人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第1回)
SIG-AM-01-03

複数の時系列データの比較に基づく グラフの言語表現生成手法

Generating Linguistic Expression of Charts Based on Comparison of Multilpe Time-Series Data

末吉 れいら [1]　松下 光範 [1*]　白水 菜々重 [2]
Reira Sueyoshi[1]　Mitsunori Matsushita[1]　Nanae Shirozu[2]

[1] 関西大学
[1] Kansai University
[2] 奈良先端科学技術大学院大学
[2] Nara Institute of Science and Technology

Abstract: This paper proposes a method for generating linguistic expressions from a time-series data. The proposed method takes differences and similarities among multiple time-series data into consideration: The method generates linguistic expressions by executes three processes sequentially. First, a characteristic such as “rise,” “drop,” and “stable” is evaluated in each data point of the data series. Second, for each data point in a data series, a weight is assigned by calculating a degree of attention, which is estimated by comparison with another time-series data. Finally, the most pertinent expression is selected.

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「ここ一週間で急落した株は?」や「価格が緩やかに上昇している商品は?」といった言語による検索要求を通じてユーザの意図や関心に合致した区間・粒度のデータを得ることは困難である。このような検索要求から、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ことができれば、ユーザの時系列データに対するアクセス性の向上が期待できる。本研究のゴールはこのような情報アクセスを可能にする技術を実現することであり、現在そのひとつのアプローチとして、時系列情報を予め自然言語表現で記述しておき、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、視覚化する手法の実現を目指している [1, 2]。

我々は、このような情報アクセス技術の実現に必要な要素技術として、(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 [1]。本稿ではこのうちの（1）に焦点をあて、時系列データの持つ解釈の多様性を考慮した言語表現の生成について検討する。

(1) に関して最も効率的・効果的な方法は、時系列データとそれを説明したテキスト（新聞記事など）を対応づけて、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であるが、このようなテキストが常に得られる保証はない。そのため、時系列データのみが与えられた状態でも、そこからその時系列データを適切に表現する言語表現を生成する技術が必要になる。

ここで注意すべきは、同じ振る舞いのデータであっても状況や文脈によって解釈が異なる場合があるという点である。例えば、ある企業の株価が変動した場合、同じ値幅の下落であっても特定の銘柄だけ下落していればその下落に注目がいくが、多くの銘柄が下落していればあまり注目に値せず、他の特徴に注目するだろう。すなわち、人は探索の文脈や状況に応じて時系列データの注目点を変えることで、適応的なデータの解釈を行なっていると言える [3]。

本研究ではこのような、ユーザが行う複数の時系列データの比較行為に着目し、それをモデル化することで、より人の直感に沿った時系列データの探索・アクセスを可能にすることを試みる。この方針の下、本稿では、状況や文脈によって変化する解釈を取り扱うた

*連絡先: 松下 光範 関西大学総合情報学部 〒569–1095 大阪府高槻市霊仙寺町 2–1–1 Tel: (072) 690–2437 Fax: (072) 690–2491 e-mail: mat@res.kutc.kansai-u.ac.jp

め、時系列データに最大値や最小値、上昇・下降・安定といった特徴を付与し、特徴の重要度に応じて重み付けを行う手法を提案する。この手法では、異なる種類のデータ、あるいは、同じ種類のデータの異なる期間のデータといった異なる複数の時系列データを比較する場合に、それぞれの特徴の差異に応じて動的に重みを変更する。

2 提案手法

前節で述べたように、複数の時系列データを比較して分析することで複数の事象に跨った包括的な知見を獲得し、より深いデータの理解が可能になると期待される。しかしこの場合、比較対象に応じて時系列データの持つ値の「意味」が相対的に扱われるため、文脈による解釈の変化が生じる。

そこで、本研究では状況や文脈によって変化する解釈を取り扱うため、値の上昇や下降、安定といった時系列データの変化傾向を特徴として捉え、特徴の重要度を算出して各言語ラベルに重み付けを行う。付与された重みは、ユーザの要求に合った複数の時系列データを比較する場合に用いる。特徴の類似性や特異性とユーザの要求への合致度を加味して動的に重みを変更することで、時系列データの相対的な評価を考慮した言語表現を生成する。

図 1 に提案手法の概要を示す。この手法では、予め時系列データに対する特徴の付与、付与した特徴に対する重み付けを行う。続いて、ユーザの要求によって絞り込まれた時系列データを対象に複数の時系列データを比較することで動的な重みの変更を行い、ユーザへ視覚的に提示する。

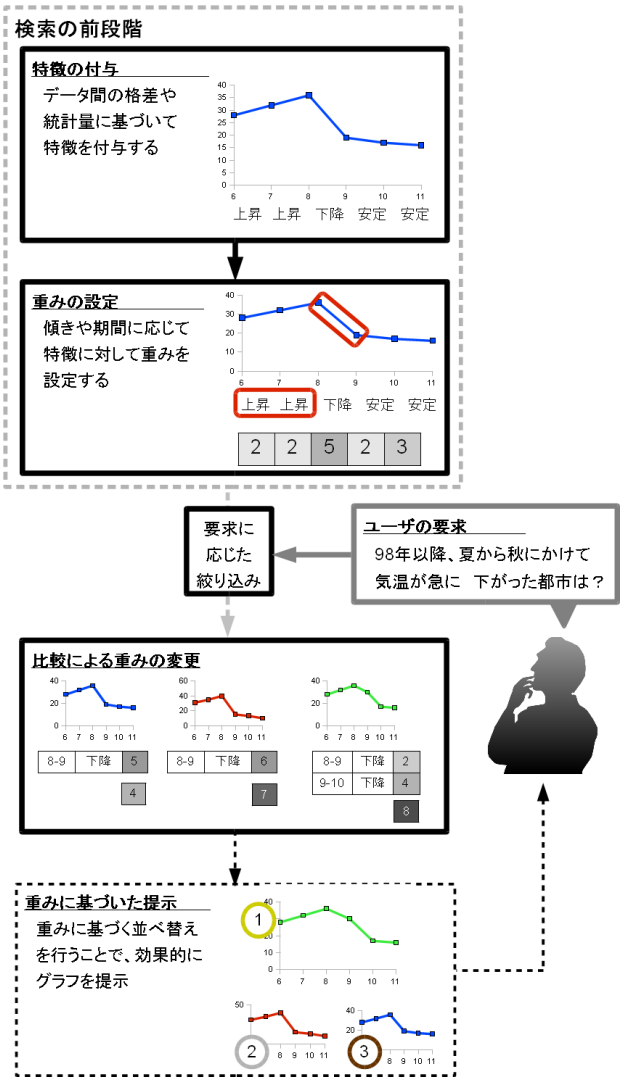


図 1: 提案手法の概要

2.1 時系列データに対する特徴の付与

時系列データに対する重み付けに際し、時系列データの数値情報から重み付けの指標や基準となる上昇や下降などの特徴を言語表現として付与する特徴や特徴の算出方法について述べる。時系列データに対して特徴の付与を行うため、まず時系列データから得られる特徴を整理した。本研究では、時系列データから得られる特徴を基礎統計量、数値間の関係性、異なる統計量比較のための指標といった 3 つの観点に着目し、整理を行った。

基礎統計量については、最大値、最小値、平均値、標本数、データ範囲、標準偏差を対象とした。

時系列データの数値間に生じる関係性については、傾斜、傾斜の寄与度、傾斜傾向、傾斜傾向の持続期間を特徴とし、傾斜の度合いと向きから得られる傾斜傾向を主な特徴とした。このうち、傾斜傾向 (gap tendency) に関しては、時系列データのとる範囲における傾斜の割合である寄与度 (contribution degree)[4] を基に算出した。時系列データ X の時点 $t \in T$ (T は時点の全体集合) における要素を $x_t \in X$ とすると、傾斜に対する x_t の寄与度 $cd(x_t)$ は式 (1) で求められる。

$$cd(x_t) = \frac{x_{t+1} - x_t}{\max(X) - \min(X)} \tag{1}$$

ここで、 $\max(X)$ は X の要素の最大値、$\min(X)$ は X の要素の最小値を各々示している。この $cd(x_t)$ に基づき、時系列データ X の時点 t における傾斜傾向 (現在の実装では「上昇」「下降」「安定」の 3 つ) を付与する。判定の基準は、閾値パラメータを $\tau(> 0)$ とすると、$|cd(x_t)| < \tau$ の場合に「安定」、$cd(x_t) \geq \tau$ の場合に「上昇」$cd(x_t) \leq -\tau$ の場合に「下降」とした。なお、現在の実装では $\tau = 0.05$ としている。

2.2 特徴に対する重み付け

次に提案手法では、2.1 節で求めた特徴を基に重み付けを行う。重み付けを行うにあたり、傾斜傾向と期間を重み付けの対象とし、傾斜や傾斜の寄与度といった特徴は対象に対するパラメータとして扱う。パラメータを元に算出されたそれぞれの重みを元に、対象とする時系列データへの重みを決定する。本研究における重み付けの段階では、比較による動的な重みの評価はユーザの要求を得た際に行うため、言語表現に対する定量的な評価として直接評価法に基づく重み付けを行う。

重み付けの手法については、標本数や傾斜の大小を用いた傾斜傾向に関する重み付けと、傾斜傾向の持続に関する期間の長短を用いた重み付けのふたつを検討している。本稿ではこのうち、前者について述べる。

傾斜や傾斜傾向を用いた重み付けの手法に関しては、以下で述べる 3 手法を検討することとした。

2.2.1 手法 1

この手法は全体的特徴を重視する手法である。すなわち、時系列データ全体において、上昇・下降・安定の各傾斜傾向ごとに標本数を算出し、傾斜傾向の標本数が多い場合に重くなるように重み付けの係数を設定する。

また、上昇・下降に関しては傾斜が大きい場合に重みを増やし、安定に関しては傾斜が 0 に近いほど重みが増加するという基準を設けた。

単一の時系列データにおいて、特徴として付与された傾斜傾向と傾斜を用いて重み付けを行った。重み付けには上昇・下降・安定の各傾斜傾向ごとに標本数に応じて設定した係数を算出し、各傾斜傾向に定められた基準に応じたポイントの付与を行い、それらを併用して重み付けを行う。具体的な方法は以下のとおりである。

係数 (coefficient) は、時系列データの各傾斜傾向の標本数に基づき、標本数の多いものについて値が大きくなるよう設定した。ある傾斜傾向 $gt_i \in GT, GT = \{$ 上昇, 下降, 安定 $\}$ に属する標本数を $num(gt_i)$ とすると、上昇・下降・安定の各傾斜傾向 gt_i に対する係数 $coe(gt_i)$ は、式 (2) によって算出される。

$$coe(gt_i) = \frac{num(gt_i)}{N} \tag{2}$$

ただし $N = \sum_{j=1}^{3} num(gt_j)$ である。 ポイントの設定では、まず、上昇・下降では傾斜が大きいもの、安定では傾斜が 0 に近いもの、という基準に応じて各傾斜傾向内で順位付けを行い、順位による得点を割り当てた。例えば、傾斜傾向「上昇」の標本数が 6 の場合、傾斜の大きなものから順に、1 位には 6 点、2 位には 5 点、3 位には 4 点といったように、その傾斜傾向の標本数に応じた得点を割り当てた。次に、各得点を傾斜傾向内における全ての得点を加算したものにより除算を行い、ポイントとした。時点 x_t の傾斜を $gap_t = x_{t+1} - x_t$ とし、x_t の属する傾斜傾向 gt_i 内の順序を考慮して与えられる得点を $gtp_i(x_t)$ とした場合、与えられるポイント $Point(x_t)$ は式 (3) によって求めた。

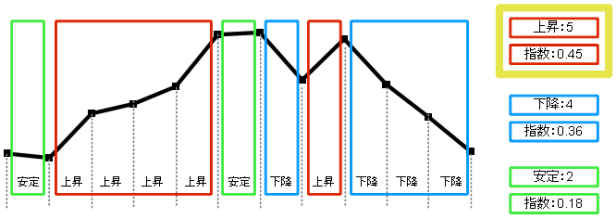


図 2: 手法 1 による重み付け

$$Point(x_t) = \frac{100}{\sum_{j=1}^{num(gt_i)} gtp_i(x_j)} \times gtp_i(x_t) \tag{3}$$

例えば、傾斜傾向の標本数が 6 の場合、傾斜傾向内における全ての得点の合計は 21 となり、傾斜傾向に応じたポイントは順位が 1 位の場合 28.57、順位が 2 位の場合 23.80 となる。

上述したように傾斜傾向ごとに算出した係数とそれぞれの傾斜傾向に与えられたポイントを積算して重み付けを行った。

図 2 に手法 1 による重み付けの例を示す。

2.2.2 手法 2

この手法は大きな変化がある局所的傾向を重視する手法である。全体的な流れや算出方法に関しては手法 1 と同様であるが、重要視する観点として上昇・下降・安定の各傾斜傾向ごとの標本数が少ない場合に重みを増やす。

この手法では、手法 1 と同様に、単一の時系列データにおいて、特徴として付与された傾斜傾向と傾斜を用いて重み付けを行う。重み付けには上昇・下降・安定の各傾斜傾向ごとに標本数に応じて設定した係数を算出し、各傾斜傾向に定められた基準に応じた得点の付与を行い、それらを併用して重み付けを行う。係数は、時系列データの各傾斜傾向の標本数に基づき、標本数の少ないものついて値が大きくなるよう設定する。したがって、傾斜傾向 gt_i に対する係数 $coe(gt_i)$ は、式 (4) のようになる。

$$coe(gt_i) = \frac{N - num(gt_i)}{N} \tag{4}$$

重みに関しては、手法 1 と同じく式 (3) によってポイントを算出し求めた。

図 3 に手法 2 による重み付けの例を示す。

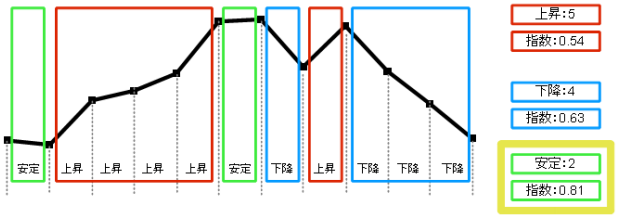


図 3: 手法 2 による重み付け

2.2.3 手法 3

この手法では、算出された乖離を傾斜傾向の重みとして用いる。すなわち、時系列データの全体的な傾向であるトレンドと個々のデータの持つ傾向の乖離が著しい箇所について顕著な特徴とし重みを設定する。

トレンドに関しては、開始点と終了点を線で結んだ 1 本の直線で求める手法や開始点・終了点・最大値・最小値の 4 点を用いて 3 本の直線で求める手法、2 次式の当てはめによる手法、単純移動平均を用いた不規則変動 (ノイズ) の除去によって求める手法などがある。本研究では、トレンドとして単純移動平均を採用し、時系列データの数値から単純移動平均を用いた乖離を算出する。

時系列データの全体的な傾向であるトレンドと個々のデータの持つ傾向の乖離が著しい箇所について顕著な特徴とし重みを設定する。

この手法では、まずトレンドの算出のために単純移動平均によってデータの平滑化を行う。平滑化には、時系列データの非系統的な誤差部分を互いに除去するために局所的に平均を取る単純移動平均法を用いる。単純移動平均法では、各時点のデータをその周辺の n 個のデータの平均によって置き換えることで平均を取る。この幅 n を「ウィンドウ幅」と呼ぶ [4]。

本研究では、対象データからグラフの概形を求める際に単純移動平均法で必要とされるウィンドウ幅を設定するため、対象データのうち、サンプル数が 12 のものと 30 のものに関しては時系列データとサンプル数を 3、5、10 で除算した値のウィンドウ幅で単純移動平均をとった場合のグラフをそれぞれ用意し、比較・検討を行った。ただし、サンプル数が 5 以下のものに関しては、3、5、10 の除算では優位性のある結果が得られなかったため、1、2、3、4 のウィンドウ幅で単純移動平均をとった場合について検証を行った。

この検証の結果、サンプル数が 12、30 のものに関してはウィンドウ幅を データ数/5 に設定することで効果的な値が得られることが判った。しかし、日経平均株価のデータでは、データ数が少なかったために全ての検証において効果的な値は得られなかった。そのため、データ数が少ない場合におけるトレンドの算出やサンプルの取り方に関して考慮する必要性がある。本研究では、上記の結果を基に単純移動平均によるトレンド算出のウィンドウ幅 n について 5 と設定した。

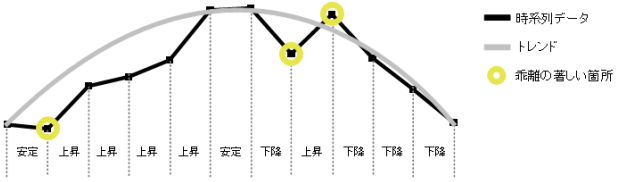


図 4: 手法 3 による重み付け

この単純移動平均により得られたトレンドを用いて、トレンドにおける傾斜を算出した。そして、トレンドの傾斜を時系列データから得られた傾斜によって除算し、傾斜傾向の乖離を求めた。

図 4 に手法 3 による重み付けの例を示す。

3 対象データ

本研究では、複数の時系列データの比較による相対的な評価を得るため、時系列データの中でも、年 (サンプル数: 12) 、月 (サンプル数: 28〜31) 、週 (サンプル数: 2〜5) の区切りで構成された 3 種類の異なるデータを対象とし、検証を行った。

年単位で構成されたデータとして、気象庁の気象統計情報より大阪・札幌の 2010 年から過去 15 年分の月別累計降水量のデータ (サンプル数: 12 件/年) を用いた。

月単位で構成されたデータとして、気象庁の気象統計情報より大阪の 2010 年から過去 3 年分の日別平均気温のデータ (サンプル数: 28〜31 件/月) を用いた。

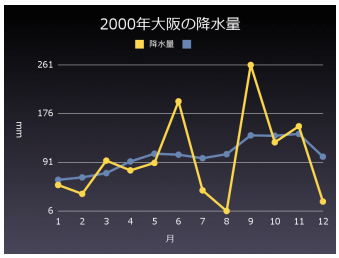
週単位で構成されたデータとして、Yahoo!ファイナンスより 2011 年の 4 月頭から 9 月末までの週別日経平均株価の始値と終値のデータ (サンプル数: 2〜5 件/週) を用いた。

検証では、上記のデータから特徴表現とグラフの生成に必要な形式に変換したテキストファイルを人手で用意した。

4 検証

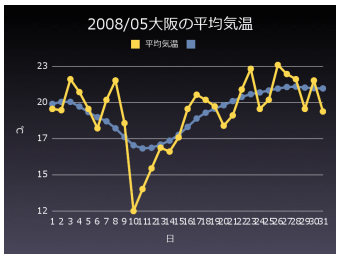
提案した重み付けの手法によって得られた重みについて、外れ値や変化点、グラフ特徴などの観点の下、評価を行った。図 5 から図 7 に各時系列データによって得られた重みを示す。

その結果、上昇・下降の傾斜傾向に対する重み付けに関して、(1) の傾斜傾向の標本数が多数で傾斜が特徴的な場合の手法では、傾斜傾向の標本数が最多であっても標本数に大きな差がない場合、傾斜の大きい傾向が優先された。しかし、傾斜傾向の標本数に大きな差



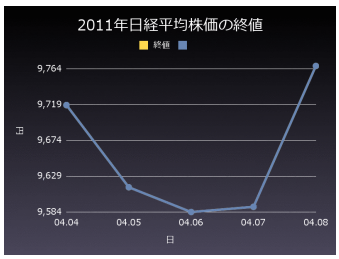
種類	内容							
文字情報	2000 年 大阪の月別累計降水量 (1 月〜8 月抜粋)							
数値情報	52.0	36.5	94.5	77.5	90.5	198.0	42.5	6.5
傾斜傾向	下降	上昇	下降	安定	上昇	下降	下降	
手法 (1)	7.27	10.91	14.55	3.64	15.58	12.99	2.60	
手法 (2)	2.27	19.09	4.55	6.36	6.06	10.82	1.01	
手法 (3)	19.75	50.58	37.33	0.61	109.22	149.30	43.06	

図 5: 年単位のデータ　サンプル数が 12 の場合 (1 月〜8 月抜粋)



種類	内容							
文字情報	2008 年 5 月 大阪の平均気温 (1 日〜8 日抜粋)							
数値情報	20.2	20.1	22.5	21.5	20.2	18.7	20.9	22.4
傾斜傾向	安定	上昇	下降	下降	下降	上昇	上昇	
手法 (1)	5.83	4.58	2.08	3.75	2.50	2.50	1.25	
手法 (2)	2.03	4.58	0.73	3.75	0.87	2.50	0.44	
手法 (3)	0.26	2.40	0.64	0.87	1.10	2.51	2.04	

図 6: 月単位のデータ　サンプル数が 30 の場合 (1 日〜8 日抜粋)



種類	内容				
文字情報	2011 年 4/4 から 4/8 までの日経平均株価 (終値)				
数値情報	9718.89	9615.55	9584.37	9590.93	9768.08
傾斜傾向	下降	下降	安定	上昇	
手法 (1)	25.0	16.67	33.33	25.00	
手法 (2)	8.33	7.14	33.33	8.33	
手法 (3)	0.00	0.00	0.00	0.00	

図 7: 週単位のデータ　サンプル数が 5 の場合

が見られた場合、傾斜の大きさに関わらず重要視されるため、傾斜が小さくても標本数が多いものが重要視された。(2) の傾斜傾向の標本数が少数で傾斜が特徴的な場合の手法では、傾斜傾向の標本数が最少となる場合、今回の対象データでは安定の傾斜傾向が最少となる場合が 9 割を占めたが、(3) の手法と比較した場合に重くすることが可能となった。この手法では、標本数の多いものが極端に軽視されるなどの問題も生じたため、係数の算出手法について検討する必要性がある。

(3) の全体的な傾向 (トレンド) との乖離が著しい場合の手法では、時系列データの傾斜とトレンドの傾斜との差をそのまま重みに用いているため、グラフ特徴と外れ値という点では最も基準に合致していた。しかし、 (3) の単純移動平均を用いたトレンドの算出に関して、月別降水量と平均気温のデータについては平均を取る幅を データ数/5 にすることで効果的な値が得られたが、日経平均株価のデータでは、サンプル数が少なかったため効果的な値は得られなかったという問題点も見られた。

以上のことから、上昇・下降に対する重み付けに関しては本研究における観点から見た場合、 (3) の手法が最も適していると判断した。また、安定に対する重み付けに関しては、 (1) 傾斜傾向の標本数が多数で傾斜が特徴的な場合と (2) 傾斜傾向の標本数が少数で傾斜が特徴的な場合について検証を行ったが、今回対象としたデータを用いた場合、安定の標本数が上昇や下降の標本数に比べて圧倒的に少なく、手法の違いによる有意差は見られなかった。このことから、上昇・下降に関しては (3) の手法による重み付けは有益であるが、サンプルの取り方と安定に対する基準の設定を考慮する必要性が明らかになった。さらに、安定に関する基準を設定していなかったため、安定に対する重み付けが低くなってしまうことも問題として挙げられる。また、単一の手法による重み付けではなく、複数の手法を併用することでより観点に沿った重み付けが可能だと考えられる。以上のような検証の結果に基づき、検索の前段階にあたる特徴の付与と重み付けを行うプログラムについて実装を行った。実行結果の例を図 8 に示す。

本システムの動作環境について、OS は Microsoft

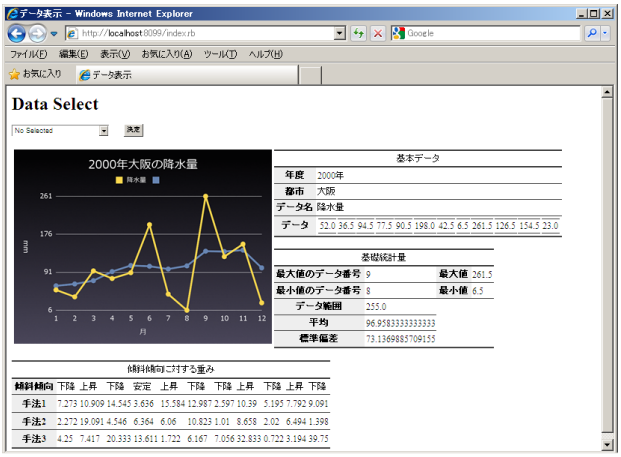


図 8: 実行結果の表示例

Windows XP Home Edition を用い、Web ブラウザは Internet Explorer 8 を用いた。プログラミング言語は Ruby1.8.7 を用いた。プロトタイプシステムの実装に際して対象としたデータは、3 章で提示したデータを用いた。

5 考察

本研究では単一の時系列データから得られる特徴を基に、その特徴に対して重み付けを行い、ユーザの要求に応じた複数の時系列データを比較した場合に要求との合致度を考慮し重みの変更を行うことで、相対的な重み付けを行う手法について提案した。

そのうち、提案手法で述べた 3 つの観点に基づき単一の時系列データから特徴を算出し、算出した特徴に対して傾きに関する重み付けの段階を対象として検証を行った。今後、提案手法で述べた期間に関する重み付けを行った場合、それぞれの重み付け手法で得られた値の取り扱いについて、単一で取り扱い 2 種の重みとすべきか、複合して 1 種の重みとすべきか検討を行う必要性がある。それぞれ独立の重みとして取り扱う場合、ユーザの要求が期間に関するものであれば期間の重みに焦点をあてることで、容易に特徴を捉えることが可能になり、要求に応じた判断が可能になると考えられる。しかし、ユーザの要求が傾斜と期間を複合したものであった場合には、それぞれ独立で重み付けられた傾斜と期間に関連性を持たせる必要がある。上記の考察を踏まえた上で、今後期間に関する重み付けの検証を行った後、検討を行う必要性がある。

今後、現段階で実装を行った傾斜傾向に対する重み付けに際する問題に関して、検証の結果でも述べたように取得するサンプル数に対する制約や重み付けの手法について検討を行い、更なる検証を行う必要性がある。加えて、重み付けの対象に期間が追加された場合、提示する情報について考慮する必要がある。また、重み付けの観点に対する整理と重み付けや提示する情報に関する被験者実験を実施し、人間の認識により合致した重み付けの値や効果的な情報提示について知見を得る必要があると考えている。

6 おわりに

本研究では、複数の時系列データの比較により変化する解釈を取り扱うため、時系列データに対する動的な重み付け手法の枠組みを提案した。提案した手法における 3 段階のうち、検索の前段階として必要である時系列データに対する特徴の付与と特徴に対する重み付けに関して検証を行い、検証により得た知見を踏まえ実装を行った。今後、本研究における検証や実装により明らかとなった問題点について、更なる検証と重み付け手法の再考や提示手法の改善を行う必要性がある。また、複数の時系列データを比較し重みを変化させることで、ユーザの要求に応じたグラフ解釈の変化に対応した方式への拡張について検討する。

7 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] 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).

[2] 末吉れいら, 田中和広, 白水菜々重, 松下光範: 比較対象に着目したグラフの言語表現の生成, 第 21 回 Web インテリジェンスとインタラクション研究会, pp. 37–38 (2011).

[3] 小泉尚之, 松下光範, 松田昌史, 馬野元秀: 言語情報と統計グラフの相互変換に関する基礎検討, 人工知能学会全国大会, 2H5-6 (2007).

[4] 熊原啓作, 渡辺美智子: 身近な統計, 放送大学教育振興会 (2007).