

特定時系列データの検索手法の提案

大根千明[†]小林一郎[†][†]お茶の水女子大学理学部情報科学科

1 研究の背景と目的

今日、Web の利用拡大に伴って膨大なデータの中から、いかに必要な情報を検索し、ユーザーに提示するかが重要な課題となっている。現在、実用化されている検索システムは主にユーザがキーワードを入力し、それを含む文書を検索する手法が主流であるが、この手法の問題点は文章で書かれた情報のみしか扱えないことである。そこで近年、画像や音楽などの文書以外の情報を検索する技術の研究も数多く報告されている [1][2]。本論文では、文章以外の情報源の一つとして、グラフとして視覚化される時系列データに着目し、特定の時系列データをその形状から検索できる手法を提案する。また、我々の手法ではグラフの部分的な形状および特徴を認識することにより、検索質問として与えた入力の時系列データに最も類似したデータを検索することを可能にする。本手法の適用対象としては、経済学者や天気予報士などが時系列データで観測された過去事例を検索・利用するなどが考えられる。

2 提案手法

2.1 提案手法の概要

本研究では複数の時系列データの中から、ユーザーの指定した時系列データに最も似ている形状の時系列データを検索し、提示する手法を提案する。図 1 に、提案手法の概要を示す。

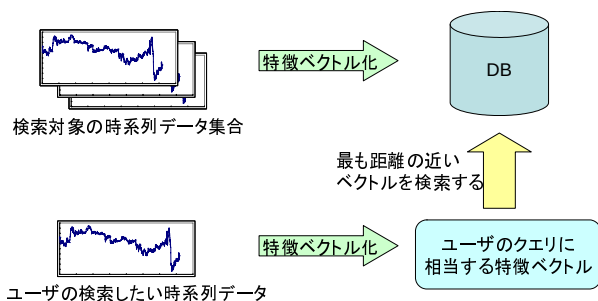


図 1: 提案手法の概要

本手法では時系列データの検索を効率的に行うため、時系列データを特徴ベクトルへと変換し、その特徴ベク

A Method of Retrieving Particular Time Series Data

Chiaki One[†], Ichiro KOBAYASHI[†]

[†]Dept. of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Ootsuka Bunkyo-ku Tokyo 112-8610
 {chiaki_o, koba}@koba.is.ocha.ac.jp

トル同士の距離を 2 つの時系列データの類似度として用いる。検索対象空間となる時系列データ集団をあらかじめ特徴ベクトルに変換した状態でデータベースに登録しておき、ユーザからのクエリを受け付けるたびにそのクエリの時系列データから特徴ベクトルを生成し、データベース内の全ての特徴ベクトルとの距離を計算する。最後に、データベース内の全ての特徴ベクトルの中で、最もユーザのクエリの特徴ベクトルに近いものを検索結果とし、対応する時系列データをユーザへ提示する。

2.2 特徴ベクトルの生成方法

入力となる時系列データは、 x 座標に時刻、 y 座標に値を持つ xy 平面上の標本点の集合 S として与えられるものとする。以後の説明のために、 $x_{max}, x_{min}, y_{max}, y_{min}$ をそれぞれ S に含まれる点の x 座標、 y 座標の最大値・最小値とする。

特徴ベクトル生成手法の概要を図 2 に示す。まず、与えられた標本点の集合 S から最小二乗法近似を用いて、 n 次多項式の近似関数 $f(x)$ を生成する。次に、この $f(x)$ の導関数 $f'(x)$ を求め、方程式 $f'(x) = 0$ を解く事によって、曲線 $y = f(x)$ 上の極大点・極小点の座標を得る。ここで得られる極の個数 M は、 $0 \leq M \leq n - 1$ となる。この点に $(x_{min}, f(x_{min})), (x_{max}, f(x_{max}))$ の 2 点を加えたものを特徴点集合 P と呼ぶ。さらに、 P に含まれる各点について、 $(x_{min}, y_{min}) \rightarrow (0, 0), (x_{max}, y_{max}) \rightarrow (1, 1)$ となるように線形変換を施し、正規化したものを P_N とする。

以上の処理により生成された正規化された特徴点集合 P_N の各要素の x 座標、 y 座標を x 座標の小さい順に並べたものを、特徴ベクトル $\vec{f} = ((x_1, y_1), (x_2, y_2), \dots, (x_{M+2}, y_{M+2}))$ と定義する。

2.3 特徴ベクトル間の距離の計算

前節で説明したように、特徴ベクトルの次元数は $[2, n + 1]$ の範囲にある整数となる。そこで、2 つの特徴ベクトル間の距離 $dist(\vec{f}_a, \vec{f}_b)$ を以下のように定義する。

$$dist(\vec{f}_a, \vec{f}_b) = \begin{cases} \infty & (\vec{f}_a \text{ と } \vec{f}_b \text{ の次元数が違う場合}) \\ \alpha & (\vec{f}_a \text{ と } \vec{f}_b \text{ の次元数が同じ場合}) \end{cases}$$

$$\text{ただし } \alpha = \sum_{i=1}^{M+2} \{(x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2\}$$

*ここでは厳密な解ではなく、微分係数の符号が逆転する点を近似解としてもちいている。

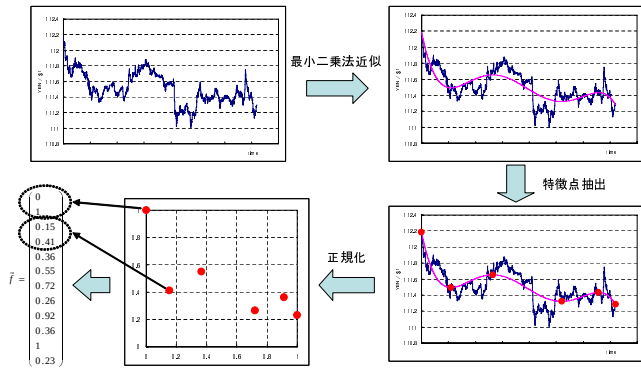


図 2: 特徴ベクトルの生成

これは、図 3 の全ての双方向矢印の長さの合計に相当する。つまり、特徴点同士の位置が近いほど時系列データ間の類似性が高いと判断されることになる。また、特徴点の個数が異なる場合は距離が ∞ となり「似ていない」と判断される。

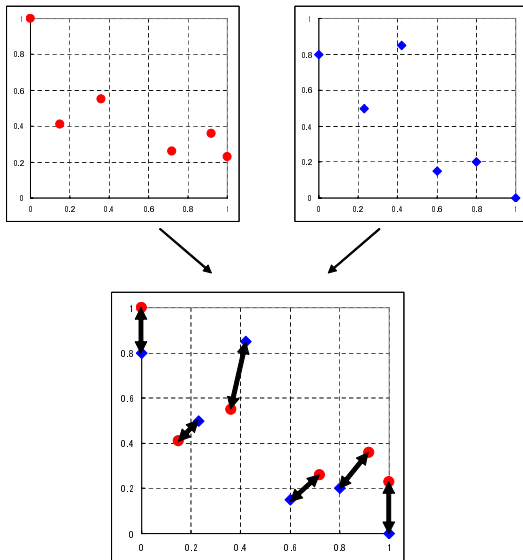
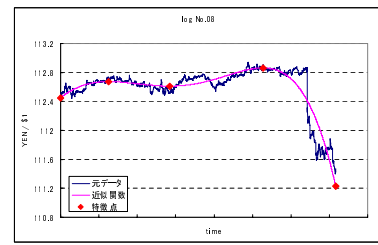


図 3: 特徴ベクトル間の距離

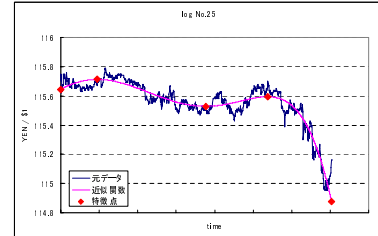
3 実装と評価

我々は提案手法を実装し、米ドル為替相場の時系列データを対象に実験を行った。対象とする時系列データは 1 分おきに Web より取得した米ドル・円の為替相場データを 79 日間分用意したものを用いた。

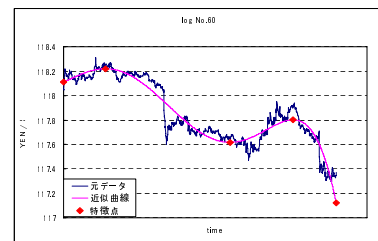
実験は 79 日のデータの中から 1 日分のデータを取り出し、残り 78 日分のデータの中から最も選択した 1 日のデータに近いものを提示する、という手順で行った。本実験での、最小二乗近似の次数 n は、 n の次数を変更してグラフの類似性を評価する予備実験により、5 と設定した。図 4 にその一例を示す。



(a) ユーザのクエリ



(b) 検索結果：最も類似していると判断された時系列データ
 $\alpha = 0.572185$



(c) 検索結果：2 番目に類似していると判断された時系列データ
 $\alpha = 0.762986$

図 4: 実験結果

4 まとめと今後の課題

本研究では、時系列データを用いて特徴ベクトルを抽出することにより、入力したデータに最も類似したデータを検索することを提案した。また、特徴ベクトル間の距離計算だけでは、正規化後の特徴ベクトルの分布によっては、マッチングの精度が下がる可能性がある。そのため、今後、更なる精度の向上には改善の余地がある。また本実験の中では、最も良い結果の近似次数は 5 次であったが、為替以外の時系列データでは何次の近似次数が適しているかについては試行する必要がある。

参考文献

- [1] David Forsyth, Jitendra Malik, Robert Wilensky, “絵の特徴から選びだす画像検索法”, 日経サイエンス, Vol. 27, No. 9, pp. 86-93, Sep. 1997. (DID. 236)
- [2] Masakazu Yagi, Tadashi Shibata, “An Image Representation Algorithm Compatible with Neural-Associative-Processor-Based Hardware Recognition Systems”, IEEE Trans. Neural Networks, Vol. 14, No. 5, pp. 1144-1161, September (2003)
- [3] 奥村菜穂子, 小林一郎, “グラフの挙動を表すテキスト生成”, 第 12 回年次大会ワークショップ「言語処理と情報可視化の接点」, pp17-18, 2006