で移動した際に，要素は互いに重なることなく円状に充填するように広がって安定する (図 5.13).

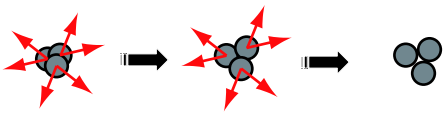

図 5.12: 要素間に働く斥力

斥力計算の疑似コードを Algorithm1 に示す．ここで，N は全要素の集合,$E_1, E_2, \cdots, E_k$ は各要素を表す．$distance(C_i, C_j)$ は C_i と C_j 間の距離を返す．$radius(E_k)$ は E_k の半径を返す．

Algorithm 1 要素同士の斥力の計算

```
loop
  for all E_i ∈ N do
    v⃗ = (0, 0)
    for all E_j ∈ N and E_i ≠ E_j do
      distance ← distance(E_i, E_j)
      if distance < radius(E_i) + radius(E_j) then
        Caluculate a vector v⃗_ij from E_i to E_j
        v⃗ ← v⃗ + v⃗_ij
      end if
    end for
    Move E_i in the direction of v⃗
  end for
end loop
```

図 5.13: 斥力によって円状に広がる様子

5.5.2 ラベルによる引力

ユーザがラベルを移動している間，要素にはラベルの方向に移動する引力が働く．引力はラベルと要素との距離が遠いほど強くなり，近いほど弱くなるように計算している．これにより，ラベルをドラッグし始めた時は要素が素早く移動し，だんだんゆっくりと移動するようになる．図 5.14 は 3 つ要素がラベルに引き寄せられる様子を表したものである．青い矢印は引力を示し，矢印の長さは引力の強さを表している．

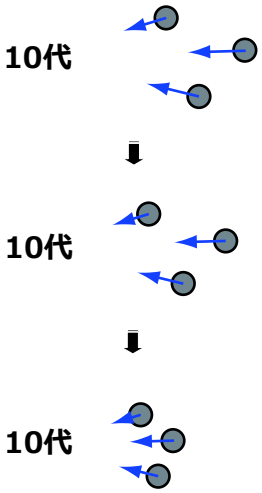

図 5.14: ラベルによる引力

引力計算の疑似コードを Algorithm2 に示す．ここで，$Label$ はユーザが操作しているラベルを表す．k は引力の強さを決定する定数である．

Algorithm 2 ラベルによる引力の計算

repeat
  for all $E_i \in N$ **do**
    $distance \leftarrow distance(E_i, Label)$
    Caluculate a vector $\vec{v}$ from E_i to $Label$
    Move E_i in the direction of $\vec{v} \times \frac{distance}{k}$
  end for
until Mousebutton is released

第6章　ケーススタディ

本ツールを使用して実際のデータを分析するケーススタディを示す.

6.1　タイタニック号の乗客乗員に関するデータ

使用するデータはタイタニック号の乗員乗客のデータである[1]．データ中の属性は表 6.1 に示す通りである.

表 6.1: タイタニック号のデータ

属性	属性の説明	カテゴリ
class	船室の等級	1st, 2nd, 3rd
sex	性別	female, male
survived	生存したかどうか	survived, died
age	年齢	整数値
age_categorized	年齢の属性を 10 歳ごとにカテゴリ化	10s, 20s, 30s,⋯, 90s
embarked	乗船した場所	
home.dest	住所/目的地	
room	部屋の番号	
ticket	チケット番号	
boat	救命ボートの番号	

まず初めに，基本的な属性である性別に着目して分析を試みる．属性パネルから「female」と「male」のカテゴリのラベルをメインパネル上に表示し，要素を 2 つに分ける．そして,「female」の要素には桃色を付け,「male」の要素には青色を付けた後，メインパネル上の全ての要素を選択してグラフを作成する (図 6.1).

図 6.1 のグラフパネルに着目する．属性「survived」に関するグラフを見ると，男性は生存者より死亡者の方が多いのに対して，女性は逆の傾向がみられることが分かる.

これをさらに詳細に分析するために,「survived」と「died」のカテゴリによって要素を分ける．そして,「survived」の要素には白色を,「died」の要素には黒色を付け，メインパネル上の全ての要素を選択して新たにグラフを作成する.

[1] http://www.statsci.org/datasets.html

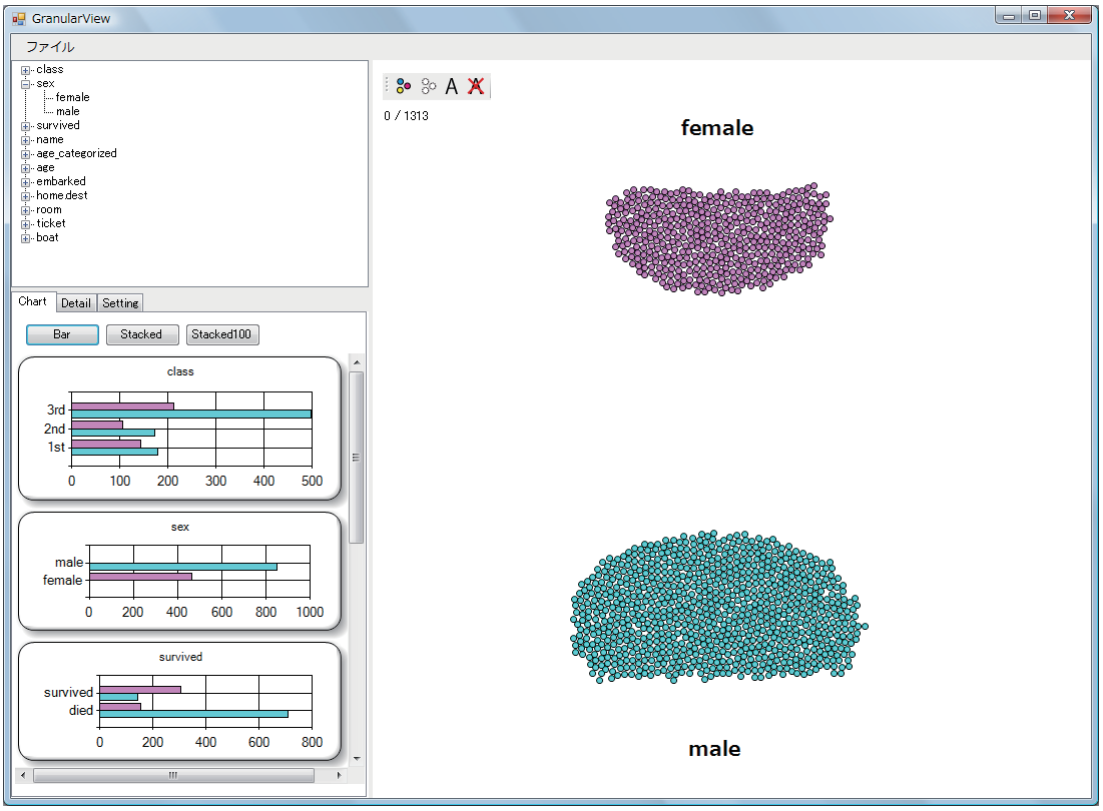

図 6.1: 属性「sex」に着目した分析

図 6.2 のグラフパネルに着目する．属性「class」に関するグラフをみると，船室の等級が上がっているほど生存率が増えていることが分かる.「3rd」の船室では 20%のみの生存率に比べて，1st の船室では 60%の生存率となっている．

属性「sex」に関するグラフをみると，ここでも性別によって明らかな傾向がみてとれる．女性は男性に比べて生存率が 3 倍以上高いこと分かる．

属性「age_categorized」に関するグラフをみると，年代によっては大きな差はみられないが,「child」のカテゴリだけは生存率が高いことが分かる．一方「70s」のカテゴリは死亡率が 100%であることがわかる．この原因として,「70s」の乗客は女性や船室の等級が高い人が多いということが考えられる．これを検証するため,「70s」のラベルによって,「70s」の要素を分ける．この状態から,「70s」の要素のみを選択し，これらの要素に関するグラフを分析することによって，局所的な傾向の分析が行える (図 6.3)．ここでは，特に女性や船室の等級が高い人が多いという傾向はないことがわかった．すなわち何か他の原因，もしくは「70s」の要素は少ないため，偶然生存率が上がったと考えられる．

以上のことから分析の結果をまとめる．船室の等級に関しては「1st」のカテゴリの要素の生存率は「2nd」と「3rd」に比べて高いことから,「1st」の船室を利用している乗客は優先して救助されていたことなどが考えられる．また，女性と子供に関しては生存率が高く，こち

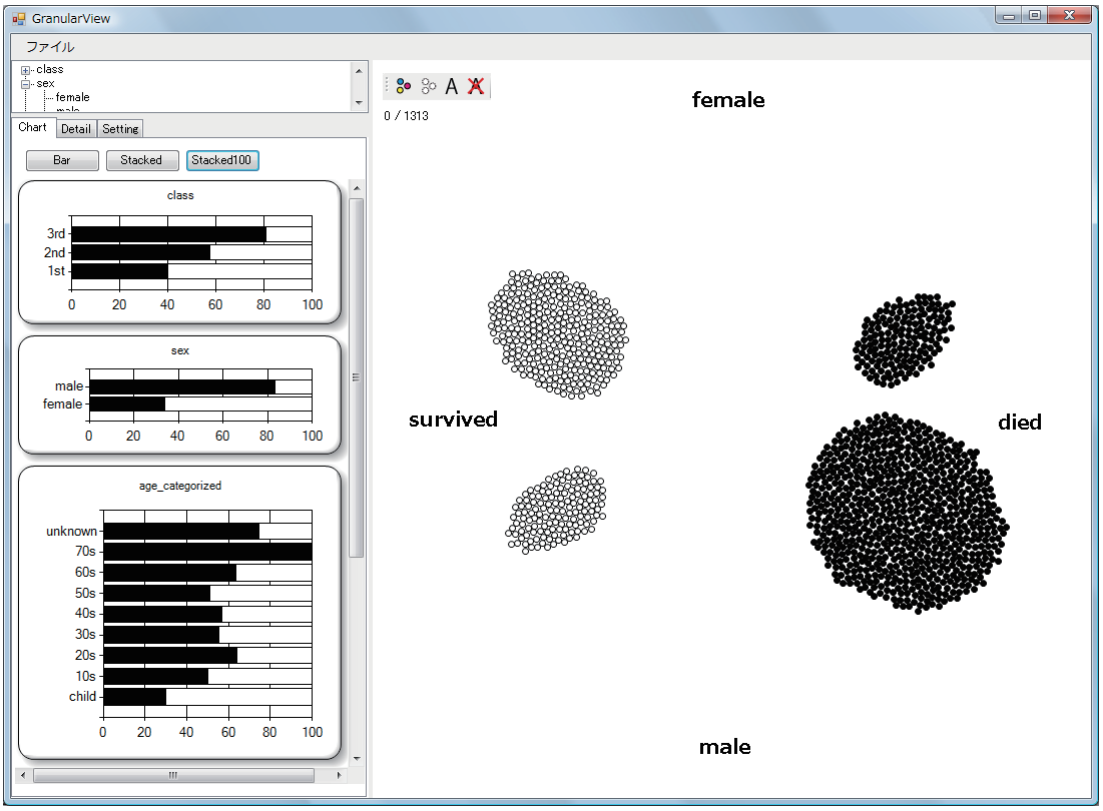

図 6.2: 属性「survived」に着目した分析

らも優先して救助されていることがわかる．一方，年代別には顕著な特徴が見られなかった．このことから，高齢者は優先されたといったことはないと考えられる．

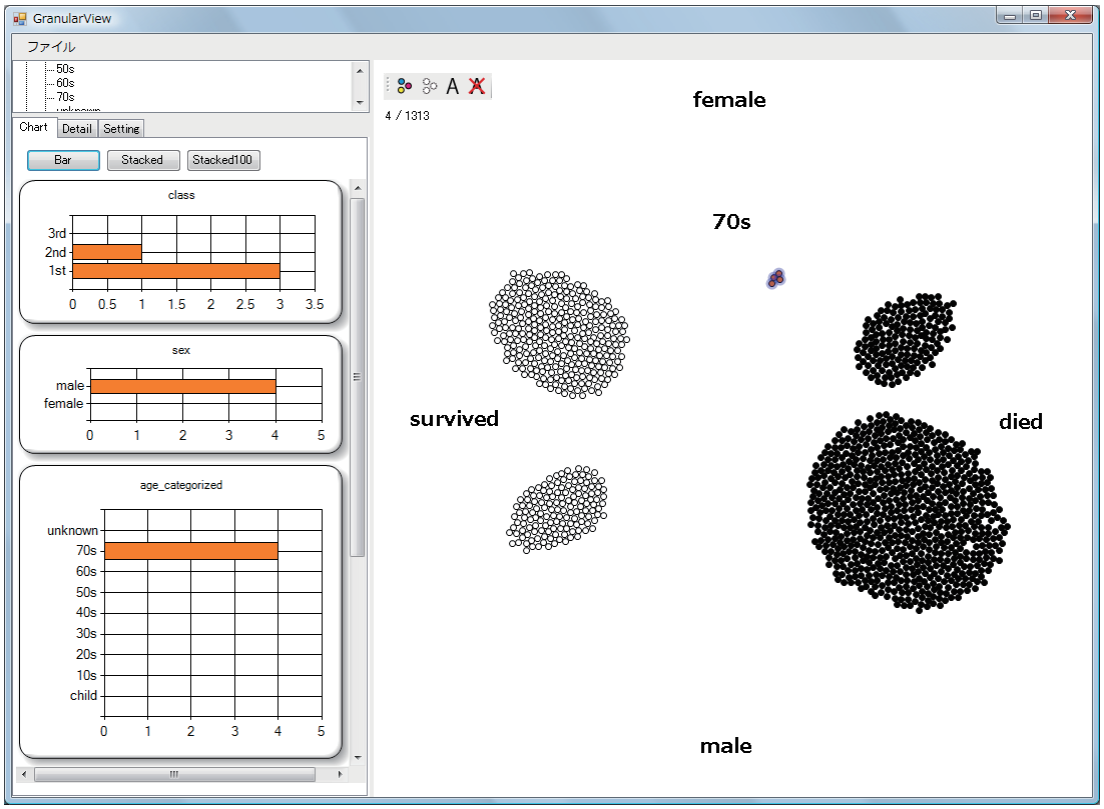

図 6.3: カテゴリ「70s」に着目した分析

6.2 携帯電話に関するアンケートデータ

使用するデータは携帯電話の利用に関して，いくつかの質問のアンケートを行ったデータである[2]．今回はこのデータをそのまま使用せずに，複数回答の属性などは現在のツールでは扱えないためいくつかの属性を削除した．また，年齢は整数値で定義されていたため，10 歳ごとにカテゴリ化を行い年代として定義し直した．また，具体的な携帯電話会社の社名は A 社，B 社，C 社のように匿名とした．最終的に使用した属性は表 6.2 に示す通りである．

表 6.2: 携帯電話に関するアンケートデータ

属性
年代
性別
結婚
職業
Q1「現在プライベートで主に使っている携帯電話はどこですか？」
Q2「現在使っている携帯電話は何台目ですか？」
Q3「あなたが前回プライベートで主に使っていた携帯電話・PHS のキャリアはどこですか？」
Q4「前回使っていた機種は，どのくらいの期間使用していましたか?」
Q5「前回使っていた機種を変更した主な理由は何ですか？」
Q6「現在使用している機種はどこで入手しましたか？」
Q7「前回使っていた機種はどうしましたか？」
Q8「携帯電話・PHS の中古品を売却・購入できる店舗があることを知っていますか？」
Q9「携帯電話・PHS の中古品販売店を利用したことがありますか？また，今後利用したいと思いますか？」
Q10「携帯電話・PHS の中古品販売店の利用について Q6 と回答した理由をお答えください.」
Q11「あなたの世帯状況をお答えください.」
Q12「あなたの年収を教えてください. 既婚の場合は夫婦合計でお答えください.」
Q13「あなたの「生活水準」は次のどれにあてはまると思いますか.」

まず初めに，アンケートの基本的な属性である性別に着目してみる．属性パネルから「男性」と「女性」のカテゴリのラベルをメインパネル上に表し，要素を「男性」と「女性」の 2

[2]株式会社ネットマイル (http://research.netmile.co.jp/) による公開調査のデータである．SPSS Japan(http://www.spss.co.jp) の SPSS データライブラリーにて無償提供されている．

つの要素に分ける.「女性」の要素には桃色を付け,「男性」の要素には青色を付けた後，全ての要素を選択してグラフを作成する (図 6.4).

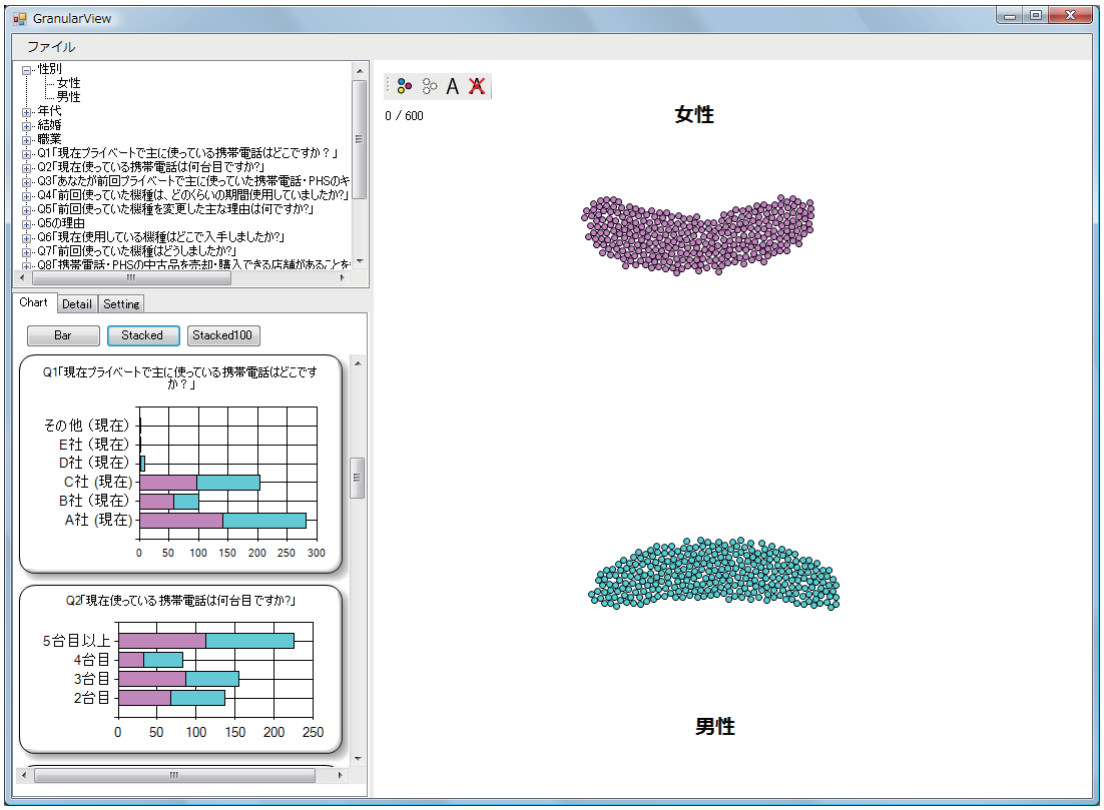

図 6.4: 属性「性別」に着目した分析

質問の一つである Q1「現在プライベートで主に使っている携帯電話はどこですか？」のグラフを見ると，A 社， C 社， B 社の順に多いことがわかる．さらに，A 社と C 社に関しては男女比率がほぼ同じであり，B 社に関しては女性の方が若干多いことが分かる．他の質問項目に関しても同様にいくつかの傾向がみられた.

次に，携帯電話会社に着目して分析を行う．Q1「現在プライベートで主に使っている携帯電話はどこですか？」の属性に関して，A 社， B 社， C 社の 3 つのカテゴリによって要素を分け，それぞれに色をつけてグラフを作成する (図 6.5)．属性「年代」に関するグラフに着目すると，C 社に関しては若い年代ほど人数が多いことがわかる．しかし，40 代に関しては例外で人数が多い．A 社に関しては 10〜30 代の年代について C 社と全く逆の傾向がみられる．すなわち，年代が上がるとほど人数が増えていることが分かる．B 社に関しては A 社と C 社のように傾向がみられず，全ての年代でおおよそ同数であることが分かる.

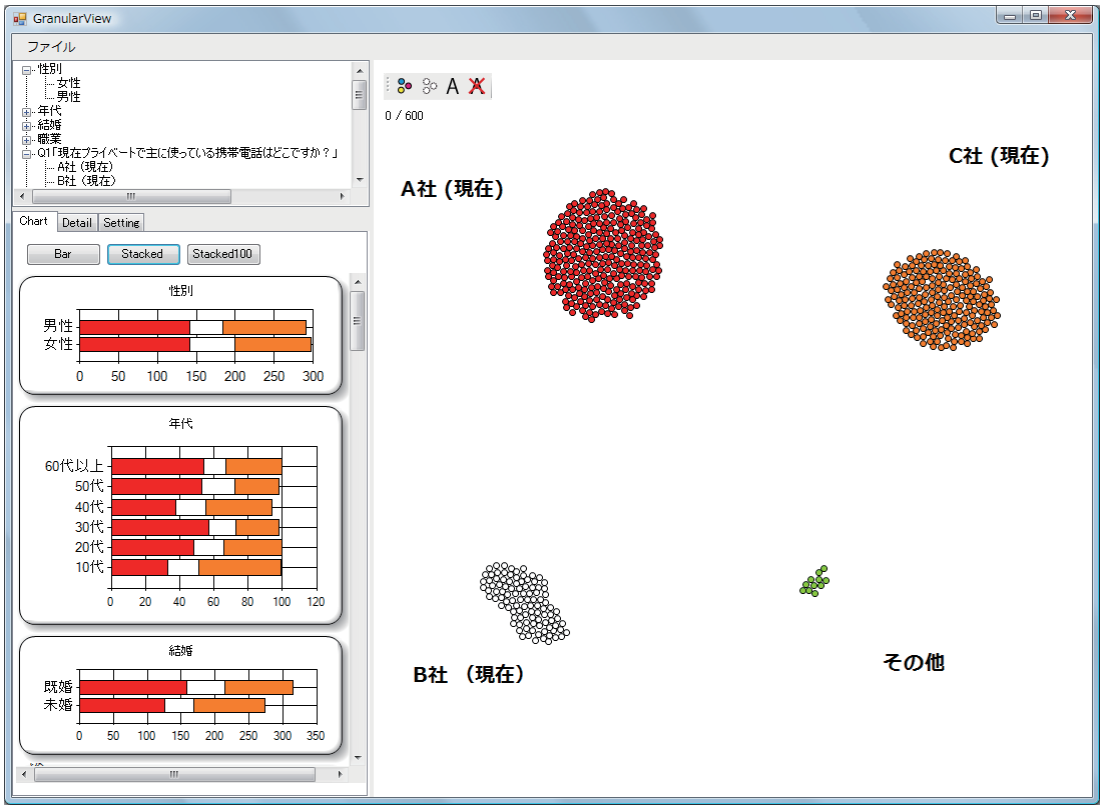

図 6.5: 携帯電話会社に着目した分析

第7章　評価実験

7.1　目的

カテゴリデータを分析する上での以下の 2 点を評価することを目的とする.

- 「つぶつぶ表現」の有効性
- 分析ツールとしての機能の有効性

7.2　概要

カテゴリデータを分析する上で必要となる操作をタスクとして設定し，開発した分析ツールと，従来のカテゴリデータ分析の基本であるピボットテーブルの両方を用いて被験者にタスクを行ってもらう．被験者はタスクごとにそのやり易さを 5 段階で評価する．その後，各ツールの使用感に関してのアンケートに回答する．ピボットテーブルには Microsoft Excel 2007 のピボットテーブル機能を使用した.

7.2.1　被験者

コンピュータサイエンスを専攻する大学生及び大学院生 6 名を対象とした．いずれの被験者もピボットテーブルを使用した経験はなく，本研究で開発したツールを使用するのも初めてである.

7.2.2　実験手順

被験者は以下の手順の通りに実験を行う.

1. ツールの使い方に関する説明を聞く.
2. 練習用のタスクを解きつつ，ツールに慣れるまで自由に触る.（最大 10 分程度）
3. 本番用のタスクを解く．タスクごとに 5 段階評価を行う.
4. 全てのタスクが完了後，ツールの使用感に関してのアンケートに回答する.

7.3　タスクの設定

実験のタスクはカテゴリデータ分析を行う上で必要となるタスクを想定した．しかし，カテゴリデータ分析の知識を要するタスクは，被験者の事前知識によって結果に影響すると考えたため，あくまで分析する上での操作に着目した．

タスクは分析する上で必要となる操作から以下の 3 つに分けた．

- 集計値を求めるタスク
- カテゴリを比較するタスク
- グループを比較するタスク

7.3.1　集計値を求めるタスク

集計表のある特定のカテゴリに該当する集計値を求めるタスクである．

(例) 以下のカテゴリに該当する人は何人いますか．

- Q1　⇒　A 社
- 年代　⇒　30 代，40 代

ピボットテーブルを使用した場合，指定された属性に関する集計表を作成した後，該当するカテゴリの集計値が答えとなる．本ツールを使用した場合，指定されたカテゴリのラベルによって要素を分けた後，要素を選択することでその数が答えとなる．

7.3.2　カテゴリを比較するタスク

カテゴリごとの集計値を比較して，指定された順に値が多いカテゴリを求めるタスクである．

(例) 以下のカテゴリに該当する人だけに着目します．Q4 に関して 2 番目に多いカテゴリは何ですか？

- Q8　⇒　知らない
- Q12　⇒　150 万〜300 万未満，300 万〜500 万未満

ピボットテーブルを使用した場合，指定された属性に関する集計表を作成した後，グラフを作成してカテゴリを比較することで答えが出せる．本ツールを使用した場合，指定されたカテゴリのラベルによって要素を分けて，選択した要素に関するグラフを作成する．指定された属性に関するグラフからカテゴリを比較することで答えが出せる．

7.3.3 グループを比較するタスク

ここでいうグループとは，あるカテゴリに該当する要素の集合である．

(例) 以下の 2 つのグループを比べます．Q3 に関して「B 社」と回答した人が多いのはグループ 1 とグループ 2 のどちらですか？

- (グループ 1) Q2　⇒　2 台目，3 台目
- (グループ 2) Q2　⇒　4 代目，5 代目以上

ピボットテーブルを使用した場合，指定された属性に関する集計表を作成した後，ピボットテーブルにおける「グループ化」の機能を利用して複数のカテゴリをグループとしてまとめる．その後，指定されたカテゴリの集計値の多い方が答えとなる．本ツールを使用した場合，指定されたカテゴリのラベルによって要素をグループとして分ける．それぞれのグループに別の色を付け，グラフを作成する．指定された属性に関するグラフからカテゴリを比較して多い方が答えとなる．

7.3.4 タスク概要

表 7.1 に示す 8 つのタスクをピボットテーブルと本ツールを用いてそれぞれ行う．両方のツールで行うタスクは同じ属性に関するタスクであるが，カテゴリは変更して行った．例えば，ピボットテーブルで行うタスクが男性の集計値を求める場合，本ツールで行うタスクは女性の集計値を求めるといったようにカテゴリを変更してある．

表 7.1: タスク一覧

タスク番号	タスクの内容
タスク 1	2 つの属性に関する集計値を求める
タスク 2	2 つの属性に関してカテゴリを比較する
タスク 3	1 つの属性に関する 2 グループを比較する
タスク 4	4 つの属性に関する集計値を求める
タスク 5	3 つの属性に関してカテゴリを比較する
タスク 6	2 つの属性に関する 2 グループを比較する
タスク 7	2 つの属性に関する 3 グループを比較する
タスク 8	4 つの属性に関して，異なる 2 つの属性の 2 グループを比較する

上記のタスクを完了後，ツールの使用感についての以下のアンケートに回答してもらった．

- 慣れるまでに時間がかかる印象があるか
- カテゴリデータを分析する上で直感的な操作で分析を行えたか

- ビポットテーブルにおいて，集計表に基づいた分析はやり易いか
- 本ツールにおいて，クエリを模した操作による分析はやり易いか
- ピボットテーブルと比べた本ツールの利点 (自由記述)
- ピボットテーブルと比べた本ツールの欠点 (自由記述)

7.4 結果

表 7.2 に各タスクに対する被験者の評価を示す．ピボットテーブルを用いたタスクには Pivot table + タスク番号のように表記し，本ツールを用いて行ったタスクには Our tool + タスク番号のように表記してある．

表 7.2: タスクの結果

タスク番号	被験者 1	被験者 2	被験者 3	被験者 4	被験者 5	被験者 6	平均
Pivot table 1	4	1	null	3	4	5	3.4
Pivot table 2	2	2	2	2	4	3	2.5
Pivot table 3	2	1	4	3	4	5	3.4
Pivot table 4	2	1	4	2	5	4	3
Pivot table 5	1	1	4	2	4	3	2.5
Pivot table 6	2	1	2	2	3	3	2.2
Pivot table 7	2	1	4	2	2	2	2.2
Pivot table 8	2	2	4	2	5	3	3
Our tool 1	4	5	4	5	5	4	4.5
Our tool 2	3	4	4	4	5	4	4
Our tool 3	3	3	4	4	5	4	3.8
Our tool 4	2	2	3	5	4	5	3.5
Our tool 5	3	2	3	4	5	5	3.7
Our tool 6	2	4	4	4	5	5	4
Our tool 7	2	2	5	4	5	4	3.7
Our tool 8	1	1	3	4	5	2	2.7

5=簡単　4=やや簡単　3=普通　2=やや難しい　1=難しい

7.5 考察

各タスクの評価平均に関するグラフ (図 7.1) を見ると，タスク 1〜7 に関しては本ツールの方がピボットテーブルに比べてタスクがやり易いと被験者は回答していることが分かる．

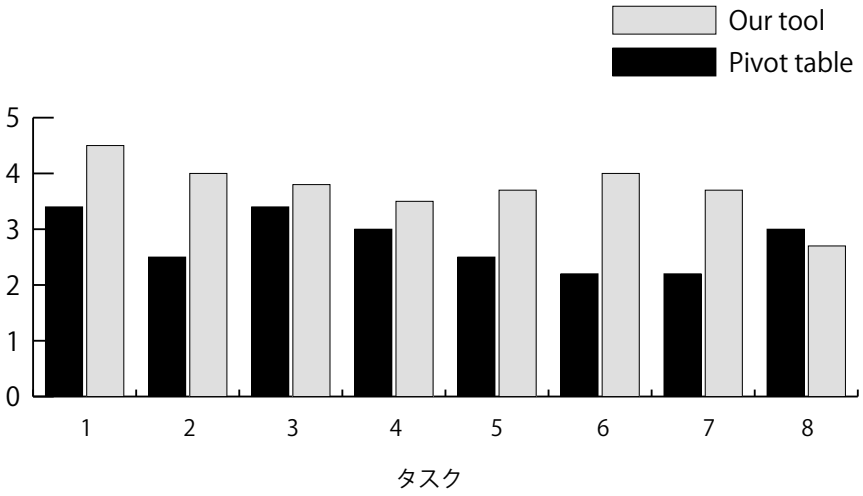

図 7.1: 各タスクの評価平均のグラフ

表 7.3: ツールの使用感に関するアンケート結果

ピボットテーブルに関する質問

質問項目	被験者 1	被験者 2	被験者 3	被験者 4	被験者 5	被験者 6	平均
慣れるまでに時間がかかる	5	3	4	4	5	5	4.4
直感的な操作で分析が可能	2	3	3	3	1	1	2.2
集計表に基づいた分析はやり易い	3	5	4	2	2	3	3.2

本ツールに関する質問

質問項目	被験者 1	被験者 2	被験者 3	被験者 4	被験者 5	被験者 6	平均
慣れるまでに時間がかかる	1	1	2	1	1	2	1.4
直感的な操作で分析が可能	4	5	4	5	5	5	4.7
クエリを模した操作による分析はやり易い	4	5	4	4	5	5	4.5

5=そう思う　4=どちらかというとそう思う　3=どちらともいえない　2=どちらかといえばそう思わない　1=そう思わない

しかし，タスク 8 に関してだけはピボットテーブルの方がタスクがやり易いという結果となった．タスク 8 はグループを比較するタスクであるが，唯一両方のグループに属する要素

が存在するタスクである．従って，両方のグループに属する要素の扱いがこの結果となった原因であると考えられる．これはアンケートデータにおける複数回答が可能な属性と同じような扱いである．現在の実装ではこのような複数のカテゴリに該当する要素の分析には対応できていない．

次に，タスク後のアンケート結果について検証する．「慣れるまでに時間がかかるか」という質問に関しては，本ツールについては全ての被験者から「そう思わない」または「どちらかというとそう思わない」の回答を得た．一方，ピボットテーブルではほとんどの被験者から「そう思う」との回答を得た．

「直感的な操作で分析が可能か」という質問に関しては，本ツールについては全ての被験者から「そう思う」または「どちらかというとそう思う」の回答を得た．一方，ピボットテーブルでは「普通」またはそれ以下の回答を得た．

ピボットテーブルに関しての「集計表に基づいた分析はやり易いか」という質問と，本ツールに関しての「クエリを模した操作による分析はやり易いか」という質問に関して各被験者の差を比べると，2 人に関しては差はないものの残りの 4 人に関しては本ツールの方が肯定的との回答を得た．

上記のアンケート回答結果から，ピボットテーブルを用いた場合は慣れるのに多少時間がかかるという欠点を各被験者は感じていると考えられる．

7.6 今後の課題

7.6.1 量的データへの対応

開発した分析ツールはカテゴリデータのみを対象としており，現在は量的データへの対応はできていない．しかし，棒グラフ以外に散布図などを導入することで量的データへも対応が可能である．また視覚的要素の色や大きさなどによっても量的データを表現できると考えられる．例えば，要素の色相を値の大きさに対応させることや，要素の大きさを値に対応させることが考えられる．

7.6.2 レイアウト計算コスト

つぶつぶ表現の視覚的要素のレイアウトを計算するコストは，現在の実装では高速化等を行っていない．現在のレイアウト計算では要素数 n に対して計算量が $O(n^2)$ で大きくなる．そのため，要素数が多くなると処理が重くなる問題がある．この解決方法として，空間分割やグラフレイアウト技術を応用することで，レイアウトの計算コストを削減することができると思われる．特に力指向のグラフレイアウト技術においては，その計算量を減らすアルゴリズムが古くから研究されており，本研究にも参考になると思われる．

第8章　結論

本研究では，カテゴリデータ分析を目的とした視覚的表現であるつぶつぶ表現を開発した．つぶつぶ表現はデータにおけるエンティティを視覚的要素として表現し，ゲシュタルト要因による色と配置によってカテゴリを表現する．

つぶつぶ表現を適用してカテゴリデータの分析ツールを開発した．既存のグラフ表現とつぶつぶ表現を統合することで，多次元データにおいても複数の属性を一覧表示することができる．つぶつぶ表現を用いて分析する上でインタラクティブなドリルダウンを支援する機能として，クエリを模したラベルによる操作を開発した．ユーザはラベルを動かすと，要素がアニメーションによって移動し，視覚的なドリルダウンが可能である．

評価実験ではカテゴリデータ分析で想定されるタスクを設定し，つぶつぶ表現及び分析ツールの有効性の評価を行った．分析ツールを使用した被験者の主観的評価では，ピボットテーブルに比べて本ツールを用いた方がカテゴリデータを分析する上でのタスクを行い易いとの回答を得た．

本研究で開発した視覚的表現により，従来の表に基づいた分析に代わり，インタラクティブな分析が可能になる．これは今後のデータ分析手法の発展の手がかりになる可能がある．

謝辞

本研究を行った 2 年間，三末和男准教授には日々の研究活動において多大なご指導を頂きました．研究が順調に進み，無事に論文執筆ができたのは先生のご指導のおかげです．本当に有り難うございました．田中二郎教授には研究室決定から相談にのって頂き，研究においても多大なアドバイスを頂きました．心から感謝しています．志築文太郎講師，高橋伸講師には研究発表の場において有意義な意見を頂き，研究を進める上で大変参考になりました．有り難うございました．

インタラクティブプログラミング研究室の皆様には公私共に大変お世話になりました．

NAIS チームの皆様には日々のゼミはもちろん，研究を進める上で有用な議論をして頂きました．皆様と過ごした日々の研究室での生活は忘れられない思い出になりました．

WAVE チームの皆様には公私共に深い付き合いをさせて頂き，大変お世話になりました．研究室での生活が充実したのは皆様が暖かく迎えて下さったおかげです．

Ubiquitous チームの皆様にはいつも暖かい励ましを頂き，有り難うございました．皆様との日々の何気ない会話は研究活動をする上での励みになりました．

学生生活最後の 2 年間をインタラクティブプログラミング研究室の仲間と共に過ごすことができて，本当に嬉しく思います．

最後に，学生生活を送る上で多大な援助をして下さった家族には大変感謝しています．家族の支えなしには充実した生活を送ることはできなかったと思います．本当に有り難うございました．

参考文献

[1] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Pub, 1999.

[2] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, Vol. 1, No. 4, pp. 69–91, 1985.

[3] Jean-Daniel Fekete Niklas Elmqvist, Pierre Dragicevic. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, pp. 1141–1148, Nov/Dec 2008.

[4] Ji Soo Yi, Rachel Melton Ponder, John Stasko, and Julie Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. In *Information Visualization*, Vol. 4, pp. 239–256, 2005.

[5] M. Sifer. User interfaces for the exploration of hierarchical multi-dimensional data. *Symposium On Visual Analytics Science And Technology*, pp. 175–182, 2006.

[6] Fabian Bendix, Robert Kosara, and Helwig Hauser. Parallel sets: Visual analysis of categorical data. In *Proceedings of the IEEE Symposium on Information Visualization 2005 (INFOVIS'05)*, pp. 133–140, 2005.

[7] Graham J. G. Upton. Cobweb diagrams for multiway contingency tables. *Journal of the Royal Statistical Society*, Vol. 49, No. 1, pp. 79–85, 2000.

[8] Michael Friendly. *Visualizing Categorical Data*. Sas Inst, 2000.

[9] Erica Kolatchm and Beth Weinstein. Cattrees: Dynamic visualization of categorical data using treemaps, 2001.

[10] Brian Johnson and Ben Shneiderman. Treemaps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of IEEE Information Visualization '91*, pp. 275–282, 1991.

[11] Matthias Schonlau. Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics*. American Statistical Association, 2003.

[12] Geoffrey Draper and Richard Riesenfeld. Who votes for what? a visual query language for opinion data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1197–1204, 2008.

[13] Ra Becker, Cleveland WS, and Shyu M-J. The design and control of trellis display. *Journal of Computational and Statistical Graphics*, No. 5, pp. 123–155, 1996.

[14] Daniel Keim, Ming Hao, Umesh Dayal, Meichun Hsu, and Julain Ladisch. Pixel bar charts: A new technique for visualizing large multi-attribute data sets without aggregation. *IEEE Symposium on Information Visualization*, p. 113, 2001.

[15] Daniel A.Keim, Ming C.Hao, and UmeshwarDayal. Hierarchical pixel bar charts. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 03, pp. 255–269, 2002.

[16] Ramana Rao and Stuart K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 318–322, New York, NY, USA, 1994. ACM.

[17] Zhicheng Liu, John Stasko, and Timothy Sullivan. Selltrend: Inter-attribute visual analysis of temporal transaction data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, No. 6, pp. 1025–1032, 2009.

[18] Wolfgang Freiler, Kresimir Matkovic, and Helwig Hauser. Interactive visual analysis of set-typed data. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 14, No. 6, pp. 1340–1347, 2008.

[19] Shixia Liu Xinghua Lou and Tianshu Wang. Fanlens: A visual toolkit for dynamically exploring the distribution of hierarchical attributes. In *Visualization Symposium, 2008. PacificVIS '08. IEEE Pacific*, pp. 151–158, 2008.

[20] Chris Stolte and Pat Hanrahan. Polaris: A system for query, analysis and visualization of multi-dimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, pp. 52–65, 2002.

[21] Matthew O. Ward. Xmdvtool: integrating multiple methods for visualizing multivariate data. In *VIS '94: Proceedings of the conference on Visualization '94*, pp. 326–333, Los Alamitos, CA, USA, 1994. IEEE Computer Society Press.

[22] Deborah F. Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. *Comput. Stat. Data Anal.*, Vol. 43, No. 4, pp. 423–444, 2003.

[23] Ben Shneiderman and Catherine Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley, 2004.

[24] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Pub, 2004.

[25] Jeffrey Heer and George Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 13, No. 6, pp. 1240–1247, 2007.

[26] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 1, pp. 1–8, 2002.

[27] S. Eick and G. Wills. High interaction graphics. *European Journal of Operational Research*, Vol. 81, No. 3, pp. 445–459, 1995.

[28] Allen Martin and Matthew Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th conference on Visualization '95*, pp. 271–278, 1995.

[29] P. A. Eades. A heuristic for graph drawing. In *Congressus Numerantium*, Vol. 42, pp. 149–160, 1984.

[30] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exper.*, Vol. 21, No. 11, pp. 1129–1164, 1991.