探索的なデータ分析のための対話的可視化手法

松下 光範 †
mat@cslab.kecl.ntt.co.jp
† 日本電信電話株式会社

加藤 恒昭 ‡
kato@boz.c.u-tokyo.ac.jp
‡ 東京大学 大学院

1 はじめに

数値データをその性質やユーザの意図に基づいて統計グラフで表示するシステムが幾つか提案されている．これらの主眼はマルチメディアドキュメントの自動生成や情報可視化など, プレゼンテーションでの利用にある．

統計グラフは蓄積された莫大な観測データを分析し特徴を見出す際に有効な手段であるため, 探索的なデータ分析にも利用される．これは, ある観点で描画されたグラフからユーザが新たな観点を想起してグラフを描き直すというプロセスを繰り返すことで, ユーザ自身がデータに対する理解を深めながら対話的に特徴を見出したり, データが示す傾向の原因を探るという利用法である．

本稿では, このような探索的なデータ分析を想定した数値データの対話的可視化手法を提案する．これは

(1) 93 年と 94 年の四国地方の降水量を知りたい

(2) 県毎に知りたい

(3) 97 年まで知りたい

といった一連のユーザ要求に応じて, 逐次適切なグラフに描き換えていく可視化手法である．

2 探索的データ分析に適したグラフの選択

本稿では二つの独立変数によって値が特定されるデータに限定して議論を進める．具体的な例として「時刻 × 場所 → 降水量」というデータを取り上げる．このデータに対して, まずユーザ要求 (1) が与えられたとする．このユーザ要求 (これを開始ユーザ要求と呼ぶ) を満たすグラフは, 例えば図 1 (a) である．このグラフに対して, ユーザが視点を変えてユーザ要求 (2) を加えた場合 (これを追加ユーザ要求と呼ぶ), 新しく描画されるグラフは図 1 (a) を県毎に詳細化したグラフであるべきで, 図 1 (b) に示すような, 棒グラフの各棒を県単位に細分化したグラフが適切であろう．

この連続したユーザ要求が全体として表しているのは

(4) 93 年と 94 年の四国地方の県毎の降水量が知りたい

に他ならないが, このユーザ要求が開始ユーザ要求として単独に与えられた場合に適切なグラフは, 図 1 (b) よりもむしろ (c) である．これはユーザ要求 (1) と (2) を逐次的に与えた場合, 四国地方全体の年間降水量に関するユーザの関心が (1) で示されているので, これを直感的に表現している (b) の方が適切だが, ユーザ要求 (4) ではその

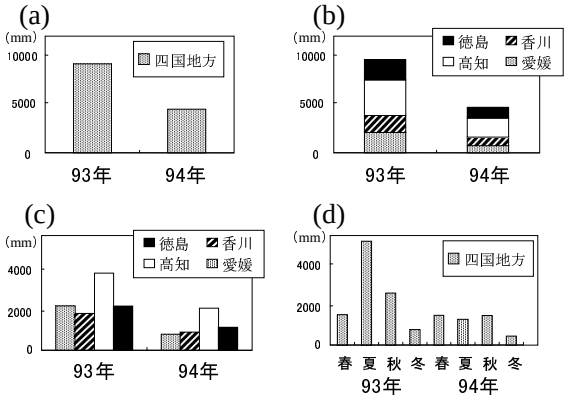

図 1: 追加ユーザ要求によるグラフ変化の違い

ような関心は示されていないので県毎の降水量が読みとり易い (c) の方が適切だからだ, と推測される．またグラフ (a) と (b) はその形状が類似しているのに対し, (a) と (c) は大きく異なるので, 描画された要素同士の対応がすぐには判断しづらい．この点でも (a) に続けて描画する際には (c) よりも (b) が適当と思われる．

このように, 自然言語処理による対話システムが行うような断片的な発話の理解によってユーザ要求 (1) を前提として (2) を解釈し, それによって得た (4) に基づいてグラフを選択する方式では適切なグラフは選択できない．従って, 従来のプレゼンテーションでの利用を前提とした可視化システムと断片発話の理解機構を単純に結合するだけでは, 探索的なデータ分析に有効な可視化システムは実現できない．

以下では, 探索的なデータ分析のための可視化手法で考慮すべき事柄について考察する．

追加ユーザ要求によって適切なグラフは異なる．例えば開始ユーザ要求 (1) に続けて

(5) 季節毎に知りたい

という追加ユーザ要求が与えられた場合, 時刻が詳細化されているので図 1 (b) のような積み上げ棒グラフよりも, (d) のように x 軸を詳細化した単純棒グラフのほうが適切であろう．

追加ユーザ要求 (2) と (5) ではどちらも同じ文末表現が用いられ, 描画されている情報の詳細化を要求しているにもかかわらず適切なグラフが異なる．これは前者が描画される対象を詳細化 (i.e., 地方を県に詳細化) しているのに対して後者は軸を詳細化 (i.e., 年を季節に詳細化) しているためである．従って追加ユーザ要求と適切なグラフとの関係を明らかにするには単に詳細化だと解釈するだけでは不十分で, 開始ユーザ要求で示されたグラフ描画のための情報に対し, 追加ユーザ要求によってどの部分にど

Interactive Visualization Method for Exploratory Data Analysis, Mitsunori Matsushita† and Tsuneaki Kato‡, †NTT Corp., ‡Tokyo Univ.

変数	ドメイン	粒度	型	制約条件
X	時刻	年	時間	$x_i \in \{93年, 94年\}$
Y	場所	地方	名義	y_i = 四国地方
Z	降水量	mm	量	

図 2: 開始ユーザ要求 (1) のフレーム表現

のような変更が生じるかを考慮する必要がある.

また同一の追加ユーザ要求であっても, 現在描画されているグラフ種によって, 新たに描画されるべき適切なグラフが異なる. 例えば描画される対象を追加するような追加ユーザ要求が与えられた場合, 単純棒グラフが描画されている場合は複合棒グラフを描画するのが適切であるが, 単純折線グラフが描画されている場合には複数折線グラフが適切であろう. 従って追加ユーザ要求によって新たに描画されるグラフの選択には, 直前に描画されているグラフ種も考慮する必要がある.

3 追加ユーザ要求からのグラフ種特定手法

既に我々は自然言語で表現されたユーザ要求を図 2 のようなフレーム形式で表現することでグラフ描画に必要な情報が獲得できることを示し, ユーザ要求からフレームを抽出する手法を提案している [1]. 図 2 では X と Y が独立変数に対応し, グラフ上ではどちらか一方が x 軸になる (x 軸になる独立変数を x 軸変数, 他方を対象変数と呼ぶ). また Z が従属変数に対応し y 軸になる.

2 章での考察から, 追加ユーザ要求から適切なグラフを描画するには, まず追加ユーザ要求によって描画されているグラフのどの部分にどのような変更が生じるかを解釈し, 次にこれと描画されているグラフの種類から描画するグラフ種を決定する, という処理が必要である.

まず, 追加ユーザ要求によって描画されているグラフに加えられる変更点を特定する.

追加ユーザ要求は描画されているグラフを前提として行われるのでユーザは変更点にのみ言及すると考えてよい. 例えば図 1 (a) のグラフが描画されている場合, ユーザは「93 年から 94 年の四国地方」という情報を省略し, 単に「県毎に見たい」と要求するものと考えられる.

追加ユーザ要求のタイプは様々考えられるが, グラフの変化を伴う要求として, 追加 (〜を追加したい), 削除 (〜を削除したい), 粒度変更 (〜毎に見たい), 軸交換 (〜を軸にしたい), 実量化 (〜の量が知りたい), 割合化 (〜の割合が知りたい), 推移化 (〜の推移が見たい), の 7 種類を対象とした.

2 章で述べたように, 同じタイプに属する要求であってもグラフのどの部分 (変数) に対する要求であるかによって適切なグラフは異なる. また, 同じ「県毎に見たい」であっても, 現在のグラフが市毎に描画されていれば粗大化であるし, 地方毎であれば詳細化である. このように追加ユーザ要求の解釈にはその表現に加えて, 現在のフレームを参照することが必要である. これを整理したものが表 1 である. 文末を主とする表現パタンで “要求パタン” が, どの変数を対象としているかで “変化する変数” が, 現在の内容との関係で “変化内容” が決定する. これらは 13 パタンに細分化できることが分かった. なお, 最右列はフレーム中で変更される箇所を示している.

表 1: 変化の種類と意味フレームの変更箇所の対応

要求パタン		変化する変数	変化内容	フレーム
追加	1	x 軸変数	追加	制約条件
	2	対象変数	追加	制約条件
削除	3	x 軸変数	削除	制約条件
	4	対象変数	削除	制約条件
粒度変更	5	x 軸変数	詳細化	粒度
	6	対象変数	詳細化	粒度
	7	x 軸変数	粗大化	粒度
	8	対象変数	粗大化	粒度
軸変更	9	独立変数	軸変数変更	変化なし
割合化	10	従属変数	割合化	型
実量化	11	従属変数	実量化	型
	12	なし	棒グラフ化	変化なし
推移化	13	なし	折線グラフ化	変化なし

次に, 描画されているグラフの種類と上記の処理で特定された変更点から適切なグラフ種を決定する.

グラフ種は上記で得られた “変化する変数”, “変化内容” と現在のグラフ種から決定される. そのため, グラフ種同士の形状の類似性や一般的な用いられ方に着目して, あるグラフから他のグラフへの変化の関係を整理した.

例えば単純棒グラフが描画されている場合, 1, 3, 5, 7 のいずれかの要求パタンに属する追加ユーザ要求が与えられた場合は, 新たに描画されるグラフとして単純棒グラフが選択されるが, 2 の要求パタンに属する追加ユーザ要求が与えられた場合は複数棒グラフが, 6 の要求パタンに属する追加ユーザ要求が与えられた場合は積み上げ棒グラフが各々選択される.

以上の処理により, 探索的なデータ分析の場面での適切なグラフ選択が実現できる. 我々はこの仕組みを可視化システム KEVIN 上で実装した.

2 章で適切なグラフ選択の基準のひとつとして, グラフに描画された要素同士の対応の容易性を挙げた. この点を重視して, KEVIN ではグラフ間の対応関係がより明確になるように各グラフ変化パタンに対応した変化アニメーションを実装し, 変化の過程を連続して見せることでユーザにそれらの対応関係を直感的に理解させようと試みている.

4 おわりに

本稿では, 探索的なデータ分析における統計グラフの満たすべき特徴について考察し, 考慮すべき事柄を明らかにした. また, これに基づいて対話的可視化手法を提案した. 今後グラフ変化に関する知識の精緻化や手法の有効性の定量的評価について検討したい.

参考文献

[1] Matsushita, M. et al: “A Frame Representation of User Requirements for Automated Data Visualization,” ECAI-2000, pp. 631–635 (2000).