1E3-03

動向情報に基づく情報可視化の基礎検討

Primary Study of Information Visualization for Trend Information

松下 光範[*1]　　加藤 恒昭[*2]
Mitsunori Matsushita　Tsuneaki Kato

[*1]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corp.

[*2]東京大学 大学院 総合文化研究科
The University of Tokyo, Graduate School of Arts and Sciences

The goal of our research is to develop a multimodal question-answering system that can provide trend information in accordance with a user's interest and intention. Since trend information is not a simple set of information but obtained by synthesis and organization of series of temporal information such as a gas-price transition and a cabinet approval rating, the system should employ as its media not only text, but also visual forms such as charts, and use the both media in a cooperative manner. This paper discusses how to generate a visual expression from trend information and clarifies challenges to be addressed.

1. はじめに

計算機の処理能力の向上やネットワーク環境の普及に伴い、ユーザが利用可能な情報は増加の一途を辿っている。そのため、ユーザの関心や興味に合致する情報に直観的かつ簡便にアクセスするための技術が求められている [2]。このような要求に応える技術のひとつとして、我々は動向情報を対象としそれらを要約・可視化する技術の研究を行っている。

動向情報と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、いくつかの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るものである。それは単に時系列データの羅列ではなく、ある観点の下で統合的に纏め上げることで得られるものである。このような動向情報は単なる一次元の時系列情報ではなく、製品のシェアのように複数の企業が関係したり地域毎の土地価格の変動のように空間的な広がりを持ったりするなど、複数主体や空間軸を含んだ多次元情報である。

我々が目指しているのは、このような動向情報に対するユーザの関心、例えば「去年から今年にかけてガソリンの価格ってどう動いていますか ?」や「去年の台風ってひどかったのだっけ ?」、「昨シーズンの大リーグのホームラン競争ってどんな経過だったのですか ?」などに、簡潔で平易な文章やグラフなどの可視化表現で、もしくはそれらを組み合わせて応答できる謂わばマルチモーダル質問応答システムの実現である。我々はこのような目的の研究を協調的かつ競争的に行うために「動向情報の要約と可視化に関するワークショップ (Workshop on Multimodal Summarization for Trend Information, 略称 MuST)」を立ち上げ、共通の素材となるコーパスを作成・配布している [3, 4]。

本稿ではこのコーパス (本稿ではこれを MuST コーパスと表記する) の仕様について簡単に説明するとともに、それを使った情報可視化研究の意義を論じ、更にそのような研究ではどのような課題が解くべき対象となるのかを具体的に見ていく。最後にこのような研究の位置づけを述べる。

連絡先: 松下 光範　日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 社会情報研究部 〒 619–0237 京都府相楽郡精華町光台 2–4 Tel: (0774) 93–5232 Fax: (0774) 93–5285 e-mail: mat@cslab.kecl.ntt.co.jp

2. MuST コーパスの概要

MuST コーパス は 1998 年から 1999 年の 2 年間分の毎日新聞を元に、ガソリン価格やパソコン出荷状況など 20 トピックについて時系列になっている記事を収集し、各トピックにつき 3 つ前後の統計量を選んで、これらの統計量の可視化に必要な要素に対して人手でタグを付与したものである。一例として原油価格の記事にタグが付与されたものを示す。

```
<unit stat="ドバイ原油価格 ">
  また、
  <name part="head"> 原油価格 ( ドバイ原油)</name>
  も、
  <date gra="月 " abs="199710"> 昨年１０月ごろ</date>
  <name part="foot"> １バレル＝</name>
  <val> 約２０ドル</val>
  をつけたのを
  <rel> ピーク</rel>
  に下落が続き、
  <date gra="旬 " abs="19980121"> 今年１月下旬</date>
  には
  <pro ref=" １バレル "> 同</pro>
  <val> 約１２ドル５０セント</val>
  まで落ち込んだ
</unit>。
```

MuST コーパスの各記事に付与されている主要なタグの意味を表 1 に示す。なお、タグの詳細な仕様に関しては文献 [3] を参照されたい。

各記事に付与されているタグは、統計量の名前や値、日付などの要素を抜き出し、値に関してはどの統計量のものか、日付に関してはその絶対表現はいつか (例えば「昨日」に対して「19990203」など) を記述したものである。これは、プレーンテキストの記事に対して統計量に関する言及であることを前提とした意味処理や省略補完などの文脈処理を行った結果に相当する。これらのタグを参照することで、プレーンテキストの記事から動向情報を抽出する際に大きな問題となる日時の省略 (例えば「3 日の〜では」という文章が指すのが何年何月の 3 日なのか分からない) [6] などを考慮しなくて良くなるため、比較的精度良く可視化表現を得ることができるようになる。

表 1: タグの一覧 (文献 [3] より抜粋)

タグ	意味
<unit>	可視化対象の統計量や出来事に言及している部分を示すタグで、言及されている統計量 (stat) や出来事 (event) が属性として付与されている。また、その情報が事実ではなく予測である場合には type="pros" が付与される。
<name>	統計量の名前を示す。複数に分かれている場合も多いため part="head" ないし part="foot" を付与することで分かれていることを示している。
<date>	時刻の表現を示す。その表現が具体的に示す日付と粒度が属性として付与される。
<val>	統計量の値を示す。
<rel>	統計量そのものではないが、その値の差や順位、比などの相対値を示す。
<pro>	参照表現を示す。
<ins>	テキスト中では省略されているが可視化に必要な要素を人手で補完した内容を示す。

3. 情報可視化から見た MuST の意義

情報可視化に関する研究にとって、前章で紹介した MuST コーパスを利用する意義は、大きく 3 つあると考えている。以下にそれらについて述べる。

(1) 自然言語処理の敷居を下げる

一般的な複数テキスト要約の流れ [12] は、① 関連するテキストの自動収集、② 重要文の抽出、③ 冗長性判定、④ 重要箇所の抽出、⑤ 書き換え、⑥ 要約表現の生成、であることを鑑みると、このコーパスは最初のふたつのステップを処理した時点の出力と見なすことができる。動向情報の可視化も広義のテキスト要約と捉えればこれと同じ処理の流れになるため、情報可視化に興味を持つ研究者にとっては、このコーパスを利用することで、自然言語処理の知識を多分に必要とする最初のふたつのステップを省き、情報可視化自体に大きく関係する後半の処理の研究に注力できるというメリットがある。

(2) 客観的な比較基準を提供する

情報可視化システムの研究は、その特徴ゆえ視覚的に訴えることに重きが置かれる傾向にある。そのため、個別のシステムを取り上げると興味深く感じられるものが多い半面、複数のシステムを比べようとした場合、扱っているデータや機能が全く異なるため比較しにくいという問題があった。 MuST では、データと目的を共有しているため、システム間の比較が行いやすくなると期待される。

(3) 自然言語処理と情報可視化に関する境界領域の研究促進に寄与する

テキストの可視化に係わる先行研究では、テキスト集合中の単語の出現頻度や共起関係に着目してテキスト同士の関連性を可視化するもの (例えば [15]) やユーザの興味や焦点に応じてテキストの易読性を向上させるもの (例えば [14]) など、内容の意味理解に踏み込まずに容易に獲得できる統計値やユーザによる指定に基づいたものが多い。それに対して、テキスト内容の意味理解に基づく可視化を行うものは、例えばタグ付き文書から概念図を生成する研究 [10] などのように豊富な背景知識を用意し極めて限られた状況下での可視化表現生成が試みられているだけで、汎用的な可視化表現生成の枠組になり得る研究はあまりなく、まだ萌芽段階にあるといえる。 MuST コーパスの利用はこの課題の難しさを低減させ、意味理解に基づいた汎用的な可視化表現生成技術を成熟させていくうえで一定の役割を果たすと考えている。

4. 解くべき課題

本章では、 MuST コーパスに付与されたタグの利用を前提として基本的な動向情報の可視化の枠組と解決すべき課題について整理する。

4.1 情報可視化の基本モデル

ユーザがデータを可視化する際の基本的なモデル (reference model) [1] を 図 1 に示す。このモデルは、元となるデータ集合 (生データ) の中から可視化に必要なデータを取り出して加工しデータテーブルに変換する data transformation と、このデータテーブルを元に描画系列や軸属性などの可視化構造を決定する visual mapping と、決定された可視化構造のパラメータを変化させて注目点の強調や表示範囲の調整などの処理により視覚効果を高める view transformation という 3 つのプロセスから構成される。ユーザは各プロセスに対してインタラクションを行うことで目的の可視化表現 (グラフなど) を得る。なお、このモデルはユーザとのインタラクションに GUI インタフェースを介した直接操作 (direct manipulation) [13] を想定しているため、各プロセスへのユーザのインタラクションを含んでいるが、動向情報の可視化は必ずしもインタラクションを必要としない[*1]。このモデルに基づいて、動向情報の可視化を行う際に解くべき課題を以下に整理する。

4.2 data transformation に関する課題

テキストの動向情報を対象とした可視化が統計データベースを対象とした可視化 (例えば [8]) と比べて大きく異なるのは、前者の場合には描画する統計量をテキスト中から抽出しなければならない点にある。そのため、動向情報を可視化する際には統計データベースを対象とした場合には生じなかった問題が生じる。 MuST コーパスでは、文の省略補完や時間表現の実時間特定など高度な自然言語処理が必要となるものに関しては省略された文要素を <ins> タグで補完する、時間表現の示す時間がいつなのかを<date> タグの属性として与えることにより問題の簡単化を図っているが、これ以外にも解決すべき課題がある。以下に主要な課題を指摘する (なお、以下の例文ではいずれもタグの箇所を除いて示している)。

(1) 統計量を必ずしも直接的に抽出できない

テキストに含まれる抽出可能な統計量の例として、次の例文を挙げる。

[**例文 1**] 1997 年 10 月頃の原油価格は 1 バーレルあたり 20 ドルだった。

2 次元の統計グラフ (折れ線グラフなど) を用いて可視化する場合、このような例文から 〈1 バーレルあたりの原油価格、1997 年 10 月、20 ドル〉 という 3 つ組が抽出できれば、これをグラフ上の点としてプロットできる。例文 1 のようなテキ

*1 自動処理のみで正確なグラフを得ることは現実的には難しい、ユーザの興味が静的なものではなく得られた可視化表現によって新たな興味が想起される状況 [9] を扱いたい、などの理由により、動向情報の可視化に積極的にインタラクションを採り入れる方向性も有り得る。ただし、 MuST の枠組に照らせば、これらの場合でもインタラクションが生じるのはシステムがグラフを提示してからになるので、インタラクションを行うこと無く one-shot のグラフ生成が可能でなくてはならない

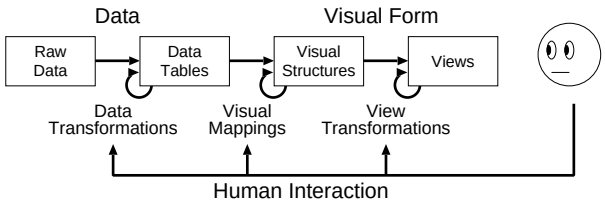

図 1: 情報可視化のモデル [1]

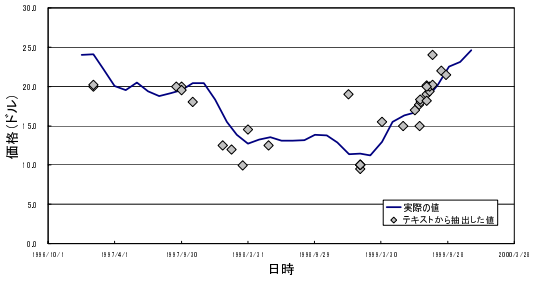

図 2: 実際の原油価格の推移と抽出した統計量との比較 (点が抽出した統計量)

ストからこれを抽出するのは比較的簡単である。しかし、テキスト中の統計量はこのような表現ばかりではない。例えば次のような例である。

[**例文 2**]1997 年 10 月の原油価格 (1 バーレルあたり) は先月より約 1 ドル上昇した。

[**例文 3**] 1998 年 10 月の原油価格は 1 バーレルあたり 14 ドルで前年の同月に比べて 30% の下落になっている。

1997 年 10 月の原油価格を抽出する場合、例文 2 では「先月 (= 1997 年 9 月)」の原油価格との比較で値が表現されているため、関連する記事など文脈情報を参照して 1997 年 9 月の原油価格 が 19 ドルであることを特定できれば 1997 年 10 月の原油価格を抽出できる。また、例文 3 では他の記事を参照する必要はないが、 1998 年 10 月の原油価格が 14 ドルで、それが前年同月 (= 1997 年 10 月) を基準とすると 30% 下落していることを理解し、 1997 年 10 月の価格を推論する必要がある。

(2) 記事間で単位が一致していない場合がある

同じ統計量を表現するにもかかわらず、記事間でその基準単位が異なる場合が存在する。例えば次のような例である。

[**例文 4**] 1998 年 5 月の原油価格は 1 リットルあたり 12 〜 13 円である。

例文 1 と例文 4 から抽出した統計量を見ると、原油の基準単位が前者では 1 バーレルあたりであるのに対し、後者では 1 リットルであるため単純に比較できない。また、原油価格も前者はドルで表現されているのに対し、後者は円で表現されているため、単位を揃える処理が必要になる*2。

(3) 記事間で粒度が一致していない場合がある

グラフを描画するには、単一の粒度で統計量を扱う必要があるが、テキストからこれを抽出する場合、必ずしも単一の粒度である保証がない。例えば次のような例である。

[**例文 5**] 原油の 1997 年第 2 四半期の平均価格は 1 リットルあたり 20 円前後だった。

例文 5 は時間の粒度が「四半期」であるが、例文 1 では「月」であった。統計グラフを描画するには同一の粒度である必要があるため、これらの例文から抽出した情報を同じグラフ上にプロットする場合には粒度を揃える処理が必要になる。

(4) 統計量の値が厳密に特定できない場合がある。

テキスト中に出現する統計量の値は、しばしば曖昧さを含んで表現される。例えば、例文 1 に見られる「約 1 ドル」という表現や例文 4 に見られる「12 〜 13 円」という範囲を伴う表現がその一例である。他にもテキスト中では「10 ドル前後」や「1 月頃」など値を厳密に特定できない表現も多用される。さらに、「10 ドル台」のように文脈によってこの表現が意味する範囲が「10.0 〜 10.9 ドル」なのか、それとも「10 〜 19 ドル」なのか変化する場合もある。そのため、これらの表現の意味するところをを解釈して正しくグラフを描画するには一連の記事内容との整合性を考慮しながら値の推定を行う必要がある。

*2 更に、この処理を正確に行うためには、この時点でのドル円相場の情報を反映する必要がある。

4.3 visual mapping に関する課題

可視化する対象として data transformation プロセスの段階でテキストから抽出されるのは主として $\langle name, date, value \rangle$ の 3 つ組のような要素の組の集合である。それらを可視化するには抽出された各要素を描画系列や軸属性等の可視化構造に対応付ける必要がある。例えば、折れ線グラフを用いて時系列データとして可視化する場合には、 x 軸に $date$ を、 y 軸に $value$ を、そして描画系列に $name$ を各々対応付けることで可視化表現が得られる。

可視化表現の種類 (グラフ種) が事前に決まっていればこの対応関係を決定するのは簡単であるが、データの種類によって適切な可視化表現は異なる [18]。そのため、汎用的な動向情報可視化の枠組を実現するには、対象となる動向情報のドメインや抽出した統計量の特徴などに応じて適切な可視化表現を選択し、その上で対応関係を決定する必要がある。

また、一連の動向情報に含まれる統計量は必ずしも一定間隔でなく、しかも疎らである。例として、 MuST コーパスの「ガソリン」カテゴリの記事 (20 記事) に含まれる「ドバイ原油価格」トピック (34 unit) を対象として、人手で統計量 (ドバイ原油価格) を抽出しグラフ化したものを図 2 に示す。この図で、実線で描かれているのが実際の原油価格の推移で点が抽出した統計量である。

この図からもわかるように、抽出した点が疎らなうえ、実際の値とは大きく異なる場所にプロットされてしまうものも存在するため、単純にこれらを結線するだけでは実際の推移と乖離したグラフになってしまう。すなわち、例文 1 〜 5 のような記述に着目して抽出した統計量だけでは必ずしも満足のいくグラフを描画できない可能性がある。この問題を低減させるひとつの手段として、テキスト中に含まれる傾向表現を利用することが考えられる。例えば次のような例を考える。

[**例文 6**] 1997 年下期の原油価格は安定傾向にあった。

[**例文 7**] 原油価格は 1997 年 10 月をピークに下落している。

例文 6 に含まれている「安定」という傾向表現に着目することで、他の記事から 1997 年下期のいずれかの時点の原油価格を抽出することができれば、 1997 年 7 月から 12 月にかけてはその値からあまり逸脱しない範囲であったと推論できる。ただし、どのくらいの価格変動範囲であれば「安定」といえるかについては多分にドメインに依存するため、注意が必要である。

また、例文 7 には価格の情報が含まれてはいないが、人はこの文から 1997 年 10 月が頂点となった凸型のグラフ形状を思い浮かべるであろう。テキスト中に直接値に言及した文はそれほ

ど多くないため、このような表現からもグラフ化の手がかりを取り出すことができれば、少ないデータからでもより精度の高いグラフ生成が可能になると期待できる。そのため、 Visual mapping においては、これらの傾向表現から取り出したグラフの概形と data transformation プロセスで抽出した統計量の値をどのように統合するかも課題となる。

4.4 View transformation に関する課題

本研究の大局的な枠組は動向情報をグラフとテキストのふたつのモダリティを用いてユーザに応答することである。これらふたつのモダリティが独立ではなく、互いに補完しあっているほうがより有益な動向情報の要約になると考えられる。そのため、可視化では、テキストの特徴的な表現をグラフ上にアノテーションを付与したり、テキストにあわせて軸の範囲を規定したりするといった協調処理が必要になる。

アノテーションに関しては、例えば、ピークの価格を注釈する、「下落傾向にある」というような傾向表現をグラフ上に付与する、などが考えられる。どのようなアノテーションがユーザの関心に沿っているかを考慮し、適切なものを選択する技術が必要である。もちろん、過剰なアノテーションはグラフの視認性を下げる結果にもなるので、これも考慮する必要がある。

5. 研究の位置づけ

グラフの自動生成に関する研究は古くから様々な研究が行われている。その多くは統計データを前提とし、ユーザの関心や興味に応じて必要十分なデータを抽出・集約し、適切な可視化表現を選択して提示するという枠組を採用している (例えば [7] や [8])。これらの研究はグラフ描画に必要十分な統計データを前提とし、データ集約や適切な可視化表現の選択を解くべき課題と位置付けている。そのため、本研究が対象とするようなテキストからのグラフ描画のように、統計データそのものを抽出しなければならないタスクにはそのまま適用することができない。

グラフを自然言語で表現する研究は幾つか行われている (例えば [16] や [5])。 MuST の課題はこの逆問題といえるが、グラフから自然言語表現の生成は系列データと基本的なパタンとのマッチングで比較的簡単に実現できる。これに対し、 MuST で目指すような可視化表現生成の問題では、利用可能な統計量が疎らであるため、推論や補完などいくつもの技術を組み合わせる必要があるため、格段に難しくなる。

また、 visual mapping に関する先行研究として、新聞記事とそれに付随するグラフの特徴に基づいて決定木を作成し、その決定木を利用してグラフ種を特定する方法 [17] や、データと描画オブジェクトの対応関係を幾つかユーザが指示し、それを参考にしてシステムが描画する方法 (draw by demonstration) [11] などが研究されている。 4.3 節で述べたように、 MuST の枠組では抽出可能な統計量のデータが疎らであるため、統計量だけでなく傾向表現にも着目してグラフを描画する必要がある。そのため、これらの先行研究だけでは課題を解決することができない。

したがって、これらの問題を解決する新しいグラフ描画技術が創出されることが、 MuST の目標となる。

6. おわりに

本稿ではガソリン価格や内閣支持率等、複数の新聞記事に時系列に出現する様々な動向情報から可視化表現を生成する技術に関する基礎検討について述べた。 MuST に関する情報はワークショップのホームページ*3 から得ることができる。

*3 http://www.kecl.ntt.co.jp/scl/workshop/must/

7. 謝辞

本研究は NTT と東京大学との産学連携共同研究、ならびに国立情報学研究所の NTT と東京大学との公募型共同研究によって支援されています。御支援をここに感謝致します。また、本稿の執筆に際し色々と御協力頂いた赤塚大典氏に感謝致します。

参考文献

[1] Card, S. K., Mackinlay, J. D. and Shneiderman, B. (eds.): *Readings in Information Visualization — Using Vision To Think —*, Morgan Kaufmann Publishers (1999).

[2] 福本淳一, 天野真家 (編): 特集 自然言語による情報アクセス技術, 情報処理, Vol. 45, No. 6, pp. 561–585 (2004).

[3] 加藤恒昭, 松下光範, 平尾努: 動向情報の要約と可視化に関するワークショップの提案, 信学技報, NLC2004-25, pp. 13–18 (2004).

[4] 加藤恒昭, 松下光範, 平尾努, 神門典子: 評価なきワークショップの試み — 「MuST: 動向情報の要約と可視化に関するワークショップ」を例に —, 言語処理学会全国大会併設ワークショップ (2005).

[5] 小林一郎: グラフ情報からのテキスト生成への一考察, 言語処理学会第 5 回年次大会, pp. 149–152 (1999).

[6] 国政美伸: 複数テキストからの動向情報の抽出と可視化, 広島市立大学情報科学部平成 16 年度卒業論文 (2005).

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

[8] 松下光範, 米澤勇人, 加藤恒昭: 表題に基づく統計データの自動可視化手法, 情報処理学会論文誌, Vol. 43, No. 1, pp. 87–100 (2002).

[9] Matsushita, M., Nakakoji, K., Yamamoto, Y. and Kato, T.: InTREND: an Interactive Tool for Reflective Data Exploration through Natural Discourse, *Proc. KES2004*, Part 2, pp. 148–155 (2004).

[10] 村山正司, 中村裕一, 大田友一: 概念図の自動生成によるタグ付文書の可視化, 信学技報, TL99-25, pp. 51-58 (1999).

[11] Myers, B. A., Goldstein, J. and Goldberg, M. A.: Creating Charts by Demonstration, *Proc. CHI'94*, pp. 106–111 (1994).

[12] 奥村学: テキスト自動要約, 情報処理, Vol. 45, No. 6, pp. 574–579 (2004).

[13] Shneiderman, B. and Plaisant, C.: Direct Manipulation and Visual Environments, in *Designing the User Interface*, Fourth Edition, Chaptor 6, Addison-Wesley, pp. 213-264 (2005).

[14] Suh, B., Woodruff, A., Rosenholtz, R. and Glass, A.: Popout Prism: Adding Perceptual Principles to Overview+Detail Document Interfaces, *Proc. CHI2002*, pp. 251–258 (2002).

[15] Takama, Y. and Hori, T.: Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization, *Proc. CIRA2003*, pp. 770–775 (2003).

[16] 馬野元秀, 篠原貴之, 瀬田和久: 言語表現に基づく時系列データの検索について, 第 20 回ファジィシステムシンポジウム, pp. 479–482 (2004).

[17] 米澤勇人, 松下光範, 加藤恒昭: 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 —, 情報処理学会第 60 回全国大会, Vol. 2, pp. 119–120 (2000).

[18] Zelazny, G: *Say It with Charts*, third edition, McGraw-Hill (1996).