2H1-OS18-7

ニュース記事クラスタリングによる取引高予測の試み

Towards Prediction of Volume of Transactions Using Clustering of Related News Articles

吉田 稔[*1] Minoru Yoshida　中川 裕志[*1] Hiroshi Nakagawa　石田 智也[*2] Tomonari Ishida　中嶋 啓浩[*2] Akihiro Nakashima　松井 藤五郎[*3] Tohgoroh Matsui　和泉 潔[*4*5] Kiyoshi Izumi
池田 翔[*5] Sho Ikeda　本多 隆虎[*6] Takatora Honda

[*1]東京大学情報基盤センター Information Technology Center, The University of Tokyo
[*2]野村證券株式会社 Nomura Securities Co.,Ltd.
[*3]中部大学生命健康科学部, 工学部 College of Life and Health Sciences, and College of Engineering, Chubu University
[*4]東京大学大学院工学系研究科 School of Engineering, The University of Tokyo
[*5]JST さきがけ PRESTO, JST
[*6]早稲田大学大学院基幹理工学研究科 Graduate School of Fundamental Science and Engineering, Waseda University

We report our on-going research to develop a system to predict volume of transactions of a brand using news articles related to the brand. Our problem is defined as a binary classification problem where the task is to predict a label (going up or going down) of each (stock, day) pair given news articles in that day. We used topic models to estimate document clusters, and used the estimated clusters and labels in training data to predict labels in test data.

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．

「テキストと株価の関係」に関する従来研究としては，例えば，小川ら [2] 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 [3] 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．張 [4] 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．取引高に関しては，前日のインターネット掲示板の投稿数や投稿内容との関連があることが指摘されている [5] が，我々は，ニュース記事という，別種の情報源からの影響を考慮することで，より予測の網羅性が高められる（投稿数が多くない銘柄に対しても対応できる，等）と考えている．

本研究の想定する応用は，「その日の新聞記事を分析し，取引高が上昇する可能性が高い場合に知らせる」システムである．これにより，取引が行い易い日を予測し，効率的な株取引に役立てるというものである．しかしながら，実際には，取引高の動きは，それ自体変動が大きく，すべての取引高上昇日をテキスト分析から予測することは困難である．このため，我々の目標は，「すべての取引高上昇日を，その日のテキストから予測する」ことではなく，「新聞記事にこのような記事があれば，その日の取引高が上がる可能性が高い」という知見を自動的に学習することである．これは，情報検索の用語で言えば，再現率ではなく，適合率を追求することに相当する．

連絡先: 吉田稔，東京大学情報基盤センター，〒 113-0033 文京区本郷 7-3-1，Tel:03-5841-0340，Fax:03-5841-2745，E-mail: mino@r.dl.itc.u-tokyo.ac.jp

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．

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

ただし，

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

[*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．

(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，

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

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），s_t の予測を行う．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，$\mathcal{S} = (s_1, s_2, \ldots)$ を返すことである*3.

3. 記事クラスタリングによる取引高予想

本研究では，特に，記事の「話題」に着目する．これは，記事には，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と，新製品紹介等の影響の小さなニュースが存在するという考え方である．

アラートとなる話題を発見することにより，その話題に属する記事を自動的に発見し，精度良く「取引高上昇」の予測を行えるようになると考えられる．これは，記事のクラスタリングを行い，クラスタを話題と見なすことによって可能になる．例えば，「決算発表の記事が，高い取引高の兆候となる」という知識を得るためには，

- 「決算発表」という同一の話題の記事をまとめ，
- その訓練データ中の取引高の上昇／下降ラベルを調べ，
- もし取引高が上昇する傾向があれば，テストデータ中の同一の話題の記事も，「取引高が上昇する確率の高い」記事と見なす

というステップを踏めばよい．

逆に，「取引高を減少させるような記事」というのは，一般的には考えづらい．「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，我々は，記事を「取引高が上昇する確率の高い順に並べる」というタスクを設定する．もしも上位に表示された記事が実際に取引高を上昇させる確率が高ければ，「取引高を上昇させるような話題」が発見できたことになる．

実際のアルゴリズムは，次のようになる．

1. 同一の話題の記事がまとまるように，記事のクラスタリングを行う．
2. 各クラスタ毎に，クラスタの取引高上昇比率（s_t が正となる記事の割合）を，訓練データから計算する．
3. 各テストデータ（記事）の取引高上昇確率を，その記事が属するクラスタの取引高上昇比率として予想する．
4. 取引高上昇確率の高い順に，テスト記事をランク付けする．

処理の単位が，問題設定において定義された記事集合 D_t ではなく，記事一つずつとなっていることに注意されたい．ある日付の取引高上昇／下降の予測は，$d \in D_t$ となる文書のうち，最も高くランク付けされた記事 d に基づいて行う．これは，その日に影響力のある記事が一つでもあれば，他の記事の内容に関わらず，その日の取引高が上昇すると考えられるためである．これにより，影響力の強い順に記事をランク付けし，適当なしきい値で，「影響力の強い記事」を取り出すことができるようになる．提案手法は，クラスタリングをランキングに用いるという点で，情報検索における Scatter/Gather 法 [8] と同様の考え方によるランキングとなっている．

3.1 記事クラスタリング

本研究で予測に用いるのは，文書のタイトルのみである．これは，タイトルは，文書の話題を端的に表現しており，記事の話題の判定に極めて有用であるという観察に基づく．

各記事のタイトルを CaboCha[12] の解析結果を利用し単語に分割する．得られた単語のうち，名詞と動詞*4を，記事の特徴量として用いる*5．このとき，特徴量に一般性を持たせるため，対応する銘柄名を含む／銘柄名に含まれる単語は除く．

しかしながら，記事のタイトルを用いる問題点として，得られる特徴量（単語数）が少ないことが挙げられる．これは，記事クラスタリングにおいては，記事間で共通する単語が少なくなり，類似する記事の発見が困難となることにつながる．特に，同義語や類義語が存在する場合，例えば，異なる記事で，「利益」と「最終益」と別々の単語が用いられていた場合，単純に単語頻度で文書を特徴付ける手法では，この二単語が完全に別の単語として扱われるため，記事の類似性を発見できない．

この問題に対処するため，我々は，**トピックモデル**に基づき，単語の類似性をモデル化する．モデルとしては，トピックモデルとして一般的に用いられる Latent Dirichlet Allocation (LDA) [7] を，文書クラスタリングに応用できるよう拡張した Dirichlet-Enhanced Latent Semantic Analysis [14] (以下，DELSA) と呼ばれるモデルを用いる．

3.1.1 Latent Dirichlet Allocation (LDA)

LDA 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**から生成されると仮定される．ここでトピックとは，単語の出現確率を表す分布（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）を与えることができる．

3.1.2 Dirichlet-Enhanced Latent Semantic Analysis (DELSA)

LDA においては，各文書に一つずつ**トピック分布**（＝各トピックの出現確率を表す多項分布）が与えられるが，このトピック分布は，単一の**ディリクレ分布**から，それぞれ独立に生成される．DELSA では，このディリクレ分布の代わりに，**ディリクレ分布を基底として持つディリクレ過程** [9] を用いて，各文書のトピック分布を生成する．ディリクレ過程は，離散確率分布を生成する確率分布（確率分布の確率分布）であり，基底分布と呼ばれる分布（この場合，ディリクレ分布）から，可算無限個のサンプル（この場合，ディリクレ分布からのサンプル，すなわち，多項分布）と，その混合比を生成することがで

*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 $s_t = 0$ となるケースは考えていない．

*3 実際には，後述の通り，取引高上昇の確率の高い順に記事を並べるというタスクになる．

*4 正確には，与えられた品詞が「名詞」「動詞」「未知語」「アルファベット」となったものを

*5 このとき，頻度がしきい値（現在は 3）未満のものは，「RARE WORD」という単一の単語として扱う．

きる[*6]．すなわち，この場合，ディリクレ分布から可算無限個のトピック分布が，その混合比とともに生成されることになる．その後，混合比にしたがって，各文書のトピック分布が選ばれる．

DELSA における文書の生成過程を，以下に示す．

1. ディリクレ過程から，可算無限個のトピック分布と，その混合比を生成．
2. 各文書毎に，1. で得られた混合トピック分布（＝混合多項分布, Multinomial Mixture）の中から，トピック分布を一つ選ぶ．
3. 文書内の各単語のトピックを，得られたトピック分布（＝多項分布）からサンプルする．
4. 各単語を，その単語に割り当てられたトピック（＝単語比率を表す多項分布）から生成する．

上記第 2 ステップで，各文書毎にトピック分布が一つ選ばれるが，このとき，同一のトピック分布が選ばれた文書は，同一の話題に属すると考えることができる．すなわち，同一のトピック分布が割り当てられた文書同士を同一のクラスタにまとめることで，文書クラスタリングを行うことができる．このように，DELSA では,「各単語に割り当てられるトピック（＝単語分布）を推定することによる単語クラスタリング」と「各文書に割り当てられるトピック分布を推定することによる文書クラスタリング」を同時に行うことができる．

事前分布としてディリクレ過程を用いる利点としては，

- ディリクレ過程は,「同一の点が複数回選ばれる確率が高い」という性質があるため [6]，文書がまとまり易い．
- 可算無限個の点を生成するため，文書クラスタの数に上限を設定する必要がない．

というものがあげられる．特に後者は，文書集合を大きくしても，文書クラスタ数を人手で調整する必要がないという点で，有利である．

以下，具体的な生成過程を示す．各文書 d 中の単語 $w_{d,n}$ を，その単語に割り当てられたトピック $z_{d,n}$ から以下のように生成する[*7]．

$$\begin{aligned} w_{d,n} &\sim Multi(\phi_{z_{d,n}}) \\ z_{d,n} &\sim Multi(\theta_d) \end{aligned}$$

ここで，パラメータ ϕ_i は，i 番目のトピックにおける単語分布を表すベクトルであり，$\sum_j \phi_{i,j} = 1$ である．各 ϕ_i は，単一のディリクレ分布 $Dir(\beta)$ から生成される．（パラメータ β は，$|V|$ 次元ベクトル．ここで $|V|$ は単語の種類数.）また，θ_d は，文書 d のトピック分布を表すベクトルであり，$\sum_i \theta_{d,i} = 1$ である．

ここで，LDA では，θ_d がディリクレ分布から生成されていたが，DELSA では，以下のような（ディリクレ過程から生成された）分布 G から生成される．（混合トピック分布から，1 つのトピック分布を選ぶ.）[*8]

$$\begin{aligned} \theta_d &\sim G \\ G &\sim DP(G_0, \alpha) \end{aligned}$$

*6 このとき，混合比（各サンプルの確率分布）は，基底分布と似た分布となる．

*7 $Multi$ は多項分布を示す．

*8 G_0 は基底分布，α は集中度パラメータと呼ばれる，クラスタの集中度を制御できるパラメータ．

以上のモデルに基づき，各単語のトピックを推定する．トピックの推定には Collapsed Gibbs Sampling[10] を用いた．ここで，DELSA においては，通常の LDA での単語トピックラベル $z_{d,n}$ に加え，文書のトピック分布ラベルを表す変数 c_d を用意し，サンプリングする必要があることに注意されたい．得られた c_d を利用し，同一の c_d を持つ文書を一つのクラスタにまとめる．

LDA をディリクレ過程で拡張したモデルとしては，Hierarchcal Dirichlet Process (HDP) [13] があるが，HDP が，**トピック**を可算無限個生成し，トピックを文書間で共有するモデルなのに対し，DELSA では，トピック数は有限のまま，文書の**トピック分布**を可算無限個生成し，トピック分布を文書間で共有する．トピック分布は各文書に一つのみ与えられるため，「同じトピック分布を共有する文書どうしをまとめる」ことで，自然に文書クラスタリングが行えるという利点がある．

4. 実験

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，s_t の予測を行った．使用した銘柄は，トヨタ自動車，本田技研工業，ソニー，東芝，三菱電機，三菱商事，三菱重工業の 7 銘柄である．

また，DELSA におけるトピック数は 100 とし，ギブスサンプリングの反復回数は 1000 回とした．

4.1 クラスタリング実行例

以下に，得られたトピックの例と，そのトピックを割り当てられた頻度上位の単語を示す．（カッコ中は頻度.）トピック番号は，システムによって暫定的に付けられた値であり，番号そのものには意味はないことに注意されたい．単語が，主要な話題ごとに，ある程度まとまっていることが観察された．

トピック 26: % (183), 円 (161), 増 (108), 利益 (93), 益 (83), 億 (80), 営業 (72), 今期 (71), 販売 (69), 最高 (66), 米 (62), 位 (56), 社 (56), ...

トピック 27: リーグ (88), ラグビー (55), トップ (46), 選手権 (31), 決勝 (31), 日本 (28), 杯 (26), 戦 (22), 開幕 (19), マイクロソフト (18), サントリー (17), 府中 (17), 全日本 (17), バスケット (15), ...

トピック 36: 液晶 (68), 携帯 (60), 円 (57), 向け (55), 半導体 (55), ＴＶ (53), 増産 (47), 工場 (45), 億 (44), ン (43), 生産 (43), 型 (43), 化 (42), メモリー (39), 割 (38), 量産 (38), 投資 (37), 開発 (36), テレビ (33), ...

トピック 55: 賃金 (42), 要求 (40), 労組 (32), 春 (31), 円 (28), 回答 (20), 賃上げ (19), 金 (17), ベア (17), 交渉 (14), 改善 (14), ...

トピック 68: 原発 (51), ＷＨ (43), 買収 (41), 米 (33), 改ざん (16), データ (14), 策 (10), 機器 (9), 出資 (9), 県 (8), 院 (8), 蒲郡 (8), 保安 (8), 合意 (8), 日立 (7), 受注 (7), ＧＥ (7), ...

トピック 76: 電池 (95), 製 (58), パソコン (36), 回収 (35), 燃料 (25), 問題 (23), 発火 (23), 車 (17), 個 (17), 万 (13), 交換 (12), デル (11), ...

表 1: 実験結果．同一銘柄・同一日の記事については，最上位の記事のみを採用．テストデータの最小値 266 日までの値．

上位 n 記事	Accuracy
10	0.563
20	0.559
30	0.537
40	0.526
50	0.518
60	0.509
100	0.495
266	0.496

トピック 93: ＤＶＤ (81), 次世代 (75), 機 (31), 発売 (29), 規格 (24), 陣営 (24), 米 (17), 統一 (17), ソニー (17), 東芝 (15), ＩＢＭ (14), ＨＤ (13), 再生 (13), ソフト (13), ブルー (11), レイ (11), ＭＰＵ (10), ＰＳ (10), ...

また，得られた文書クラスタで，予測取引高が高かったものとしては,「決算記事クラスタ」のほか,「決算以外の金銭関係記事クラスタ」（投資，時価総額，社債等）,「事件記事クラスタ」（談合，事故，リコール等）があった．

4.2 取引高予測

各銘柄 1 つずつをテストデータ，残りの 6 銘柄を訓練データとした 7 分割交差検定で，s_t の予測正解率を測定した．3 年間 7 銘柄のうち，記事の抽出できた日付は延べ 2999 日であった．そのうち，取引高上昇日は 1348 日，下落日は 1651 日であり，上昇日の比率は 44.9 %である．また，テストデータの数は，最大がトヨタの 675 日，最小が三菱電機の 266 日であり，平均は 428.4 日であった．

7 銘柄それぞれで，取引高上昇確率の順に，上位から記事ランキングを出力し，取引高上昇予測の精度をランキングの各点で計算した．予測は，取引高上昇確率が 0.5 より大きければ上昇（$s_t = 1$），小さければ下降 ($s_t = -1$) と予測し，実際の s_t との値との比較で正解率を算出する．ギブスサンプリングは乱数によって毎回結果が変わるため，試行は 10 回行い，その平均を出力した．結果を表 1 に示す．ランキング上位（すなわち,「取引高上昇」の確信度が高い記事）のほうが高い精度となっていることがわかる．すなわち，本研究の目的である,「取引高上昇となりやすそうな時にアラートする」というシステムに向けて，提案手法が有用であると考えられる．予測結果を人手で確認したところ，ランキング上位の多くの記事が決算関係の記事であり，それ以外の記事では，事件記事や株価に関する記事がやや目立つものの，それほど明確な傾向は見られなかった．

5. おわりに

本稿では，銘柄の取引高予測に新聞記事のクラスタリングによる話題推定を用いる手法を提案した．トピックモデル DELSA を用いることにより，単語クラスタリングと記事クラスタリングを同時に行い，記事のデータスパースネスを回避する手法を提案した．実験の結果，取引高上昇について確信度の高い記事を，ある程度選別できることがわかった．今後の予定としては，より多様な銘柄の分析，DELSA によって得られた結果のより効果的な利用の検討を考えている．また，決算の話題が取引高に影響している傾向は見られたものの，その他の話題についてはそれほど明確な傾向が明らかにできなかったため，様々な話題に関する検証を行うことも今後の課題である．また，他のランキング手法との比較等も今後進める必要がある．

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] 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)

[3] 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)

[4] 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)

[5] 丸山健, 梅原英一, 諏訪博彦, 太田敏澄, インターネット株式掲示板の投稿内容と株式市場の関連性, ファイナンスにおける人工知能応用研究会（第 2 回）, pp.51–58, (2008)

[6] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Annals of Statistics, vol. 2, no. 6, pp. 1152–1174, (1974)

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

[8] Douglass R. Cutting and David R. Karger and Jan O. Pedersen and John W. Tukey, Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, SIGIR'92, pp. 318–329, (1992)

[9] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics, Vol. 1, No. 2, pp. 209-230, (1973)

[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)

[13] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei. Hierarchical Dirichlet Processes. JASA 101(476): pp. 1566–1581, (2006)

[14] K. Yu, S. Yu, and V. Tresp, Dirichlet Enhanced Latent Semantic Analysis, AISTATS-05, (2005)