

経済専門辞書を用いたテキストマイニングによるインターネット株式掲示板の投稿分析に関する研究

Analysis of Stock BBS Postings by Text Mining using Economic Dictionary

◎和田 英一¹, 諏訪 博彦¹, 太田 敏澄¹
Eiichi WADA, Hirohiko SUWA and Toshizumi OHTA

¹電気通信大学 University of Electro-communication

Abstract We propose the text mining using economic dictionary for stock bbs posting and investigate the difference between PCA based on economic dictionary and PCA based on general dictionary. Text Mining using Economic dictionary have two advantages. First, we can pick out economic word like PER. Second, we can filter out noises like mental abuse. As a result, we find the difference between TextMining using economic dictionary and TextMining using general dictionary.

キーワード インターネット株式掲示板、形態素解析、主成分分析、自然言語処理、

1. 研究背景

近年インターネットの普及によって、SNS、blog、掲示板などを通じて個人が情報を発信できるようになり情報量が増え続けている。しかし、すべての情報をみるのは時間的に不可能であり、せつかくのテキストデータが十分に活用されていない現状がある。そこでテキストマイニングという形式化されていないテキストから有用な知識、情報を抽出する方法が発展してきた。

例として、オンラインチャットの活発さから、本の売上を推定する研究[1]やブログの感想、意見から映画の売上を推定する研究[2]がある。

また、金融分野に関してもテキストマイニングを用いた研究がされてきた。松井ら[3]は新聞記事を対象にSVMを用いてテキスト変化と市場の変化の関係を学習し、東証株価指(TOPIX)の騰落を予測している。その結果、56.6%の騰落を予測している。Johanら[4]はTwitterから世の中の雰囲気を得て、その情報からダウ工業株30種平均の騰落を86.7%予測している。和泉ら[5]は、日本銀行の金融経済月報に対して形態素解析と主成分分析を用いることで、実際の金融市場(株式、国債、円ドルレート)をどの程度説明しているのか検証を行っている。主成分分析の結果、30個の主成分で累積寄与率が60%となった。主成分得点を説明変数として回帰分析を行った結果、第1に国債2年・5年物に関しては約85%の確率で、日経平均株価に関しては約77%の確立でトレンドの予測が可能であること、第2にトレンド予想の期間は最大でも三ヵ月程度であることを報告している。このようにテキスト情報を用いて株価を予測する研究が行われている。我々は、インターネット株式掲示板に着目し、経済専門辞書を用いることで、より精度を高めることを目指す。

2. 関連研究

2.1 金融に関する研究

株式市場では、市場参加者や景気の動向、投資先の企業の将来を予測しながら投資活動が行われている。このため自分以外の投資家の動向や意見は、投資活動において重要な情報であると考えられる。また、投資家が投資判断を下す際、数値化された情報(会計データ、株価)だけではなく、新聞やヘッドラインニュース、アナリストの意見など非数値的な情報も利用している。Vam Bommelの研究[6]によると、株価は、インターネット上の投資家の噂にも影響されている可能性が指摘されている。この点からも、他の投資家の意見を直接取得できるCGMは有効な判断材料である可能性がある。株式に関するCGMは古くから存在しており中でも株式掲示板は他の投資家の生の声を膨大に蓄積している点から、これまで実証分析の対象となってきた。本研究では、日本で古くから存在しており議論が活発であるYahoo!株式掲示板を分析対象とする。

2.2 Yahoo!株式掲示板を分析対象とした研究

Antiweiler & Frank[7]は米国株式市場で、CAPMおよびFama & Frenchの3ファクターモデルに対して、Yahoo!株式掲示板の投稿数を基とした投稿数ファクターを加え株式との関連性を検証している。結果、投稿数ファクターは有意な働きをしている可能性を示唆している。

丸山ら[8]は、日本株式市場でYahoo!株式掲示板の投稿数Top50を2005年から2006年の期間でメッセージを分析した。掲示板より取得した投稿データを、自然言語処理(形態素解析、ノイズ除去)と機械学習により、メッセージを「弱気投稿」「強気投稿」「中立」の3種類に分類している。結論として、第1に投稿数と株式リターンの関連の可能性は低いが、投稿内

容を考慮した場合、弱気投稿数が多いと翌日の株式リターンが負になる傾向があること。第2に投稿数がボラティリティ、出来高と関連していること。第3に強気比率が株式リターンやボラティリティ、出来高と関連している可能性があることを報告している。

諏訪ら[7]は、Yahoo!株式掲示板の投稿に付与されている「強気」、「弱気」といったタグを投稿内容ファクターとしてCAPMおよびFama & Frenchの3ファクターモデルに加え、株式リターンを説明するか検証を行っている。その結果、投稿内容は株式リターンを説明する可能性が高いことを示唆している。しかし、投稿内容ファクターは投稿内容の全てを捉えているのではなく、残差の存在が明らかとなっている。また、諏訪らの研究では投稿内容を「強気」という一軸でしか分類していない。関連研究によるとテキストデータはいくつかの軸で分類できることが確認されており、諏訪らの研究も同様に投稿内容を「強気」以外の複数の軸で分類できる可能性がある。

清水ら[8]は、投稿内容を「強気」「弱気」以外の軸で分類を試み、主成分得点を利用したPCAファクターを定義し、CAPMに加えることによって株式リターンとの関連を検証している。その結果、PCAファクターは株式リターンを説明する可能性が高いことを示唆している。しかし、清水らの研究では形態素解析の際に、一般辞書を用いて株式掲示板から単語を抽出しており、また抽出する単語は出現頻度でしかフィルタリングしていない。このため誹謗中傷のような株取引に関係がないと思われる単語も多数抽出されている。

そこで本研究では、清水らの方法を基に形態素解析に経済専門辞書を用いることによって、優先株、PERといった一般辞書では抽出できない経済用語を取得し、投稿に含まれる誹謗中傷といったようなノイズと思われる単語を除去する。このことにより、投稿内容から株式リターンの動きを、より精度よく説明できると考える。本稿では、経済専門辞書と一般辞書の差異について検討する。

3. 分析方法

3. 1. 分析対象

本研究では分析対象掲示板としてYahoo!株式掲示板を利用する。対象銘柄は東証一部上場銘柄1501社、分析期間は2003年1月1日から2008年12月31日までである。この分析期間のTOPIXの推移を図2に示す。

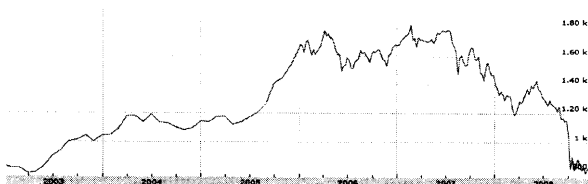


図1 2003～2008年におけるTOPIXの推移

図1から分析期間は、上昇、下降、保ち合い相場を含んでいることがわかる。

3. 2. 分析の流れ

投稿内容の分析は以下の手順で行う。分析の流れを図2に示す。

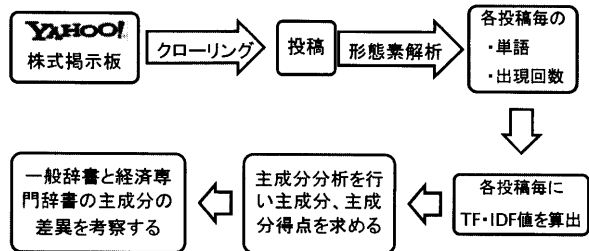


図2 分析の流れ

①投稿取得

Yahoo!株式掲示板からの投稿取得は、プログラムによって自動的に行われる。取得できるデータは、(1)題名 (2)投稿時の気持ち (3)投稿ID (4)投稿日時 (5)投稿No (6)本文 (7)返信かどうかである。図3にトヨタ自動車に対するYahoo!株式掲示板の投稿例を示す。

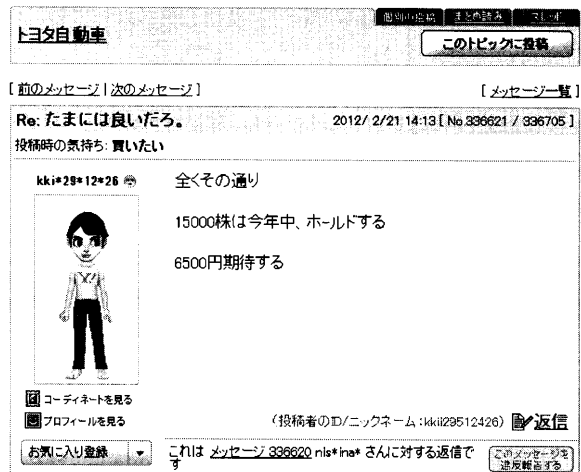


図3 Yahoo!株式掲示板投稿例

②形態素解析

各投稿に対して経済専門辞書を用いて形態素解析を行い、単語を抽出する。経済専門辞書は日経シソーラスを用いる。日経シソーラスは日本経済新聞デジタルメディアが1982年から作成している新聞記事検索のための用語集であり、収録名詞数は約15,000語である。収録語には社会的事象も含まれる。また清水ら[10]が用いた辞書は、NAIST-jdicである。NAIST-jdicは奈良先端科学技術大学院大学が発行している辞書であり収録名詞数は約255,000語である。形態素解析には、McCabを用いる。

③TF・IDF 値の算出

TF・IDF 値を算出し、単語の特徴量を求め単語に重み付けを行い、多くの投稿に出現する単語の重要度を下げる。また、同一の単語が複数の投稿で出現する場合は、各投稿毎に求めた TF・IDF 値を合計することで全投稿における単語の特徴量とする。

$$TF \cdot IDF = \text{単語 } t \text{ の出現回数} \times \frac{\text{投稿の数}}{\text{単語 } t \text{ が出現する投稿の数}} \dots (1)$$

本研究においては、TF・IDF 値の合計が大きい上位 700 単語を主成分分析に使用する。

④主成分分析

株式リターンを説明する株式掲示板に基づくファクターを作成するために、主成分分析を行う。TF・IDF 値の算出によって求めた 700 単語を説明変数として主成分分析を行い単語のグループ分けを行う。各主成分において、主成分負荷量の絶対値が大きい単語を調べること、その主成分を構成する単語がどのような傾向を持っているのかを判断する。

4. 結果

①投稿取得

2003 年 1 月 1 日から 2008 年 12 月 31 日までの期間における取得可能な東証一部上場銘柄 1501 社の投稿を取得した結果、2,950,002 件の投稿を取得した。

②TF・IDF 値の算出

TF・IDF 値を算出し、全投稿における単語の重要度を求めた。そして、主成分分析において使用する上位 700 単語を決定した。表 1 に一般辞書と経済専門辞書における TF・IDF 値の合計が高い単語 15 単語を表 1 に示す。

表 1 を見ると一般辞書で抽出されていた「うう」「Afa」といったようなノイズを思われる単語が除去されていることが確認できた。

③主成分分析

主成分分析を行い、経済専門辞書を用いた主成分と一般辞書を用いた主成分を比較する。表 2 は、経済専門辞書を用いた主成分のうち第 10 成分までを示している。同様に表 3 は一般辞書を用いた主成分のうち第 10 成分までを示している。経済専門辞書を用いた結果、社会で話題になっているニュース記事に関する主成分が抽出できていることを確認できた。また、一般辞書で抽出されている「おたく」「じじい」といった直接的には社会事象に関係ない用語が削除されていることが確認できた。

5. 考察

経済専門辞書を用いた結果、ニュース記事に関する主成分が抽出できた。このような主成分が抽出された理由としては、今回用いた経済専門辞書が日経ソー

表 1 TF・IDF の合計が大きい上位 15 単語

順位	経済専門辞書	一般辞書
1	年齢	平均年齢
2	年収	平均年収
3	結婚	帝王
4	株式会社	うう
5	水素	利益狙い
6	利益	申し込み
7	公務員	Afa
8	亜鉛	ぴん
9	喫煙	保有割合
10	副大臣	作戦
11	周波数	保有時価総額
12	検査	容疑
13	電池	被告
14	離婚	大笑い
15	企業情報	最適

ラスであり、経済専門辞書とはいえ社会ニュースに関する用語を含んでいるためと考える。今後、株式リターンとの相関を確認し、今回抽出した主成分が株価を説明する可能性があるか検討したい。

また、その結果に基づいて、より経済用語に特化した辞書を作成すべきか検討したい。

6. 結論

本稿では、Yahoo!株式掲示板に対する2,950,002件の投稿に対して、経済専門辞書に基づく形態素解析と主成分分析を行い、一般辞書を用いた場合との比較を行った。その結果、経済専門辞書を用いることにより社会ニュースに関連する主成分を抽出できることを確認している。

今後は、抽出した主成分と株式リターンとの相関を確認し、株式リターンを説明する成分の抽出精度向上を図ることが課題である。

参考文献

- [1]Daniel, Gruhl. R, Guha. Ravi, Kumar. Jasmine, Novak. Andrew, Tomkins. (2005): "The predictive power of online chatter", The eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp.78-87.
- [2]Gilad, Mishne. (2006): "Predicting movie sales from blogger sentiment", In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.
- [3] 松井藤五郎, 石田智也, 中嶋啓浩, 和泉潔, 吉田稔, 中川裕志, (2010): "新聞記事を対象とした時系列テキスト分析による市場予測", 第7回 人工知能学会 ファイナンスにおける人工知能応用研究会.
- [4] Johan, Bollen. Huina, Mao. Xiao-jun, zheng. (2011): "Twitter mood predicts the stock market",

Journal of Computational Science, Vol.2, No.1, pp.1-8.

- [5]和泉潔, 後藤卓, 松井藤五郎, (2010): “テキスト情報による金融市場変動の要因分析”, 人工知能学会論文誌, 25 巻, pp.383-357.
- [6]Van, Bommel. (2003): “Rumors” Journal of Finance, Vol.58, No.4, pp.1499-1519.
- [7]Antweiler, W. M, Z, Frank. (2002): “Internet Stock Message Boards and Stock Returns”, University of British Columbia Working Paper.
- [8]丸山健, 梅原英一, 諏訪博彦, 太田敏澄, (2008): “インターネット株式掲示板の投稿内容と株式市場の関

係”, 証券アナリストジャーナル, Vol.46, No.11-12, pp.110-127.

- [9]諏訪博彦, 梅原英一, 太田敏澄, (2012): “ファクターモデルによるインターネット株式掲示板の投稿と株式リターンの分析”, 情報処理学会論文誌, Vol.53, No1, pp.117-125.
- [10]清水直樹, 諏訪博彦, 梅原英一, 太田敏澄, (2011): “インターネット株式掲示板の投稿内容に基づく新たな指数の作成”, 第17回社会情報システム学シンポジウム学術講演論文集, pp.67-72.

表2 経済専門辞書を用いた主成分

主成分1		主成分2		主成分3		主成分4		主成分5	
年齢	0.915	結婚	0.754	利益	0.762	売上高	0.73	副大臣	0.784
年収	0.912	選挙	0.658	結果	0.738	営業利益	0.71	大卒	0.754
背任	0.775	女性	0.560	げた	0.699	経常利益	0.68	企業再生	0.614
不買運動	0.752	離婚	0.544	銘柄	0.536	純利益	0.46	離婚	0.486
政治家	0.484	妊娠	0.472	資金	0.452	営業利益率	0.35	結婚	0.291
純利益	0.333	北海道	0.272	信用	0.282	一株利益	0.22	東京	0.144
放送	0.317	政治家	0.251	情報	0.115	営業収益	0.16	事実婚	0.144
電力	0.243	政治	0.238	利回り	0.079	決算	0.14	北海道	0.139
自動車	0.229	事実婚	0.215	投資	0.039	修正	0.11	国会議員	0.132
家電	0.183	訴訟	0.206	企業	0.028	中期経営計画	0.11	弁護士	0.118
主成分6		主成分7		主成分8		主成分9		主成分10	
恐喝	0.757	鑑定	0.726	女子大学	0.753	放火	0.794	太陽	0.861
売春	0.644	出産	0.673	事実婚	0.735	入院	0.735	発電	0.764
逮捕	0.494	妊娠	0.637	学歴	0.575	窃盗	0.508	電池	0.436
東京	0.475	子供	0.351	大学	0.400	犯罪	0.265	風力発電	0.281
父親	0.352	週刊誌	0.218	出向	0.261	懲役	0.264	補助金	0.137
出産	0.162	殺人	0.216	東京	0.139	週刊誌	0.227	共同開発	0.135
妊娠	0.157	結婚	0.199	女性	0.121	警察	0.205	電力	0.130
事件	0.135	監視	0.160	航空	0.109	病院	0.194	増産	0.121
個人情報	0.131	電話	0.135	離婚	0.095	個人情報	0.147	家庭	0.113
国会議員	-0.107	恐喝	0.134	心理学	0.090	医療	0.093	世界	0.099

表3 一般辞書を用いた主成分

主成分1		主成分2		主成分3		主成分4		主成分5	
びん	0.939	貸借値段	0.936	平均年収	0.933	賞賛	0.748	保有時価総額	0.820
作戦	0.927	貸借比率	0.912	平均年齢	0.927	人格障害	0.725	時価総額	0.714
最適	0.910	回転日数	0.871	おたく	0.777	親戚	0.720	累計	0.659
日間	0.868	差引	0.841	不買運動	0.746	他者	0.717	残高	0.622
前日終値	0.862	申し込み	0.728	既得	0.740	嫉妬	0.621	取引	0.536
基準	0.830	決済	0.694	市場価格	0.733	夫婦	0.599	前週	0.487
利益狙い	0.792	不足	0.640	じじい	0.708	自作自演	0.536	返済	0.285
過去	0.729	融資	0.614	癒着	0.698	自己	0.511	新規	0.272
実施回数	0.694	約定	0.546	純利	0.490	自分	0.413	融資	0.211
買値	0.615	残高	0.528	国民	0.285	脅迫	0.403	履歴	0.189
主成分6		主成分7		主成分8		主成分9		主成分10	
中略	0.773	売上	0.685	経済	0.703	保有割合	0.794	不倫	0.728
妊娠	0.735	経常利益	0.644	金融	0.656	共同保有	0.742	国会議員	0.694
結婚	0.722	営業利益	0.643	株式	0.597	受付	0.714	中絶	0.513
週刊	0.395	前年同期	0.542	投資	0.569	報告	0.709	子供	0.467
選挙	0.330	前期	0.424	投稿	0.528	提出	0.135	父親	0.464
売春	0.325	従来予想	0.346	商業	0.476	証券取引	0.075	週刊	0.318
愛人	0.315	四半期	0.331	製造	0.430	金融	0.042	離婚	0.286
離婚	0.310	利益	0.328	不動産	0.179	株券	0.041	妊娠	0.222
政治	0.280	当期	0.273	保険	0.120	総数	0.029	結婚	0.214
中絶	0.277	部門	0.217	通信	0.118	機種変更	0.028	捏造	0.158