5N－8

数値データ可視化のためのユーザ要求理解手法

松下 光範　米澤 勇人
{mat,hayatoyo}@cslab.kecl.ntt.co.jp
NTT (株) コミュニケーション科学基礎研究所

1 はじめに

我々はデータベースに蓄積された膨大なデータから効率よく情報を取り出すために、ユーザ要求に基づいて必要なデータを自動的に選択・集約し、効果的に可視化するシステムの実現を目指している [1]。

このようなシステムの実現には、計算機がユーザ要求を理解し、適切なデータの詳細度や可視化形式を判断する必要がある。そこで本稿では、ユーザ要求を「何を描画するか」と「どのように描画するか」という二つの観点で捉え、これらを過不足なく表現するためにはユーザ要求からどのような情報を取り出す必要があるかについて検討する。そしてそれらの情報を計算機上で形式的に取り扱うために、二つの意味フレームを提案する。

2 ユーザ要求の意味フレーム表現

我々は、膨大なデータ群を有益な情報にまとめ上げる過程を (1) 必要なデータの種類や範囲の分析、(2) 適切な詳細度でのデータ集約、(3) 効果的な可視化方法の選択、という一連のプロセスとして捉えた。この考え方に基づき、我々のシステムではまず元となるデータからユーザ要求に即したデータテーブルを再構成し、それから適切な可視化手法を用いて表示することにした。

我々のシステムでは、例えば次のようなユーザ要求を想定している。

(1) 1998 年の近畿地方における降水量は京都府が特に多かった。

(2) 1997 年の四国の各県における男女別の人口は？

ユーザ要求には「何を描画するか (描画対象)」と「どのように描画するか (描画方法)」という情報が含まれている。ユーザ要求例 (1) の場合、「1998 年の近畿地方における降水量」が描画対象を表している。また「京都府が特に多かった」がユーザの注目点を示しており、描画方法を決定する重要な手がかりとなる。これらをフレーム形式で表現し、描画対象を表現する意味フレームに基づいてデータテーブルを再構成する。そしてこのデータテーブルと描画方法を表現する意味フレームに基づいて効果的な可視化方法を決定する。

Understanding Method of User Requirement to Visualize Numerical Data, Mitsunori Matsushita and Hayato Yonezawa, NTT Communication Science Laboratories, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

2.1 描画対象を特定するための意味フレーム

まず、描画対象を表現する意味フレーム (話題項目フレーム) について検討する。元となるデータからテーブルを再構成するためには、まずユーザが話題とする対象 (話題項目) を判断しなければならない。

例えばユーザ要求例 (1) では “1998 年”, “近畿地方” という条件の元で “降水量” について言及されている。これらは “時刻”, “場所”, “降水量” という話題項目に整理できる。話題項目は、データテーブルを再構成する際にデータベース中のどのデータカラムを選択するかを判断するのに用いる。従って話題項目となり得るのはデータベースのカラム名として登録されている語である。

また、データテーブルを再構成するためには各々の話題項目について、各話題項目の性質 / 性格 (属性)、記述の詳細度 (粒度)、とり得る範囲 (制約条件) を特定する必要がある。従って、話題項目フレームにはこれらの情報が記述されなくてはならない。

属性は話題項目の特徴を抽象化し取り扱うために用いる。すなわち、ある話題項目に適用可能な演算規則は、その話題項目の属性に基づいて決定される。我々は属性の種類を名義属性、量属性、時間属性、順序属性に分類した。例えば名義属性の話題項目は上位 / 下位概念による包含関係を持つことはあるが、和差演算などの数学的演算は適用できない。それに対して量属性の話題項目は和差演算の適用や割合・累積値への変換が可能である。

粒度とは話題項目の記述単位のことである。例えば、“京都府” は日本を都道府県の単位で記述した際のラベルであり、その属する粒度は県である。粒度はデータをどの詳細度に集約するかを決定する際に用いられる。

元となるデータからデータテーブルの再構成に必要な部分を選択するには、各話題項目の制約を把握しなければならない。例えばユーザ要求 (1) では、“場所” には近畿地方という制約があり、“時刻” には “1998 年” という制約がある。制約条件は要素の指定 (y_1 = 京都府 など)、上位概念の指定 ($y_i \in$ 近畿地方 など)、範囲の指定 ($0 < y_i < 100$ など) などの記述方法が可能である。

これらを考慮するとユーザ要求 (1) および (2) に対す

変数	話題項目	属性	粒度	制約条件
X	時刻	時間	年	$x_1 = 1998$年
Y	場所	名義	県	$y_i \in$ 近畿地方
Z	降水量	量	mm	

図 1: ユーザ要求 (1) の話題項目フレーム

変数	項目名	属性	粒度	制約条件
W	時刻	時間	年	$w_1 = 1997$年
X	場所	名義	県	$x_i \in$ 四国
Y	性別	名義	性	$y_1 =$ 男性, $y_2 =$ 女性
Z	人口	量	人	

図 2: ユーザ要求 (2) の話題項目フレーム

る話題項目フレームは図 1、2 のようになる。

2.2 話題項目間の項目関係

データテーブルを再構成するには、話題項目間の従属関係を判断する必要がある。 この関係を項目関係と呼ぶ。項目関係は、話題項目のうち複数の要素をとるものに着目することで導出することができる。例えばユーザ要求例 (1) の場合には 3 つの話題項目が存在するが、"時刻" は要素数が 1 のため、変数とはならない。従って図 1 に示す話題項目間には

$$f(\text{場所})|_{\text{時刻}=1998\text{年}} \longrightarrow \text{降雨量} \qquad (1)$$

という項目関係が成立する。現在は数値データの可視化を対象としているため、量属性の話題項目を従属変数、複数の要素を取る他の属性の話題項目を主変数とするようになっている。データテーブルを再構成する際には、主変数となる話題項目が表頭・表側となる。これにより表 1 に示すテーブルが再構成される。

また、ユーザ要求例 (2) の場合には、"場所", "性別", "人口" の各話題項目が複数の要素を取り得る。従って図 2 に示す話題項目間には

$$f(\text{場所}, \text{性別})|_{\text{時刻}=1997\text{年}} \longrightarrow \text{人口} \qquad (2)$$

という項目関係が成立する。これにより表 2 に示すデータテーブルが再構成される。

2.3 描画方法を特定するための意味フレーム

次に描画方法を表現する意味フレーム (比較フレーム) について検討する。ユーザ要求には注目すべき点について言及されている場合がある。例えばユーザ要求例 (1) では京都府の降水量が注目点である。このようにユーザ要求の中で注目点となる部分を主題部位と呼ぶ。また主題部位を対比するための対象は、(1) では近畿の他の県の降水量である。このような比較の対象となる部分を比較対象と呼ぶ。主題部位と比較対象は、グラフを描画する際に強調部位を決定するのに必要になる。更に、ユーザ要求の中には "急激に伸びた" のように主題部位の比較対象に対する様子を表す表現がある。同じデータが対象であっても、この表現が "わずかに伸びた" の場合と "急激に伸びた" の場合では、ユーザの意図が異なるのでその差異を明確にするためにグラフを描き分ける必要がある。このような主題部位の様子を表す程度表現を比較内容と呼ぶ。これらを考慮すると、描画方法を決定するための比較フレームは図 3 のようになる。

表 1: 表 1 のフレームにより再構成されたテーブル

滋賀	京都	奈良	大阪	和歌山	兵庫
2330	1200	600	300	1310	920

表 2: 表 2 のフレームにより再構成されたテーブル

	徳島	香川	愛媛	高知
男性	395000	494000	711000	382000
女性	436000	534000	793000	431000

主題部位	$Y_k =$ 京都府
比較対象	$Y - Y_k$
比較内容	多い

図 3: ユーザ要求 (1) の比較フレーム

3 従来研究との比較

ユーザ要求に基づいてデータを可視化する研究として Green らの研究 [2] が挙げられる。Green らは限定量化一階論理 (RQFOL) を用いてユーザ要求を表現する言語を提案している。この記述言語はユーザ要求が複雑であっても表現することができるが、述語などの記述要素はドメインに依存したものであり、そこからの一般化が議論されていないので処理の汎用性に疑問が残る。また、ユーザ要求に適したデータの詳細度を考慮していないため、そのままでは様々なデータに対応することができないと考えられる。

これに対して我々の手法では、ユーザ要求からデータテーブルを作成し、そのデータテーブルに基づいて可視化を行なう。データテーブルはデータのドメインに関わらず話題項目、粒度、制約条件により作成できるので、Green らの手法に比べるとドメイン依存性が低いといえる。また、ユーザ要求に合致した詳細度でデータテーブルを再構成するため、様々なデータに対応可能である。

現時点では複合棒グラフの描き分けなど Green らの手法が表現力の高い部分もあるが、今後、グラフの描き分け規則や比較フレームの改良によりこれらにも対応していく予定である。

4 終りに

本稿では、数値データをユーザ要求に基づいて適切なグラフ形式で可視化するシステムを実現するための意味フレームについて提案した。

再構成されたデータから適切な表示手法を選択するアルゴリズムについては別稿にて示す。

参考文献

[1] 米澤、松下、牧野、加藤 : 「ユーザ要求を反映した数値データ可視化手法」, 第 13 回人工知能学会全国大会, pp. 191 - 194 (1999).

[2] Green, N. *et.al.*: A Media-Independent Content Language for Integrated Text and Graphics Generation, *CVIR '98*, pp. 69–75 (1998).