言語表現による時系列データ検索システムの提案

A System for Retrieving Time-Series Data Based on Linguistic Expression

蓮井大樹 [1*]　　松下光範 [2]
Daiki Hasui[1]　　Mitsunori Matsushita[2]

[1] 関西大学大学院 総合情報学研究科
[1] Graduate School of Informatics, Kansai University
[2] 関西大学 総合情報学部
[2] Faculty of Informatics, Kansai University

Abstract: This paper proposes a system for retrieving time-series data based on a linguistic query given by a user. Our proposed system uses a line chart as a query. The system generates a linguistic query by verbalizing the chart first, then retrieves similar charts by using the obtained linguistic query.

1　はじめに

現在、インターネットを介してさまざまな時系列データや統計データを得ることが出来るようになってきた。しかしグラフの検索を行う際に、どのようなグラフを求めているのかを言語化し、それに適したグラフを得ることは難しい。

例えば、「全体的に凸型のグラフ」のように、頭の中には具体的なグラフの形が浮かんでいるが、それを適切に言語化することができないために曖昧な表現しかできない場合には、システムとユーザがインタラクションを繰り返しながら徐々にユーザ要求を明確化し、探索的に適切なグラフを見つける必要がある。また、「昨年 8 月頃に特徴的な変化をした株は?」「他の商品にくらべて価格変動が緩やかな商品は?」といったように、他と比較した値の変化の特徴を利用して探索的に条件を満たす対象を見つけるような場合では、ある対象が全体集合の中でどのような位置づけにあるかを理解し、それを考慮してユーザの意図や関心に沿った区間や粒度を特定する必要がある。

本研究のゴールは、様々な時系列データに対する柔軟な情報アクセス手段の提供である。そのために、ユーザが与えた検索要求から、その条件に見合った変動をしている時系列データを特定したり、特定の時系列データから該当する時期を見つけたりするための検索機構の実現を目指している。

現在そのひとつのアプローチとして、時系列情報に予め自然言語表現を付与しておき、それとユーザの検索要求とのマッチングによって適切な範囲・粒度の時系列情報を特定し、視覚化する手法について検討を進めている [6]。この手法では、(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、という 3 段階の枠組を想定している。本稿ではこのうちの (1) と (3) に焦点をあて、時系列データを対象に、あらかじめ用意したグラフから検索したいグラフと類似している形状や傾向の部分を指定することでグラフ自体をクエリとし、そのグラフを言語化することで検索を行うシステムを提案する。

*連絡先：関西大学総合情報学部
大阪府高槻市霊山寺町 2-1-1
E-mail: mat@res.kutc.lansai-u.ac.jp

2　関連研究

時系列データなどの数値情報の集合を直感的に理解・把握するために、自然言語を用いて表現する手法について、これまで様々な研究が行われている。

例えば、グラフの解析に関する研究として、Ahmad ら [1] や小林ら [5] の研究が挙げられる。Ahmad らは Wavelet 解析を行って変極点や変動サイクルなどのグラフ特徴を抽出し、それらを元にテキストを生成する手法を提案している [1]。小林らは SAX 法を用いて記号化し、複数のグラフを比較することで時系列データ間の関連性を捉え、その結果を言語化する手法を提案している [5]。

また、時系列データからのテキスト生成に関する研究として Kukich [3] や 小林ら [4] の研究などが挙げられる。Kukich は入力として与えられた数値情報列を元に、予め用意したドメイン知識ベースを参照してメッセージ集合を生成し、それを統合することで概況を生成する手法を提案している [3]。また、小林らは与えら

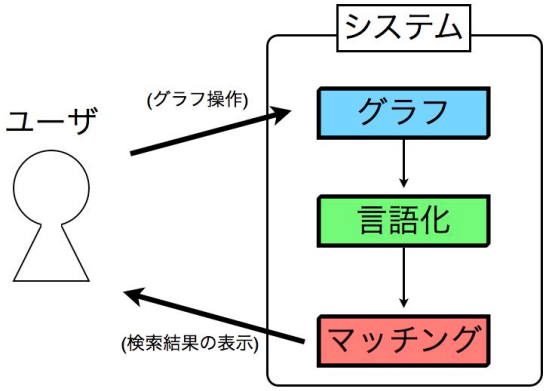

図 1: システムの構成

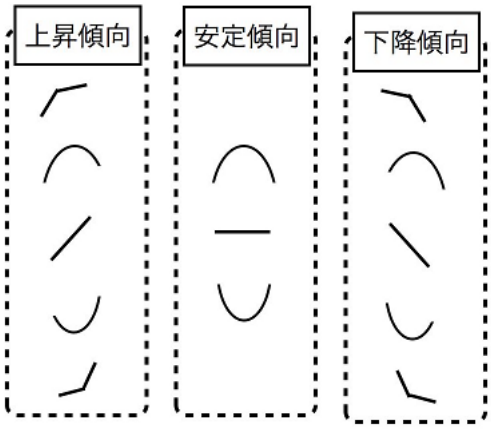

図 2: グラフの概形

れた焦点に基づいて生成する文書のパタンを決定し、そのパタンに応じたテンプレートを用いて説明文テキストを生成する手法を提案している [2, 4]。

3 提案手法

本章では、提案システムで用いたグラフの言語化と言語同士のマッチング手法について述べる。

大まかなシステムの構成を図 1 に示す。このシステムでは、検索の例となるグラフの一部分をユーザが指定し、その箇所を言語化して、それと予め蓄積された時系列データ集合に対応する言語表現とのマッチングにより、適切なグラフを見つけ出す。ユーザはこれを繰り返すことで探索的に統計データにアクセスする。以下では、各段階での処理について説明する。

3.1 グラフの言語化

本節ではまず、時系列データに対する言語表現特徴の付与方法について述べる。

本研究ではグラフに言語表現を付与するために、グラフの変動、変化の度合い、グラフの概形、の 3 つの特徴に着目している。グラフの変動には「上昇」「下降」「安定」の 3 つの表現を用いる。ユーザが指定したグラフの範囲の終点から始点の値を引き、その差がグラフ全体の上限と下限の差の 1/10 以下であれば「安定」とし、そうでない場合で差の値が正ならば「上昇」、負ならば「下降」と判断することとした。

変化の度合いは、終点から始点を引いた差の値を、指定した範囲の期間で割った値から導出している。傾きの値が 2.0 以上なら「大きく」変動している、傾きの値が 1.0 以下なら「小さく」変動しているとし、それ以外の傾きの場合は「なだらかに」変動していると判断した。

グラフの概形は、始点、中間点、終点の各位置関係をもとに導出している。

グラフの概形の分類は「山型」「谷型」「前半が急で、後半が緩やか」「前半がゆるやかで、後半が急」「直線形」の 5 つを基本とし、それぞれグラフの変動が、上昇、下降、安定の場合に分けることで 13 通りになる。すべてのグラフパターンを図 2 に示す。

それぞれのグラフの分類方法としては、中間点と始点の差と、終点と中間の差の方向がある範囲を越えて異なっている場合は、「山型」か「谷型」のどちらであるかを判断する。システム上でそこにグラフの変動を加えることで、例えば上昇傾向の山形のグラフの場合は「前半は上昇しているが、後半は下降している」と表示し、逆に上昇傾向の谷型のグラフの場合は「前半は下降しているが、後半は上昇している」と表示している。下降傾向の谷型のグラフの場合も、上昇傾向の谷型のグラフと同じように「前半は下降しているが、後半は上昇している」と表示されるが、システム内では別の概形として扱っている。グラフの変動が安定であった場合は「山形に変化するが、最終的に安定している」あるいは「谷型に変化しているが、最終的に安定している」と表示している。

前半と後半のグラフの変動が同じ場合で、前半あるいは後半の差が、全体の 1/8 より大きく、もう一方が全体の 1/12 よりも小さい場合には「前半が急で、後半が緩やか」あるいは「前半がゆるやかで、後半が急」だと判断している。システム上では、そこにグラフの変動を加えることで、例えば上昇傾向の前半が急で、後半が穏やかなグラフなら「前半は大きく上昇しているが、後半はあまり上昇していない」と表示し、上昇傾向で前半が穏やかで、後半が急なグラフなら「前半はあまり上昇していないが、後半は急に上昇している」と表示している。

以上の条件に当てはまらない場合は、「直線形」であ

ると判断している。システム上では、グラフの変動と変化の度合いを組み合わせることで例えば上昇傾向で、大きく変化しているグラフなら「全体的に大きく上昇している」と表示し、下降傾向で、小さく変化しているグラフなら「全体的に小さく下降している」と表示している。グラフの変動が安定であった場合は、変化の度合いに関係なく「全体的に安定している」と表示している。

3.2 マッチング手法

ユーザが指定したグラフの範囲をもとに、対象データで同じ範囲のグラフの言語化を行いマッチングを行う。マッチングでは、グラフの変動、変化の度合い、グラフの概形のそれぞれを比較し、0%–100% の間で一致度を導出している。

具体的には、グラフの変動が「安定」、「上昇」、「下降」のいずれかで一致しているかを調べる。このときお互いの傾向が一致しなかった場合、一致度は 0% となる。

一致していた場合は、変化の度合いとグラフの概形での一致度を調べる。それぞれ最大で 40%と 60%の一致度を割り当ててあり、どちらも完全に一致していた場合のみ 100%の一致度となる。今回は変化の度合いよりも概形が似ているグラフの方がユーザが似ていると感じると考えたため、形の方に一致度を多く割り振っている。また、複数の範囲で検索を行った場合は、それぞれに 100 を範囲の数で割った値を一致度の最大として割り当てて計算し、最後に合計する。

概形の一致度を求めるにあたり、「前半が急で、後半が穏やか」なグラフと「前半が穏やかで、後半が急」なグラフでは、グラフの変動が「上昇」であるか「下降」であるかによってグラフの形状が大きく異なっているため、グラフの変動を加味した計算を行っている。上昇傾向の前半が急で、後半が穏やかなグラフと山形のグラフでマッチングを行った場合、類似していると考え、一致度は高めに設定している。逆に下降傾向の前半が急で、後半が穏やかなグラフと山形のグラフでは、一致度は低く設定している。しかし、谷型のグラフとは類似していると考え、高めの一致度としている。前半が急で、後半が穏やかなグラフ同士であっても、グラフの変動が「上昇」と「下降」で異なっている場合は一致度を低くし、上昇傾向の前半が急で、後半が穏やかなグラフと、下降傾向で前半が穏やかで、後半が急なグラフとの一致度は高くしている。

傾きの一致度の導き方を表 1 に、概形の一致度の導き方を表 2 に示す。

表 1: 変化の度合いでのマッチング

度合い	大きい	普通	小さい
大きい	**40**	20	10
普通	20	**40**	20
小さい	10	20	**40**

4 実装

本章では実装したプロトタイプシステムについて述べる。

4.1 対象データ

対象となるデータは統計局ホームページ[1]から収集した。統計局ホームページ内の総合統計データ月報から主な気象官署別の平均気温や降水量などを対象に 2009 年 3 月から 2012 年 2 月までの 4 年分のデータを収集した。

これらのデータはスケールを合わせるために、値が 0 から 400 程度になるように修正している。

4.2 デザイン指針

はじめに、提案システムでは時系列データへのアクセス手段として、言語を用いて検索を行う手法ではなく、グラフの形を提示することでそれを言語化し、類似したグラフを検索する手法を取ることとした。

そのために提案システムでは、見本となる様々な形状のグラフを用意し、検索したいグラフと類似しているグラフの場所を指定することで検索をしたり、探したいグラフの範囲と、その範囲で上昇や下降しているといった傾向を指定することで類似するグラフを検索する機能を設けることとした。

次に、ユーザが指定したグラフはどのようなグラフなのか、グラフの特徴を言語化する。提案システムでは、グラフの指定された範囲を、上昇傾向のグラフか、下降傾向のグラフか、どちらでもない安定傾向のグラフかといった増減傾向と、グラフが全体的に同じ増減傾向か、途中で増減が切り替わっているか、増減の度合いが変化しているかといった、グラフの概形の二つの要素から言語化することとした。

最後に、ユーザが指定したグラフと検索対象となるグラフのマッチングを行う。本システムではさきほど述べたグラフの増減傾向とグラフの概形の二つの要素でマッチングを行い、一致度が高い順に結果を提示する機能を実装することとした。

[1]http://www.stat.go.jp/index.htm

表 2: 概形のマッチング

	山	急→穏やか (上昇)	穏やか→急 (下降)	直線	急→穏やか (下降)	穏やか→急 (上昇)	谷
山	**60**	50	50	30	20	20	20
急→穏やか (上昇)	50	**60**	50	30	20	20	20
穏やか→急 (下降)	50	50	**60**	30	20	20	20
直線	30	30	30	**60**	30	30	30
急→穏やか (上昇)	20	20	20	30	**60**	50	50
穏やか→急 (下降)	20	20	20	30	50	**60**	50
谷	20	20	20	30	50	50	**60**

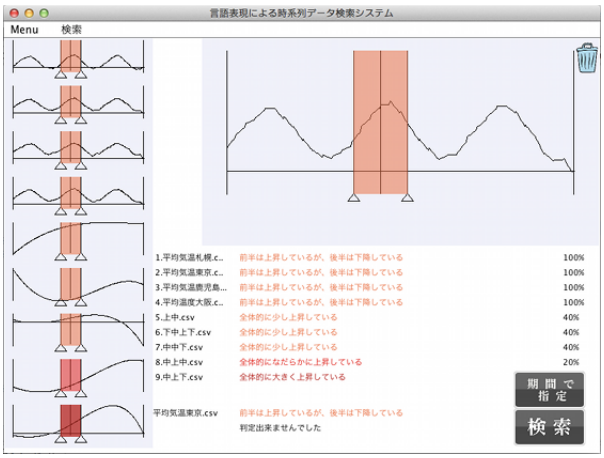

図 3: プロトタイプシステムのインタフェース

4.3 操作方法

図 3 が作成したシステムのインタフェースである。

右側に検索を行う際の見本となるグラフが表示され、左側には検索結果のグラフが上から一致度の高い順に表示される。左上のメニューを選択することでメニュー画面が表示され、グラフ名を選択することで見本のグラフを変更することが出来る。見本のグラフ上をドラッグすることで、検索を行うグラフの範囲が指定出来る。グラフの範囲は始点と終点となる縦線を掴むか、その縦線の下に表示されている三角形の部分を掴むことで左右に調整することが出来る。指定している範囲のグラフを言語化したものが下に表示される。右上のゴミ箱にドラッグすることで指定した範囲は消すことが出来る。

始点と終点の間は色が塗られ、その色はグラフが上昇しているか、下降しているか、安定なのか、によって変化し、それぞれその増減の度合いによって色の濃淡が決定される。詳細を表 3 に示す。

範囲の始点と終点以外の場所でドラッグすると、複数の範囲を指定することが出来る。複数の範囲が重なると始点と終点が掴みにくくなるので、掴んだ状態で上下にドラッグすることで下に表示されている三角形の位置をずらすことが出来る。

表 3: グラフの変動と度合いに対応する色

	小さい	普通	大きい
安定	薄黄	黄	濃黄
上昇	薄赤	赤	濃赤
下降	薄青	青	濃青

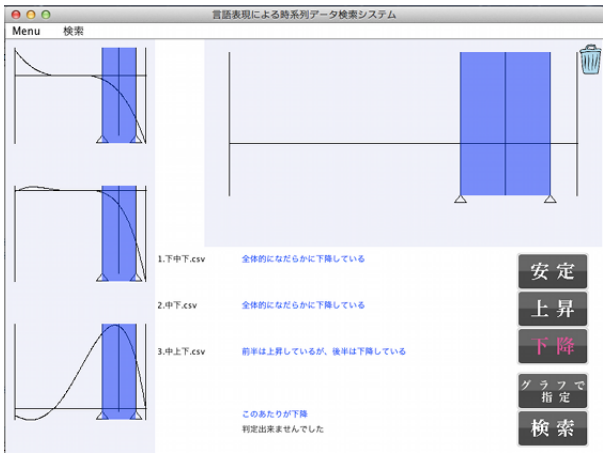

図 4: 期間と変動タイプを指定して検索するモードのインタフェース

右下にあるボタンをクリックすることで、グラフの期間とグラフの変動を指定することで検索を行うモードに切り替えることが出来き、もう一度クリックすることで元のモードに戻すことが出来る。このモードの状態を図 3 に示す。右下にあるボタンを上昇、下降、あるいは安定ボタンをクリックすることで探したいグラフの変動を選択することが出来る。ボタンを選択した状態で、右側の期間指定画面をドラッグすることでどの期間がどのような変動のグラフなのかを指定するこ

とが出来る。

範囲を指定している状態で右下にある検索ボタンをクリックすることで、類似するグラフの検索が行える。

5 おわりに

本稿では折れ線グラフの全体傾向や局所的特徴を言語化して、言語による検索を可能にするシステムの実現に向けて、クエリとなるグラフを言語化し生成された言語表現に基づいて検索を行うシステムを提案した。今後、被験者実験を通じてこのシステムの有用性を明らかにし、より的確な検索が行えるシステムへの改良につなげたい。

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] Saif Ahmad, Paulo C F de Oliveira, Khurshid Ahmad: Summarization of Multimodal Information, *Proc. 4th International conference on Language Resources and Evaluation*, pp.1049–1052 (2004)

[2] Kobayashi, I.: A Study on Text Generation from Non-verbal Information on 2D Charts, *Proc. Computational Linguistics and Intelligent Text Processing 2nd International Conference*, pp.226–238 (2001)

[3] Kukich, K.: Design of a Knowledge-based Report Generator, *Proc. 21 st Annual Meeting on Association for Computatonal Linguistics*, pp. 145–150 (1983)

[4] 小林 一郎, 渡邉 千明, 奥村 奈穂子: グラフとテキストの協調による知的な情報提示手法: 日経平均株価テキストとグラフの提示を例にして (ヒューマンインタフェース基礎,< 特集 > インタラクション技術の原理と応用), 情報処理学会論文誌, Vol. 48, No. 3, pp.1058–1070 (2007)

[5] 小林瑞希, 小林一郎: 複数の時系列データの関連性発見に基づく言語化の一考察, 情報処理学会第 *74* 回全国大会講演論文集, No.4, pp. 629–630 (2012)

[6] 松下光範, 末吉れいら: 言語表現による時系列データ検索のための基礎検討, 第 *19* 回 *Web* インテリジェンスとインタラクション研究会, pp.31–32 (2011)