6K-03　数値データ可視化のためのグラフ判別知識の構築 — 新聞記事中のグラフに基づく分析 —

米澤 勇人　　松下 光範　　加藤 恒昭
{hayatoyo, mat, kato}@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

数値データから自動的にグラフを描画する技術の実現を目指している。そのためには、データの性質や特徴に基づき適切なグラフ種を選択するための知識 (以下、グラフ判別知識と呼ぶ) を構築する必要がある。

APT[1] や PostGraphe[2] 等の可視化システムでは、グラフ種の選択はヒューリスティクスに頼っているが、このように構築された知識では、網羅性や妥当性の保証が困難である。そのため、我々は新聞や白書等で実際に使われているグラフから統計的手法によりグラフ判別知識を自動構築することを目指している。

本稿では、まずグラフ種の決定要因の候補について述べる。次に軸とその性質の扱い方について述べ、最後にこれらの要因を元に自動構築したグラフ判別知識のグラフ判別能力を評価する。

2　グラフ種の決定要因

一般にグラフ種は描画するデータの特徴や性質を考慮して決定される。この特徴や性質を判断基準にすれば、グラフ種を決定できると考えられる。

まず、従来手法で用いられているような、ヒューリスティクスによって構築された知識を参考に、グラフ種を決定する要因の候補を “軸属性”、“属性数”、“依存関係”、“100%割合”、“強調部位” の 5 つに絞った。

“軸属性” とは各軸の持つ性質のことであり、量、割合、名義、順序、区間、時間、無のいずれかの値を取る。これら軸の持つ性質は描画対象のデータが持つ性質を反映したものである。しかし描画されるデータ系列が複数ある場合、そのデータ系列間の性質は軸として表現されない。これを表現するために、グラフを三次元に展開して扱った。　例えば図 1(a) のグラフは、図 1(b) に示すように三次元に広がるグラフの合成として捉えた。これにより、データ系列が複数ある場合、データ系列間の性質は V_z 軸の性質として表現できる。

“属性数” とは描画されるデータ系列が単系列か複数系列かを区別するものであり、軸属性を有する軸の数を値とする。例えば図 1(a) では、3 軸共に軸属性が意味をなすため 3 となるが、東京のみの単純棒グラフでは、V_z 軸は意味をなさないため 2 となる。

Knowledge Construction Method to Choose Appropriate Graph-types Hayato Yonezawa, Mitsunori Matsushita and Tsuneaki Kato, NTT Communication Science Labs., 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

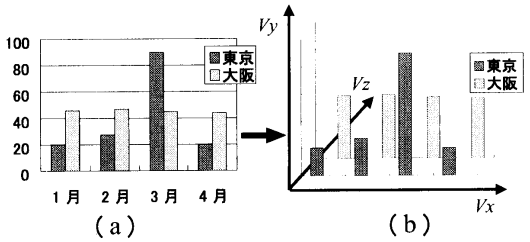

図 1: グラフの分解

```
( ID0001
  ( グラフ種      複合棒グラフ )
  ( 属性数        3 )
  ( 依存関係      無 )
  ( 100%割合      無 )
  ( 強調部位      無 )
  ( Vx 軸         時間 )
  ( Vy 軸         量 )
  ( Vz 軸         名義 ) )
```

図 2: サンプルデータ

“依存関係” とは描画されるデータ系列が複数ある場合に、その系列間の包含関係の有無を区別するものである。例えば人口を図 1(a) のような複合棒グラフで描画する場合、東京と大阪を比較したグラフでは、各棒の長さがそれぞれの都市の人口を示すため、両者の間には包含関係が存在しない。しかし東京と関東を比較したグラフでは、東京は関東の一部であるために、両者の間に包含関係が存在する。

“100%割合” とは軸属性が割合である時に、更に合計が 100%であるかどうかを表す。

“強調部位” とはグラフ中の強調された部位がどの軸に存在するかを表すものであり、V_x, V_y, V_z, 無のいずれかの値を取る。

以上の決定要因を用いると、図 1(a) のグラフは図 2 のように表現できる。

3　軸とその性質の扱い方

グラフ描画において、縦軸や横軸の割り当ては重要な問題である。しかし、グラフによっては軸が入れ替わっても意味を損なわないものがあり、その選択はデザイン的要因によると考えられる場合もある。例えば図 1(b) のグラフの場合、V_x 軸と V_y 軸が入れ替わると、グラフ化の意味が損なわれるが、V_x 軸と V_z 軸が入れ替わった場合には、グラフ化の意味を損なわない。

このような、軸とその属性の相関を調べるために、軸の扱いに関しては次の3パターンについて検討した。

パターン 1: 軸と属性との対応を考慮する場合
(V_x, V_y, V_z) = (時間, 量, 名義)

パターン 2: 軸の属性の組み合わせを考慮する場合
軸属性 = { 時間, 量, 名義 }

パターン 3: 軸に関係なく軸の属性を考慮する場合
量 = 1, 割合 = 0, 名義 = 1, 時間 = 1, ...
(数値はその属性の数を示す)

パターン 1 では各軸と "軸属性" との個々の対応が判断基準になるのに対して、パターン 2 では "軸属性" の組み合わせがグラフ種の判断基準となる。また、パターン 3 では、例えば "軸属性" に量が存在するかどうかが判断基準となる。

4 評価実験

上述したグラフ決定要因と軸属性の扱い方を考慮し、99年6, 7月の7社分の新聞記事と97年度通信白書で用いられているグラフから、265個のサンプルデータを作成した。このときの内訳は、折線グラフ81、棒グラフ68、円グラフ44、複合棒グラフ30、積上げ棒グラフ18、二重円グラフ13、構成比率棒グラフ11個であった。これらを132個の訓練データと133個のテストデータにランダムに二分し、グラフ決定要因を分類属性、記事のグラフを分類クラスとして、ID3 を用いて訓練データから決定木を作成した。

作成した決定木の葉ノードのうち、複数の候補を持つものが、平均して2, 3割程度存在する。そのため、図3に示すようにそれらの候補に順位を付けた。この優先順位は、その葉ノードに辿り着いた訓練データのグラフの数により決定した。

このように作成した決定木を用いて、テストデータによる評価実験を行った。各パターンについて200回の実験を行なった時の、平均正答率を表1に示す。ここで、"一意" とは優先順位が一番目の正しいグラフ種候補に決定したものであり、"候補" とは優先順位に関係なく葉ノードに正しい候補が存在していた場合である。

いずれのパターンの場合にも、全訓練データのうち70%以上が一意に正しいグラフ種を決定できている。また正しく候補が選択されたものは、全体の90%前後の正答率となっている。人間がグラフ種を判断する際にも、少なくともデータの特徴や性質からは適切なグラフを一意に決定できない場合があることを考慮すれば、グラフが完全に一意に決定できないのは、妥当であると思われる。また一意に決定できないもののうち、20個は候補として選ばれており、今回選択したグラフ種の決定要因は妥当なものと考える。

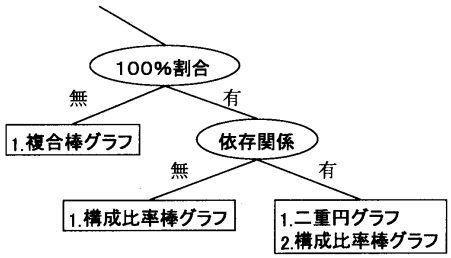

図 3: 優先順位付き決定木

表 1: 実験結果

	パターン 1	パターン 2	パターン 3
一意に決定	74.0%	72.1%	76.9%
候補が決定	89.6%	88.0%	92.8%

軸とその属性値の対応は、グラフ種の決定に大きな影響を与えていると思われたが、実際には、軸とは無関係に属性を扱った方が、正答率が良くなることが確認できた。これは、グラフ種の決定には、どのような軸属性を取るか、どのような軸属性の組み合わせであるかよりも、どの軸属性を含んでいるかということが重要であることを示唆する。

また作成された決定木を分析したところ、パターン1と2の場合は軸に関する属性がトップノードに用いられているケースが多く、パターン3の場合には "100%割合" がトップノードになっているケースが多かった。"100%割合" ということは、軸の属性のうちいずれかが "割合" であるを意味する。これらのことから、グラフ種決定には、軸属性が他の要因よりも大きな影響を与えているといえる。

5 終わりに

グラフ種決定要因の候補について述べた。それに基づき作成した決定木の実験により、グラフ種決定には軸の持つ属性が大きな影響を与えていることがわかった。また、グラフ種決定には、各軸がそれぞれどのような属性を持っているかではなく、どのような属性が存在するかということが重要であることがわかった。

謝辞 本研究を進めるにあたり、プログラム作成、実験に協力して頂いた大阪府立大学総合科学部の修士学生の松本 裕二氏に感謝します。

参考文献

[1] Mackinlay, J. D.: Automating the Design of Graphical Presentations of Relational Infomation, *ACM Trans. on Graphics*, Vol. 5, No. 2, pp. 110–141 (1986).

[2] Fasciano, M. *et.al.*: Automatic generation of statistical graphics, *CMC '95*, pp. 303–305 (1995).