テキストマイニングを利用したノイズトレーダーの行動予測に関する研究

岡田克彦（関西学院大学）　羽室行信（関西学院大学）　森田裕之（大阪府立大学）

- **研究背景と動機**
  - 株式価値は企業の将来期待キャッシュフローの現在価値であり、将来に関する情報はランダムに発生するため、株価動向の将来の予測は不可能。
  - 株式投資家の中にはこうした構造を理解せず、存在しない法則性を信じて売り買いを繰り返すnoise traderと合理的に情報だけで投資するrational traderが存在する。
  - 平常時、noise traderの動向はお互い相殺されているが、市場のセンチメントがある方向に傾くと、その動向も一方向に集中
  - Finance 分野のliteratureによって、株価を長期的に決めるものはファンダメンタルズであるが、短期的にはnoise traderによって振れることが知られている。
  - 市場のセンチメントを指数化できれば、noise traderの行動を予測する一助となる。
  - 市場関係者の間に広く読まれている日経新聞記事をテキストマイニングし、市場のセンチメント指数を開発する。

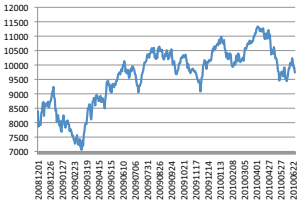
利用データ

・新聞記事の取得：日経テレコン21を利用
　☞2008年12/1～2010年6月25日の約1年半の朝刊
　　1～10面までの記事全てを取得

・株価データ：日経平均225の日別終値

	2008/12～2010/6 1年半データ
記事数	27,757
文章数	333,966
同定対象文節数	380,189
単純エントリ種類数	15,762
複合エントリ種類数	347,586

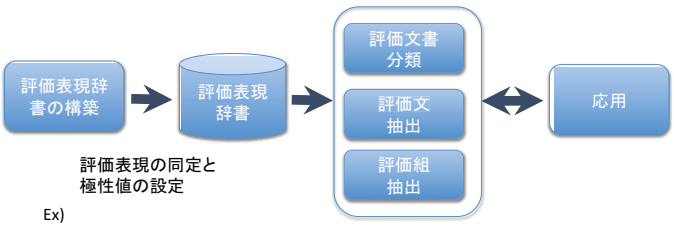
解析対象とした新聞記事に関する統計



日経平均225の推移

テキストを対象とした評価情報の分析に関する研究動向（乾,奥村 2006）

評価情報：個人の評価に関する情報
評価極性：評価情報の良い/悪いに関する軸（肯定と否定）
評価表現：肯定極性か否定極性をもつ評価情報がテキスト内で記述された表現



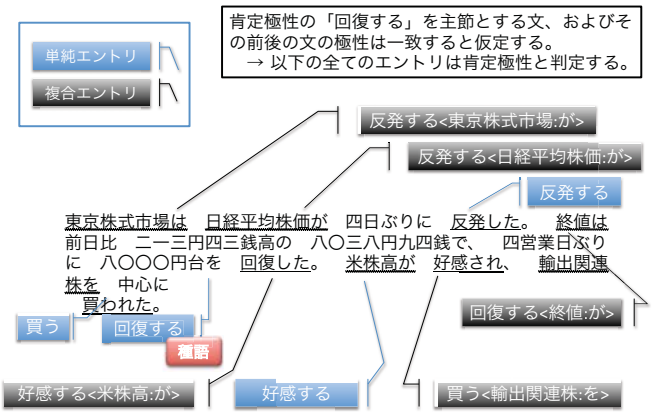
評価表現の同定と極性値の設定

Ex)
回復する：肯定極性
下落する：否定極性

周辺文脈法(那須川・金山2004)における評価表現辞書構築の流れ

1) 種語を初期辞書として用意する(どちらかの極性の言葉一つでもよいし、両極性複数用意してもよい)
2) 評価表現を同定する。
3) 種語の周辺文脈から極性付きで評価表現候補を抽出する。
4) 評価表現候補が評価表現であるかどうかを判定する。
5) 評価表現と評価された語を辞書に追加する。
6) 追加する語がなければ終了する。
7) 辞書に登録された語を種語として2)の手順に戻る。

単純エントリと複合エントリの実例



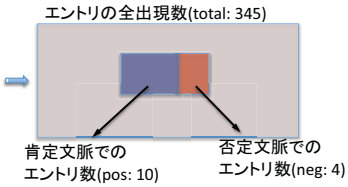
評価表現候補の抽出

業績が一時的に悪化している。しかし、資産算定の甘さが指摘されており、信頼は回復していない。会社側は赤字の子会社を整理する計画である。

	同定用言	単純エントリ	複合エントリ	属性	辞書極性	文節極性	用言極性
業績は						+1	
一時的に						+1	
改善している。	○	改善する	改善する<業績:が>			+1	+1
しかし、				逆接		-1	
資産算定の						-1	
甘さが						-1	
指摘されており、		指摘する	指摘する<甘さが:が>			-1	-1
信頼は						-1	
回復していない。	○	回復する	回復する<信頼:が>	否定	+1	-1	+1
会社側は						-1	
赤字の						-1	
子会社を						-1	
整理する		整理する	整理する<子会社:を>			-1	-1
計画である。	○					-1	

候補表現の判定

エントリ	pos	neg	total
一巡する	3	0	41
上昇	10	4	345
備える	0	2	350
上回る	12	8	1244
:	:	:	:



条件1. エントリの出現数が一定以上
total ≧ *minSupp*

条件2. 極性の割合が一定以上

pos/(pos+neg) ≧ *minPolRate* or neg/(pos+neg)≧ *minPolRate*

条件3. 二項分布B(total,π)においてpos(もしくはneg)が十分に大きいかどうか

πの推定: 辞書に登録されたエントリの平均誤り率
肯定極性エントリにおいては neg/total
否定極性エントリにおいては pos/total

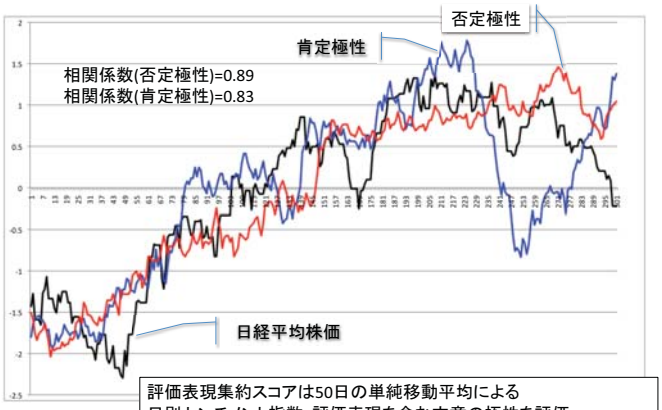
* ただし、今回の実験では、π=0.05に固定

Pr(x>pos)≧α （0.01,0.05,0.1で実験）

得られたエントリー覧 （青字は種語）

	肯定極性 (76エントリ)	否定極性 (70エントリ)
SE	上方修正する、反発する、回復、終える、更新する、好感する、買う、上昇する、緩む、上方修正、回復する、付ける、続伸、優勢	下方修正、下方修正する、悪化する、漂う
CE	公表する<日銀:が> なる<黒字:に> 更新する<高値:を> 上昇する<円:が> 上回る<市場予想:を> 反発する<日経平均株価:が> 広がる<中心:に> 広がる<安心感:が> 終える<上海総合指数:が> なる<優勢:と> 付ける<高値:を> なる<終値:が> 推移する<ダウ平均:が> 上昇する<株価:が> なる<安値:と>	目立つ<不振:が> 拡大する<赤字幅:が> 漂う<先行き:に> よる<雇用統計:に>

日別センチメント指数と株価推移



評価表現集約スコアは50日の単純移動平均による
日別センチメント指数：評価表現を含む文章の極性を評価
(肯定極性は+1、否定極性は-1)し、その日の出現数を計算したもの。

収益率を目的変数としたVARモデル

$$Nikkei_t = \alpha_1 + \beta_1 \cdot L4(Nikkei_t) + \gamma_1 \cdot L4(neg_t) + \delta_1 \cdot L4(Vlm_t) + \lambda_1 \cdot dummy1_{t-1} + k_1 dummy2_{t-1} + \varepsilon_{2t}$$

$$neg_t = \alpha_2 + \beta_2 \cdot L4(Nikkei_t) + \gamma_2 \cdot L4(neg_t) + \delta_2 \cdot L4(Vlm_t) + \lambda_2 \cdot dummy1_{t-1} + k_2 dummy2_{t-1} + \varepsilon_{2t}$$

但し,
◆Nikkeiは当日の日経平均株価指数の日次収益率
◆L4はラグオペレーターであり, L4(Xt)=[Xt-1,Xt-2,Xt-3,Xt-4]
を意味する.
◆Dummy 1 = 外生変数の1月ダミー
◆Dummy2=外生変数weekend dummy
◆Neg t = t日における全記事から導かれる負の極性値の合計値
◆Vlm t= t日における株式市場の出来高

実証結果

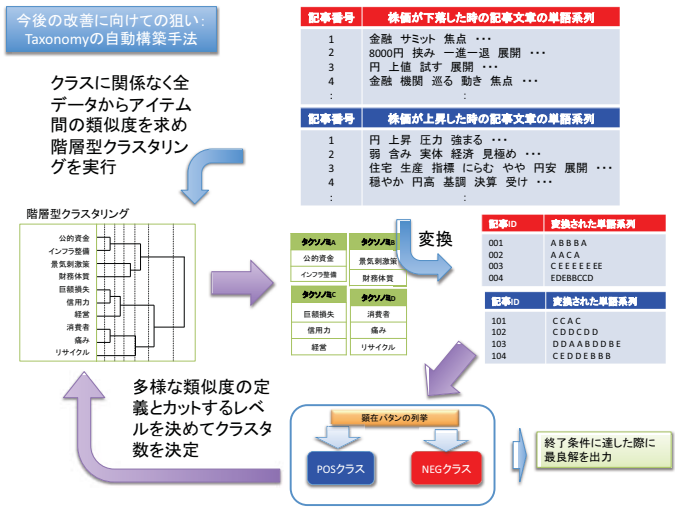
Model 1	Nikkei	t-value
Nikkei(-1)	-0.0567	-0.81
Nikkei(-2)	0.0386	0.54
Nikkei(-3)	-0.0512	-0.75
Nikkei(-4)	0.0017	0.25
Volume(-1)	0.0037	0.52
Volume(-2)	-0.0094	-1.19
Volume(-3)	-0.0017	-1.19
Volume(-4)	0.0003	0.04
neg(-1)	-0.0018	**-2.60**
neg(-2)	0.0017	**2.36**
neg(-3)	0.0003	-0.52
neg(-4)	-0.0002	-0.67
c	0.1534	1.05
dummy1	-0.0013	-0.39
dummy2	-0.0078	-1.39

Model 2	neg	t-value
Nikkei(-1)	6.965	-1.066
Nikkei(-2)	2.176	0.33
Nikkei(-3)	4.441	0.68
Nikkei(-4)	-14.689	-1.52
Volume(-1)	-3.392	**-2.15**
Volume(-2)	1.567	1.23
Volume(-3)	-0.069	-0.05
Volume(-4)	1.365	1.07
neg(-1)	0.256	1.31
neg(-2)	0.203	**1.81**
neg(-3)	0.041	1.49
neg(-4)	0.198	**2.27**
c	8.692	0.33
dummy1	1.416	**4.48**
dummy2	-0.784	**-2.50**

11

ー考察

- 市場価格と記事極性の相関を見た場合、正の記事極性の増減よりも、負の記事極性の増減の方が高いことがわかった。これは心理学の知見であるところの人間のnegative biasと整合的な結果である。
- 負の記事極性(neg)の時系列値をVector Auto Regression Modelで推定したところ，1日前の負の極性記事の係数が有意になることがわかった．これは前日の朝刊記事をマイニングすることで，翌日のリターンの予測が出来る可能性を示唆している．
- 負の記事極性(neg)の株式リターンに対するインパルス応答関数によると，負の極性記事に1標準偏差のショックが加わると，その効果は2日後まで残存するが，3日後にはほぼ元の水準に戻った．これはnoise traderのセンチメントに呼応する動向が一時的に株価に影響を与えるが，ファンダメンタル価値からの乖離はrational traderによって修正されるという仮説と整合的である．



新聞記事おける顕在(系列)パターンの例

記事番号	単語系列
1	円 上昇 圧力 強まる …
2	弱 含み 実体 経済 見極め …
3	住宅 生産 指標 にらむ やや 円安 展開 …
4	穏やか 円高 基調 決算 受け …
:	:

株価が上昇した時の記事

記事番号	単語系列
1	金融 サミット 焦点 …
2	8000円 挟み 一進一退 展開 …
3	円 上値 試す 展開 …
4	金融 機関 巡る 動き 焦点 …
:	:

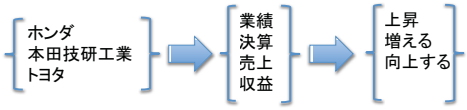
株価が下落した時の記事

顕著な単語の出現パターンの違いを列挙する。

パターンXが当該クラスに顕著であるかどうかを増加率GRで判断する。
GRとは各クラスにおけるパターンXのサポート比のこと。

$$GR_{D_{neg}\to D_{pos}}(X)=\frac{supp_{pos}(X)}{supp_{neg}(X)}$$

最終的には以下のようなtaxonomyのシーケンスが列挙される事を期待する。



Taxonomyの自動構築におけるZDDの利用