中国株式市場に関する株式価格情報を用いた金融極性辞書の作成

Building a Financial Polarity Dictionary for News Analytics Using Stock Price Information on Chinese Stock Markets

瞿雪吟[1]　菅愛子[1]　高橋大志[1]

Xueyin Qu[1], Aiko Suge[1], and Hiroshi Takahashi[1]

[1] 慶應義塾大学大学院経営管理研究科

[1]Graduate School of Business Administration, Keio University

Abstract: The paper proposes a method of building a sentiment dictionary using news and stock prices in China markets by textual analysis in finance. In order to obtain the amount of word polarities, we associated the frequencies of the news' one-hot wordlist to the abnormal change rate of stock prices on the publish date which is calculated by the method of the event study. We conducted the support vector regression (SVR) and build a sentiment dictionary with polarity data from learners. Furthermore, we attempt to predict the news polarities by using the sentiment dictionary.

1. はじめに

近年、新聞の紙面やインターネット上には、株式に関連したニュースが多く存在しており、自然言語処理を通じ、資産価格変動とニュースの関連性を明らかにする多くの実証研究がなされている。

例えば、Tetlock［1］は、文章の評価に心理学をベースとした辞書を用い、内容分析を行っている。また、Loughran/ McDonald［2］は、金融分野のテキスト分析のための辞書の提案を行っている。

本研究では、中国株式市場の日次データ、及びニュースデータを用い、中国語の極性辞書の構築を試みる。具体的には、中国本土の上海市場A株と深セン市場A株を対象とし、ニュースから現れた銘柄の株式価格変動から単語のポジティブ／ネガティブの程度（極性）を算出することを試みる。

2. 関連研究

極性辞書に関する分析は数多いが、そのうち金融分野に特化した報告として、五島/高橋［3］が挙げられる。当研究では、ニュースと株式価格のデータから、キーワードリストの作成を行っており、作成した金融辞書により、将来のニュース記事や、異なるメディアのニュース記事の分類を行っている。

また、関/柴本［4］は、個別銘柄の株価など対象とする金融指標によって異なる可能性があるため、金融指標の短期変動に関する語を収録した辞書を作成している。

3. データ

本研究においてワードリストを作成するため，ニュースは、中国金融情報サイト「和讯首页」[1]で掲載している 2013 年 1 月から 4 月のニュース（合計 6,590 個、関連銘柄数 14,401 個）を使用した。図 1 が示す通り、記事で言及された銘柄は、後ろに証券コードが付与されている。つまり、ニュースと関連する主要銘柄の情報はニュースに含まれている。

这部分个股中，涨幅最多前7只股票均为重组股或借壳股，其中多只个是因停牌早于上证综指攀升至6124高点，意味着这些股票或多或少的缺席了2007年的大牛市，重组后麻雀变凤凰，加之正常补涨，涨幅惊人，如三安光电(600703,股吧)（600703）、广弘控股(000529,股吧)（000529）、中福实业(000592,股吧)（000592）棱光实业(600629,股吧)（600629）等。

図 1　「和讯首页」掲載新聞の一例

それらに加え、「和讯首页」より「ニュース記事の配信日付時間」、「ニュースの見出し」を抽出した。

ワードの極性評価を行うために、各銘柄の株式リターンとリスクファクター・リターンのデータを用いた。各銘柄の株式リターンとして、ファイナンシャルデータオンラインサービスシステム「同花順」から、株式価格に関する日次データを取得した。また、リスクファクター・リターンのデータは、中国

[1] 和讯首页;http://stock.hexun.com/stocknews/index.html

中央財経大学[2]が提供しているデータを用いた。

4. 作成方法

本研究における、学習用データセットの作成は、大きく二つの部分に分かれる。作成過程の概略を、図 2 に示す。

(1) 教師スコアの算出

はじめに、ニュース配信時間の調整を行った。本分析では、中国証券取引場の営業時間に合わせて、15 時以降に配信されたニュースをその翌日に調整し、また週末に配信したニュースはその次の月曜日に調整した。

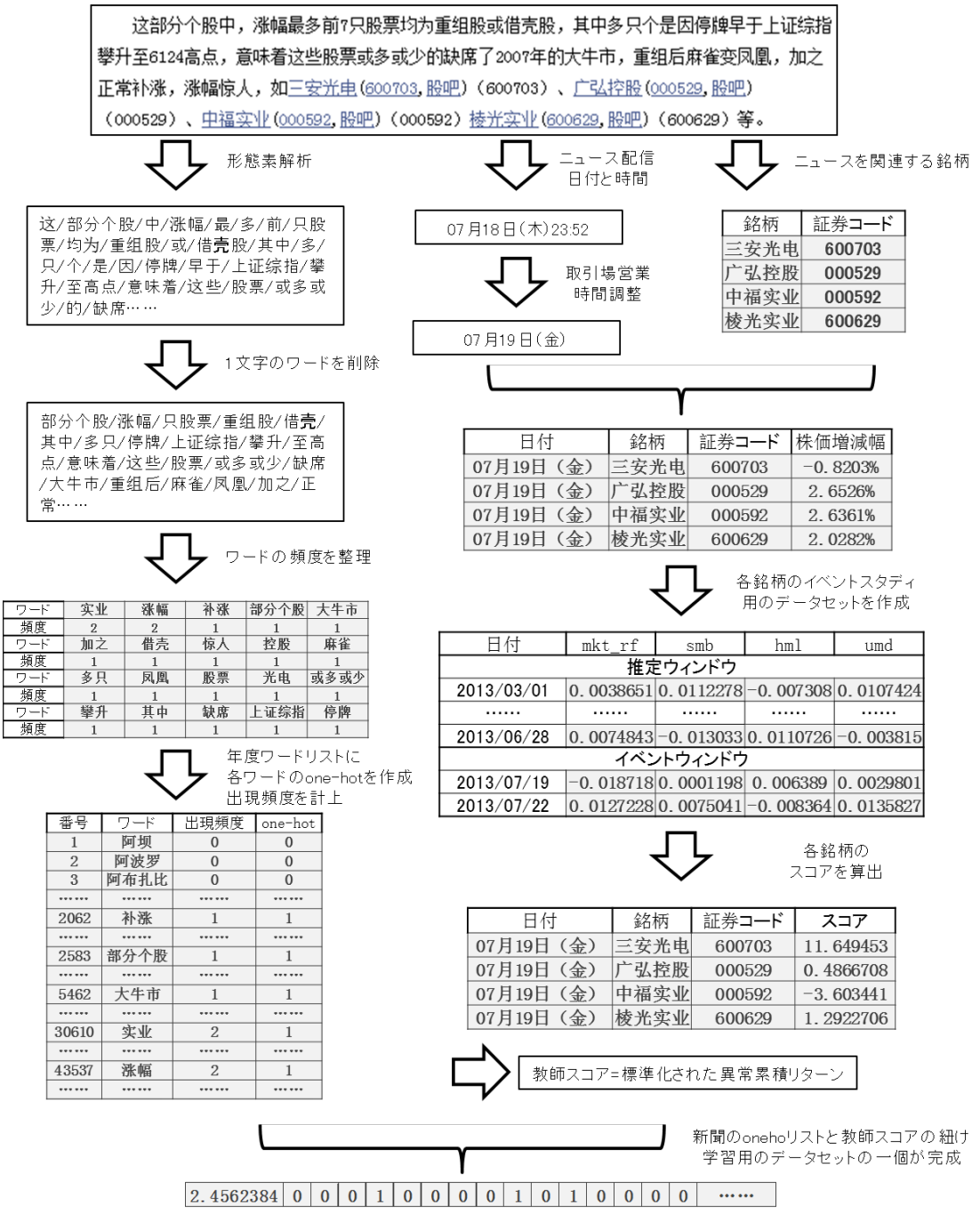

図 2　学習用データセットの作成過程の概略

[2] 中国中央財経大学;
http://sf.cufe.edu.cn/kxyj/kyjg/zgzcglyjzx/zlxzzq/98178.htm

次に、ニュースと関連する主要銘柄の情報を取得し、イベント・スタディにより各銘柄のリスク調整後のリターンを算出した[5]。ここで、イベント・スタディとは、企業の活動に関する情報の発表が、その企業の資産価格に与える影響を分析する手法である。ニュース配信日と株式価格変動の概念図を、図3に示す。

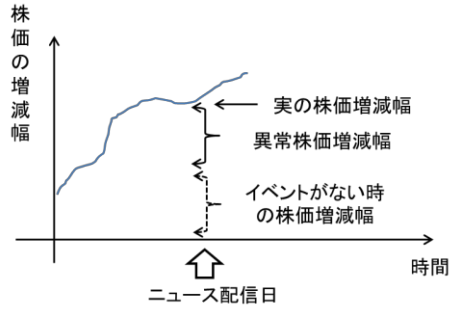

図3　ニュース配信日と株式価格変動の
の概念図

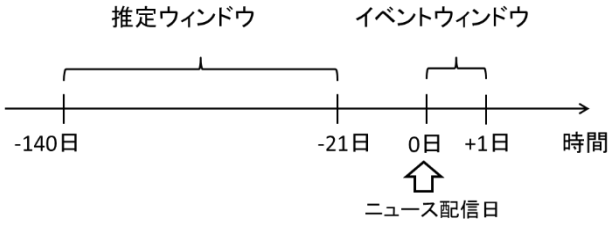

図4　ウィンドウの設定

イベント・スタディにおいては、ニュースごとに推定ウィンドウとイベントウィンドウを設定する。本分析では、推定ウィンドウはニュース配信日の140日前から21日前までの120日間、イベントウィンドウはニュース配信日当日からその翌日までとした。推定ウィンドウにおいて、Fama-Frenchの3ファクターモデルにより、パラメータを推定した[6]。

(2) ワードリストの作成

次に、2013年度のニュースからワードリストを作成する。本分析では、Rの Rwordseg パッケージ、形態素解析ツール Ansj を用いた。金融経済分野のワードを分析するため、証券、経済、金融分野のセル辞書を導入した。セル辞書は中国で有名な中国語入力システム（IME）である Sogou[3]が提供しているものである。

本分析では、1年分のワードリスト 47,378 語の TF-IDF 値を算出し、それらのうち、上位 10,000 語を対象として分析を行った。

5.　分析結果

本研究では、2013年1月から4月のニュースを対象とし、分析を行った。同期間において出現した企業上位15社の内訳は、金融証券会社12社と、酒造業2社、電子部品メーカー会社1社であった。

本分析においては、単語の TF-ID の算出、単語極性の推計、推計した単語極性を用いたニュースの分類を行った。本分析にて得られたニュースの分類精度は、60%から70%の水準であった[4]。

6.　まとめ・今後の課題

本研究では、中国株式市場の日次データ、及びニュースデータを用いて、中国語の極性辞書の作成を試みた。本分析では、サンプル数を増加した分析、異なるデータソースを対象とした分析等は、今後の課題である。

参考文献

[1] Tetlock, P.C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market, Journal of Finance Vol.62, No.3, pp.1139-1168.(2007)

[2] Loughran, T. and B. McDonald : When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance, Vol.66, No.1, pp.35-65. (2011)

[3] 五島 圭一, 高橋 大志.: 株式価格情報を用いた金融極性辞書の作成, 自然言語処理, Vol.24, No.4, pp.547-577, (2017)

[4] 関和広, 柴本昌彦: 銘柄固有の金融極性辞書の構築, 第18回人工知能学会, 金融情報学研究会(SIG-FIN), (2017)

[5] Campbell, J.Y., Lo., A.W., and MacKinlay, A.C.: The econometrics of financial markets, Princeton, NJ: princeton University press, 1997.

[6] Fama, E. F. and French, K. R.: Common Risk Factors in Returns on Stock and Bonds. Journal of Financial Economics, Vol.33 No.1, pp. 3-56. (1993).

[3] Sogou: https://pinyin.sogou.com/?r=shouji

[4] 本分析にて用いたサンプル数は限定的なものであることから、より大規模なデータを用いた分析等、詳細な分析は今後の課題として挙げられる。