論文 / 著書情報
Article / Book Information

論題	データの可視化・文章化技術を複合させた大規模表データ探索システム
著者	細川 夏生, 有本 昂平, 脇田 建
出典	日本ソフトウェア科学会 第37回大会講演論文集
発行日	2020, 9

データの可視化・文章化技術を複合させた大規模表データ探索システム

細川 夏生　有本 昂平　脇田 建

近年，自然言語生成と可視化を組み合わせたデータ分析システムが増えてきている．ユーザーがデータに埋もれている重要な事実を理解するのに，文章という媒体は可視化単体に比べて特別な知識が必要ないという点で優れている．しかし，データが大きいと当然統計的特徴も多くなってしまうため，大規模なデータを自動的に文章で要約しようとしても長大になってしまう．本研究ではこのスケーラビリティの問題に対処するため，多くの大規模データが項目と属性の両方に顕在的，潜在的に持っている階層構造を利用して，第一にユーザーが興味を持っている部分にのみ焦点を絞って文章化と可視化を行い，第二にユーザーの興味の移り変わりに応じてインタラクティブに焦点を移動させることで，一度に提示する文章の量を抑制しつつもデータ全体をスムーズに探索可能なシステムのアイデアを提案する．また，実際にシステムを実装し，このアイデアが有効であることを示す．

In recent years, data analysis systems that combine natural language generation and visualization are increasing. The medium of text is superior to visualization in that it does not require special knowledge for users to understand the important facts buried in data. However, since larger data contain more statistical features, it is difficult to automatically summarize large amounts of data in the text. In this study, we address this scalability problem by using the hierarchical structure that many large data sets have explicitly or implicitly, in both items and attributes. First, the system focuses text generation and visualization only on the areas that users are interested in. Secondly, it interactively shifts the focus as the user's interest shifts. In this way, we propose an idea for a system that allows users to explore the entire data smoothly while limiting the amount of text presented at once. We also implement the system and show that this idea is effective.

1 はじめに

近年，データ分析のためのツールとして，探索的可視化の技術とデータを説明する文章を自動生成する技術を組み合わせる研究が盛んである [2] [8] [10] [11] [12] [15] [19] [20]．探索的可視化はユーザーの自由な情報探索行動を促し，文章の自動生成技術はデータマイニングの結果得られた統計的事実をわかりやすく提示する．この両者を組み合わせることで，ユーザーがデータを迅速に理解するために必要な機能を補完しあうことができる．このアプローチは，可視化されているデータの特徴を文章によって説明することによってデータへの理解を深めることができることから，特に可視化リテラシーが十分でないユーザーが正しくデータを理解するのに有効である [19]．

このようなアプローチのひとつの課題は，データセットの増大に対応したスケーラビリティである．過去の研究で扱われたデータセットでは属性数が数個に留まっている [15] [10] [11]．これまでに提案された方式を素朴に属性数が多いデータセットに適用した場合，

An Explorative System for Large-scale Table Data That Combines Data Visualization and Text Generation Techniques

Natsuki Hosokawa, 東京工業大学情報理工学院, School of Computing, Tokyo Institute of Technology.

Kohei Arimoto, 株式会社帝国データバンク 総合研究所, 東京大学大学院情報学環, Teikoku Databank, Ltd., Center for TDB Advanced Data Analysis and Modeling; The University of Tokyo, Graduate School of Interdisciplinary Information Studies.

Ken Wakita, 東京工業大学情報理工学院, 帝国データバンク先端データ解析共同研究講座, School of Computing, Tokyo Institute of Technology; Teikoku Databank, Ltd., Center for TDB Advanced Data Analysis and Modeling.

属性数の多項式倍のオーダーの文章が生成されることが予想される．たとえば，国連加盟国を比較する 51 個の基礎指標を有する Social Progress Index [18] に対して文章を合成したところ，391,273 単語からなる長大な文章が生成された．

本研究の目標は，属性数が多い大規模表データにおいて探索的可視化と自動文章生成を組み合わせたシステムを実現することであり，具体的な課題として，すでに述べたスケーラビリティのほか，概要の提示と興味深い領域の焦点化の機能，探索行動の支援，詳細情報の提示と比較，そして協調作業の支援を挙げた．

本研究では，多くの大規模表データが属性とデータ項目の両方に潜在的，顕在的に階層構造を持っていることに着目し，階層構造を活用したデータ要約と文章要約を行うアプローチをとった．具体的には，属性方向，データ項目方向のふたつの軸を利用した**ふたつの焦点化の軸**に沿った文章要約技術と，それらの軸に沿った**二次元的ドリルダウン**の機能を提供することによる問題解決を図った．

また，本提案にもとづいて前述の Social Progress Index データセットを視覚的分析するためのシステム SPIViewer[†1] を実装した．この SPIViewer を用いて本提案の有効性を，単語数の計測に基づく文章のスケーラビリティの評価と，ユーザー実験による定量的・定性的評価によって示した．

2 関連研究

この章では，目標という面で本研究に近い自動洞察システムと，それに関わりの深い自然言語処理に関する研究についてまず紹介する．さらに，本提案の主要な構成要素である自然な文章の生成と探索的可視化に関する重要な研究について紹介する．

2.1 自動洞察システムと自然言語処理

近年，大きな統計データに統計分析を適用することでデータの重要な性質，すなわち**データファクト** [21] を抽出する手法への関心が高まっている．データファクトとは，ここでは外れ値解析に代表される統計的計算によってデータから抽出される統計的事実のこととする．データファクトの量はデータセットの大きさに依存し，大きなデータセットではたくさんのデータファクトの中からユーザーにとって特に重要なものを抽出する必要がある．Tang らは統計データから k 個の重要なデータファクトを抽出する技術を確立した [21]．Demiralp らの Foresight は，さらにデータファクトを探索的に分析するためのシステムである [4] [5]．これらの自動洞察システムは，ユーザーが大きく複雑なデータを概観する目的で開発されている．

Microsoft Power BI [14] や Google Sheets [6] は前述の自動洞察システムと自然言語処理，そして情報可視化を組み合わせた機能を提供している．この手法では，自動的に抽出された統計的特徴の内容を説明する図と短文を生成する．これにより，ユーザーは統計データの重要な特徴を視覚的にも文章からも理解を深めることができる．また，Quill [3], Wordsmith [13] はこのようなシステム上で自然言語生成の機能を強化するプラグインで，テンプレートベースの NLG によって文章を生成し，ユーザーが可視化を使って統計的特徴を理解するのを助ける．

2.2 自然な文章の生成

前述のシステム群が採用した自然言語処理では，データファクトごとに短文を生成した．しかし，それぞれの文は独立していて一つの文章としての繋がりを持っていないため，データセット全体の把握にあまり向かない．Latif らは論文の共著関係を対象とする VIS Author Profiles [11] や地理的に分布している事象を表現した 2 属性データセットを対象とする interactive Map Reports (iMR) [10]，プログラムコードに関して可読性や保守性を評価したデータを対象とする視覚的分析システム [15] を提案している．これらは人間が執筆したデータ分析結果の文章に基づいて設計された文書テンプレートを用いることで，自然で読み易い文章を生成する．

2.3 探索的可視化

本研究のもっとも重要な課題は，項目と属性がともに大きな表的データセットを対象としたデータ可

†1 SPIViewer: `https://smartnova.github.io/spi/`

視化と文章生成を組み合わせたシステム構築である．表的データセットとは，表計算ソフトなどで扱われるような，行がデータの項目を表し，列が属性を表し，それぞれのセルが自らの属する行が表すデータ項目が持つ，列が表す属性についての値を持っているデータである [16]．探索的検索システムについて White と Roth は，ファセットとメタデータベースの検索結果フィルタリングの提供，洞察と意思決定をサポートする可視化の提供，協働分析活動などの支援を含む，8 つの要件を指摘している [24]．

自動洞察システム [5] [9] では，ユーザーは必要な情報をシステムが提示した k 個のデータファクトから得られる一方，探索的可視化はユーザーが主体的に必要なデータを探索するという点で異なる．ユーザーごとに置かれた状況や好みは異なっており，必要な情報は異なる．したがって，特定のデータのプロフェッショナルに限られない，様々な状況にある利益関係者がユーザーになりうる場面においては，ユーザーが主体的にデータを探索して自らが必要とする情報を見つけられるシステムが有効である．探索的可視化は自動洞察システムに比べて必要な操作が多くなる代わりに，ユーザーの置かれた状況や好みなどのニーズにより柔軟に対応できるという利点がある．

3 システム

この章では，前述した課題を克服するシステムのための抽象的なアイデアを 3.1 節で，より具体的なアイデアを 3.2 節で紹介する．残りの節ではこれらのアイデアに基づいて実装された，SPI データセットを分析するためのシステム SPIViewer が実際にどのように実装されているのか，具体的な説明を行う．

3.1 要件分析

大きな表的データセットを対象としたデータ可視化と文章生成を組み合わせたシステムの構築にあたって必要な要素として，以下の六つの要件 (design considerations) を挙げる．

DC1. ユーザー主導の分析 — 本研究ではユーザーが主体的にデータの分析を行うシステムを提案する．ユーザーが全てを調べ尽くすことが困難なほど多くのデータファクトがあるデータセットを分析しなければならないとき，重要なデータファクトだけを抽出する必要がある．重要なデータファクトの抽出には，自動で行う方法と，ユーザーの分析による方法が考えらえる．ユーザーが自ら分析を行う方法は，自動で行う方法と比べて時間や手間といったユーザーの負担が多い代わりに，ユーザーの置かれた状況や好みを反映してデータファクトを抽出できるため，特に多様な利益関係者をユーザーとする場合に有効である．

DC2. スケーラビリティ — 本研究では，データセットを分析するユーザーの視覚的探求活動 [24] を補助するために，そのデータセットを説明する図と文章を自動生成するアプローチを取る．注目に値するデータファクトの数は，データ項目数と属性数の増加に伴い組合せ論的に増加する．従来の手法の素朴な応用では，図も文章も表現しなければならないデータファクトの数が増大することが課題である．可視化が表示する要素が増加する問題に対しては，過去に精細な描画手法，データクラスタリング，データの多層化，インタラクションの採用などさまざまな研究がなされてきた．一方，説明文の生成については研究が浅く，データの複雑化に伴う文章量の抑制方法はあまり扱われていない．

DC3. 概観タスクと注目点タスク — 大規模なデータ分析には，データを概観し要約すること，全体を部分構造に分離すること，パターンを発見するなどといった**概観タスク**と，ユーザーが関心を持っている特定の領域を詳細に分析する**注目点タスク**がある．いずれのタスクについても，ユーザーにとって有益なインターフェイスを提供することが求められる．

DC4. 情報探索の指針 — 未知のデータについての探索的な分析においては，まず概観を把握し，得られた発見に応じて，焦点となる部分構造についてさらに詳細に分析することが求められる．この自然な繰り返し作業の一般的な指針として Shneiderman らが唱える Information seeking mantra は広く受け入れられている [17]．

DC5. 詳細分析と比較分析 — 単一のデータ項目に

ついての情報を詳しく分析するタスクと，複数のデータ項目を比較するタスクの両方が重要である．特定の項目や項目の集合について分析するとき，単に注目している項目の詳細について知りたい場合と，類似，あるいは対照的な別の項目との違いについて知りたい場合とがあるからである．属性についても同様に詳細分析と比較分析が重要である．

DC6. 協働活動の補助 — 生成した文章と可視化を他人と共有する機能が必要である．共有機能は，複数の探索者で作業を分割して，重要な情報を相互に共有することで作業時間を短縮するのに役立つ．また，分析して得られた情報を，分析には直接関与しない利益関係者に情報共有し，内容を説明することにも使うことができる．

3.2 設計の概要

前述したように表的データの分析において，項目数と属性数についてのスケーラビリティ (DC2)，概観的視点と絞りこまれた視点の提供 (DC3)，そして探索的分析の支援 (DC1, DC4) が求められている．本研究では多くの大規模表的データにおいて，しばしば項目，属性それぞれに階層構造がある点に着目した．スケーラビリティへの対処においては，分析者が項目，属性それぞれの階層について，自ら焦点を設定できる「ふたつの焦点化の軸」を採用し，階層構造における焦点を文脈とした文章を生成することにより文章量の爆発を避けている (DC1, DC2)．また，二種類の階層性を扱うために，システムの階層的な探索機能は「二次元的ドリルダウン」という概念に沿って設計した (DC1, DC3, DC4)．このふたつの概念を表したものを図 1 に示す．分析作業においては，分析者が注目するデータ項目や属性，あるいはそれらの階層の情報を提示する機能を提供する．さらに，データ項目や属性の組，あるいは項目や属性からなる階層の組を比較する機能も提供する (DC5)．データ分析を実施する任意の時点で分析状況を保存し，他者に共有する機能も提供する (DC6)．

3.3 データファクトの抽出

自動洞察システムはデータファクトを文章化することによって，データセットを説明する．本研究は Latif の iMR [10] の提案に沿って，極値（最小値，最大値など），一次元外れ値，二次元外れ値の三種の統計量を元に文章を生成する [1]．ただし，巨大なデータセットを概観している場合，この方法を素朴に適用すると自動生成される文章が膨大になる．そこで本研究では，焦点となっている部分のみに注目して詳細に記述し（図 1 の赤い領域），さらに焦点となっている範囲でも，低い階層から得られるデータファクトは集約的に説明する（図 1 の黒い太枠の粒度で説明する）ことで冗長性を排除している．これにより，人間が選択した階層にふさわしい抽象度の文章を提示することができる．

一次元外れ値の抽出においては，Hoaglin [7] が推奨し，iMR でも採用している第 1，第 3 - 四分位数の幅を基準として判定している．二次元外れ値については，相関係数の絶対値が十分に大きい (> 0.7) 属性対を対象とし，主成分分析における第二主成分が一次元外れ値となるものとした．

ふたつのデータ項目の属性値を比べて論じることはしばしばある．このとき，データの分布を無視し，属性値の多寡にのみ注目すると統計的なバイアスに陥る．このため，属性値の差が統計的に意味があるかを見極めることは重要である．本研究では，各属性についてデータセットの全項目について線形回帰分析を施し，ふたつのデータ項目の属性値の差が回帰直線の傾きと比較して大きく偏位する場合のみを抽出している．SPIViewer では，データ項目の順位に対する属性値の差が回帰直線の傾きの 3 倍以上のものをデータファクトとして扱っている．

ふたつの焦点と二次元的ドリルダウンの概念を実現するには，階層構造の各ノード（末端ノードを除く）の概要を集約した代表値を算出するための抽象化機構が求められる．各ノードの概要をまとめるための算出方法はデータセットの文脈に依存し，単純に平均をとることが常に適切というわけではない．本研究で扱った SPI は予め属性の階層構造と代表値が与えられており，SPIViewer ではこれをそのまま利用した．

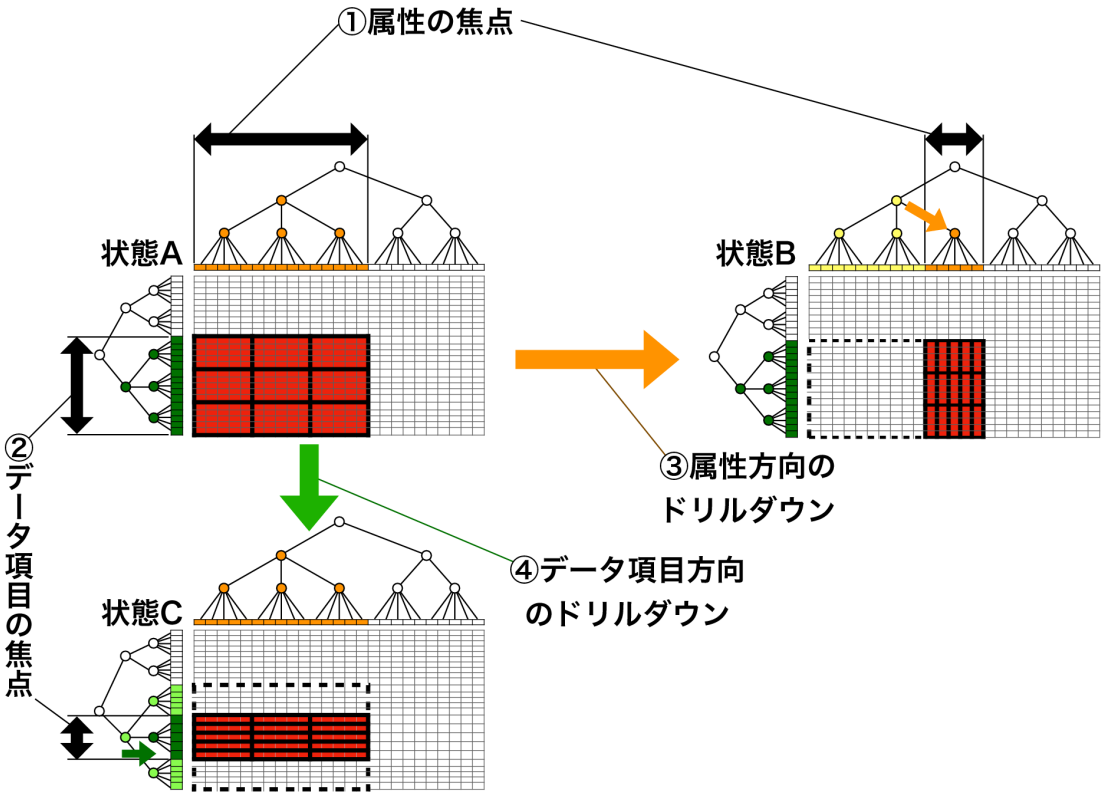

図 1　「ふたつの焦点化の軸」と「二次元的ドリルダウン」の概念図．ここには例として A, B, C の３つの状態が描かれていて，状態はユーザーの探索行動によって遷移し，システムはそれぞれの状態の時に赤い領域についてのみ，黒い太枠の粒度の詳細度で可視化と文章生成を行う．①はユーザーが属性について焦点化の軸に沿って焦点を絞ったときの焦点の範囲を示していて，同様に②は焦点を絞ったデータ項目の範囲を示している．この属性の焦点化の軸とデータ項目の焦点化の軸をまとめて「ふたつの焦点化の軸」と呼ぶ．③は状態 A のときに属性方向にドリルダウンを行なって状態 B に遷移したときの例を示している．状態 B に遷移したとき，属性の焦点が絞られ，システムは属性に関して状態 A のときよりも詳細な情報を示す．同様に，④はデータ項目方向にドリルダウンを行なったときの A から C への状態遷移を示している．この，ふたつの次元におけるドリルダウンをまとめたものが「二次元的ドリルダウン」である．

一方，データ項目についての地域的階層については，国連が定める地域階層 [22] を用い，国と地域の人口 [23] についての重みつき平均によって算出した．この算出方法を採用した根拠は，SPI が国や地域が生む富や経済規模の指標ではなく，基本的人権，人生における満足度，社会への参画の度合いのような，個人ごとの人生の充実に関する指標であるため，人についての平均が相応わしいと考えたためである．

3.4　可視化

可視化については，データセットの中で，ふたつの焦点化の軸によってユーザーが焦点を絞った範囲について文脈に応じて図示すればよい．データセット全体に占める焦点の当たった範囲を位置付けるために，全データを図示し，そのうち焦点内のデータに視覚的ハイライトを施すことも有効である．

SPIViewer において，地域的焦点については地図を表示している（図 2）．地図上に地域的焦点に含ま

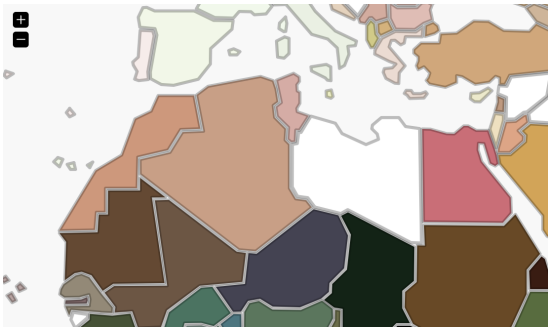

図 2　SPIViewer において，地域的な焦点が北アフリカとなっていたときに表示される地図．

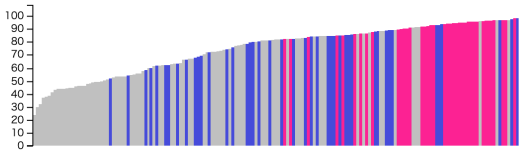

図 3　SPIViewer において，属性の焦点が生活必須要素であるときに表示される棒グラフ．ここではヨーロッパとアジアの比較をすることがデータ項目の焦点となっており，ヨーロッパがピンク，アジアが青でハイライトされている．

れる国や地域を示す (DC2) とともに，選択された国や地域の概要を SPI データセットにおける総合評価を構成する 3 つの主要属性（**生活必須要素**，**社会的環境**，**社会進出の機会**）のそれぞれに RGB 表色系の各色成分を割り当てることで国と地域の全体的な傾向を観察できるようにしている (DC3)．たとえばエジプトに与えられた特徴的な赤みがかった色■から，エジプトは赤■で代表される**生活必須要素**が，他の属性（**社会的環境**■と**社会進出の機会**■）に比べて飛び抜けて優れていること，そして，北アフリカ諸国と比較してこの傾向が顕著なことが読み取れる．

また，属性の焦点に応じて棒グラフを生成し，焦点となっている国々をハイライトすることで，焦点でない国々との比較をできるようにする (DC2)．同様に，二つの属性を比較しているときは棒グラフの代わりに二次元散布図を生成する (DC5)．さらに，棒グラフと散布図の両方について，二つの地域や国を比較する際には，二つの色を使ってハイライトする (DC5)．

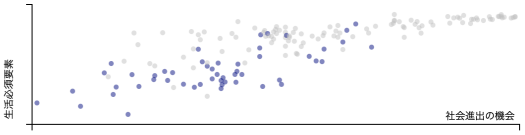

図 4　SPIViewer において，社会進出の機会と生活必須要素について比較しているときの散布図．ここではアフリカが青くハイライトされている．

図 3, 4 に SPIViewer でのグラフの表示方法を示す．

3.5　文章生成

ふたつの焦点化の軸に沿ってユーザーが定めた範囲について，説明する文章を生成する必要がある．SPIViewer では，説明の出力にあたって，焦点の範囲として (1) 特定のデータ項目が選択されている場合，(2) データ項目の階層構造におけるある中間的なデータ項目群が選択されている場合，(3) 二つのデータ項目群が選択されている場合に応じて三種の文章テンプレートを提供している．以下では，SPI データセットを例としてそれぞれの場合について説明する．

(1) 特定のデータ項目が選択されている場合

データが意味的に適切な階層化がなされていれば，データ項目階層の根から選択されているデータ項目に至るパス上の要素は，根 $\supset$ 中間階層のノード$_1$ $\supset$ ... $\supset$ 中間階層のノード$_n$ $\ni$ 選択されているデータ項目，という形で表せる意味的な包含関係がある．

この包含関係の成立するデータ項目群のそれぞれと，選択されている末端のデータ項目を比較することによって，選択されているデータ項目の特徴がより広い文脈のなかで明らかになる．たとえば SPI データセットにおいて，ドイツが焦点に選ばれているとき，世界におけるドイツの特徴，ヨーロッパにおけるドイツの特徴，そして西ヨーロッパにおけるドイツの特徴を順次調べることによって，ドイツという国をさまざまな地域的な広がりの文脈で特徴づけることができる．このことを利用し，SPIViewer では，参照されているデータ項目群の属性値を文脈として，選択されたデータ項目が極値を取る属性の紹介，一次元，二次元外れ値となる属性（の組）の紹介をする文章を生成

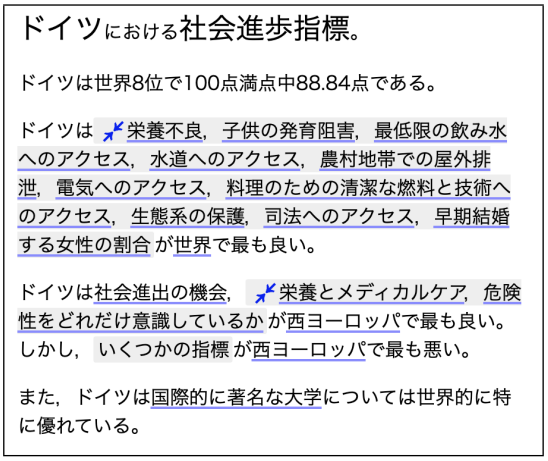

図 5　日本語版 **SPIViewer** において，データ項目の焦点がドイツのときに生成される文章．世界一である属性は，ヨーロッパでも西ヨーロッパでも当然一位なので省略されている．西ヨーロッパで最も悪い指標について「いくつかの指標」という表現を使っているが，この薄灰色になっている部分はクリックして展開することができ，指標の一覧をみることができる．焦点としてドイツを選択したあと何も操作をしなければ，世界一，西ヨーロッパーの指標についても同じように省略されているが，この図はクリックして展開したときの様子である．展開後は薄灰色部分の左端のアイコンをクリックすることで，省略した状態に戻すことができる．

する．

ただし，この方法を単純に適用すると地域の包含関係から自明なデータファクトを文章化してしまう．たとえば，ドイツは栄養不良の人の割合が世界一少ない国の一つであるなど，10 の指標について世界最高値を得ている．世界最高値なのでドイツがヨーロッパと西ヨーロッパにおいても，これらの指標について一番となることは明らかである．したがって，これらの自明な事実を下位階層の文脈で繰り返し文章化することは冗長である．本システムは上位階層で言及された属性について，下位階層での言及を抑制することで，簡潔な文章を生成している（図 5）．

生成される文章の中で，言及したい属性が多数ある場合もある．すべてを列挙すると冗長な文章が生成されてしまうため，本研究では属性の階層性に着目し，

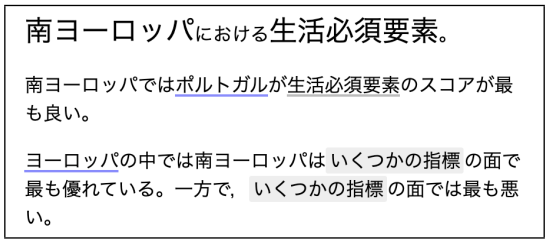

図 6　日本語版 **SPIViewer** において，データ項目の焦点を南ヨーロッパとしたときに生成される文章．

ある中間層についての説明を出力した場合に，その中間層に包含される下位階層の属性についての説明は省略する．これもまた，下位階層の属性は意味的に上位階層の属性に包含されていると考えられるからである．

また，文章の冗長性を減らすもう一つの工夫として，焦点となっている属性の階層よりも二つ以上低い階層の属性についてはデフォルトでは列挙を省略するようにしている．これは意味的な包含に関係のない省略なので，分析者が必要に応じて省略された属性群を表示できるようなインタラクションを提供している（図 5）．

これらの属性群の表示の冗長性を低減する工夫は (2), (3) の場合にも行なっている．

(2) データ項目の階層構造におけるある中間的なデータ項目群が選択されている場合

SPIViewer ではまず，選択された要素が代表するデータ項目群のうち，総合評価の最も高い項目を紹介する．例えば，地域と指標の階層についてそれぞれ南ヨーロッパと生活必須要素が選択されると，南ヨーロッパで 生活必須要素 の値が最も高い国がポルトガルであることを説明する文章を生成する．次に，データ項目の階層において，南ヨーロッパと同レベルに位置する他の地域，すなわち東，西，北ヨーロッパとの比較から南ヨーロッパが優れている指標と，劣っている指標を紹介する説明が出力される（図 6）．

(3) 二つのデータ項目群が選択されている場合

SPIViewer は，**比較モード**ボタンをクリックすることで二つのデータ項目群（A 群と B 群とする）を選択でき，この二つを比較した文章を生成する．こ

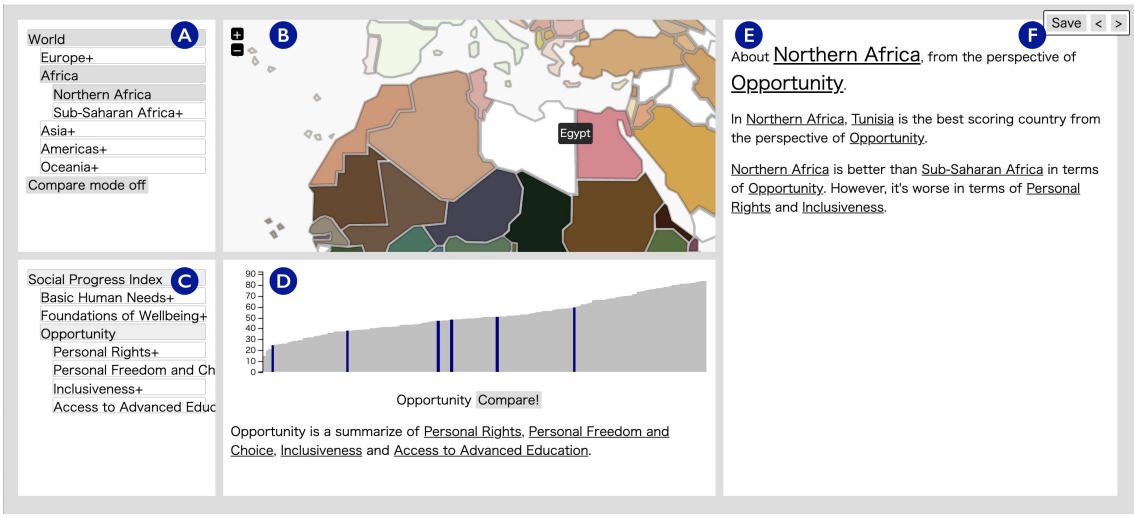

図 7　SPIViewer の全体図．地域選択ペインⒶと地図ペインⒷで国や地域を選択し，属性選択ペインⒸとデータ可視化ペインⒹで属性を選択すると，選択した国や地域，属性に応じてデータ可視化ペインⒹには可視化が，説明ペインⒺにはデータを説明する文章が生成される．また，地図ペインⒷは色を使ったデータ全体を俯瞰する可視化も兼ねている．サポートペインⒻは探索結果の共有と探索履歴上の移動をサポートする．

のとき全属性を，A 群が目立って優れている属性群，B 群が目立って優れている属性群，A と B の値に大きな違いが見出せない属性群（以降，**僅差属性群**と呼ぶ）の三種に分類する． SPIViewer では，文章のテンプレートとしてこの 3 種の比率に応じた 5 種類のパターンを用意している．(i) 全属性で一方が優れている場合（したがって僅差属性がない場合）には，その優れた群がもう一方を圧倒することを紹介する．(ii) 半数以上の属性で一方が優れており，残りの属性すべてが僅差の場合，その優れた群について「多くの属性で優れている」と述べる．(iii) 一方の群が半数未満の属性について優れ，残りの属性が僅差の場合は，一方が部分的に優れていることを述べ，その属性を列挙する．(iv) 両群がそれぞれに優れている属性を持つ場合は，双方それぞれが優れている属性を列挙する．(v) 全ての属性が僅差属性の場合は両群に目立った差がないことを述べる．このように適度に属性の列挙を避けて要約することで可読性を向上し (DC2, 3)，かつ自然な文章を生成する．

(1), (2), (3) で生成する文章中で極値について述べるにあたって，SPI データセットのデータ中にある，必ずしも値が大きいほど良い訳ではない属性について事前に処理する必要があった．値が小さいほうが良いと思われる属性に関しては，単に文章の内容を反転させた．指標 “Gender parity in secondary enrollment” については大きくても小さくても良いとは言えず，SPIViewer の実装にあたっては，この指標を省いた．

3.6　ユーザーインターフェイス

本研究のアイデアに基づいて，SPI データセットを視覚的に分析するためのウェブアプリケーションとして実装した SPIViewer の全体図を図 7 に示す．画面は主に**地域選択ペイン**Ⓐ，**地図ペイン**Ⓑ，**属性選択ペイン**Ⓒ，**データ可視化ペイン**Ⓓ，**説明ペイン**Ⓔの五つに分割されている．地域選択ペインと地図ペインはデータ項目の階層の中で，注目する領域の指定に用いる (DC2)．属性選択ペインとデータ可視化ペインでは属性階層における属性の選択ができる (DC2)．ペインⒶⒷⒸⒹを操作すると，選択内容に応じて適切なデータの説明文章が説明ペインⒺに表示される．**サポートペイン**Ⓕでは探索結果の可視化と文章の共有や，検索履歴をたどることができる．

地域選択ペインは地域的階層についてのツリー

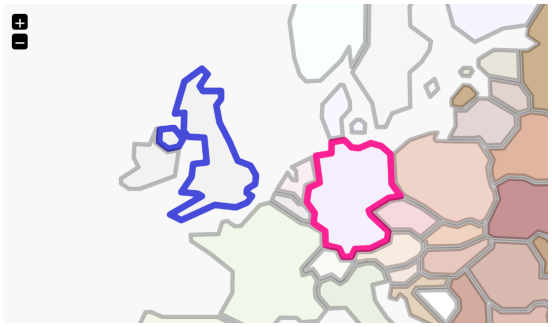

図 8　地域的な焦点としてドイツとイギリスを比較しているときの地図ペイン．イギリスを青，ドイツをピンクで縁取っている．

ビューを提供している．ツリービューは，分析者が世界の国々のなかから関心のある地域を地域的階層に沿って絞り込むことを促す (DC4)．地域選択ペインでの操作に応じて，地図ペインは選択された地域を自動的に拡大表示する．分析者は地図ペインから国や地域を選択することもできる．選択されたものは青く縁取られる．地域選択ペインで選択される地域，あるいは地図ペインで選択される国や地域は，地域的焦点を設定する．地域的階層の焦点が選択されるごとに，説明ペインに表示される内容は該当地域の説明に更新される．地域選択ペインには**比較モード**トグルスイッチを提供している．比較モードでは，事前に選択されていた国や地域と，その後，選択された国や地域の統計データを比較し，その結果を説明ペインに出力する．ふたつの国を比較する場合には，地図ペインでは一方が青く縁取られ，もう一方がピンクで縁取られる（図 8）．このとき，説明ペインでも同様の配色で下線が引かれ，データ可視化ペインでも同じ配色で棒グラフか散布図がハイライトされる．

属性選択ペインでは，地域選択ペインと同様のツリービューによって，SPI データセットに含まれる 51 個の属性とそれらをまとめた集約属性 16 個からなる上位階層から構成される階層構造を閲覧し，階層を辿って属性的焦点を絞りこめる (DC4)．すべての国と地域について，選択された属性の値が棒グラフとしてデータ可視化ペインに表示され，選択されている地域に該当する国と地域が青くハイライト表示される．図の下には，該当する統計データの出典と簡単な説明，そして情報源へのリンクが表示される．データ可視化ペインでも地域選択ペインと類似の比較機能を利用できる．トグルスイッチを利用することで，分析者は二つの異なる統計量を散布図によって観察し，相関分析を実施できる．

最後に，閲覧作業の補助をするサポートペインについて説明する．本システムは二次元的ドリルダウンについての履歴管理をしている．分析者が地域的焦点，あるいは属性的焦点を変更するたびにシステムの履歴に新しい焦点の情報が追加される．“<”と “>”のアイコンは，記録された閲覧履歴に沿って後退，前進を指示する．Save ボタンは分析画面のスナップショットを保存するための機能を提供する．ボタンを押すと現在の画面の状態をそのまま再現できる URL が生成され，その URL を新規タブで開く．新規タブが開いた時点で，この URL はウェブブラウザの閲覧履歴にも記録され，ウェブブラウザを終了してもその状態を回復したり，ウェブブラウザのブックマークにも保存できる．また，この URL を共同分析者と共有しあうことで協同分析作業の効率化が達成できる．さらに，ソーシャルネットワーク上で共有し，一般社会の人々に分析結果を共有することも容易である．

4　評価

この章では，本論文のアイデアを実装した SPIViewer を用いた SPI データセットの分析についてのユースケースを示し，スケーラビリティに関する文章量についての定量評価，ユーザー実験による既存の表計算ソフト（Google スプレッドシートを使用した [6]．以下，GSS と略記する）との比較と定量的，定性的評価について論じる．また，ここでは地理的焦点が A で属性的焦点が B のときの状態を (A, B) と表現することとする．

4.1　ユースケース

SPIViewer を起動したときに地図ペインに表示される世界全図のなかで，北アフリカにある特徴的な赤い色の国に関心を持ったとしよう．この国をクリックすると焦点は（全世界，社会進歩指標）から，(選択し

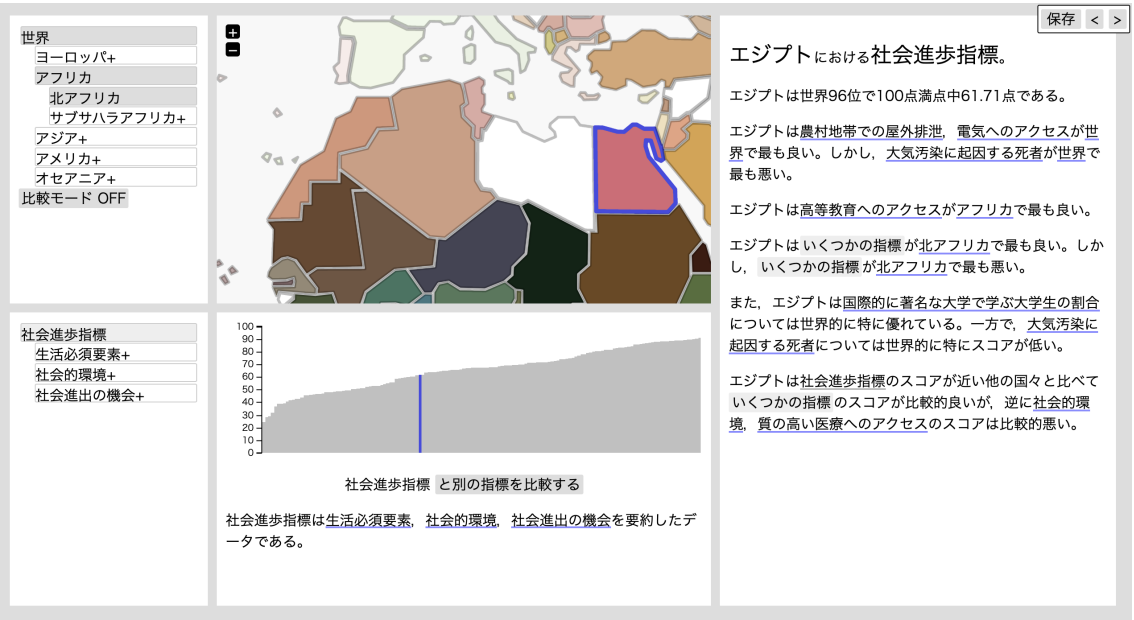

図 9　日本語版 SPIViewr で，システムの初期状態からエジプトを選択した時の様子．

た国，社会進歩指標）に遷移する．ただし，ここで**社会進歩指標**とは SPI の全属性に関する総合評価データである．この結果，地図ペイン，データ可視化ペイン，そして説明ペインにはここで絞り込まれたデータ項目に合わせた内容に更新される（図 9）．説明ペインに表示される文章とデータ可視化ペインに表示される**社会進歩指標**の分布から，選択した国，すなわちエジプトの世界全体における順位とその良し悪しがわかる．説明ペインに提示された文章から，エジプトがアフリカにおいて**高等教育へのアクセス**のスコアが高く，**大気汚染に起因する死者**の割合については世界最悪なことがわかる．ここで説明ペインの**大気汚染に起因する死者**の文字をクリックして選択すると，焦点は（エジプト，大気汚染に起因する死者）に移動する．更新したデータ可視化ペインでは確かにエジプトがこの属性について最悪値をとることだけでなく，その深刻度がずば抜けていることも読み取れる（図 10）．ここで戻るボタンを利用して焦点を戻し，今度はエジプトがアフリカで最高値を誇る**高等教育へのアクセス**をクリックしてみると，説明ペインの文章からエジプトがこの指標について世界第 39 位であることがわかる．指標についての比較機能を利用して，こ

図 10　大気汚染に起因する死者（10 万人あたり）．地理的な焦点をエジプトとしているので右端のエジプトがハイライトされている．

の指標を**社会進歩指標**と比較すると，これら 2 指標の散布図において，エジプトが全世界的傾向のなかで突出して優秀な高等教育を提供していることがわかる．

ここで，エジプトだけでなくアフリカ全体の傾向に興味が移る．データ項目選択ペインでアフリカを選択すると，焦点は（アフリカ，高等教育へのアクセス）となる．この結果から，アフリカの大部分の国々がこの指標の成績が悪い反面，エジプト，南アフリカをはじめとする優秀な国々が少数あることがわかった．

アフリカについての調査を続けるために，属性選択ペインで**社会進歩指標**を選択した．説明ペインの文章からモーリシャスがアフリカにおける**社会進歩指標**トップの国であることがわかる．文章中のモーリ

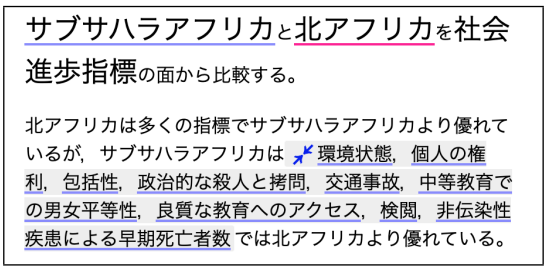

図 11　サブサハラアフリカと北アフリカを比較した時に説明ペインに表示される文章.

シャスという単語をクリックし，この国の概要を読むと，この国がサブサハラアフリカに属することがわかった．こののち，地域の比較機能を利用してエジプトが属す北アフリカとサブサハラアフリカを比較した．その結果，前者の優位的傾向が見られるものの，後者は自然環境等の指標について優れていることがわかった（図 11）.

以上の探索から，エジプトが大気汚染に起因する死者に代表されるように，福祉面に課題を残すものの，アフリカでは基本的な生活環境が整い，高等教育が進んでいることが読み取れた．また，アフリカ全体ではサハラ砂漠を挟んだ南北で社会の発展の度合いに大きな差があることがわかった．ここで，今回の調査の発端となった（エジプト，社会進歩指標）の時点に立ち戻り，Save ボタンとブラウザのブックマーク機能を利用して今回の調査の記録とした．この URL を友人に送ったところ，彼は私が分析した画像や文章が見るだけでなく，その分析画面を起点に南アフリカの調査を担当してくれた.

4.2　スケーラビリティ

本研究のもっとも重要な目標はデータサイズの大きさや，ユーザーの視点の広さによらず，読者のために適正な分量の文章を生成することである．本研究では，分析者が設定する焦点のひとつ下の層から得られるデータファクトを集約して文章を合成する．この場合，生成される文章量は選択した二次元的焦点のひとつ下の階層にあるデータ項目数と属性数の量に応じることが期待される．この方針の重要な利点は，文章の量がデータセットに与えられた階層の高さに依存しないことである．評価のベースラインとして，SPI データセットに含まれる全データファクトをまとめて文章化したものを用意した．具体的には，この方法はすべての国々と属性の組合せから得られる有意義なデータファクトを列挙し，それを文章化する方法である．この方法から生成された文章は 367,476 単語からなり，実用性は全くなかった.

我々は SPIViewer において選択可能な，全てのデータ項目と属性の組合せについて文章を生成し，生成された文章量を比較した．生成される文章は平均 24 語であり，最長となったのは焦点に（エジプト, 社会進歩指標）を設定した場合の 122 語であった．逆に最短の 11 語となった事例が 3,234 件ある．これらの著しく短い文はデータ欠損が原因である．表 1 にいくつかの特徴的な組合せの結果を提示する.

4.3　検索タスク実験

SPIViewer を用いたユーザー実験を行なった．タスクベースの実験で，GSS と比較してどのような長所と短所があるのかを検証した.

4.3.1　被験者と使用言語

3 人の学士課程の学生と 4 人の修士課程の学生に協力してもらった．7 人は全員情報系の学生で，SPIViewer に触れたことはないが，ヴィジュアルアナリティクスの知識はある．また，全員 20 代前半の日本人である．被験者には次節で説明する理由からグループ A とグループ B に別れてもらった．学士課程でグループ A の 2 人は BA1, BA2, 学士課程でグループ B の 1 人は BB1, 修士課程でグループ A の 2 人は MA1, MA2, 修士課程でグループ B の 2 人は MB1, MB2 と呼称することにする．このうち，被験者 BA2 のみ女性である．また，被験者 BB1 は赤と緑の色を見分けることが難しい.

はじめに制作した SPIViewer が生成する文章や図中のキャプションは全て英語であり，4.2 節では英単語の数を基準に評価したが，英語の習熟度によって実験結果に差が生まれることを避けるため，日本語版 SPIViewer[†2] を制作し，検索タスク実験は日本語

†2　日本語版 SPIViewer:
`https://smartnova.github.io/spi/jp/`

表 1　**SPIViewer** において選択可能な全ての焦点の組み合わせについて文章生成し，単語数でランク付けをした．この表はその抜粋である．

順位	国，地域	指標名	単語数
1 位	Egypt	Social Progress Index	122 単語
1914 位	Austria	Personal Freedom and Choice	31 単語
3490 位	Hungary	Discrimination and violence against minorities	30 単語
6093 位	Kenya	Access to piped water	22 単語
9447 位	Argentina	Traffic deaths	21 単語
11865 位	Romania	Globally ranked universities	20 単語
12079 位	Greenland	Social Progress Index	11 単語

版 SPIViewer のみを使って行なった．

4.3.2　実験手法

被験者には SPIViewer と GSS を使って，表 2 に記した 8 つのタスクに取り組んでもらった．グループ A の被験者には GSS を使ってタスクに取り組んでもらったのち，SPIViewer を使って同じ 8 つのタスクに取り組んでもらった．逆にグループ B の被験者には先に SPIViewer を使ってタスクに取り組んでもらい，その後，同じタスクを GSS を使って取り組んでもらった．被験者をグループ A とグループ B に分けたのは，同じタスクに 2 回取り組むという都合上，SPIViewer と GSS のどちらを先に使うかによって実験結果が変わってしまう可能性を考慮したからである．

被験者が GSS を使ってタスクに取り組む際には，Social Progress Imperative が提供する表データ [18] をあらかじめ記入しておいたスプレッドシートを渡した．ただし公平を期すため，タスクに取り組むのに必要で，かつ SPIViewer ではあらかじめ処理されている「地域ごとのデータ」と「数値が大きければ良いという訳ではない指標」の 2 つについては，あらかじめ処理を行なった．「地域ごとのデータ」については，SPIViewer と同様の手法で国連の人口データ [23] を利用して地域ごとのデータを用意し，専用のシートにまとめた．「数値が大きければ良いという訳ではない指標」については，まず，“Gender parity in secondary enrollment”についてのデータは SPIViewer と同様に取り除いた．また，数値が小さいほうが良い指標 16 個については，値に $\times -1$ を施すことで，値が大きいほうが良いという状態にし，GSS の MAX 関数等を使いやすくした．

データそのものについての説明 †3，SPIViewer の使い方 †4，GSS の使い方 †5 についてチュートリアルとなるビデオを作成し，被験者には実験の前にあらかじめ視聴してもらった．また実験直前にも双方のシステムを実際に使ってもらって使用方法を確認した．一般的な表計算ソフトの関数を使い慣れていない被験者には，最低限の知識として IF 関数と MAX 関数の使い方を教えた．

実験はビデオ会議システムを通して一人一人行い，被験者がシステムを使っている様子を分析者が確認できるよう，被験者の画面を分析者が見られるようにした．また画面は録画し，被験者には考えていることを口に出すことを心がけるよう伝え，被験者の声を録音した．

1 つのタスクに 5 分以上かかった場合は，そのタスクを諦めて次のタスクに取りかかってよいこととし，逆に諦めずにそのタスクに取り組み続けてもよいこととした．取り組み続ける選択をした場合には，詰まっている点についての助言を与えた．また，被験者 BB1 は SPIViewer の色使いの認知に困難があったため，図の読み取りに関して問題が生じた場合には助言を与えた．

表 2　被験者に取り組んでもらったタスク

タスク番号	タスク内容
タスク 1	マレーシアの「水道へのアクセス」のスコアはいくつか答えよ．
タスク 2	スペインが 1 位となっている指標を全て挙げよ．
タスク 3	ポルトガルが南ヨーロッパで 1 位となっている指標を挙げよ（ただし世界で 1 位，ヨーロッパで 1 位となっている指標は除く）．
タスク 4	南ヨーロッパが東西南北ヨーロッパの中で最も優れている指標を挙げよ．
タスク 5	アフリカで「基本的知識へのアクセス」が一位の国を見つけよ．
タスク 6	ドイツとフランスの「生活必須要素」に分類される指標を比較してドイツがフランスより優れている点を挙げよ．
タスク 7	アフリカとアジアの「社会進出の機会」の分布を比較したとき，全体的にどちらの方がスコアが高いか答えよ．また，どうしてそう考えたのか説明せよ．
タスク 8	アフリカとアジアの「社会進出の機会」のスコアの差は，アフリカとアジアの「生活必須要素」のスコアの差と比べて大きいか小さいか答えよ．また，どうしてそう考えたのか説明せよ．

4.3.3　実験結果

7 人の被験者が 8 つ全てのタスクを終えるのにかかった時間を，GSS と SPIViewer で比較すると図 12 のようになった．図 12 から，SPIViewer を使った時の方が概ね 2 倍ほど速くタスクをこなす傾向がみられた．

タスク 1 つ 1 つを見ていくと，正答，誤答，ギブアップ，5 分以上かかった例は表 3 のようになった．タスクに取り組んだ所要時間を見ると，1 つのタスクに 5 分以上かかった割合は，56 例（7 人 8 タスク）のうち，GSS で 22 例，SPIViewer で 4 例であった．ただし，SPIViewer の 4 例は一番長くても 6 分 6 秒であり，SPIViewer を使った方が速く容易にタスクをこなせると言える．一方，誤答数では差は出なかった．また，SPIViewer ではギブアップとなる例は発生しなかったものの，GSS を使った場合もギブアップは 3 例しか発生しなかった．

所要時間に注目すると，タスク 2, 3, 4, 5, 6, 7 では被験者 7 人中，6 人または全員が SPIViewer を使った時の方が時間がかからなかった（図 13）．一方，タスク 1, 8 では 7 人中 5 人が SPIViewer を使った方が

†3 https://vimeo.com/442918454/a0e3e713af
†4 https://vimeo.com/442918453/15ca566adc
†5 https://vimeo.com/442918455/cd88e21704

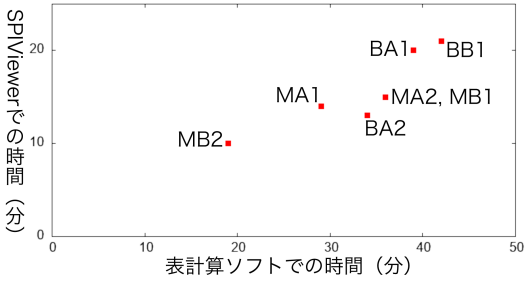

図 12　横軸を GSS を使った時の全てタスクを終えるのに必要とした時間，縦軸を SPIViewer を使った時の全てのタスクを終えるのに必要とした時間として，全ての被験者についてプロットした．SPIViewer を使用した時の方が GSS を使用した時よりも 2 倍程度速くタスクをこなせていることがわかる．

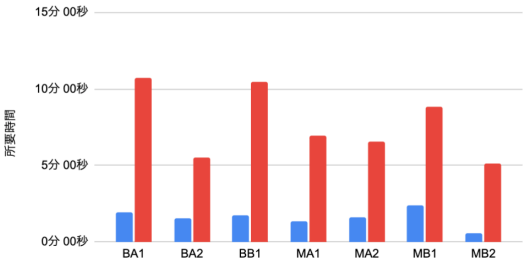

図 13　タスク 3 での各被験者の所要時間．青が SPIViewer，赤が GSS を表している．

タスク		1	2	3	4	5	6	7	8
BA1	SPIViewer	0分 44秒	0分 42秒	1分 56秒	5分 21秒	0分 45秒	4分 56秒	1分 47秒	3分 41秒
BA2		0分 53秒	0分 55秒	1分 33秒	1分 00秒	0分 47秒	1分 31秒	3分 14秒	3分 26秒
BB1		1分 14秒	0分 55秒	1分 45秒	1分 04秒	1分 11秒	5分 11秒	4分 45秒	4分 51秒
MA1		0分 41秒	0分 37秒	1分 20秒	0分 26秒	0分 35秒	1分 37秒	2分 33秒	6分 06秒
MA2		0分 54秒	0分 49秒	1分 36秒	0分 50秒	0分 33秒	1分 58秒	2分 21秒	5分 45秒
MB1		0分 50秒	1分 10秒	2分 25秒	1分 00秒	1分 05秒	1分 01秒	2分 38秒	4分 48秒
MB2		0分 34秒	0分 40秒	0分 34秒	0分 50秒	0分 43秒	0分 50秒	1分 20秒	4分 10秒
BA1	Google スプレッドシート	0分 33秒	3分 12秒	10分 42秒	2分 31秒	4分 06秒	7分 04秒	4分 49秒	6分 16秒
BA2		1分 05秒	3分 39秒	5分 29秒	2分 53秒	2分 52秒	5分 09秒	7分 05秒	5分 20秒
BB1		2分 40秒	5分 59秒	10分 29秒	7分 17秒	2分 07秒	6分 02秒	5分 36秒	3分 20秒
MA1		0分 39秒	2分 22秒	6分 58秒	1分 27秒	3分 36秒	3分 20秒	7分 17秒	2分 34秒
MA2		0分 31秒	4分 03秒	6分 35秒	4分 46秒	2分 48秒	6分 18秒	6分 03秒	5分 05秒
MB1		0分 45秒	5分 24秒	8分 49秒	3分 54秒	4分 53秒	4分 10秒	5分 50秒	3分 08秒
MB2		0分 27秒	2分 08秒	5分 07秒	2分 59秒	0分 42秒	1分 44秒	3分 40秒	1分 35秒

表 3　全てのタスクについて所要時間をまとめた表．黄色く塗りつぶされた箇所は 5 分以上かかったことを表し，白い箇所は時間内に回答したことを示す．赤く囲まれた箇所は誤答したことを表し，青く囲まれた箇所はギブアップして無回答だったことを表し，それ以外は正答したことを示す．

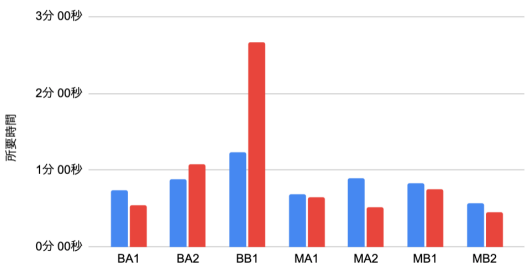

図 14　タスク 1 での各被験者の所要時間．青が **SPIViewer**，赤が **GSS** を表している．

時間がかかった (図 14)．タスク 1 がこのような結果になった要因は，SPIViewer では特定の国について調べたいとき，検索機能がないため，その国の地理的な位置を知らないと選択できないことにあると考えられる．一方で Google スプレッドシートには検索機能があるため，ある国が表の中でどこに位置するかは瞬時にわかる．タスク 8 については，GSS の実験のときに地域ごとの重み付き平均だけを比べてタスクを終わらせた被験者は，GSS の方が早くタスクを完了している．SPIViewer は地域ごとの比較を文章で行うことはできるが，重み付き平均の値自体は表示しないため，GSS より時間がかかったのだと考えられる．

実験後に SPIViewer を使った感想を聞いたところ，「画面が分割されているから（可視化結果，文章化結果を）見ながら操作できるので使いやすい.」(MB1),「スプレッドシート上で探索するより遥かに楽でした.」(MB2),「スプレッドシートより断然使いやすかったです.」(BA1) など好意的な意見を多くもらった．また，SPIViewer をどのようにしたら改善できるか意見をきいたところ，全員共通して，データ項目と属性の検索機能があった方が良いという意見を述べた．実際，GSS では検索機能が使えるため，タスク 1 に関しては過半数の被験者が GSS の方が時間がかからなかった．SPIViewer では検索機能を採用しなかったが，被験者の意見からも需要が大きいことは明らかなので，次期システムでは改善したい．

4.4　自由探索実験

ユーザーに明確な目的を与えずに自由な探索を許した場合の行動を観察するために (DC1)，SPIViewer を使ってもう 1 つ実験を行なった．

4.4.1　実験手法

検索タスク実験の中で，特に SPIViewer に興味を示した被験者 BA2, MB1, MB2 に協力してもらった．3 人には SPIViewer を使って表 4 に示した 2 つのタスクにそれぞれ 20 分を目安に取り組んでもらった．実験はビデオ会議システムを通して分析者と一対一で行い，録画，録音を行い，被験者には考えていることを口に出すことを心がけるよう伝えた．また，実験中は分析者にいつでも質問をできるようにした．

また，被験者が簡単にメモをとれるよう，実験は

表 4　被験者に取り組んでもらったタスク

タスク番号	タスク内容
タスク 1	トルコについてあなたが興味深いと感じた点を文章でまとめよ.
タスク 2	サブサハラ・アフリカについてあなたが興味深いと感じた点を文章でまとめよ.

SPIViewer にメモ機能を追加したシステム †6 を使って行なった.

4.4.2　実験結果

3 人の被験者，2 タスク全ての分析は完了していない．ここでは BA2 がタスク 1 に取り組んだ録画を分析した結果を述べる．BA2 は 4 段落 375 字で文章をまとめ，12 の指標に直接言及した.

BA2 はまず焦点として（トルコ，社会進歩指標）を選択し，説明ペイン (図 7-Ⓔ) を見て概要を把握した．その後，SPI データセットの総合指標 (社会進歩指標) を構成する 3 つの主要な指標を順に見てゆき，それぞれの指標でドリルダウンを行なって詳細に調べていった.

例えば，焦点として（トルコ，社会的環境）を選択したのち，属性選択ペイン (図 7-Ⓒ) で子ノードの指標を次々と選択していった．**社会的環境**の子ノードの指標は**基本的知識へのアクセス**，**情報とコミュニケーションへのアクセス**，**健康度**，**環境状態**である．BA2 はこれらの指標を表示されている上から順に,「基本的知識へのアクセスが微妙，まあ普通，なのかな」,「情報とコミュニケーションへのアクセスは別に」,「健康度？健康度は普通に健康だし」,「環境状態．環境状態が... 悪い.」と発言しながら次々と選択していった．すると次に，トルコのスコアが特に低かった**環境状態**の子ノード指標に注目し，説明ペインが言及している**生態系の保護**を選択した．説明ペインにトルコが世界 217 位のスコアであることが表示されると，BA2 は驚いたようで,「217って... え？」と発言した.

このように，BA2 は指標を順に選択し，文章や可視化を確認しつつ，スコアが特に良いか悪い指標に出会うとさらに詳細な指標を選択してドリルダウンをする，という行動を繰り返した．このとき，子ノードの指標を全て順に見ていくという行動をとった．SPIViewer が生成する文章には，ある指標のスコアが悪かった時に，子ノードの指標が全て悪いのか，どれかが飛び抜けて悪いのか，といった指標に関する分析はなく，子ノード指標に関する可視化もない．そのため，BA2 はすべての子ノード指標を順に見ていかなくてはならない場面が多かったのだと考えられる．その悪影響として，BA2 はトルコの**個人の権利**のスコアが低い理由を調べる際，全ての子ノード指標を網羅し損ねたために，**政治的権利**のスコアが低い為だ，という誤った結論をだしてしまっている.

また，BA2 は度々，なぜそのスコアになったのか，という疑問を抱いていた．例えば，焦点を（トルコ，生態系の保護）としたときに,「生態系の保護，なんでこんなに悪いのか，逆にめっちゃ調べてみたい」と発言している．SPIViewer が，探索作業を始めた時点では本人すら思いもしなかったような，新しい疑問の発見の機会を与えたと言える．これは，自由な探索行動の結果，さらなる調査への動機づけを得ることができ，SPI データセットのデータファクトを，SPI データセットに留まらない新たな知識の獲得へと繋げる可能性を示している.

BA2 は 3 つの主要な指標を一通り詳しく調べた後,「どっかと比べてみよっかな」と発言し，トルコと他の国との比較を始めた．属性を（トルコ，社会進歩指標）にした後，比較モードのボタンをクリックし，データ可視化ペイン（図 7-Ⓓ）にカーソルをホバーさせた．トルコと**社会進歩指標**のスコアが近い国の中にタイがあることを発見すると,「タイと比べてみよっ」と言って棒グラフの中のタイを示すバーをクリックして，焦点を（トルコとタイの比較，社会進歩指標）に切り替えた．ここで開始から 20 分が経過し，BA2 も終了を希望したため，タスク 1 の実験はこれで終了した．そのため，BA2 が最終的にまとめた文章ではタイについて触れられなかったが，このことについて実験後にインタビューを行なった.

インタビューの結果，BA2 はタイのドラマを見る

†6　メモ機能付き日本語版 SPIViewer: `https://smartnova.github.io/spi/jp/textarea.html`

機会があり，馴染みがあったためにタイを選んだことがわかった．また，タイとトルコを教育関連の指標について比較するつもりだったようだ．BA2 はその時点までの探索の結果としてトルコは教育関連の指標が際立ってよいことがわかっており，さらにタイのドラマで大学がよく登場したことで，教育関連での比較に興味を持ったようだった．

BA2 はこのような探索を通して，新たな知識を得ただけでなく，トルコの生態系の保護のスコアが低いのはなぜなのか，といった自発的な調査に繋がる興味を持った．さらに，トルコとタイの教育面での比較という，タスクを設定した分析者の想像を超えた新たな切り口での調査を見出した．SPIViewer がユーザーの自由な探索によって，SPI データセットが持つ膨大なデータファクトの中から，ユーザーの知識や好み，置かれた状況に応じて，新しい知識や，新しい知識に繋がるような興味を持たせるデータファクトを抽出してユーザーに提供できることが示された．

5 議論

この研究を通してテンプレートを用いた文章生成技術における，文章テンプレートを作成するための人的コストも課題となることがわかった．本システムでは，テンプレートエンジンを用いることで比較的小さなコストでテンプレートを作成した．生成される文章の質を高めるためのテンプレート編集は比較的単純であるものの，面倒な作業の連続である．生成された文章に対する修正内容からテンプレートを修正する技術や，データから文章を生成する自然言語処理 AI の応用が望まれる．

本研究は大規模なデータセットにおいて，データ項目と属性のそれぞれに階層構造が与えられていることを前提としている．データセットによっては，このような階層構造が事前に用意されていない場合も考えられ，用意されていたとしても，3.3 節に述べたノードごとの代表値が提供されていない場合も多いだろう．このような階層構造はデータセットの文脈に応じた選択が必要であり，本研究の成果を階層構造が与えられていない大規模表的データに応用するにはさらなる研究を要する．

提案するアプローチで生成された文章は焦点によって絞られたデータの範囲に含まれるデータファクトについて，属性に関して特徴的な点を簡潔に説明する．このため文章を通してデータ項目についての理解は深めやすい．さらに，この説明文とデータファクトを視覚化した画像の連携も容易である．一方，属性を中心に据え，その属性について特徴的なデータ項目を説明する文章は生成していない．このため属性についての理解を深めることは難しい．今回の試みとは逆に，属性を主体としてデータ項目を用いて説明する形式の文章を生成することはさほど困難ではないだろう．この場合，多数の属性を扱う平行座標法に代表される多次元データ可視化技術の採用が考えられる．さらに重要な観点として，ユーザーの操作履歴に応じて二種類の文章のスタイルを適宜切り替える手法も考えられる．

Mumtaz ら [15] は探索的可視化システムの使いやすさを向上させるためのいくつかの考え方を提唱している．彼らの主張に照らすと SPIViewer について改善可能な点が見えてくる．例えば，文章の生成に用いた統計処理や計算手法についての説明を提示することによって，データファクトの説明をより充実できる．また，説明ペインの文章中にデータファクトの説明を補助する小さな可視化を埋め込めば，探索的可視化を強化できる．

地図ペインへの着色方法には検討の余地がある．SPI Viewer では，属性的焦点の変化によらず，国と地域について一環して SPI 総合評価値を表す配色を施している．この設計を採用したのは焦点が変化するごとに配色を変化させることがユーザーのメンタルマップを破壊することを配慮したためである．一方で，文脈に応じた色チャネルを活用していない点は批判を招くかもしれない．技術的な面として，文脈における属性数が 3 を越えた場合の次元削減や利用者のメンタルマップの保持の手段などの研究の余地がある．

データファクトの抽出方法については，本研究では過去の研究提案に沿って，極値，一次元外れ値，二次元外れ値の 3 種を扱った．しかし，ほかにも有用な統計指標がありそうだ．たとえば，ユースケースの調査はエジプトの色を視認したところから始まる．しかし，今回，採用した外れ値分析はいずれも，こ

の SPI の属性値の分布の差についての興味深い事実を発見できなかった．

SPIViewer では，説明ペインに登場する属性について知りたいときにその属性をクリックすると，焦点がクリックした属性に変わってしまう．これは，その属性の定義を見たいだけのときやその属性についてグラフを見たいだけの時には，クリックしたあとに履歴機能を使って戻る操作をしなければならない煩わしさがあるほか，画面全体の表示が変わってしまうため，ユーザーのメンタルマップを破壊してしまう可能性がある．また SPIViewer の制作過程でも，説明ペインを読んでいるときの誤クリックで焦点が変わってしまうということが度々あった．これについてはインターフェイスの改善が必要である．

また Save ボタンについては，ブックマークに保存するにしても SNS 等で共有するにしても，一度 Save ボタンをクリックしてから，さらにシステム外でユーザーの操作が必要で，ワンクリックで共有できるようにするといった改善の余地があった．

6 まとめと今後の課題

情報可視化技術を利用することでデータをわかりやすく表現するだけでなく，文章生成技術と組み合わせることで，分析中の文脈を文章として要約するシステムが増えてきている．本研究の貢献は，従来よりも大規模なテーブルデータに対して，このような情報可視化と文章を組み合わせたシステムを構築するための問題点を明らかにし，多くの大規模テーブルデータに備わる階層構造を利用することによってその課題を克服する方法を示した点にある．提案した概念を Social Progress Index データセットに適用した視覚的分析システム SPIViewer を実装し，ユースケース，文章量の計測，ユーザー実験を通して，その有効性を明らかにした．

今後は，ユーザー実験の定性的分析，他のデータセットへの応用，時系列分析の支援，探索的調査のためのインタラクションの充実，ソーシャルメディアとの連携などを考えている．

謝辞 帝国データバンク先端データ解析共同研究講座からの研究助成に深く感謝いたします．

参 考 文 献

[1] Boukerche, A., Zheng, L., and Alfandi, O.: Outlier Detection: Methods, Models, and Classification, *ACM Comput. Surv.*, Vol. 53, No. 3(2020).

[2] Bryan, C., Ma, K., and Woodring, J.: Temporal Summary Images: An Approach to Narrative Visualization via Interactive Annotation Generation and Placement, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 23, No. 1(2017), pp. 511–520.

[3] Chen, J.: Quill - Your powerful rich text editor, `https://quilljs.com/`.

[4] Demiralp, Ç., Haas, P., Parthasarathy, S., and Pedapati, T.: Foresight: Rapid Data Exploration Through Guideposts, *Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS 2017*, 2017.

[5] Demiralp, c., Haas, P. J., Parthasarathy, S., and Pedapati, T.: Foresight: Recommending Visual Insights, *Proc. VLDB Endow.*, Vol. 10, No. 12(2017), pp. 1937–1940.

[6] Google: Google Spreadsheets: Free Online Spreadsheets for Personal Use, `https://www.google.com/intl/en/sheets/about/`.

[7] Hoaglin, D. C., Mosteller, F., and Tukey, J. W.: *Understanding Robust and Exploratory Data Analysis*, Wiley, 2000.

[8] Kong, N., Hearst, M. A., and Agrawala, M.: Extracting References Between Text and Charts via Crowdsourcing, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, ACM, 2014, pp. 31–40.

[9] Kucher, K., Paradis, C., and Kerren, A.: The State of the Art in Sentiment Visualization, *Computer Graphics Forum*, Vol. 37, No. 1(2018), pp. 71–96.

[10] Latif, S. and Beck, F.: Interactive map reports summarizing bivariate geographic data, *Visual Informatics*, (2019).

[11] Latif, S. and Beck, F.: VIS Author Profiles: Interactive Descriptions of Publication Records Combining Text and Visualization, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 25, No. 1(2019), pp. 152–161.

[12] Latif, S., Liu, D., and Beck, F.: Exploring Interactive Linking Between Text and Visualization, *EuroVis 2018 - Short Papers*, Johansson, J., Sadlo, F., and Schreck, T.(eds.), The Eurographics Association, 2018.

[13] Ltd, L. A. S.: WordSmith Tools home page, `https://www.lexically.net/wordsmith/`.

[14] Microsoft: Data Visualization — Microsoft PowerBI, `https://powerbi.microsoft.com/en-us/`, 2020.

[15] Mumtaz, H., Latif, S., Beck, F., and Weiskopf, D.: Exploranative Code Quality Documents, *IEEE Transactions on Visualization and Computer*

Graphics, Vol. 26, No. 1(2020), pp. 1129–1139.

[16] Munzner, T.: *Visualization: Analysis and Design*, AK Peters Visualization Series, A K Peters/CRC Press, November 2014.

[17] Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations, *Proceedings 1996 IEEE Symposium on Visual Languages*, Sep. 1996, pp. 336–343.

[18] Social Progress Imperative: Social Progress Imperative, `https://www.socialprogress.org/`. Accessed: 2020-01-28.

[19] Srinivasan, A., Drucker, S. M., Endert, A., and Stasko, J.: Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 25, No. 1(2019), pp. 672–681.

[20] Strobelt, H., Oelke, D., Kwon, B. C., Schreck, T., and Pfister, H.: Guidelines for Effective Usage of Text Highlighting Techniques, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 22, No. 1(2016), pp. 489–498.

[21] Tang, B., Han, S., Yiu, M. L., Ding, R., and Zhang, D.: Extracting Top-K Insights from Multi-Dimensional Data, *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, New York, NY, USA, Association for Computing Machinery, 2017, pp. 1509–1524.

[22] United Nations: Standard country or area codes for statistical use (M49), `http://www.iso.org/iso/home/standards/country_codes.htm`.

[23] United Nations: World Population Prospects - Population Division - United Nations, `https://population.un.org/wpp/`. Accessed: 2020-01-28.

[24] White, R. W. and Roth, R. A.: Exploratory search: Beyond the query-response paradigm, *Synthesis lectures on information concepts, retrieval, and services*, Vol. 1, No. 1(2009), pp. 1–98.