2I1-4

Web ニュースの主成分を用いたアルゴリズムトレード

Algorithm Trading Using Principal Component from Web News

宮崎 邦洋　　大知 正直　　松尾 豊
Kunihiro Miyazaki　　Masanao Ochi　　Yutaka Matsuo

東京大学工学系研究科
School of Engineering, University of Tokyo

In this paper, we analyzed the relation between Web news and future index S&P. We used principal component analysis (PCA) to extract the topic from news, then we got the correlation between PCA and the volatility of S&P. As a result, we got the tendency that the longer window we have, the less correlation we got.

1. はじめに

インターネット上には膨大な量のニュースが掲載されている．それらの情報はアクセスが手軽で，多様な情報に触れることができ，また紙媒体のニュースよりも格段に早く手に入るので，多くの人がインターネット上のニュースを利用している．また，それら Web 上の情報は企業の株価にも影響する[高橋 2007] ．

最近では Web 上の情報をテキストマイニングの技術を利用して取引を行う手法が提案されている[Zhang 2010, Bollen 2010]．これは，Web 上に集まるメディアや掲示板の様々な情報と株価の関係を分析することで，意思決定の判断に利用する方法である．そしてこの方法が有効であると報告されている．

上記の研究は日毎のデータを使うことで，経済全体の動きを抽出しようという試みであるが，本研究では日に約 1000 本のニュースという大規模なコーパスと分刻みの取引価格のデータを使うことで，より詳細なニュースと値動きの関係を明らかにすることを目的とする．

そこで，本研究では最終的な自動売買取引を目的に置き，Web 上のニュースから主成分分析によって主成分を抽出し，ニュースが指数先物に与える影響を検証した．従来研究では単語と取引価格の関係を分析する手法が多いが，本研究では長期間の取引価格との相関を分析するために文章のより深い意味を抽出できる主成分分析法を利用する手法を提案する．
提案手法は，各ニュースに主成分に基づいたスコア付けをし,時間平均を取ることである．

実験では，指数先物の分刻みのデータを用い，ニュースはロイターの過去の記事を用いた．提案手法と指数先物のボラティリティの相関をとることで，ニュースと銘柄の関係性を検証した．

分析の結果，相関を計算するためのデータの取得期間を 5~50 分で 5 分おきに検証したところ，5 分と設定した際に，ニュースの銘柄に対する影響が顕著に現れることがわかった．これは，多くのニュースのうち，実際に取引価格の変動に影響を与えるニュースは少ないことを示している．

2. 主成分の時系列データの作成

2.1 主成分分析によるニュースの主成分抽出

主成分分析は機械学習において次元削減に繁く使われる分析手法である．テキストマイニングの文脈においては特に，潜在的意味解析の方法として，文章の実際の意味を主成分として抽出する目的で使われる．

例えば，「オバマ」という単語の背後には経済や政治など様々な文脈があることが予想されるが，単語のみからはそれを判別することができない．また，「オバマ」と「アメリカ大統領」は同じ意味であるが，これも単語のみからは判別できない．そこで，単語と文章の行列を作成し，その次元削減を行う主成分分析によって，その文章全体の主成分が抽出され，単語の裏にある意味を理解できるようになる．

この主成分分析によって抽出される意味を，実際の先物の価格の変動の相関をとることで，ニュースと先物価格の関係性を分析する．

2.2 抽出された主成分による時系列データ の作成

今回は主成分分析を用いて一定期間のニュースの主成分空間を作成した．まず，主成分空間を作成する期間を設定し，その期間内の記事を単語に分解した後，その空間から主成分空間を新たに作る．その際，The や if などの文脈に関係のない単語はストップワードとして省いた．そして，その空間内にそれ以降に新たに発行されるニュースを配置した．このようにして，Web 上のニュース一つ一つに主成分得点をつけ，その時点での値とすることで，時系列データ を作成した(Fig.1)．

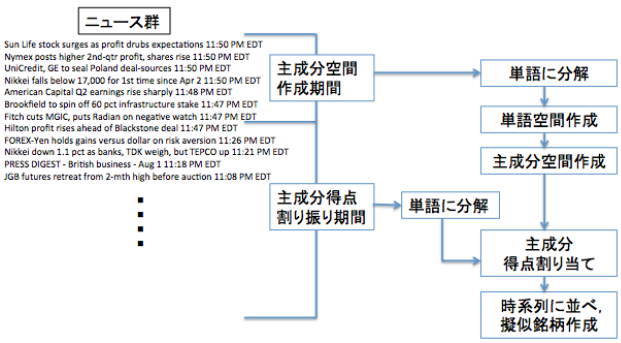

Figure 1: 擬似銘柄の作成手順

3. ニュースの銘柄に対する影響度の分析

3.1 用いたメディアと銘柄

今回主成分を取り出したメディアは，ロイターの英語ニュースのアーカイブ(http://www.reuters.com/)から取得できる 2007～2011 年のニュース記事である．ロイターは 2007 年から一日約 1000 本の記事をネット上に公開しているおり，一つ一つのニュ

連絡先：宮崎 邦洋，東京大学工学系研究科，miyazaki@weblab.t.u-tokyo.ac.jp

ース記事に点数をつける時系列データではその変動が評価しやすい.

また, 銘柄としては指数先物を用いる. 中でも今回は S&P を用いた. S&P はアメリカの景気と共に連動する重要な指数であり, 株価の値動きにはあらゆる情報を含んでいることから, 実際にニュースと値動きの関係を調べる分析に適している. 今回は 2007 年～2011 年の分毎の終値の値動きを示すデータを用いた.

3.2 分析

(1) 手法

今回の研究では, 主成分と銘柄の変動の関係を分析した. これはニュースの主成分と銘柄の時系列データを用意し, それぞれ一定時間内の価格を分毎に取得し, そしてそれらの相関係数を計算するものである. その際, 2つのデータの時間をずらすことで, 片方の銘柄がもう片方の銘柄とどのように連動しているかを観察する. ここでは, 主成分で作った時系列データを先行に置き, そして遅行に S&P のボラティリティを置くことで, ニュースの銘柄に対する影響をその連動のし具合から分析する (Fig.2).

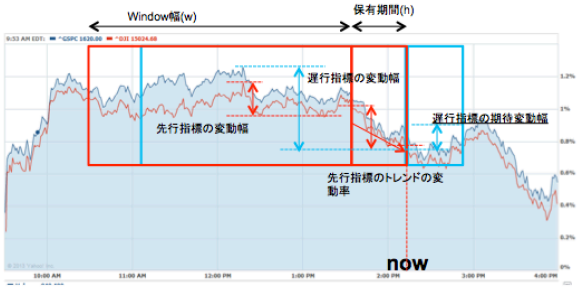

Figure 2: 相関分析のイメージ図

(2) パラメータ

上記の相関分析において, 主に以下のパラメータを使う.

- ウィンドウ幅
- ディレイ
- 主成分空間作成期間
- ボラティリティ計算期間
- 相関分析期間
- 主成分空間作成に用いる単語数

ウィンドウ幅は相関を計算するために使用する銘柄の価格のデータを取る期間である. また, ディレイは銘柄間のウィンドウのずれを指す. この2つが銘柄間の影響の機微を観察する上で重要である.

ボラティリティ計算期間は, 銘柄の現在時刻からどれほど遡って変動を計算するかのパラメータである. ここでボラティリティは標準偏差を用いている.

また, その他にも主成分空間の作成に使用する記事データの取得期間や, 時系列データの全体の期間, 主成分空間を作成するために使用する単語の数がある. 今回単語の選定方法は, 主成分空間を作成する期間内において頻度の高い単語を使用した.

4. 実験と結果

今回, ディレイは 5 分として, ウィンドウ幅の値を変えていき, 2ヶ月間の相関の平均と取った. また, 相関分析期間は各年 8 月, 9 月の二ヶ月間とした. 単語数は頻度の高い単語上位 100 単語を用いて主成分空間を作成した. その空間に, 新しく発行された記事を外挿し, 主成分得点を取り出してそれを主成分銘柄とした. ボラティリティの計算は現在時刻から5分遡った値を用いた.

また, 今回使用した主成分は第一主成分のみであり, それを S&P のボラティリティに先行する主成分時系列データとして扱った.

5 年分の平均の結果は以下のようになった(Fig.3). また, ロイターの影響度との比較のために, S&P と同等の指数先物である NASDAQ のボラティリティを先行とした場合も図に載せている.

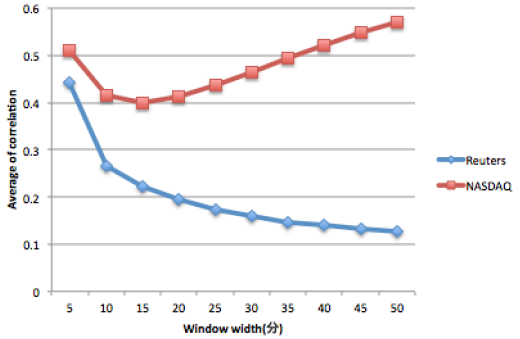

Figure 3: ウィンドウ幅と相関の関係

NASDAQ と S&P はアメリカの大きな銘柄なので, 常に相関の絶対値は安定して 0.5 付近にとどまっている. それに対し, ニュースを先行とした時の場合, 相関の平均値は window を長くすればするほど下がる傾向が見られた.

ウィンドウ幅を変えた場合結果に変化が生じたのは, 実際に株価に影響を与えているニュースの数は, 発行されるニュースの量に比べて少ないことを示している. つまり, 突発的な大きな事件などのニュースは銘柄に影響を与えるが, 通常のニュースは比較的小さな影響力しか持たない. その大きな影響を持つニュースも, ウィンドウが長ければ長いほど, その期間に様々なニュースが混在し, 影響が現れにくくなるからであると考えられる.

5. まとめと今後の展望

本研究では詳細な Web ニュースと値動きの関係を明らかにした. 分析には, ロイターの過去のニュースと指数先物の分刻みのデータを用いた. 本稿では, 主成分分析法を利用して各ニュースを時系列にスコアリングする手法を提案した. 実験では提案手法で得たデータと指数先物のボラティリティとの相関を調べた. その結果, 単体のニュースが実際の値動きに影響を与えていることがわかった.

今後の展望としては, 主成分分析は外れ値の影響を受けやすいため, よりロバストな主成分を作ることがある. また, 最終的にはトレードにつなげていきたいと考えている.

参考文献

[高橋 2007] 高橋悟, 高橋大志, and 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 10 (2007).

[Zhang 2010] Zhang, Wenbin, and Steven Skiena. "Trading Strategies to Exploit Blog and News Sentiment." ICWSM. 2010.

[Bollen 2010] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.