2B1-R-3-1

学習データの拡充による評価主体検知性能の改善

Improvement of Detection Performance for Evaluation Objects by the Expansion of Learning Data

櫻井 茂明[*1] 牧野 恭子[*1] 松本 茂[*1]
Shigeaki Sakurai Kyoko Makino Shigeru Matsumoto

[*1] 東芝ソリューション(株) IT 技術研究所
Advanced IT Laboratory, Toshiba Solutions Corporation

This paper proposes a method expanding amount of learning data in the analysis method of complex sequential data. Here, the data is composed of numerical sequential data and text sequential data. The method applies a topic dictionary into learning data to extract more evaluation objects from the data. The dictionary stores keywords related to specific stock brands whose prices change. This paper applies the expansion method to a prediction task of attractive stock brands. We collect experimental data from 5 sites distributing news headlines and a site storing stock information. Lastly, this paper shows that the proposed expansion can improve the detection performance of evaluation objects through numerical experiments.

1. はじめに

コンピュータ及びネットワーク環境の発展、センサーの小型化及び低価格化に伴って、多様な時系列データが簡便に収集されるようになってきている。これら多様な時系列データの中には、人の生活をよりスマートにする知見が埋もれていると考えられており、これらデータを総合的に分析するニーズが、近年益々高まっている。一方、これらデータの分析は、その目的や対象とするデータの特徴に依存しており、唯一無二の分析法を確立することはできない。このため、その目的やデータの特徴に応じた分析法を開発、適用していくことが不可欠である。

このようなデータ分析法のひとつとして、我々は現段階において、データ収集が比較的容易であり、予測といった定量的な評価も実施しやすい、特定の評価主体に関連した、複合時系列データを対象とした分析法の確立に取り組んでいる[Sakurai et al. 2012]。提案する分析法では、評価主体に関連するテキスト時系列データと数値時系列データから、数値時系列データの次期における変化を代表するトレンドルールを発見する。また、発見したルールを新たに収集されたテキスト時系列データに適用することにより、次期において、大きな数値の変化が期待できる評価主体を、注目すべき評価主体として抽出することを可能としている。本論文では、提案における予測性能の向上を目的として、トレンドルールの発見において、学習データを拡充する方法を検討する。具体的には、注目すべき株式銘柄の予測問題を題材として、特定の株式銘柄の株価の変化と関連するキーワードを格納したトピック辞書の活用を通して、学習データの拡充を図る。加えて、各株式銘柄の日別の株価系列及びニュース配信サイトから配信されるニュースヘッドラインからの実データを用いた数値実験への適用を通して、学習データ拡充の効果を検証した結果を報告する。

以上により、本論文の残りを次のように構成する。2 節では、[Sakurai et al. 2012]に提案した複合時系列データの分析法を簡単に説明し、3 節で、学習データの拡充法を説明する。4 節では、学習データを拡充した効果を、数値実験を通して検証した結果を報告する。最後に、5 節で、本論文のまとめと今後の研究課題を述べる。

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: 042-340-6415, Fax: 042-340-6013, E-mail: Sakurai.Shigeaki@toshiba-sol.co.jp

2. 複合時系列データの分析法

本節では、先に提案した複合時系列データの分析法を簡単に紹介する。提案法はトレンドルールを発見するフェーズと、発見したトレンドルールに基づいて、注目すべき評価主体を予測するフェーズのふたつからなっている。以下においては、各フェーズについて、順にその概略を説明していく。

2.1 トレンドルールの発見

トレンドルールの発見フェーズでは、テキスト時系列データと、数値時系列データの 2 種類の時系列データから、トレンドルールを発見する。ただし、テキスト時系列データを構成するテキストには、評価主体に関連した表現が記載されているとし、各評価主体に対応する数値時系列データが与えられているとする。本フェーズでは、最初に、各テキストに、形態素解析を適用することにより、名詞表現を抽出する。このとき、抽出された名詞表現が組織を示す固有名詞であるとすれば、当該名詞表現が、評価主体であるかどうかまでを識別する。一方、評価主体以外の名詞は、属性として識別される。あるテキストから評価主体が抽出されたとすれば、当該テキストに付随する時間情報を確認し、当該テキストに対応する数値時系列データの中から、当該時間と次期の時間に対応する数値を抽出する。このふたつの数値から、数値の変化率を算出し、算出された変化率の大きさに従って、そのクラスを識別する。抽出された評価主体及び属性は、テキストごとにまとめられ、各テキストに対応する学習トランザクションが生成される。次に、クラスごとにまとめられた学習トランザクションに対して、頻出パターンの発見法[Pei et al. 2004]を適用し、最小支持度以上の頻出パターンを発見する。発見された頻出パターンはクラスを代表する評価主体や属性の組み合わせとなっているため、この頻出パターンとクラスの組が、トレンドルールとして抽出されることになる。

2.2 注目評価主体の発見

トレンドルールを用いた注目評価主体の発見では、指定された期間に新たに収集されたテキスト時系列データに対して、クラスごとのトレンドルールを適用することにより、次の期間において注目すべき評価主体を抽出する。具体的には、テキスト時系列データの各テキストの中から評価主体と属性を抽出する。この抽出処理は、トレンドルールの発見フェーズにおける処理と等価な処理となっている。注目評価主体の発見フェーズでは、各

テキストから抽出した評価主体と属性をまとめることにより、評価トランザクションを生成する。本フェーズでは、評価トランザクションと、最小アイテム数以上のアイテムからなるトレンドルールを個別に比較することにより、評価トランザクションが、トレンドルールを完全に含んでいるかどうかを識別する。また、トレンドルールを完全に含んでいるとすれば、当該評価トランザクションに対して、トレンドルールのクラスに対応した重みとして 1 を加算する。このような比較処理を、評価トランザクションごとに、すべてのトレンドルールに対して実施し、各評価トランザクションに対応するクラスの識別を行う。すなわち、評価トランザクションの元となったテキストによって、割り当てられているクラスの影響が与えられることになる。ただし、どのトレンドルールも完全には含んでいないような評価トランザクションに対しては、クラスの割り当ては行われないことに注意する必要がある。このような評価トランザクションに対応するテキストは、クラスに影響を与えていないものと解釈されることになる。本フェーズでは、次に、評価トランザクションに含まれる評価主体に着目し、その評価主体の対応するクラスの重みを 1 加算する。最終的には、評価主体に対して与えられているクラスの重みを評価することにより、評価主体が注目すべき評価主体であるかどうかの判断を行うことになる。

なお、本論文で対象としている注目株式銘柄の予測タスクにおいては、翌日における株価の値が変化すると考えられる株式銘柄を注目株式銘柄として判定している。このような問題設定にしているのは、株価が上がるか下がるかといった問題はかなり難しい問題である一方、上がり下がりを正しく予測できないとしても、その変化を知らせることも、株式のトレーダーにとっては、意味がある知見になるためである。このため、クラスを無視した重みの合計値、すなわち、評価主体に割り当てられたクラスを識別されたトランザクションの数の総数が、指定したしきい値(最小トランザクション数)以上となる評価主体を、注目すべき評価主体として抽出されることになる。

3. 学習データの拡充

[Sakurai et al. 2012]の評価実験により、評価主体が割り当てられないテキストが多数存在することが判明している。一方、機械学習法にとっては、学習時に利用できるデータの数が多い程、より妥当な学習を行うことが期待できる。このため、トレンドルールの発見においても、より多くの学習トランザクションを利用できるとすれば、より妥当なトレンドルールを発見することが期待できる。そこで、本節では、従来は評価主体を抽出することができなかったテキストから評価主体を抽出し、学習トランザクションとするための方法を検討する。

円高になると輸出企業の企業業績が悪化することが懸念され、株価が下がる傾向があるように、テキスト時系列データを構成するニュースヘッドラインに、具体的な株式銘柄が出現していないとしても、株価の変化に結び付くニュースヘッドラインは存在している。そこで、このようなニュースヘッドラインからも学習トランザクションを生成するために、トピック辞書の活用を検討する。ここで、トピック辞書とは、特定のトピックと特定の評価主体との関係を記述した 3 階層からなるシソーラスのことである。各階層には、トピック、サブトピック、評価主体が、それぞれ記載されている。表 1 は、トピック辞書の例を示しており、為替、レアメタル、エコポイントといったトピックに関連して、そのサブトピックや、影響を受ける評価主体が記載されている。本論文では、http://www.asset-alive.com/thema/に与えられている、トピックと株式銘柄との関係を記述した知識を、理想的なトピック辞書とみなして利用する。

表 1:トピック辞書の例

第1階層:トピック	第2:階層:サブトピック	第3階層:評価主体
為替	円高	A社
為替	円安	B社
・・・	・・・	・・・
レアメタル	ニッケル	C社
レアメタル	タングステン	D社
・・・	・・・	・・・
エコポイント	エアコン	E社
エコポイント	冷蔵庫	F社
・・・	・・・	・・・

一方、トピック辞書の利用方法としては、トピック辞書の効果をまずは簡便に見極めるため、提案法における現在の枠組みを、あまり変えずに、トピック辞書を利用することを試みる。すなわち、株式銘柄の上位 2 階層に与えられている表現の中から、形態素解析によって、名詞表現だけを抽出する。また、当該の名詞表現が、ニュースヘッドラインから抽出される名詞表現と合致する場合に、当該名詞表現の株式銘柄への読み替えを行う。現状の枠組みの変更を少なくするといった観点では、株式銘柄の割り当てといった方法も考えられるが、割り当ての場合には、割り当てた評価主体とそのトピックとを含む頻出パターンが極端に増えることが予想されたため、読み替えを行うことにしている。加えて、予備実験を実施したところ、通常複数の評価主体がひとつのトピックに対応している影響で、複数の評価主体を含んだ頻出パターンが多数発生する現象が確認された。そこで、複数の評価主体を含んだ頻出パターンを、頻出パターンとしては出力しないようにするといった制限が、頻出パターンの発見時に課されている。

このようなトピック辞書の活用により、大幅な枠組みの変更を実施することなしに、クラスの割り当てが行われなかったニュースヘッドラインからも、学習トランザクションを生成することが期待できる。

4. 数値実験

本節では、学習トランザクションの拡充による効果を検証するために実施した数値実験について説明する。実験に利用したデータ及び評価基準、実験方法をまずは説明し、実験結果を提示して、その効果を議論する。

4.1 実験データ

Excite、Goo、Infoseek、Livedoor、Yahoo の 5 つのサイトから適宜配信されているニュースヘッドラインを、我々のグループでは収集している。本ニュースヘッドラインを、テキスト時系列データとして利用する。現在、3 期間にわたって収集したデータが実験において利用可能であるが、今回の実験では、第一期間(D1:2010/8/28～2011/1/31)に収集されたニュースヘッドラインを学習用に利用する。また、第二期間(D2:2011/2/1～2011/4/6)に収集されたデータのうち、2011/3/11 に発生した東日本大震災の前までのデータ(D2a)を、注目株式銘柄の予測用に利用する。第二期間の後半のデータ(D2b)及び第三期間(D3:2011/4/7～2011/5/22)のデータを今回実験に利用していないのは、東日本大震災の影響によって、その前後では株価の変動に関するルールが変化したものと考えられたためである。表 2 は 5 つのサイトから収集されたニュースヘッドラインの件数を示している。表から分かるように約 100 万件のニュースヘッドラインが学習用に利用されており、約 27 万件のニュースヘッドラインが予測用に利用されている。

表 2:テキスト時系列データ

		Period		
		D1	D2a	
Site	Excite	132,878	38,761	
	Goo	143,062	30,593	
	Infoseek	240,141	62,407	
	Livedoor	233,773	66,740	
	Yahoo	253,619	70,184	
	Total	1,003,473	268,685	単位:件

数値時系列データとしては、http://www.geocities.jp/sundaysoftware/csv/keiretu.html にて公開されている株価情報を利用する。当該サイトでは、250 営業日における日次の株価情報を株式銘柄ごとに csv 形式で公開している。各株価情報には、銘柄コード、日付、始値、終値、高値、安値、出来高が記録されている。今回の実験では、テキスト時系列データの収集期間を含むよう、数値時系列データの収集が行われている。

評価主体としては、東証 1 部に加えて、東証 2 部、マザーズ、札証、大証、名証、福証に上場しているすべての株式銘柄を対象とする。具体的には、各証券市場のホームページから個別に株式銘柄を収集しており、収集した株式銘柄をマージして重複を取り除くことにより、対象とする株式銘柄のリストを生成している。

4.2 評価基準

対象とする証券市場を拡大したこともあり、株式銘柄の総数は 3,951 社にも及んでいる。多数の株式銘柄がシステムによって、推薦されたとしても、トレーダーは多数の株式銘柄に関する判断を行うことはできない。このため、注目に値する株式銘柄の一部が見逃されたとしても、トレーダーにとってはあまり大きな問題とはならない。これに対して、推薦された株式銘柄の中に、注目に値しない株式銘柄が多数含まれているとすれば、システムに対して、悪い印象を持つと考えられる。このため、推薦された株式銘柄は、高い確率で、実際に注目に値する株式銘柄になる必要がある。そこで、このような特徴を評価する基準として、式(1)に定義する適合率を重視する。

$$適合率 = \frac{抽出された注目株式銘柄数}{抽出株式銘柄数} \cdots (1)$$

ただし、本実験においては、各日の株価とその翌日の株価の差の相対的な割合である変動率が、上昇(Rise)方向では予め指定されている最小変化率より大きい場合に、真の注目株式銘柄であると判定している。また、下降(Drop)方向においては、最小変化率以下となる場合に、真の注目株式銘柄であると判定している。このような判定は、評価期間にわたって実施されており、その積算値によって適合率が算出されている。

上述した観点によって適合率を重視するとしても、何らかの推薦を行っているといった観点では、適合率とトレードオフの関係にある再現率や、適合率と再現率のバランスをとった総合的な指標である F 値も、提案法の効果を確認する上では意味ある指標となる。このため、式(2)、式(3)によって、定義される再現率と F 値も評価基準に加えた評価を、以下の実験では行っていくことにする。

$$再現率 = \frac{抽出された注目株式銘柄数}{注目株式銘柄数} \cdots (2)$$

$$F値 = \frac{2 \times 適合率 \times 再現率}{適合率 + 再現率} \cdots (3)$$

ただし、注目株式銘柄数の算出においては、抽出処理において評価主体として抽出された株式銘柄を各日においてまずは求めている。この株式銘柄の集合のうち、翌日の株価変動に対して、上昇あるいは下降と判断された株式銘柄を各日における真の注目株式銘柄と判定している。また、この株式銘柄数を評価期間にわたって積算した値を、注目株式銘柄数とみなして、再現率を算出している。このような評価主体の限定を行っているのは、ニュースヘッドラインからは読み取ることができない要因によって株価が変動している場合、注目評価主体として抽出できないのは当然の結果だからである。このような真の注目株式銘柄の影響を排除し、提案法の効果を評価するため、評価主体の限定が行われている。

4.3 実験方法

[Sakurai et al. 2012]における実験結果との比較を行うために、学習フェーズ、予測フェーズにおけるパラメータの設定としては、同じ設定を利用する。すなわち、学習フェーズにおいて、クラスを識別するための変動率を 0.050 とする。また、頻出パターン発見時に参照する最小支持度として、0.005、0.010、0.020、0.030 の 4 種類を利用する。一方、予測フェーズにおいて、真の注目評価主体であるかどうかを判定するために利用する変動率として、0.001、0.020、0.025、0.030、0.050 の 5 種類を利用する。また、注目評価主体として抽出するために利用する、トレンドルールに含まれる最小アイテム数を 2、最小トランザクション数を 1、3、5 の 3 種類設定する。

上述したパラメータ設定での実験を行い、[Sakurai et al. 2012]における実験結果との比較を行う。

4.4 実験結果

図 1～図 4 に実験結果の一部を示す。各図においては、学習時におけるパラメータである、変動率及び最小支持度を 0.050、0.005、予測時におけるパラメータである、変動率、最小アイテム数、最小トランザクション数を 0.020、2、1 と設定した場合の結果を示している。

このとき、図 1 は、抽出される学習トランザクションの数が変化する様子を示している。図 1 においては、縦軸が学習トランザクションの数を表しており、横軸が手法の違いを表している。ただし、Sakurai は、[Sakurai et al. 2012]における結果であり、東証 1 部の株式銘柄のみ(1,680 社)を対象とし、トピック辞書を活用しなかった場合の結果となっている。一方、No_topic 及び Topic が全証券市場を対象とし、トピック辞書を活用しなかった場合とした場合の結果を示している。各棒グラフにおいては、上昇、下降の別に学習トランザクションの数が積み上げられている。

図 2 は、抽出された学習トランザクションから発見されるトレンドルールの数を示している。図 2 においては、縦軸がトレンドルールの数を表しており、横軸は図 1 と同じ意味を持っている。また、図 1 の場合と同様に、各棒グラフにおいては、上昇と下降別に、トレンドルールの数が積み上げられている。

図 3 は発見されたトレンドルールによって注目すべき評価主体として抽出される評価主体の数を示している。図 3 においては、縦軸が抽出される評価主体の数を表しており、横軸は図 1 と同じ意味を持っている。

最後に、図 4 は抽出された評価主体の妥当性を示す評価値が変化する様子を示している。図 4 においては、縦軸が適合率、再現率、F 値といった評価値を示しており、横軸は他の図と同様の意味を持っている。各棒グラフにおいては、個々の評価値が証券市場ごとに、左から右に、適合率、再現率、F 値の順に並べた結果を示している。

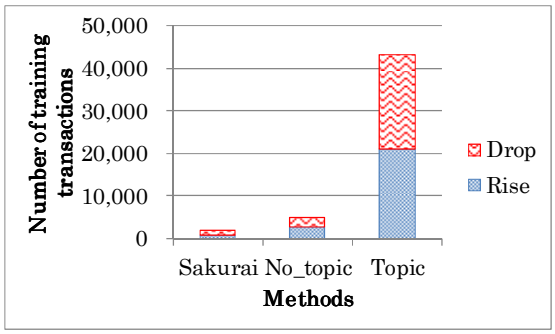

図 1:学習トランザクション数の変化

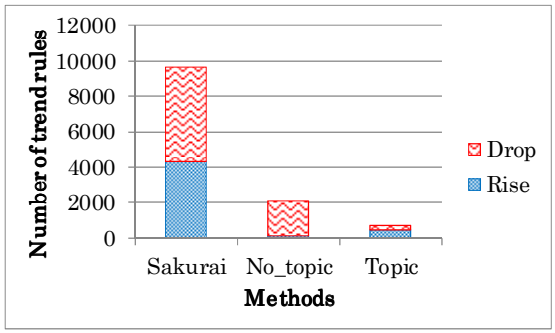

図 2:トレンドルール数の変化

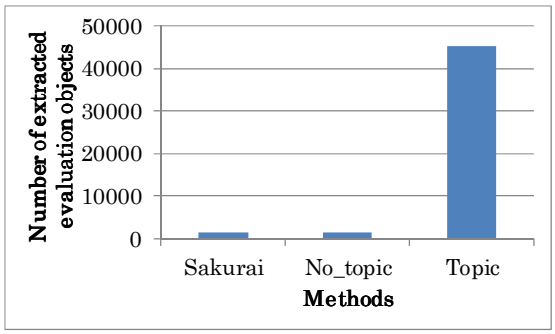

図 3:抽出評価主体数の変化

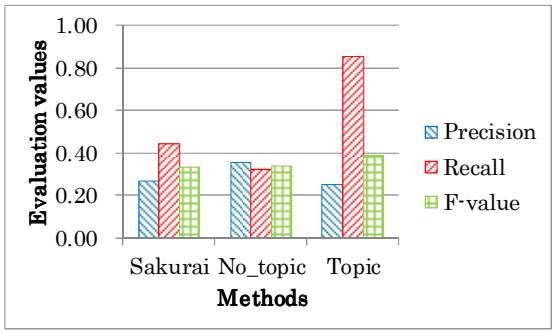

図 4:評価値の変化

4.5 考察

図 1 に示すように、上昇、下降に割り当てられる学習データの件数が、[Sakurai et al.]の場合と比較して、約 25 倍、トピック辞書を活用しなかった場合と比較しても、約 8.7 倍増えており、学習トランザクションの大幅な増加を確認することができる。一方、図 2 に示すように、発見されるトレンドルール数は、大幅に減少している。学習トランザクションが増えたことにより、アイテムの種類が増え、各アイテムの相対的な出現頻度が低下したために、各アイテムが頻出しにくくなったことがその原因であると考えられる。他の実験結果から、トレンドルールの数をある程度確保した方が予測精度が高いといった結果も得られているため、アイテム数にあまり左右されることなしに適切な数のトレンドルールを発見する方法を今後検討する必要があるかもしれない。

次に、抽出銘柄数に着目してみると、図 3 に示すように、トピック辞書の活用によって、抽出数が大幅に増加している。評価トランザクションの作成における属性の株式銘柄への読み替えによって、多くの株式銘柄に多数のトランザクションが割り当てられ易くなったために、このような大幅な抽出数の増加が起こったものと考えられる。一方、今回のトピック辞書の適用では、トピック辞書によって割り当てられた株式銘柄と、元々割り当てられていた株式銘柄の区別を行っていない。後者の株式銘柄の方が、評価トランザクションのテキストと深く関連しているものと考えられるため、トレンドルールに基づいた評価トランザクションの評価においても、この違いを考慮した評価主体の評価を行うことも予測精度の向上に役立つ可能性があり、今後検討していきたい。

最後に、予測精度を比較してみると、図 4 に示すように、トピック辞書を活用しなかった場合に比べて、適合率は低下する傾向にあるものの、[Sakurai et al. 2012]における結果と同程度の水準となっている。一方、再現率に関しては、トピック辞書を活用した場合に、全体として大幅に改善している。評価トランザクションの作成における、属性の株式銘柄への読み替えが、多数の株式銘柄の抽出につながり、再現率の向上に寄与したものと考えている。加えて、F 値に関しても、その値が向上しており、[Sakurai et al.]との比較で 14.3%、トピック辞書を活用しなかった場合との比較で 12.7%向上している。この結果は、適合率の低下を超える、大幅な再現率の改善がなされたためであり、適合率をある程度重視するとしても、この傾向は維持できるものと考えられる。従って、トピック辞書の活用方法にまだ改善の余地があるとしても、トピック辞書の活用は効果があるものと考えられる。

以上の考察により、学習トランザクションの拡大は、より妥当なトレンドルールの発見及び評価主体の抽出に、効果があるものと考えられる。

5. まとめと今後の課題

本論文では、評価主体に関連する数値時系列データとテキスト時系列データからなる複合時系列データの分析法において、より妥当なトレンドルールの発見を目指した学習データの拡充法を提案した。その結果、トピック辞書の適用により、予測性能を向上できることが明らかになった。今回の実験では考慮していないものの、現在研究開発中のトピック辞書の自動構築法では、各トピックに確信度に関する情報が付与されている。このため、この確信度の利用の効果を今後見極めていく予定である。提案した複合時系列データの分析法は、注目株式銘柄の予測タスクに限定した手法ではなく、適用対象を広げ、横展開を図っていく予定である。特に、スマートコミュニティ分野及びヘルスケア分野への適用の可能性を検討していく。

参考文献

[Pei et al. 2004] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu: Mining Sequential Patterns by Pattern-Growth: the PrefixSpan Approach, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp.1424-1440, 2004.

[Sakurai et al. 2012] S. Sakurai, K. Makino, and S. Matsumoto: A Discovery Method of Trend Rules from Complex Sequential Data, Proc. of the Workshop in AINA2012, 2012.