文書中の数値的特徴を用いた情報可視化

太田 彰† 福本 淳一‡

†立命館大学大学院 理工学研究科 ‡立命館大学情報 理工学部 メディア情報学科

E-mail: †a_ota@nlp.is.ritsumei.ac.jp, ‡fukumoto@media.ritsumei.ac.jp

1. はじめに

近年，文書中から数量表現を抽出し，それらを表やグラフを用いて視覚的に表す情報可視化が注目を集めており，2005 年から「動向情報の要約と可視化に関するワークショップ(MuST)」[1]が開催され多くの発表がなされている．MuST においては，ある商品の価格や売上高，ある会社の業績，内閣や政党の支持率の推移など，統計量に関する時系列データを基にして，その変化を通時的に捉えて纏め上げ，グラフや表として出力する手法が多く提案されている．

一般に，グラフには様々な種類があり[2]，文書中の数値情報のタイプや書き手が情報をどのように伝えたいのかによって使用するグラフは異なる．折れ線グラフは時間経過による数量の変化を視覚的に表示するときに適している．例えば，日経平均株価の変動や内閣支持率の推移など，日付によって価格や支持率が変わる場合に使用する．棒グラフは数量の比較を視覚的に表示するときに適している．メーカ毎の出荷台数や地域ごとの地価の比較をしたい場合は，棒グラフを用いることで，数量の比較を明示することができる．円グラフは割合を使用して数量の比較を行なう場合に適している．メーカ別出荷量シェアや商品購入者の年代別割合などを示したいときに用いる．

可視化する対象となる数値情報を適切なグラフを用いて出力するためには，数値情報のどの要素によって適切なグラフが決定されているのかを明らかにする必要がある．山口らは，数値情報のタイプに合わせて自動的にグラフの種類を選択するため,プロットの数,単位の有無などからグラフの選択を行なっているが，グラフの選択に関する具体的な手法については示されていない[3].

我々は複数文書から数値情報を抽出し，抽出された数値情報に適したグラフの種類を自動的に選択し，可視化をする手法を提案する．そのための抽出した数値情報からグラフの種類を決定する基準を示す．本研究では，可視化の対象として，MuST コーパス中の新聞記事に対して統計量の可視化に必要な要素に対して付与されたタグを基に，グラフを生成するために必要な価格や日付といった情報の抽出を行ない，抽出したデータからグラフの要素となるものを選択する．そして，選択したデータのタイプによって適切なグラフを生成する．

以下，次節では，グラフ作成の元になる 4 つ組データの抽出方法について述べ，第 3 節ではグラフの要素となるデータの選び方について述べる．第 4 節では，適切なグラフの選択基準について述べる．

2. グラフ生成に必要な情報の抽出

グラフを作成するためには，

(1) 統計量が何について言及しているものなのかを示す属性名
(2) 属性名に複数の項目がある場合それを指す対象
(3) 属性名や対象についての数量を表す属性値
(4) 属性名が何時の値なのかを示す日付表現

の 4 つの情報が必要である.これらグラフを描くために必要な要素の組を本稿では 4 つ組と定義し，MuST コーパスに付与されているタグを基に抽出を行なう．以下,それぞれの情報を抜き出すために使用したタグやパターンについて記載する．

・属性名

属性名の抽出は MuST タグの stat 属性を利用して行なう．stat 属性は，指定された統計量や出来事に言及している部分に付与される<unit>タグ，統計量の名前に付与される<name>タグ,<val>タグに付く属性であり，どの統計量について言及しているのかを表しているものである．例えば，属性名の対象としては,“国内出荷台数”,“ドバイ原油価格”などがある．

抽出にこの属性を利用するのは，コーパス中で属性名を表す語は統一がされてないことが多いため，どの統計量について述べているのかが分かりにくいからである．以下に同じ統計量について言及している MuST コーパス中の例を示す．

<name part="head">国内出荷台数</name>を見ると，</ins> <date gra="月" abs="199810-199907">1999冷

凍年度（９８年１０月～９９年９月）の７月</date>までの<name part="foot">累計</name>は<val>５９３万９０００台</val>

<date gra="年" abs="1999">今年</date>は<date gra="年" abs="1998">前年度</date><name part="foot">実績</name>の<val>６５５万台</val>を下回る

この例では国内出荷台数と実績はどちらも同じ統計量について言及しているが，これらが同じものを指していると判断することは難しい．以上の理由から stat 属性を属性値の対象とすることで属性名の決定を行う．

・対象

属性名について複数の項目がある場合，対象の要素とする．対象の要素となるものは，メーカ名，人名，地名などがあげられ，例えば，“NEC”，“サミー・ソーサ外野手”などがあげられる．抽出には<par>タグ，<topic>タグを使用している．<par>タグはパラメータ，出来事をおこした主体やその場所等，出来事の一部となる事物やシェアにおける会社名や台風の号数等に付与される．このうち，利用者の関心として指定されたものには<topic>タグを使用している．

・属性値

属性値の抽出には統計量に振られている<val>タグを使用して抽出を行なう．“前後”，“以下”などの概数表現や範囲表現も属性値に含まれる．これらの表現を含んでいると属性値が一意に決まらないので，“１５ドル前後”，“１０ドル台”などの概数表現を含む場合は「15 ドル」，「10 ドル」と概数表現を取り除いて抽出を行なう．“１２～１３ドル”などの範囲表現は，範囲表現の最小値と最大値の平均を取り「12.5 ドル」として抽出を行なう．また“１２ドル５０セント”，“１万７９３９円１２銭”などのドルやセントは同じ数量について書かれているが，大きさの単位が異なる．このような数値は「12.5 ドル」，「17939.12 円」として抽出を行う．

・日付表現

日付表現には<date>タグが付けられている．<date>タグには abs 属性として，その表現が指示する日付が 4 桁から 8 桁の数字で示されている．日付表現は，この数字の抽出を行なう．抽出結果は”19980214”，“199803”の様に年/月/日で表される．また，”前年同月比”，”同５%増”，”８月”と比べてなどの比較として使用されている表現は予め登録し，それらを日付表現の要素としては抽出しないようにしている．

次に抽出したものを 4 つ組としてまとめる方法について述べる．4 つ組みとしてまとめるため属性値を中心となる要素とする．これは属性値が無い場合，グラフの作成が不可能であるためである．抽出されたデータのうちまず属性値を選び，その属性値を保持する他の要素を選び出すことによって 4 つ組みを構成する．まず，属性値から<unit>タグの開始場所までの間に<date>タグがあれば，属性値から最も近いものを日付表現として抽出する．属性値より前に<date>タグが無い場合，属性値と同文中で属性値より後ろに<date>タグがあれば日付表現として抽出する．日付表現が見つからない場合は 4 つ組の要素として抽出をしない．属性名より前にある日付表現を優先的に抽出しているのは，MuST コーパス中の属性値と対応する日付表現を分析した所，日付表現は属性値より前に出現することが多かったからである．属性名は抽出した属性値に付与されている<val>タグに stat 属性があれば，それを属性名として抽出する．<val>タグに stat 属性がなければ，属性値と同文にある<name>タグの stat 属性を抽出する．<name>タグにも無ければ，属性値を囲っている<unit>タグの stat 属性を属性名として抽出する．更に抽出した属性値の同文中に<par>タグ，もしくは<topic>タグがあればそれを対象として抽出する．4 つ組の抽出において，対象以外の情報が欠けた場合はグラフの要素として抽出を行なわない．

3. 抽出された情報のまとめ上げ

4 つ組として抽出された情報を用いてグラフ作成を行うためには，表示すべき情報を選択し，まとめる必要がある．そのため，トピックごとに抽出した 4 つ組のデータについて属性値の単位が同じものを一つの集合として扱う．これは，単位が等しいもの同士の集合でなければ一つのグラフとして生成することができないためである．また，同じ属性名についての日付による値の変化や，ある日付の会社別の売り上げ比較を行うといったグラフを描くため，属性名，対象，日付表現のうち一つが異なり，残りが等しい 4 つ組の集合である必要がある．日付表現だけが異なるもの，属性名だけが異なるもの，対象だけが異なるものを 4 つ組集合から集める．

日付表現が異なり，他が等しい 4 つ組集合を表 1 に，属性名が異なり，残りが等しい 4 つ組集合を表 2 に，対象が異なり，残りが等しい 4 つ組集合を表 3 に示す．

表 1:日付表現が異なる 4 つ組集合

属性名	対象	属性値	日付表現
ドバイ原油価格		12.5 ドル	19980121
ドバイ原油価格		12 ドル	19980214
ドバイ原油価格		9.98 ドル	19980317
ドバイ原油価格		10 ドル	19980317
ドバイ原油価格		12.5 ドル	19980527

表 2:属性名が異なる 4 つ組集合

属性名	対象	属性値	日付表現
主要メーカーの国内生産台数	日産自動車	96038	199908
主要メーカーの輸出台数	日産自動車	44842	199908
主要メーカーの海外生産台数	日産自動車	81497	199908

表 3: 対象が異なる 4 つ組集合

属性名	対象	属性値	日付表現
メーカー毎の出荷シェア	NEC	30･9%	1997
メーカー毎の出荷シェア	富士通	24･1%	1997
メーカー毎の出荷シェア	アップルコンピュータ	7･8%	1997

以上の基準に沿って集める場合，1 トピックについて複数の 4 つ組集合ができる場合がある．現在は複数の候補がある場合,その全てを出力することとしている．

4. グラフの種類の選択

4 つ組集合の特徴から，適切なグラフの種類を選択する手法について述べる．グラフの種類は，属性名，対象，日付表現のうちどの要素が異なるのかによって決定される．円グラフについては，属性値の単位や全体の数量が分かっているかどうかで決定することができる．グラフの種類を分類する基準を表 4 に示す．

表 4： グラフ種類の分類基準

属性名	対象	日付表現	その他	グラフの種類
等しい	等しい	異なる		折れ線グラフ
等しい	異なる	等しい		棒グラフ
異なる	等しい	等しい		棒グラフ
等しい	異なる	等しい	･単位が割合を示している ･全体の数量が分かる	円グラフ
異なる	等しい	等しい		円グラフ

属性名と対象名が等しく，日付表現が異なるものは,日付の変化によって属性値が変化をしているので，推移を表す折れ線グラフが適切である．集まるデータが 3 つ以下の場合には折れ線グラフを生成しない．表 1 のケースではドバイ原油価格の値が日付によって変化しているので折れ線グラフになる．表 1 から作成したグラフを図 1 に示す．

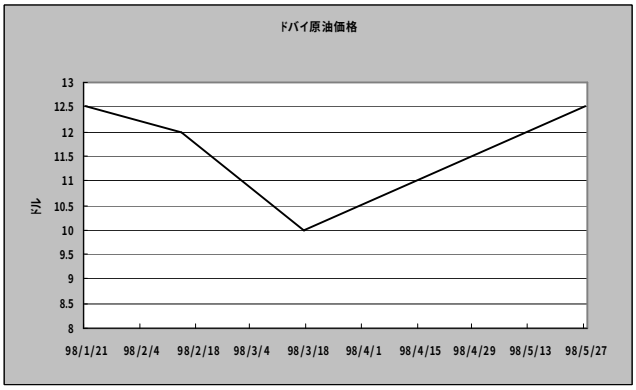

図 1:折れ線グラフの出力例

図 1 において，横軸は日付となり，最も古い日付から新しい日付までが値域になる．途中の日付が抜けていても構わないが，グラフの外形が崩れてしまうため正しい間隔でプロットをする必要がある．縦軸は属性値になり，0 から属性値の最大値よりもやや高い値を値域とする．下限を 0 以外の値に取ることで変化を大きく見せ,属性値の推移を捉えやすくすることもある．下限として適切な値は，属性値の変化の大きさや最小値，最大値によっても変わるが，今回は最小値の半分の値を下限とした．グラフの外形が崩れるので抽出した属性値の最小値と最大値の間を省略することはしない．

日付表現が等しく属性名，もしくは対象のどちらか片方が等しいものを集めた場合，属性名同士，あるいは対象同士の属性値についての比較を行なっているので棒グラフになる．表 2 のケースでは日産自動車の国内生産台数，海外生産台数の比較を行なう棒グラフが適切である．表 2 から作成したグラフを図 2 に示す．

属性名や対象のうち，他と異なっている部分が横軸の要素となる．上の例では，国内生産台数，海外生産台数，輸出台数といったものが横軸の要素になる．縦軸の要素と地域は折れ線グラフと同じであるが，値の比較を見せやすくするために，属性値と属性値の間を省略することがある．

棒グラフになるための条件を満たしていて，更に全体の数量が分かっているか,属性値が%や割合で表されている場合は円グラフが適切である．表 3 から作成したグラフを図 3 に示す．

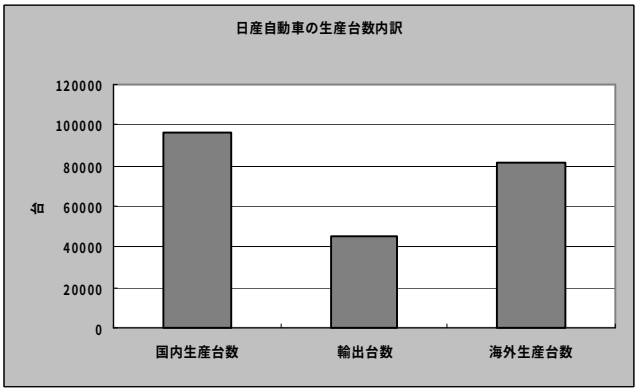

図 2:棒グラフの出力例

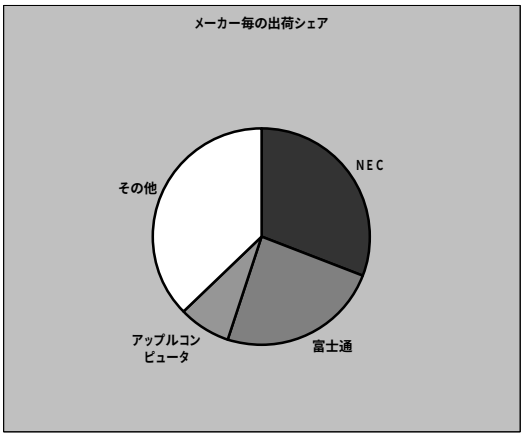

図 3:円グラフの出力例

円グラフの場合，項目名は属性名として割合は属性値を要素とする．同様に全体の数量が分かっている場合は，個々の値の合計が全体の数量を超えている場合は円グラフを出力しない．円グラフは項目が多いものには向いてない．本稿では項目が 5 つ以上の場合は，上位 5 項目以外はその他の項目に表示を行なう．

属性値が変化する関連性のある 4 つ組の集合を複数集めた場合は，1 つの集合に融合させ複数の折れ線グラフを用いて表すことが可能である．日付表現の期間が揃っていること，ドルと円のように単位が同じ数量について述べていることが融合させる基準となる．実際に 2 つのデータを融合させた例を表 5 に示す．

2 つの関連した折れ線グラフを 1 つに表すことで，2 つの数値の関連性が読み取ることができる．

5. おわりに

本稿では，記事から抽出されたデータの特徴を基に適切なグラフの種類を決定し，グラフの作成を行う手法について述べた．この手法を適用することで，MuST コーパスから抽出した 4 つ組に対して折れ線グラフ，棒グラフ，円グラフについて適切な分類をすることができた．しかしながらグラフの描画については本報告ではエクセルの機能を用いているが，描画ツールの実装を行い，与えた文書集合からグラフの作成までを自動化する必要がある．

グラフの分類において，推移を表す場合でも棒グラフを使用する場合や，属性値の単位が「%」でも棒グラフを使用する場合などグラフの種類を一意に決められない場合もある．しかしながら，属性値の種類によってグラフのタイプを決定できる場合や，特定の数値にはある種のタイプのグラフは用いることはないなど，グラフの決決定に数値のタイプなどを利用するなど手法を拡張することが今後の課題である．さらに，今回はグラフの種類として 3 つのグラフに限定をしたが，今後は他のグラフへの適応や，数字が大きい場合で変化の部分だけを強調したい場合に属性値の間を省略する処理を自動で行えるようにすることも今後の課題である．新聞記事や Web サイトなどのタグが付与されていないコーパスへの適応も行いたい．

表 5:2 つのデータの融合例

属性名	属性値	日付表現
ドバイ原油価格	20 ドル	199710
ドバイ原油価格	10 ドル	199812
ドバイ原油価格	17 ドル	199907
ドバイ原油価格	20 ドル	199908
レギュラーガソリンの全国平均店頭価格	100 円	199710
レギュラーガソリンの全国平均店頭価格	94 円	199804
レギュラーガソリンの全国平均店頭価格	94 円	199907
レギュラーガソリンの全国平均店頭価格	94 円	199908

参考文献

[1] 加藤恒昭，松下光範，神門典子:“動向情報の要約と可視化 －その研究課題とワークショップ－”，知能と情報， vol.17，No.4，pp.424-431 (2005).

[2] 内田治:“グラフ活用の技術 : データの分析からプレゼンテーションまで”，PHP 研究所(2005).

[3] 山口智由，関洋平，青野雅樹:“要素の出現系列に着目した動向情報の抽出と可視化”，動向情報の要約と可視化に関するワークショップ 第一回成果進捗報告会予稿集，pp.19-22(2005).