取引高とニュース記事の関連性の分析

吉田 稔[1*]　中川裕志[1]
松井 藤五郎[2]　和泉 潔[3]　石田 智也[4]　中嶋 啓浩[4]

[1] 東京大学情報基盤センター
[1] University of Tokyo
[2] とうごろう機械学習研究所
[2] Tohgoroh Machine Learning Research Institute
[3] 産業技術総合研究所
[3] National Institute of Advanced Industrial Science and Technology
[4] 野村證券株式会社
[4] Nomura Securities Co.,Ltd.

Abstract: This paper reports our research in progress on relations between volume of transactions and related news articles. Our current goal is to make a system to predict volume of transactions of a brand from the news articles related to the brand. Our algorithm clusters news articles by LDA and predicts whether the volume of transactions at a target day will increase or decrease.

1　はじめに

本稿では，現在我々が研究を進めている，テキストと株の取引高の関連分析について紹介を行う．

テキストと数値的情報の関連については，これまで，評価記事と評価値の関係 [4]，記事とアクセス数の関係 [6] の分析が提案されており，株価に関しても，将来の株価予測への応用を目指した「テキストと株価の関係」に関する研究が盛んに行われるようになっている．例えば，小川ら [5] は，新聞記事をルールベースでテーマ分類し，テーマが株価動向にどのような影響を及ぼすかを解析した．高橋 [7] らは，ヘッドラインニュースを情報源とし，Naive Bayes 法により分類されたニュースの Good/Bad のラベルと，ニュース配信時の株価リターンとの関連を調査し，有意な関連があったと報告している．また，和泉 [1] らは，日本銀行の金融経済月報を題材として経済市場分析を試みている．単語共起関係抽出ツール KeyGraph[3] を用い，抽出された共起パターンと月末における金利の関係を主成分分析によって解析し，金融経済月報が市場金利に対して，一定の説明力を持つ可能性が高いことを示した．また，[2] においては，国際金融情報センターの発行する市場解説記事を自動分類した結果を，人工市場の分析に利用する試みを行っている．張 [8] らは，株価の変動を記事や語句の評価値の推定に用い，係り受け関係を使うことで良好な結果を得ている．

本研究では，個別銘柄と，それに関する新聞記事の関係を分析することを目的とするが，テキストからの株価予測に関しては，高い精度で実現することの難しさが既存の研究でも指摘されている [11]．このため，本研究では，最初の目標として，比較的予測が容易と考えられる,「取引高」に着目し，新聞記事が与えられたときに，その日の取引高の高低を予測するシステムの実現を目指す．

2　問題設定

入力として，

D_t: ある銘柄に関する，日付 t の記事集合

v_t: ある銘柄の，日付 t の取引高

が与えられるとする．ここで，取引の無い日付については，D も v も定義されず，t は，日付そのものではなく，取引の有った日付を古い順に並べた順番を表すものとする．[1]

ここで，取引高 v_t に対し，前日 N 日と比較しての増加傾向，減少傾向を表す値 y_t を，以下で定義する．

$$y_t = \frac{v_t}{a_t}$$

*連絡先：東京大学情報基盤センター
〒 113-0033 文京区本郷 7-3-1
E-mail: mino@r.dl.itc.u-tokyo.ac.jp

[1] 取引の無い日の記事は，翌取引日の t に対応づけられるものとする．

ただし，

$$a_t = \frac{\sum_{t-N \leq t' \leq t-1} v'_t}{N}$$

（現在は，$N = 5$ を設定している.）本研究の目的は，D_t が与えられたときに y_t に関する予測を行うことであるが，問題設定としては，値の大小を 2 値分類することを考える．y_t は比率のため，1.0 より大きければ取引高の「増加傾向」，小さければ取引高の「減少傾向」を表していると考えられる．すなわち，

$$z_t = sign(y_t - 1.0)$$

を定義し（ここで $sign(x)$ は，$x \geq 1.0$ ならば +1，さもなくば −1 を返す関数とする．[2]），z_t の予測を行う．

すなわち，本稿の提案システムのタスクは，入力 $\mathcal{D} = (D_1, D_2, \ldots)$ と $\mathcal{V} = (v_1, v_2, \ldots)$ が与えられたときに，$\mathcal{Z} = (z_1, z_2, \ldots)$ を返すことである．

3 予測手法

本研究では，SVM による二値分類を $\mathcal{D}$ を入力として行うことにより，$\mathcal{Z}$ の予測を行う．以下，具体的な手法について述べる．

3.1 使用素性

各記事のタイトルと本文を CaboCha[12] の解析結果を利用し単語に分割する．現在，記事の素性としては，本文からは，単語ペア素性を用いるが，タイトルからは単語素性を抽出している．これは，タイトルは単語数が少なく，単語ペア素性を用いることがデータスパースネス問題を起こす一方，タイトルは内容を簡潔に表しており,「タイトル中のある単語の有無」が十分に話題を特定できるという観察に基づく措置である．

3.2 素性選択

取引高の予測において，我々は,「アラートとなる記事」の存在を仮定する．すなわち，記事の中にも，決算発表や事件等，取引高を増加させるような大きなニュース記事と，新製品紹介等の影響の小さな記事が存在すると考える．このことは，記事素性の中にも，取引高を増加させるような素性（「黒字」「リコール」等）と，それ以外の素性が存在するということになる．

しかしながら，逆に,「取引高を減少させるような素性」というのは，一般的には考えづらい.「影響の小さな単語」が記事中に存在していたとしても，記事中の他の単語が「影響の大きな単語」であれば，記事の影響は大きくなるためである.「投資家を様子見させるようなニュース記事」の存在は有り得るものの，一般的にはその存在は少ないと考えられる．このため本研究では，記事（記事タイトル）中の各単語 w を選別し，アラートなる可能性の高い単語のみを素性として選択する．

訓練データにおいて，w の登場する記事の，プラス記事・マイナス記事の数を $plus(w), minus(w)$ とする．このとき，

$$\frac{plus(w)}{(plus(w) + minus(w))} > \alpha$$

となる w だけを残すこととする．α は，素性選択のパラメータであり，現在は $\alpha = 0.7$ としている．

さらに残った w を記事頻度でスコア付けし，スコア上位 M 個の w のみを残す.（現在は，$M = 10$ としている.）

最終的に残った w のみを記事特徴量として用いる．このとき，どの w も含まない記事は「特徴のない記事」として 0 ベクトルになる．

3.3 記事クラスタリング

本研究では，記事をクラスタリングし，各クラスタ内で取引高の予測を行う．これには，以下の 2 つの目的がある．

精度の向上 ある記事に関する取引高を予測する際，学習に用いる記事として，対象記事以外のすべての記事を用いることも考えられるが，例えば工場建設に関する記事の中で,「稼働」という単語が取引高へ影響することがわかったとしても，その知見が決算発表の記事に関して同様に使えるとは考えづらい．このため，予め類似の話題の記事をクラスタにまとめ，同一クラスタ内の記事のみを学習に用いることにより，速度や精度の向上を目指す．

可視性の向上 例えば決算発表のように，取引高への影響が大きいと考えられる話題と，新製品紹介のように，影響が小さいと考えられる話題について，実際に取引高の予測精度を比べることにより，どんな話題が「取引高への影響」を大きく持つかについてある程度の知見を得ることができる．

分類アルゴリズムとしてはSupport Vector Machine を用いる．カーネルとしては RBF カーネルを使用した．クラスタリングには，K-means 法を用いた．

[2]定義から，y_t が厳密に 1.0 となる可能性は低いため，ここでは $z_t = 0$ となるケースは考えていない.

3.3.1 Latent Dirichlet Allocation によるトピック推定

クラスタリングの際の文書素性としては，単語頻度を用いるが，本研究ではさらに，各単語のトピックを推定し，このトピックの比率についても文書の素性として用いる．例えば同じ「増加」という単語が用いられていても,「賃金の増加」か,「売上高の増加」か,「採用数の増加」かで，意味は異なる．文書のトピック推定を用いることにより，このような複数の話題に跨る単語の影響により，複数の話題が同一クラスタとなることをある程度防ぐことが期待できる．

Latent Dirichlet Allocation (LDA) [9] は，文書等の，スパースなベクトルを効率良くモデル化するための生成的確率モデルである．LDA においては，文書中の各単語は,「トピック」（の混合分布）から生成されると仮定される．ここでトピックとは，単語に対する多項分布であり，例えば，決算に関するトピックは「赤字」や「今期」といった単語に高い確率を与える分布，新製品に関するトピックは「発売」や「価格」といった単語に高い確率を与える分布として表現される．LDA を用いることにより，文書中の各単語に対し，その単語がどのトピックから生成されたかの推定値（トピックの番号）が与えられる．我々は文書毎にこの番号を集計し，各文書のトピック分布とする．なお，トピックの推定には Collapsed Gibbs Sampling[10] を用いた．

4 実行例

対象となる銘柄の取引高と，各銘柄名を見出しに持つ新聞記事（2005-2007年日経新聞の記事）を，日付で対応付け，記事の存在する日付について，z_t の予測を行った．学習は leave-one-out 法で行った．すなわち，日付 t について z_t の予測を行う際，t 以外の全ての日付を学習データとして用いた．対象銘柄「ソニー」と「本田技研工業」[3]について，結果の例を示す．

K-means 法のクラスタ数は 20，LDA のトピック数は 30 とした．表 1 に，ソニーに関する記事から推定されたトピック（それぞれ，確率値の高い上位５単語を表示），表 2 に，文書クラスタと，各クラスタにおける取引高推定の結果を示す．決算発表（表中「%」）や電池回収問題（表中「電池」）等のニュースにおいて F-value が高く，新製品紹介等のニュース（表中「ニューフェース」）においては F-value が低い傾向があった．

また，同様に、表 3 と表 4 に，ホンダに対する結果を示す．こちらも，決算発表（表中「円」）に関して F-value が高くなる傾向が観察された．

[3]記事検索のキーワードは「ホンダ」とした．

表 1: トピック（ソニー）

番号	上位５語 (句点やハイフンは除いて表示)
0	デル, 用, プレーヤー, 強,
1	円, 億, 今期, 不振, 利益,
2	型, 工場, 半導体, 生産,
3	会長, 経営, 陣,
4	液晶, 開発, サムス, ン,
5	ン, エリク, 拡大, テレビ,
6	会社, 人事, イーエムシーエス, 保険, コンピュータエンタテインメント,
7	発売, ＰＳ, 社, ソフト,
8	マン, ウォーク, 投資, 技術, 活用,
9	テレビ, パソコン, 台,
10	音楽, 事業, 携帯, 売却,
11	強化, ネット, 営業, 率,
12	人, 企業, 取締役, 株主,
13	東芝, 米, 次世代, 日, 株,
14	%, マーケティング, 利益, 営業,
15	万, 機, ゲーム, 欧州,
16	銀, 金融, フラッシュ, 銀行, 提携,
17	ニューフェース, 対応, ハイビジョン, ＩＣ,
18	回収, 米, パソコン, 電池,
19	製, 電池, 万, 発火, 問題,
20	化, デジタル, 世界, 割, 増,
21	九州, セミコンダクタ, 松下, 円,
22	向け, 子会社, コン, 可能,
23	ＴＶ, 化, 薄型,
24	部門, 販売, 映像,
25	氏, 社長, 聞く, ＣＦＯ,
26	氏, 改革, 視点, 採点,
27	配信, 系, 米, 映画, カード,
28	デジカメ, 中国, 回復, 市場,
29	次世代, ＤＶＤ, 取得, 方式,

5 おわりに

取引高とテキストの関係分析について，現状の報告を行った．クラスタリングと分類を組み合わせることにより，取引高と関係の強い話題，関係の弱い話題があることが確認された．今後の予定としては，より多様な銘柄の分析，クラスタリングや素性選択による分類精度の変化の観察等を行う予定である．

参考文献

[1] 和泉潔，後藤卓，松井藤五郎. テキスト情報を用いた金融市場分析の試み. 人工知能学会第 22 回全国大会 (2008)

[2] 和泉潔，松井宏樹，松尾豊．人工市場とテキストマイニングの融合による市場分析．人工知能学会誌, Vol. 22, No. 4, pp. 397-404 (2007)

[3] 大澤幸生, ネルス E. ベンソン, 谷内田正彦．KeyGraph. ：語の共起グラフの分割・統合によるキーワード抽出．電子情報通信学会論文誌 D-1, Vol.J82-D-1, No.2, pp.391-400, (1992)

表 2: 記事クラスタと分類精度（ソニー）

タイトル高頻度語	記事数	Accuracy	Precision	Recall	F-measure	素性の例
%, 円, 利益,	183	0.60	0.62	0.74	0.68	利益:営業, ○:し, 五:円, テレビ:液晶,
ニューフェース, 対応,	81	0.46	0.31	0.29	0.30	モニター:液晶, 万:ＣＣＤ, ＤＳＣ:Ｔ,
配信, 映画, 系, 米,	33	0.61	0.64	0.44	0.52	東京:港, する:配信,
氏, 社長, 中鉢, 聞く,	60	0.50	0.54	0.47	0.50	する:企業, し:社長, 九:八, さ:せ, 売上:高,
音楽, 携帯, 事業,	44	0.56	0.58	0.52	0.55	アップルコンピュータ:米, 億:十,
ブルー, レイ, 万,	41	0.62	0.65	0.65	0.65	五:%, 利益:営業, 億:円, し:の, 三月:年,
経営, 会長, 陣,	42	0.58	0.57	0.50	0.53	会長:兼, ○:二, ストリンガー:会長,
会社, 人事, ...	103	0.47	0.11	0.02	0.04	エレクトロニクス:役,
金融, フラッシュ, 損保,	19	0.32	0.00	0.00	–	する:ソフト, さ:れる, ロサンゼルス:猪瀬,
電池, 製, 回収, パソコン,	53	0.51	0.61	0.71	0.65	い:し, 製, し:する, いる:し, 製:電池,
ＢＭＧ, ダイジェスト, 米,	29	0.48	0.22	0.20	0.21	万:千, 六:同期, 七:比, 三:千, 一:万,
円, 億, 上場,	31	0.62	0.70	0.80	0.74	し:発表, 億, ○:二, ○:年, 億:百,
技術, ウォーク, マン, ...	26	0.31	0.00	0.00	–	億:円, ○:し, 以来:年, し:ブランド,
銀, 黒字, 行, 金融,	37	0.60	0.58	0.65	0.61	二:専業, 五:億, し:経常, 二:四, 五:年,
研究所, 開発, 水, 知能,	27	0.62	0.50	0.40	0.44	二:二, し:二, 三:九, 五:千, し:今回,
次世代, ＤＶＤ, 東芝, ＰＳ,	80	0.60	0.59	0.64	0.61	五:百, 三:百, ロサンゼルス:猪瀬,
液晶, ＴＶ, 工場,	167	0.57	0.50	0.37	0.42	二:六, 一:二, 二:八, テレビ:用, する:家電,
氏, 視点, 人, 採点,	45	0.69	0.55	0.43	0.48	人:外国, する:事業, トップ:人,
人事, 会社, ...	46	0.51	0.40	0.10	0.15	優:加藤, 日:本部, 取締役:敏明, 敏明:石野,
取締役, 人, 株主,	46	0.59	0.67	0.43	0.53	最高:者, 者:責任, 最高:責任, 経営:者,

[4] 岡野原 大輔, 辻井 潤一., "レビューに対する評価指標の自動付与", 自然言語処理. 14(3). pp. 273–295, (2007)

[5] 小川 知也，渡部 勇．株価データと新聞記事からのマイニング．情報処理学会 自然言語処理研究会研究報告 2000-NL-142-19 (2000)

[6] 沢井 康孝, 山本 和英. 文書に対する大衆の興味の強さの推定. 自然言語処理, Vol.15, No.2, pp.101-136 (2008)

[7] 高橋悟，高橋大志，津田和彦．ヘッドラインニュースに対する株価の反応について．第 6 回行動経済学ワークショップ. (2007)

[8] 張 へい，松原 茂樹，株価データに基づく新聞記事の評価，第 22 回人工知能学会全国大会論文集, (2008)

[9] D.M.Blei, A.Y.Ng, M.I.Jordan, "Latent Dirichlet Allocation" JMLR, vol.3, pp.993-1022 (2003)

[10] Griffiths, T. L., Steyvers, M. A probabilistic approach to semantic representation. In Proceedings of the. 24th Annual Conference of the Cognitive Science Society. (2002).

[11] Moshe Koppel and Itai Shtrimberg. Good News or Bad News? Let the Market Decide. Computing Attitude and Affect in Text: Theory and Applications, 297-301 (2006)

[12] Taku Kudo and Yuji Matsumoto, Japanese Dependency Analyisis using Cascaded Chunking, Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002), pp.63–69 (2002)

表 4: 記事クラスタと分類精度（ホンダ）

タイトル高頻度語	記事数	Accuracy	Precision	Recall	F-measure	素性の例
円, %, 億, 最高,	44	0.59	0.63	0.88	0.73	二:日, 五:千, 円, 七:十, 万:千,
工場, 埼玉, 着工, 株主,	15	0.79	0.80	0.67	0.73	○:一, する:一, 三:千, 一:万, 六:十,
取得, 設定, 枠, 株,	15	0.20	0.29	0.22	0.25	万:八, 億:六, 万:円, 万:百, 二:六,
車, 生産, 輪, 工場,	23	0.48	0.20	0.11	0.14	十:日, 二輪車:生産, 七:万, 二:工場, 十:年,
台, 万, リコール, 生産, 計画,	15	0.67	0.60	0.86	0.71	なる:恐れ, リコール:生産, 修理:国土,
工場, 米, 企業, シビック,	21	0.48	0.44	0.40	0.42	さ:れ, 万:六, 三:年, 一:九, ○:し,
販売, 系, 販社,	26	0.61	0.42	0.71	0.53	い:れ, 情報, い:さ, ピックアップ, 二:減少,
系, 子会社, 会, 企業,	16	0.60	0.00	0.00	–	する:進出, 一:億, する:九月, 億:千,
工場, 生産, 部品, 熊本,	41	0.41	0.32	0.38	0.34	し:建設, 三:工場, 七:千, 二:町, 可能:性,
リコール, 社, 車種, 乗用車,	11	0.18	0.00	0.00	–	九:六, シオン:二, 七:九, 四:昨年, 九:十,
車, トヨタ, 中国,	90	0.57	0.39	0.23	0.29	五:六, 五:四, し:市場, 自動車, さ:れる,
ジェット, 機, 米,	17	0.47	0.38	0.43	0.40	し:開発, し:設立, 小型:米国, 機:開発, し:し,
賃金, 春, 停止,	15	0.64	0.71	0.62	0.67	三:二, ・:三, 十:日, 十:四, ・:二,
万, 台, 生産,	121	0.55	0.46	0.28	0.35	五:六, 八:日, 性能:燃費, する:メーカー,
中国, 現代, 外資, 材料,	21	0.25	0.14	0.10	0.12	五:百, 二:五, 五:年, 億:百, する:二,
人事, 会社, 東, 鈴鹿,	12	0.67	0.00	0.00	–	三月:連結, 億:円, し:二, ○:年, 二:十,
ニューフェース, トヨタ, ...	17	0.82	–	0.00	–	し:トヨタ, ○:年度, ○:し, 六:年, ぶり:年,
社長, 氏, 福井, 威夫,	25	0.43	0.54	0.50	0.52	し:し, し:する, メーカー:自動車,
車, 電池, 燃料,	24	0.46	0.14	0.12	0.13	し:導入, サービス:始める, し:今年,
リオ, 分売, 東, 外, ク,	11	0.82	–	0.00	–	コンテスト:年生, ロボット:全国, 三:作品,

表 3: トピック（ホンダ）

番号	上位 5 語 (句点やハイフンは除いて表示)
0	技術, 福井, 環境, 企業,
1	勢, 日本, 急ぐ, 挑戦, 一,
2	倍, 株式, 分割, 前期,
3	電池, トヨタ, 部品, 新型,
4	二輪車, 体制, 増強, 各社,
5	万, 台, 自動車, 輸出, 計画,
6	機, 製作所, 小型, ジェット, 参入,
7	取得, 整備, 研究, 設定,
8	開発, 化, 子会社, 系, 産業,
9	ニューフェース, 会社, 用,
10	生産, 車, 燃費, 輪,
11	日産, ．, 金, 最高, 検討,
12	社長, 氏, 供給, 威夫, 者,
13	発表, 軽, フィット, 安,
14	円, 億, 最高, 株, ぶり,
15	%, 生産, 年, 中国, 販売,
16	販売, 国内, 鈴鹿,
17	工場, シビック, 米, 稼働,
18	車, 日本, 攻勢, 日産,
19	年, 車, ハイブリッド, 割,
20	工場, 向け, 増産, 熊本, 生産,
21	車, 燃料, 情報, 社,
22	販売, 販社, し, カーズ, 権,
23	増, 北米, 対応,
24	埼玉, 拠点, 高, 県, 拡充,
25	トヨタ, 小型車, インド, 市場, 新車,
26	リコール, 自動車, 乗用車, 寄居, 首位,
27	中国, 事業, 開始, 現代, 力,
28	中, 進出, 聞く, 改善,
29	人, 日, 米, 削減, 物流,