2G2-01

複合時系列データに基づいた評価対象のランキング

Ranking of Evaluation Targets based on Complex Sequential Data

櫻井 茂明*1 Shigeaki Sakurai
牧野 恭子*1 Kyoko Makino
鈴木 裕之*1 Hiroyuki Suzuki
正岡 良規*1 Yoshinori Masaoka

*1 東芝ソリューション(株)
Toshiba Solutions Corporation

This paper proposes a method that ranks evaluation targets from complex sequential data related to them. The data is composed of numerical sequential data and textual sequential data. The method notes change ratio of the numerical one and occupation ratio of the textual one. It calculates attractiveness of the evaluation targets based on these ratios. The attractiveness represents how degree the evaluation targets are attractive. This paper applies the method to the complex sequential data in the economical field. This application regards each evaluation as a company, the numerical data as stock information, and the textual data as news headlines. It verifies the effect of the method by numerical experiments based on the data collected from a stock information collection site and news distribution sites.

1. はじめに

各種センサーの小型化及び低価格化に伴って、各種センサーを簡便に実世界に埋め込むことが可能となりつつある。これらセンサーによって構成されたネットワークを介して、大量の時系列データが収集されると予想されており、これら時系列データを分析することにより、我々の生活がより豊かになることが期待されている。

このような背景の下、本論文では、複合イベント処理(Complex Event Processing)技術に着目し、本技術の活用形態のひとつとして、特定の評価対象に関連した数値時系列データと、特定の評価対象に関連したテキスト時系列データといった異なる形式の時系列データを扱う方法を提案する。また、提案法の実世界における応用を指向して、数値時系列データとしての株価情報、テキスト時系列データとしてのニュースヘッドラインに着目した分析を実施し、その評価結果に基づいて、提案法の効果を検証する。

2. 時系列データの分析法

2.1 分析方針

本論文で扱う時系列データは、永続的に与えられる時系列データである。このため、データ収集開始時点からの、すべてのデータを保存しておくことはできない。また、リアルタイム性が要求されるため、データ間の多様な組合せを考慮した分析を実施することは困難である。そこで、比較的処理負荷の低い、データ集計に基づいた分析法を検討する。

一方、多様な情報源から情報が収集されることを鑑みて、複数の異種の時系列データを扱うことにする。そのような異種の時系列データの組合せには多様なものが考えられ、その組合せに応じた手法が必要になる。本論文では、従来の我々の研究との親和性が高い、数値時系列データとテキスト時系列データを対象とした、分析方法を検討することにする。

ここで、数値時系列データとテキスト時系列データの特徴に着目してみると、数値時系列データはテキスト時系列データに比べて、データ量が少ない場合が多く、データの処理時間も通常短いものとなる。このため、数値時系列データは、テキスト時系列データに比べて、より頻繁にデータを収集することができる。また、数値時系列データの場合、その数値時系列データを構成する個々の数値は比較的変動が大きい。これに対して、テキスト時系列データの場合、個々のテキストの中から特定の情報を抽出する必要があるものの、抽出される内容全体としては、その変化は大きくなく、短期的には大きな変動がないものと考えられる。そこで、数値時系列データに関しては、前の数値と現在の数値との変化に着目し、テキスト時系列データに関しては、抽出された内容の総量に対する特定の内容の割合に着目することにする。

2.2 分析法

2.1 節に検討した分析方針に基づいて、本節では数値時系列データ及びテキスト時系列データを対象としたデータ分析法を提案する。提案する方法では、評価対象ごとに与えられた数値時系列データと、評価対象に関する話題が記載されたテキストの集合で構成されたテキスト時系列データが与えられるとする。また、このような時系列データから、多数の評価対象の中から注目すべき評価対象を抽出することを考えることにする。

すなわち、各評価対象 t に対して、数値時系列データとして、$v^t(=\{v^t_1, v^t_2, \cdots, v^t_n\})$が与えられているとする。ここで、1,2,・・・,n はデータを収集する単位時間とし、その値が小さい程、時系列的には古い時間であるとする。また、n はデータを格納する期間とする。時系列データは、単位時間が 1 進むごとに、切れ目なく収集されることになるため、新たなデータを収集した場合に、収集済みの各データの単位時間 1 のデータが削除されるとともに、他の単位時間の値が 1 だけ小さくなる。新規に収集されたデータは、単位時間 n の値として格納されることになる。

一方、評価対象に関する内容が記述されているテキスト時系列データとして、$s(=\{s_1, s_2, \cdots, s_n\}, a_j \in s_i)$が与えられているとする。ここで、$a_j$は i 番目のテキスト集合 s_iに含まれる j 番目のテキストを表すとする。

このようなテキスト時系列データを分析していくには、テキスト時系列データから何らかの数値を抽出することが必要となる。本論文では、各テキストに評価対象に関する記述がなされている

連絡先:櫻井茂明, 東芝ソリューション(株), 〒183-8512 東京都府中市片町 3-22, Tel: 042-340-6364, Fax: 042-340-6013, E-mail: Sakurai.Shigeaki@toshiba-sol.co.jp

かどうかを判定し、評価対象に関する記述を含んでいるテキストの件数をテキスト集合ごとに積算することによって、数値化を試みることにする。ただし、特定の評価対象に関する記述が記載されているかどうかを、1、0 によって判定するのではなく、その程度を[0,1]の実数値として与えることも可能であることに注意する必要がある。しかしながら、ここでは、テキスト処理の簡便化を優先させて、1、0 によって判定することにする。従って、テキスト時系列データから、各評価対象 t に応じた数値時系列データ $u^t(=\{u^t_1, u^t_2, \cdots, u^t_n\})$を得ることができる。

以上により、評価対象ごとの v^t 及び u^t に基づいて、評価対象ごとの評価値 q^t を算出し、この評価値に従って注目すべき評価対象を決定する。ここで、前節の分析方針に沿って考えてみると、v^t は前の値との変化に着目し、u^t は抽出内容の総量に対する割合に着目している。そこで、v^t 及び u^t から、式(1)、式(2)に定義される新たな時系列データを生成することができる。

$$r^t = \{r^t_1, r^t_2, \cdots, r^t_{n-1}\}, r^t_i = \frac{|v^t_{i-1} - v^t_i|}{v^t_{i-1}} \quad (1)$$

$$w^t = \{w^t_1, w^t_2, \cdots, w^t_n\}, w^t_i = \frac{u^t_i}{\sum_x u^x_i} \quad (2)$$

式(1)の時系列データの場合、前の値に対する変化率に着目しているため、構成要素の数が当初の n から n −1 個に減っていることに注意する必要がある。また、テキストの内容をより詳細に解析することにより、評価対象に対して正方向のインパクトがあるか負方向のインパクトがあるかを、評価できる可能性はあると考えられる。しかしながら、このような評価は、比較的処理負荷の高いものとなるため、テキスト時系列データから生成される u^t に関しては、方向性を加味した評価は行っていない。このため、u^t から生成される w^t に対しても方向性を加味することはできない。一方、数値時系列データから生成される r^t に関しては、比較的低い処理負荷で方向性を加味することができるものの、数値時系列データだけに方向性を加味したとしても、数値時系列データとテキスト時系列データを組み合わせたものにおいては、方向性を加味することはできない。このため、式(1)の時系列データにおいては、絶対値を基準とした評価を行うことにする。

そこで、w^t 及び r^t に基づいて、各評価対象を評価するモデル化を実施する。このようなモデル化には多様なものを考えることができる。このような時系列データにおける自然な仮定では、直近のものの方が、過去のものよりも、現在の評価対象の状態に対して、より大きなインパクトを与えていると考えられる。また、入力として与えられる時系列データ及びテキスト時系列データに、特段の非線形性を仮定する明確な理由は存在していない。このため、本論文では、シンプルなモデル化をまずは試みることにする。すなわち、式(3)により、注目すべき評価対象であるかどうかを評価する評価値(以下、注目度)を、評価対象に対して定義することにする。

$$q^t = (1-\beta)\sum_{i=1}^{n-1} \alpha^{n-i+1} w^t_i + \beta \sum_{i=1}^{n} \alpha^{n-i} r^t_i \quad (3)$$

ここで、$\alpha \in [0,1]$は過去のものをどの程度重視するかを決定するパラメータであり、$\beta \in [0,1]$は数値時系列データとテキスト時系列データをどの程度の割合で重視するかを決定するパターメータである。以下においては、前者を減衰率、後者を表現重視率と呼ぶことにする。また $\alpha=0$ の場合に 0^0 になりうるが、$0^0=1$ とみなすことにする。

本モデル化によって、評価対象に対する注目度を算出することができ、注目度にしたがって、注目すべき評価対象を抽出することが可能となる。このようなモデル化は、複数、異種の情報源に基づいてモデル化を行うことができるため、より精密なモデル化を行うことを期待することができる。

3. 数値実験

提案する分析法の効果を検証するために、数値時系列データとして、株価情報、テキスト時系列データとして、ニュースヘッドラインに着目する。このような、経済関連の数値時系列データと、Web 上で収集可能なニュースや記事との間の関係を分析する研究は、近年活発化している[Bollen et al. 2010] [Peramunetilleke and Wong 2006]。本論文では、株価情報とニュースヘッドラインに基づいて、評価対象である会社(銘柄)を注目すべき順に、ランキングすることを試みる。以下においては、実験に利用したデータ、実験方法、評価方法、実験結果を順に説明し、提案法の効果を考察する。

3.1 実験データ

http://www.geocities.jp/sundaysoftware/csv/keiretu.html から株価情報をダウンロードする。本サイトの場合、銘柄コードごとに、最大 250 日間の株価情報を、csv 形式で保存している。各ファイルには、銘柄コード、日付、始値、高値、安値、終値、出来高といったデータが、行単位に格納されている。一方、goo、ITmedia、毎日新聞社などのニュース配信サイトから、ニュースヘッドラインをダウンロードする。今回の実験の場合、2010/02/03～2010/03/02 の期間に配信された 94,523 件のニュースヘッドラインを実験対象とする。各ニュースヘッドラインにおいては、ニュースヘッドラインの他に、サイト名、ニュースヘッドラインの内容が属する分野、配信社名、配信日、配信時刻などの関連情報が付随しており、各サイトからダウンロードすることができる。

3.2 実験方法

1 日を単位時間とすることにより、数値時系列データ及びテキスト時系列データを生成する。このため、各ニュースヘッドラインに付与されている時刻を無視することにより、テキスト時系列データを構成するテキスト集合を生成することができる。また、注目すべき順に、銘柄名をランキングすることを目指すため、株価情報への影響はほとんどないと考えられる、「スポーツ」、「芸能」、「恋愛」などの分野に割り当てられているニュースヘッドラインを、実験データから削除することにする。

一方、株価情報の場合、株式市場がオープンしている日のデータだけを収集することになるため、土日、祝日などの株式市場がクロースしている日のデータを収集することができない。これに対して、数値時系列データでは、変化率を算出する必要があるため、直前の値が存在しない場合には、その値を補間する必要がある。本論文では、その日より前で、最も近い日のデータで、その値を補間することにする。

ニュースヘッドラインにおいて、評価対象に関する内容が記載されているとしても、各ニュースヘッドラインを、逐次人が読んで、評価対象に関する内容が記載されているかどうかを判定することはできない。このため、何らからの機械的な処理を可能とする必要がある。このような処理を簡単に実現する方法として、東京証券取引所のサイトに記載されている銘柄名を、まずはダウンロードし、本銘柄名そのものが、ニュースヘッドラインに含まれているかどうかを判定する方法が考えられる。しかしながら、ニュースヘッドライン中では、銘柄名というよりは、その略称で記載されていることも多く、銘柄名の一致を判定に利用した場合

には、評価対象に関する内容が記載されているとしても、記載されていないと判定される可能性が高くなる。このため、銘柄名を一度形態素解析し、その中から固有表現を取り出して、当該固有表現がニュースヘッドラインに含まれているかどうかを判定することによって、ニュースヘッドラインに評価対象に関する内容が記述されているかどうかを判定することにする。

3.3 評価方法

提案するモデルのパラメータである減衰率を 0.0、0.5、1.0、表現重視率を 0.0、0.25、0.5、0.75、1.00、データ収集期間を 5、10、15、20 とした数値実験を実施する。また、株価情報としては、出来高と高値を指定した実験を実施する。本実験では、翌営業日における、出来高及び高値の変化率の絶対値を、実際の注目度とみなすことにする。

一方、評価基準としては、実際の注目度に基づいて各評価対象に付与された順位と、提案法に基づいて算出された注目度の高い順に割り当てられた順位の差を利用する。すなわち、式(4)で定義される積算順位絶対誤差率を評価基準として利用する。ただし、n を評価対象の数、p_iを実順位 i 位の評価対象の予測順位、|・|を絶対値を計算する演算とする。

$$\begin{cases} \dfrac{\sum_{i=1}^{2k} |i - p_i|}{2k^2}, (n = 2k) \\ \dfrac{\sum_{i=1}^{2k+1} |i - p_i|}{2k(k+1)}, (n = 2k+1) \end{cases} \quad (4)$$

本値は、各評価対象の予測順位と、評価対象の予測順位の差を、最大の順位の差(最大順位絶対誤差)で割った値として定義されている。このとき、最大順位絶対誤差は、予測順位が実順位の逆順位として与えられた場合の値であり、評価対象数が偶数、奇数の場合で、$2k^2$、2k(2k+1)と与えられる。実験においては、提案法によって個々の評価対象の注目度を算出した場合に、その翌日の数値時系列データの値との絶対変化率を算出することにより、その大きい順に与えられる順位を、実順位とみなすことにする。また、複数の評価対象が、同順位にならないようにするために、絶対変化率が同じ場合には、銘柄コード順に順位が与えられるとする。

3.4 比較対象

提案法の性能を検証するために、ランダムに順位を決定した場合に、積算順位絶対誤差率がどの程度の値となるかを評価する。ランダムにその順位を決定した場合、実順位 i 位の評価対象に対して、評価対象の数が偶数の場合には、1位から 2k 位、奇数の場合には、1 位から 2k+1 位のうちの任意の順位が、等しい確率で、評価対象に割り当てられることになる。

次に、評価対象の数が偶数の場合に着目してみると、実順位が i(≦k)位であるとすれば、予測順位 1 位、2 位、・・・、2k−1位、2k 位に対して、順位絶対誤差は、i−1、i−2、・・・、1、0、1、・・・、2k−i−1、2k−i と与えられる。同様に、実順位が 2k+1−i 位であるとすれば、順位絶対誤差は、2k−i、2k−i−1、・・・、1、0、1、・・・、i−2、i−1 と与えられる。このため、実順位が i 位の場合における順位絶対誤差を積算した値(積算順位絶対誤差)は、実順位が 2k+1−i 位の場合における積算順位絶対誤差と一致している。また、実順位が i−1 位の順位絶対誤差は、i−2、i−3、・・・、1、0、1、・・・、2k−i、2k−i+1 と与えられている。このため、実順位 i 位の積算順位絶対誤差は、実順位 i−1 位の積算順位絶対誤差より大きくなる。一方、実順位が k 位の場合の順位絶対誤差は、k−1、k−2、・・・、1、0、1、・・・、k−1、k と与えられる。従って、i 位の実順位における積算順位絶対誤差を z_i とおけば、式(5)に示す関係が成立している。

$$z_{i+1} = \begin{cases} 2\sum_{j=1}^{k} j - k, (i = k-1) \\ z_i + 2(k-i), (1 \le i < k-1) \end{cases} \quad (5)$$

このため、積算順位絶対誤差の期待値は、式(6)に示すように与えられる。

$$\frac{2\sum_{i=1}^{k} z_{k-i}}{2k} = \frac{(2k+1)(2k-1)}{3} \quad (6)$$

以上により、評価対象の数が偶数の場合における期待順位絶対誤差率は式(7)に示すように与えられ、k の値が十分大きい場合に、その値は 2/3(≒0.67)となる。

$$\frac{\frac{(2k+1)(2k-1)}{3}}{2k^2} = \frac{2}{3}\left(1+\frac{1}{2k}\right)\left(1-\frac{1}{2k}\right) \quad (7)$$

次に、評価対象の数が奇数の場合について検討する。先に検討した偶数の場合と同様に考えることにより、i 位の実順位における積算順位絶対誤差を z_i とおけば、式(8)に示す関係が成立している。

$$z_{i+1} = \begin{cases} 2\sum_{j=1}^{k} j, (i = k) \\ z_i + 2(k-i) + 1, (1 \le i < k) \end{cases} \quad (8)$$

このため、積算順位絶対誤差の期待値は、式(9)のように与えられる。

$$\frac{2\sum_{i=1}^{k} z_{k+1-i} + 2\sum_{j=1}^{k} j}{2k+1} = \frac{4k(k+1)}{3} \quad (9)$$

以上により、評価対象の数が奇数の場合における期待順位絶対誤差率は式(10)に示すように与えられ、k の値に関係なく、その値は 2/3(≒0.67)となる。

$$\frac{\frac{4k(k+1)}{3}}{2k(k+1)} = \frac{2}{3} \quad (10)$$

実験においては、評価対象の数が 1,690 存在し、偶数であるものの、k の値は十分大きいため、ランダムに順位を選択した場合の期待順位絶対誤差率は 0.67 とみなすことができる。

3.5 実験結果

実験条件を変えて実施した実験結果の一部として、出来高及び高値を利用した場合の結果を、図 1、図 2 に、それぞれ示す。ただし、紙面の都合上、データ収集期間は、20 の場合に限定されている。各図は、表現重視率を 0.00 から 1.00 に変えた場合における順位絶対誤差率の、各日における値の平均値の推移を示している。各グラフにおいては、藍色、桃色、黄色の各ラインが、減衰率を表した結果を示している。また、水平軸が表現重視率、垂直軸が順位絶対誤差率を示している。

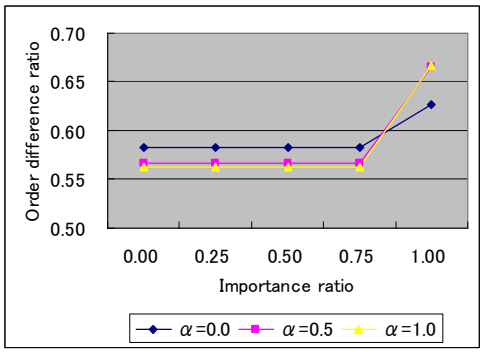

図 1:実験結果(出来高の場合)

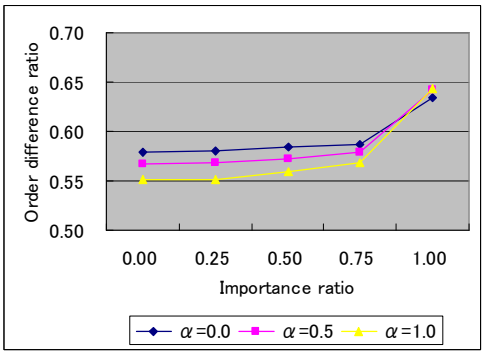

図 2:実験結果(高値の場合)

3.6 考察

以下に示す 3 つの観点から実験結果を考察する。

(1) 順位絶対誤差率

パラメータの違いや、利用している数値時系列データの違いによって、結果に違いはあるものの、表現重視率を 0.00 にした場合には、順位絶対誤差率の平均値を 0.56 程度に抑えることができ、ランダムに順位を決定した場合における期待値である 0.67 に比べて、順位絶対誤差率を改善することができる。株価やその出来高の変動及び、ニュースヘッドラインにおける話題の変遷には、ある程度の連続性があると考えられるため、過去の時系列データに基づいて、次の単位時間における評価対象の注目度を評価する提案法は、次に注目すべき評価対象をある程度、妥当に予測できたものと考えられる。利用する時系列データに依存するものの、提案法は、次に注目すべき評価対象のモデル化に、ある程度適用可能と考えられる。

(2) パラメータの影響

減衰率に着目してみると、紙面の都合上、図示していないものの、減衰率が大きい程、順位絶対誤差率の平均値が、比較的小さくなっている。ここで、減衰率が大きくなる程、評価対象ごとの注目度のモデル化に、過去のデータを反映させる効果があることに着目すれば、この結果は、過去のデータをなるべく利用した方がよいことを示している。また、データ収集期間に着目してみると、データ収集期間が長い方が比較的順位絶対誤差率の平均値が小さくなっており、データ収集期間の面からみても過去のデータをモデル化に反映させた方がよいといえる。

一方、表現重視率に着目してみると、表現重視率が小さい程、順位絶対誤差率の平均値は小さくなっている。テキスト時系列データを加えることにより、数値時系列データだけでは説明できない現象を、説明できることを期待していたものの、今回の実験からは、そのような効果を明確に確認することはできなかった。しかしながら、表現重視率を 1.00 にした場合であっても、ランダムに順位を決定した場合よりは、順位絶対誤差率の平均値は改善されており、ニュースヘッドラインと株価情報との間に何らかの関連性はあると考えられる。今回の実験では、株価情報取得の問題から、日単位でのモデル化となっているが、株価情報をより細かい粒度で収集することも可能であり、細かい粒度でのモデル化によって、より明確な関連性を見出せる可能性はある。また、本論文で利用した株価情報以外の株価情報も存在しており、それらとの間に、関連性を見出せる可能性もある。この他、ニュースヘッドラインからの評価対象の抽出法を改良することによって、関連性を見出せる可能性はあると考えられる。

(3) センサーネットワークへの適用可能性

今回のモデル化では、減衰率を予め決定しておくことにより、データ収集期間分の時系列データと、現時点における数値時系列データだけから得られる注目度と、テキスト時系列データだけから得られる注目度を記憶する。これにより、指定された表現重視率に対応する評価対象の注目度を算出することができる。このため、評価対象の数の定数倍の記憶容量を確保するだけで、注目度を算出することができる。一方、新規に入力されたデータと最も古くから存在するデータを参照することにより、数値時系列データだけから得られる注目度とテキスト時系列データだけから得られる注目度を更新することができる。このため、低い処理負荷でモデルを更新することができる。センサーネットワークへの適用を考えた場合、多数の評価対象を、実時間で評価することが必須と考えられるため、このようなモデル化はそのニーズにあったモデル化になっていると考えられる。

4. まとめと今後の課題

本論文では、評価対象に対して与えられた数値時系列データとテキスト時系列データから、次期における注目すべき評価対象をモデル化する方法を提案した。また、提案法を実際の株価情報とニュースヘッドラインに適用し、その効果を検証した。その結果、過去のデータを利用することにより、ある程度、次期において注目すべき評価対象を決定できることを確認した。

今後の課題としては、考察のところでも触れたが、数値時系列データとテキスト時系列データの相乗効果を、明確に確認するまでには至らなかったため、その効果が得られるように、提案法を改良することを検討していきたい。また、他の期間のデータも収集して、提案法に適用することにより、より詳細な効果検証を行っていきたい。一方、今回、数値時系列データとしては、株価情報、テキスト情報としてはニュースヘッドラインを入力としたが、本枠組は他の分野へも適用可能である。例えば、点検保守においては、機器から得られた計測データを数値時系列データ、保守員による点検記録メモをテキスト時系列データとみなすことにより、注目すべき機器の特定をある程度予測できるのではないかと考えている。この他、提案する枠組みの、さらなる適用先を検討するとともに、実際の CEP システムに組み込んだ評価を行なっていく予定である。

参考文献

[Bollen et al. 2010] J. Bollen et al.: Twitter Mood Predicts the Stock Market, http://arxiv.org/PS_cache/arxiv/pdf/1010/1010.3003v1.pdf, 2010.

[Peramunetilleke and Wong 2006] D. Peramunetilleke and R. K. Wong: Currency Exchange Rate Forecasting from News Headlines, Proc. of the 13th Australasian Database Conference, vol. 5, pp. 131-139, 2002.