言語表現による時系列データ検索のための基礎検討

Toward a Time-series Data Retrieval with Qualitative Linguistic Expression

松下 光範　　末吉 れいら
Mitsunori Matsushita　　Reira Sueyoshi
関西大学総合情報学部
Faculty of Informatics, Kansai University

1. はじめに

折れ線グラフは時系列データの経時的特徴を理解する上で有効な表現の一つであり、人は折れ線グラフで描かれた情報を読み解くことで、ある統計量の推移を把握したり特徴的な振舞いをした時期を同定したりすることができる。我々はこのような折れ線グラフの全体的傾向や局所的特徴を言語化して、言葉による検索を可能にするシステムの実現を目指している [1]。例えば、「全体的に穏やかに値上がりしている株は ?」や「電力量が急激に増加する時期は ?」などの質問から、そのような変動をしている時系列データを特定したり、該当する時期を見つけたりすることができれば、時系列データに対するアクセス性の向上が期待できる。

時系列データを言語で表現する場合、必ずしも定性的な表現だけで行われるわけではなく、具体的な値を含めた表現（e.g.,「140 円を超えたあたりから値上がり傾向が鈍化」）や特徴的な事由と絡めた表現（e.g.,「最安値を付けた後の緩やかな改善」）も用いられる。よりユーザにとって利用しやすい問い合わせシステムを実現するには、単に全体的傾向や局所的特徴を用いた言語表現だけでなくこれらの表現を含めた言語表現を取り扱える枠組みが必要になると考える。

このような表現パタンを収集・分類するために、本研究では動向記事の要約文に着目した。要約された動向記事と折れ線グラフはどちらも、対象とする統計量のある期間の要約として捉えることができる。生成された要約文に含まれる時系列データの言語表現を分析することで、人がどのように統計データの動向を捉えているかが明らかになり、時系列データに対する検索質問パタンの整理やユーザに提示する応答の生成に役立つと考える。本稿ではこのような観点から、ある統計量に関する一連の新聞記事を要約したテキストを収集し、そこに含まれる表現の分析を行った。

2. 言語表現による時系列データの検索

本研究で目指すのは、様々な時系列データのリポジトリからユーザが与える言語表現（質問）に基づいて、その質問に合致するデータの種類や期間を特定する検索機構の実現である。そのための要素技術として (1) 時系列データに基づく言語表現の生成、(2) 自然言語で表現された質問の解釈、(3) これらふたつのマッチング方法の定式化、が必要であると考えている。この方式の特徴は、ユーザ質問と生の時系列データ（raw data）とのマッチングを行うのではなく、時系列データに関して予め言語化されている表現とマッチングを行う点にある。これは、同じ振舞いのデータであっても状況や文脈、ドメインによってその解釈が異なるので、それらを考慮した検索を可能にするためである。現在は特に (1) に力点を置いて研究を進めている。

時系列データに基づく言語表現の生成において、人の解釈に沿った表現を生成するには、時系列データとそれを説明したテキスト（新聞記事など）とを対応づけ、時系列データの特徴を適切に表現している文を抽出することが最も効率的・効果的であろう。しかし、このようなテキストが常に得られる保証はないため、その時系列データの全ての利用場面を網羅した記事集合を集めておくことは困難である。そのため、その補完として機械的に時系列データから表現を生成する枠組みが必要になる。

3. 要約文の収集

本稿では被験者実験を通じて時系列データに関する要約文の収集を行った。実験は人材派遣会社によって集められた 10 代から 50 代の男女 24 名を対象として行われた。実験は時間順に並べられた関連する新聞記事集合から要約文とグラフを作成するもので、被験者には筆記具としてシャープペンシルと消しゴムの他に 2 色のマーカペンが与えられた。

被験者に提示された新聞記事は 1998 年 3 月から 1999 年 7 月にかけてのレギュラーガソリンの小売価格に関する記事 14 件で、各被験者には、記事を読み 100 文字から 300 文字程度の要約文と当該期間のグラフの概形を生成するという課題が課せられた。また、与えられたマーカを用いて記事原文の中でグラフや要約文生成の根拠として用いた個所にマーカでラインを引くように指示が与えられた。このとき、色の使い分けに関する指示は与えられなかった。

4. 結果の分析

加藤らは文献 [2] において、テキスト中に含まれる統計量の言及に関する情報として、「100 円」や「20 リットル」など、ある時点での統計量の具体値を表す情報（時点数値情報）、「ピーク」や「値上がり」など値の解釈や変化に関する情報（変化情報）、「イラク情勢の緊迫化

で」や「原油はだぶつき気味」など状況変化の理由や状況に関する情報（状況情報）の 3 つがあることを指摘している。本稿では、この分類のうち時点数値情報と変化情報に着目し、これらの表現の利用の特徴を被験者の生成した要約文に基づいて整理する。

実験で被験者が生成した要約文は平均 264.6 文字（最大 431 文字、最短 120 文字）で、そこから時点数値情報を示す表現 105 個、変化情報を示す表現 226 個の計 331 個（1 人あたり平均 13.8 個、最大 24 個、最小 4 個）を人手で抽出した。

4.1 時点数値情報

要約文中に含まれていた時点数値情報に関する表現は 105 個で、その内訳は、具体値 80 個、概数（e.g.,「100 円台」）9 個、相対値（変動幅など）16 個であった。このうちの具体値に着目し、その使われ方を以下の 5 つのパタンに分類した（括弧内は各パタンに該当した表現数）。

パタン 1 ある期間の極大/極小値 (36)
「90 円の最安値」、「125 円をピークに」など

パタン 2 目安となる値への到達 (3)
「100 円」を割り込んだ」、「95 円を回復」など

パタン 3 値の継続 (5)
「6 ヶ月連続して 92 円」など

パタン 4 変化の始点/終点 (14)
「90 円まで下げた」など

パタン 5 値のみ (22)
「8 月には 94 円に」など

ただしパタン 2 に関しては、表現上は「95 円の底値」と記述されているものの、その後も値を下げたために実際の当該期間の底値（90 円）とは異なる値になっているものが 10 個含まれていた。これは、その時点までの底値が 95 円でその額に到達したという意味合いが強く、表現上の区別はつかないものの意味としてはパタン 2 に相当するものである（これを考慮するとパタン 1、パタン 2 に分類される表現は各々 26 個、13 個になる）。

このように、値単体での記述（パタン 5）は 1/4 程度に留まり、値に対する解釈を伴ったり（パタン 1）、変化情報を補足したり（パタン 2、3、4）する場合に利用されるケースが多いことが分かる。

4.2 変化情報

要約文中に含まれていた変化情報に関する表現は 226 個で、その内訳は、極大/極小 38 個、傾き 108 個、継続（「〜し続ける」）33 個、基準値（「〜を下回る」）11 個、傾向変化（「〜に転じる」）28 個、その他 8 個であった。

極大/極小に属する表現は「92 年末の 125 円をピークに」のように、時点数値情報の具体値と併せて出現する確率が高かった（38 個中 35 個）。一方、傾きを表す表現では具体値を伴うのは 108 個中 29 個に留まり、それ以外は具体値との関係を把握するには文脈から判断しなくてはならなかった。その代わり、傾きを表す表現は多くの場合、時期に関する表現を伴っていた。ただし、それらの時期表現は年の粒度や月の粒度が混在したり、文を跨いで出現したりするため、その対応付けを機械的に行うことは難しいと考える。

傾きを表す表現には「緩やかな」や「急激な」などの修飾語が付与されるケースが見られた。グラフの変化率と対応付けてみると、その傾斜角について「緩やかな下落」<「急激な下落」は概ね成立するが「緩やかな下落」<「下落」は成立するとは限らないことが分かった。

5. 関連研究

渡邉らは日経平均株価のデータを対象として、ユーザが興味の下で特定した範囲に該当するテキストの中から、重要度の高い文を抽出して当該期間の折れ線グラフと併せて表示する方法を提案している [3]。この手法は日経平均株価に特化した表現をテンプレートとしているため、他のドメインにそのまま適用することは難しい。

Ahmad らは、ロイターの配信記事を対象とし、テキストとグラフを用いたマルチモーダル要約を生成する方法を提案している [4]。この研究では Wavelet 解析を用いてグラフの変動サイクルや変極点を特定し、グラフとともにユーザに提示する方法を提案している。Ahmad らの手法は、変動サイクルや変極点などグラフのトレンドや特徴を理解するうえで有効な手法であるが、グラフに対する人の解釈を考慮した要約生成手法ではないため、利用者の意図と一致しない可能性が残る。

6. おわりに

本稿では、被験者実験によりガソリン小売価格に関する一連の新聞記事を要約したテキストを収集し、そこに含まれる時点数値情報と変化情報に関する表現の用いられ方の分析を行った。今後、この知見に基づき、時系列データに関する人の解釈の嗜好を考慮した要約生成手法の検討を進める。

謝辞

本研究は科学研究費補助金基盤研究（C）（課題番号: 22500209）の助成を受けた。記して謝意を表す。

参考文献

[1] 小泉, 松下, 松田, 馬野: 言語情報と統計グラフの相互変換に関する基礎検討, 第 21 回人工知能学会全国大会, 2H5-6 (2007).

[2] 加藤, 松下, 神門: 時系列情報の値と変化に関する言語表現コーパスの構築 —動向情報の情報編纂に向けて—, 人工知能学会誌, 25(**5**), pp.637-650 (2010).

[3] 渡邉, 小林: 動向情報を表すテキスト生成, 第 21 回人工知能学会全国大会, 2H5-4 (2007).

[4] Ahmad, S., de Oliveira, P. C. F., Ahmad, K.: Summarization of Multimodal Information, *Proc. LREC2004*, pp.1049–1052 (2004).