5J－05

時系列データのXMLフォーマットとその類似検索方法

植田智広　川越 恭二
立命館大学理工学部

1 はじめに

インターネットの普及、情報技術の発展に伴い、画像や動画、音声のインターネットでの利用だけでなく株価や気温など、時間の変化に伴い値が変化する時系列データに対するニーズも増えている。しかし、時系列データを扱うためのフォーマットが存在しない上に類似検索機能も提供されていない状況である。

そこで、本稿では時系列データを扱うためのフォーマット(TIM ファイル)の提案と、そのフォーマットを基本にグラフ的に類似した形状を持つ時系列データを検索する類似検索方法の提案を行う。

2 TIM ファイル

TIM ファイルのフォーマットは XML 形式である。XML 形式を採用することで、1)タグを自由に設定できる、2)文書をデータとして扱うことができる、3)WWW との親和性が高いなどの特徴を利用できる。TIM ファイルは<time>,<from>,<to>,<body>,<t>の5つの要素のみによって記述される。以下の例を用いて各要素についての説明を行う。なお、DTD(Document Type Definition)は省略する。

```
<time name = Kabuka>
<from>19960710</from>
<to>19990910</to>
<body>
    <t>19960710</t>903,
    <t>19960711</t>901,
    <t>19960712</t>904,
    ……
    <t>19990909</t>492,
    <t>19990910</t>477,
</body>
</time>
```

図1　TIM ファイルのフォーマットの例

XML Format to descrive Time series data set and its Similarity Searching
Tomohiro UEDA ,Kyoji KAWAGOE.
Ritsumeikan University
1-1-1,Noji-Higashi,Kusatsu,Shiga,525-8577,Japan

図1において、<time>内の name 属性は時系列データの種類を示す。すなわち、株価、気温、会社の売り上げなどが name 属性値として用いられる。次に、from 要素ではファイル内の時系列データの開始年月日を示す。西暦4桁、月2桁、日2桁を連続して並べた8桁の数字として扱う。図1の例では最初のデータが 1996 年 7 月 1 日の株価であることを表している。一方、to 要素ではファイル内の時系列データの終了年月日を示す。from 要素と同様、8桁の数字として扱う。最後に、body 要素は実際の時系列データを示す。時系列データを日付の早いものから順番に","で区切って記述する。なお、body 要素内の t 要素はそのデータの日付(時刻)を示すがこの要素は省略可能である。

3 時系列データの類似検索方法

3.1 類似検索方法の特徴

現在、時系列データの類似検索方法としては離散フーリエ変換や離散コサイン変換、ユークリッド距離を用いた検索などの方法が提案されている。ユークリッド距離を用いた検索は時系列データの各要素を順番に比較していく方法で、直感的には最もわかりやすく、検索の精度も高い。しかし、すべてのデータの比較が必要となるため計算量が多くなり、データ数が多い時系列データの場合は現実的な計算時間では結果を求めることができない。一方、離散フーリエ変換や離散コサイン変換を用いた検索方法では時系列データを別の特徴空間に変換を行うことで時系列データの特徴を近似的に求め、その特徴を用いて検索を行う。このため検索時の計算量は少なくなるが、特徴を変換により近似するため検索の精度が低くなるという問題点がある。

そこで、離散フーリエ変換とユークリッド距離、ウェ

ーブレット変換の3つ方法を組み合わせる類似検索方法を提案する。まず、離散フーリエ変換による検索を行い、その結果に対してウェーブレット変換による検索を更に行うことで、検索の精度の向上をはかる。そして、その結果に対して最後に、ユークリッド距離を求めていくことで結果の順位付けを行い、類似度の高いものを求める。

本方法は、従来の方法と比べて、ウェーブレット変換を用いて検索結果に含まれるノイズ成分を除去することを狙っている。ウェーブレット変換は一般的に画像検索で用いられる検索方法である。この変換を時系列データに適用できるように、もとの時系列データを近似関数と差分関数を用いて表現し、その2つの関数から特徴を抽出する方法を提案する。

3.2 離散フーリエ変換

離散フーリエ変換は

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi kn}$$

で表され、N は時系列データファイルの要素数を、$x(n)$は時系列データファイル内の n 番目の要素の値を、k は離散フーリエ変換の次数を X(k)は離散フーリエ変換後の値を表す。本方式では、離散フーリエ変換の次数は3次まで求め、各次数の実数項、虚数項ごとの2乗和を比較して検索を行う。

3.3 ウェーブレット変換

ウェーブレット変換では基底関数が必要となるが、本稿では Haar 基底を使用している。時系列データを近似関数と差分関数を用いて表現し、その係数である近似係数 F・差分係数 G は

$$F_k=[\frac{F_{k-1}(0)+F_{k-1}(1)}{2},\ldots,\frac{F_{k-1}(N_{k-1}-1)+F_{k-1}(N_{k-1})}{2}]$$

$$G_k=[\frac{F_{k-1}(0)-F_{k-1}(1)}{2},\ldots,\frac{F_{k-1}(N_{k-1}-1)-F_{k-1}(N_{k-1})}{2}]$$

で表され、k は次数、Nk は次数 k における近似係数の N 番目のデータを表す。ここで、k=0 の場合の近似係数は検索を行う時系列データの集合を表している。本方法では、ウェーブレット変換の次数は4次まで求め、各次数で近似係数、差分係数ごとの2乗和を比較して検索を行う。

3.4 ユークリッド距離

ユークリッド距離では2つの任意の時系列データ A={a1,a2,…,an}、B={b1,b2,…bn}を考えたとき、A・B の類似度 R は

$$R=\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2}$$

で表す。R の値が小さいものほど類似度が高いことを表している。

4 実験

ウェーブレット変換を併用する本方法を用いることによって、離散フーリエ変換による検索の結果のうち、形状が異なる時系列データをどれだけ取り除けるかを測定した。時系列データファイルとして約3年分の株価データ 370 件を用いた。離散フーリエ変換とユークリッド距離を組み合わせた検索方法と離散フーリエ変換とウェーブレット変換とユークリッド距離を組み合わせた本方法の2つの方法に対して、表1に示す結果を得た。ここで、適合率とは検索対象のデータと類似していると考えられるデータ(類似データ)が検索結果に含まれる割合、誤りデータ率とは類似データとは明らかに異なっているデータが検索結果に含まれる割合である。各々、10 回の検索を行った結果の平均を示している。表1に示すように従来の方法に比べてノイズを除去できることが得られた。

	離散フーリエ変換	本方法
適合率	67%	76%
誤りデータ率	19%	8%

表1 実験結果

5 終わりに

今後は TIM ファイルフォーマットの改良、本方法の検索精度の向上、株価データ以外の時系列データへの適用を予定している。

参考文献

[1]C. Faloutsos:"Multimedia Databases by Content", Proa. of DMIB.1997
[2]小早川他:"ウェーブレット変換を用いた対話的類似画像検索と民族資料データベース",情報処理学会論文誌.1999