株価情報とニュース記事の統合的検索・分析システム

吉田 稔[1*]　廣川 敬真[2]　浦 信将[2]　山田 剛一[2]　増田英孝[2]　中川裕志[1]

[1] 東京大学情報基盤センター
[1] University of Tokyo
[2] 東京電機大学 未来科学部
[2] Tokyo Denki University

Abstract: This paper presents our system for mining relations between texts and stock prices. We developed databases for news texts, stock prices, and company names, whose data were extracted from various sources including web articles and Wikipedia. We implemented the interface called MarketSearcher on these databases, which helps users to search and analyze texts related to various types of trends of stock prices.

1　はじめに

本稿では、我々が現在開発を進めている、新聞記事や Web ニュースと対応する株価情報を統合的に検索・分析できるシステムの紹介を行う。

現在に至るまで、株価動向を予測するアプローチとして、様々な方法が考えられて来た。そのアプローチは大きく、テクニカル分析とファンダメンタルズ分析の 2 つの分析方法 [5] に分けられる。投資家は株式の売買をする際の指標として、それら 2 つの分析方法を合わせて活用している。

テクニカル分析では、企業のニュースや業績などは考慮せずに、チャートによる株価動向のみに着目する。過去のチャートと現在のチャートの変動から今後のチャートの変動を予測したり、平均移動線などを用いて、チャートの変動を解析することで、将来の株価動向を予測する分析方法である。

一方、ファンダメンタルズ分析では、企業の価値そのものを分析する。例えば、新商品の発売などの企業の一般的なニュースや上方修正、下方修正、リストラなどの経済情報や景気などそれらに関連する情報を活用することにより、株価動向を予測する。ファンダメンタルズ分析においても、株価収益率 (PER) や株価純資産倍率 (PBR) などの投資指標を用いた数学的な予測が用いられることが多い。

ニュース記事は、一般的な個人投資家にとって、比較的入手が容易な情報源として、非常に重要である。新聞記事には、”新発売”、”開発”、”謝罪”、”暴言” などの株価動向に影響を及ぼすと考えられる情報に溢れている。しかし、これらの表現は、数学的な処理に直接利用できないため、新聞記事のテキスト情報は、株価動向の予測にはほとんど活用されていない。そのため、近年、将来の株価予測への応用を目指し、「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 [1] 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 [2] 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 [3] 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[9] 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、[4] 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。

我々は、このような「テキストと株価の関係に関する研究」における基盤環境として、ニュース記事およびそれに関連する株価情報を収集したデータベースの構築と、その上で実際に株価とテキストの関係をインタラクティブに分析するためのツールの開発を進めている。本稿では、このデータベース及び分析ツールの紹介を行う。

2　システム概要

本システムは、新聞記事および株価情報を格納するデータベースと、企業名に関するデータベース、及び、それを利用し分析を行うツール MarketSearcher により

*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

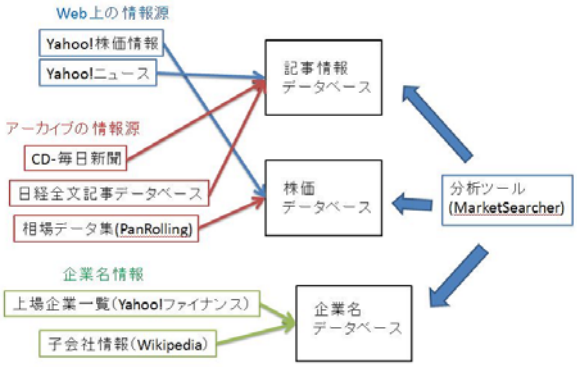

図 1: システム全体図

構成される。(図 1) データベースに格納する新聞記事としては、新聞社各社から CD-ROM を通じて提供される電子データアーカイブのほか、Web から定期的に収集するニュース記事も含む。これにより、質の高いアーカイブ情報と、速報性の高い Web 情報の両者を使用することが可能となり、より網羅性の高い分析が行える。MarketSearcher (図 2) は、これらのデータベースから動的にデータを取得し、株価とニュース記事およびその分析結果を同時に表示することで分析を補助するツールである。現在は、Web 記事に対する分析しか行えないが、将来的にはアーカイブ情報も含めた全データベースを対象とする予定である。

2.1 新聞記事データベース

新聞記事アーカイブとしては、現在、日経全文記事データベース CD-ROM 1998-1999 年版および CD-毎日新聞データ集 1998-1999 年版を利用している。新聞記事においては、見出しや第 1 段落が重要な情報を伝えていることが多いため、見出し・第 1 段落・本文全体を分けて格納を行っている。その他、日付および朝刊・夕刊の別などの情報等も格納している。

また、Web ニュースとして、Yahoo!JAPAN ニューストピックス内の全記事を対象とし、2008 年 6 月より収集を行っている。具体的には、RSS により XML ファイルを取得・更新日時を抽出し、記事が新しいものを判別してその文書本体を取得する。その後、フォーマット解析を行い、本文を取得する。記事データベースには 1 日で約 600 件の新着記事が追加される。総記事数は、2008 年 12 月の時点で、約 90,000 記事である。

データベースには、日付情報のほか、ディレクトリ構造によって得られる記事のカテゴリを活用するため、「経済」「国際」等のカテゴリも同時に格納してある。

2.2 株価データベース

本研究で使用する株価データは、PanRolling の相場データ集・国内相場版 [11] である。株価データとしては、日付・始値・高値・安値・終値・出来高を格納している。

また、Web からの株価情報として、Yahoo!株価情報からの抽出も行う。対象とするのは、東証 1 部、2 部、マザーズに上場している銘柄で、2,381 銘柄である。収集しているのは、2007 年 1 月 1 日から現在までの始値、安値、高値、終値、出来高であ る。

2.3 企業名データベース

株価データとテキストの関係を分析するには、どのテキストがどの株価に関係しているかの判定が必要となる。一般的には、例えば、「見出しに含まれている企業名」や、「第 1 段落に含まれている企業名」等を抽出し、それと名前の一致する株価データを取得する、といった手順で対応付けを行う。すなわち、

- 見出しや段落から、企業名を抽出する
- 抽出された企業名と、データベース中の企業名の対応付けを取る

という 2 つの処理が必要となるが、この際、「どの名詞が企業名なのか」という知識や、「実際の企業名と、ニュースに登場する企業名の違い」を吸収するための知識が必要となる。このため、本研究では、企業名に関するデータベースを構築し、検索に利用する。次節では、この企業名知識データベースについて説明する。

3 企業名データベース構築

企業名データベースは、「企業名」と、「会社の親子関係」を主に格納するデータベースである。これらは、以下のような処理に利用される。

テキストからの企業名の抽出においては、固有名詞 (Named Entity, NE) 抽出器を用いる (本研究では、構文解析ツール CaboCha[6] 内の NE 抽出器を利用している) ことになるが、通常、NE 抽出器においては、名詞が「組織」であることまでは判別できるが、「企業名」かどうかまでは判別しない。そのため、NE 抽出器の辞書ファイルに、企業名辞書を追加することを行う。この際に、企業名のデータが必要となる。

また、ニュース中に出てくる企業名がそのまま上場企業名になっていない場合もある。例えば、野村証券の社員が起したインサイダー取引の事件で影響を受けるのは、上場している親会社の野村ホールディングス株式会社である。このため、親会社と子会社の関係に

関する知識を獲得し、データベースに格納する必要がある。

次節では、このデータベースに格納する情報のうち、会社の親子関係の抽出手法について述べる。

3.1 Wikipedia を利用した子会社名抽出

Yahoo ファイナンスより取得した東証上場企業の一覧から、その子会社を Wikipedia[10] を利用し取得する。Wikipedia は、ユーザ参加型の百科事典であり、多くの企業について記事がある。

企業に関する記事の多くには、InfoBox と呼ばれる表形式が付随しており、所在地や資本金といった情報が記載されている。この InfoBox は、機械処理が容易であるため、ここから自動的に企業に関する情報を獲得することができる。本研究では、InfoBox 内の「主要子会社」の項目を抽出対象とした。[1]

3.2 抽出企業数

東証 HP のデータベースによれば、 1 部、 2 部、マザーズを含む企業数は 2,389 である。このうち、Yahoo より取得した上場企業数は 2,384 であった。また、Wikipedia に企業のページとして存在するものは 1,920 社であり、最終的に、そのうち 1,904 社の社名を抽出できた。[2]また、Wikipedia から子会社を抽出した結果、2,185 社の子会社が得られた。

3.3 データベース化

子会社、親会社を含めて取得した企業情報をデータベース化する。表 1 に生成しているデータベースの例を示す。code は銘柄コード、name は企業名、category は上場している市場、parent はその企業の親会社（自身が親なら同一名が挿入される）である。

表 1: 企業名データベースの一部.

code	name	category	parent
1301	極洋	東証 1 部	極洋
	極洋水産		極洋
	極洋食品		極洋
	極洋海運		極洋
	キョクヨー秋津冷蔵		極洋
1332	日本水産	東証 1 部	日本水産
	日水製薬		日本水産
	ホウスイ		日本水産

4 MarketSearcher

新聞記事と株価の関係を調査するためのツール MarketSearcher を作成した。対象記事は、現在、Web ニュースのみとなっている。

図 2 に、スクリーンショットを示す。MarketSearcher は、ユーザによって指定された銘柄の株価を表示する。銘柄入力のテキストボックスには入力補完機能があり、途中まで銘柄名を入力すると候補を表示する（図 3）。株価は一日単位でプロットとして表示される。銘柄ごとに色、形の異なるプロットで表示され、日付は範囲を指定することが できる。

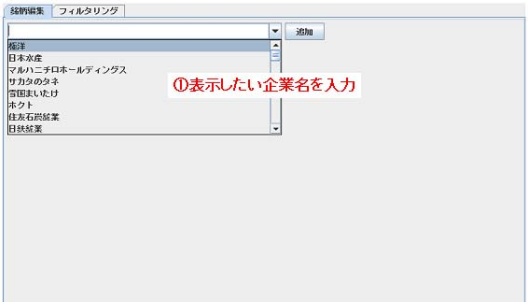

図 3: MarketSearcher 説明 1

また、プロット上にカーソルを置くとプロットの表す内容が画面右上に表示される (図 4)。プロットをクリックすると、その日付の記事中でタイトル、または本文に銘柄名 を含む記事を一覧表示する（図 5）。[3]一覧表示した記事は、クリックして選んだ各日の東証営業時間前、営業時間中、 営業時間後に時間ごとに区切られている。タイトル上にカーソルをのせると、記事本文、記事キーワード、記事のカテゴリと配信社を表すプロパティの 3 つがタブで切り替え可能なパネルで出現する。

記事キーワード 記事キーワードとは、記事のタイトル、本文を形態素解析器 ChaSen[8] を用いて単語 単位に分割し、TF-IDF 法で単語をランク付したものである。ここでの TF-IDF 法で は、1 記事に対し多く出現した単語ほど価値が高くなり、かつ全記事中で出現頻 度が低いほど価値が高くなる。複合語の抽出には、用語抽出ツール TermExtract[7] を利用している。図 6 に、抽出例を示す。

記事カテゴリ YAHOO!JAPAN ニューストピックスにおいて記事が分られている分類。記事表示ま で

[1]実際は、区切り記号「・」「／」による列挙への対応等、いくつかの追加ルールを適用して抽出を行っている。

[2]企業名の表記の揺れのため、全てを抽出することはできなかった。

[3]Web ニュース中で対象として書かれる企業・銘柄に は数に差があるため、クリックしても一致する記事が 0 という場合もある。

図 4: MarketSearcher 説明 2

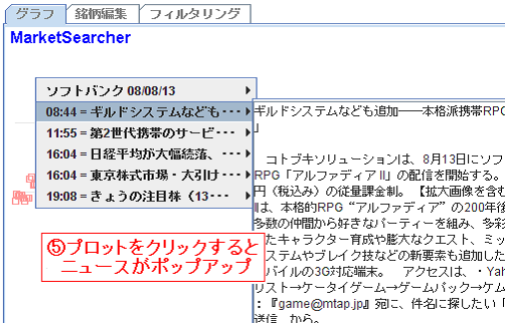

図 5: MarketSearcher 説明 3

に辿っていくリンクの名前がカテゴリに値する。最下層のカテゴリは、Web 上 で記事タイトルのすぐ上に付いている分類。のちに、株価に影響を与えやすいカ テゴリが存在するかを調査するために付加し、プロパティとする。図 7 に、抽出例を示す。

フィルタリングタブ (図 8) では、表示するプロットを各変動率、比によって絞り込む設 定を行う。自銘柄に対し、前営業日終値と次の営業日終値の変動率でフィルタを かける場合と、日経平均の変動率と前述の前日比の差をとる日経比があり、併用 することも可能である。日経比を使うことにより、市場全体の動き以上の動きが あった日付を抽出することができる。

$$\text{前日比変動率}\,(\%) = \text{前営業日終値}/\text{次営業日終値} - 1 * 100$$

$$\text{日経比}\,(\%) = \text{対象銘柄前日比変動率} - \text{日経前日比変動率}$$

例えば、図 2 から、５%の増加率でフィルタリングを掛けると、グラフは図 9 のようになる。

また、分析タブを選択することにより、記事キーワード表示で利用している TF-IDF による単語のランク付

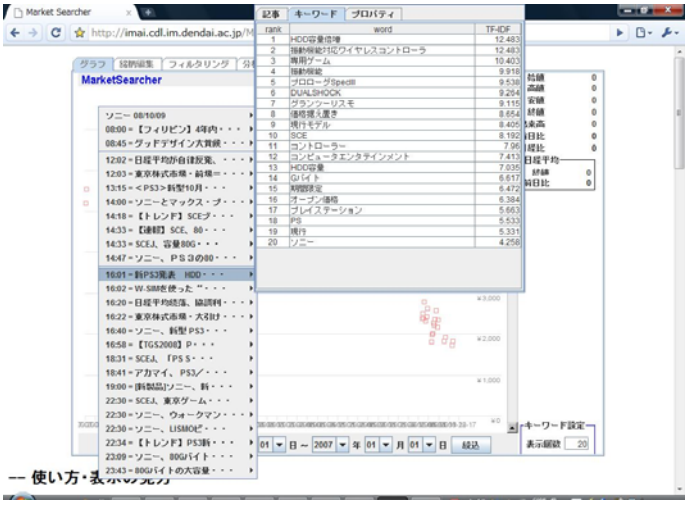

図 6: キーワード表示例

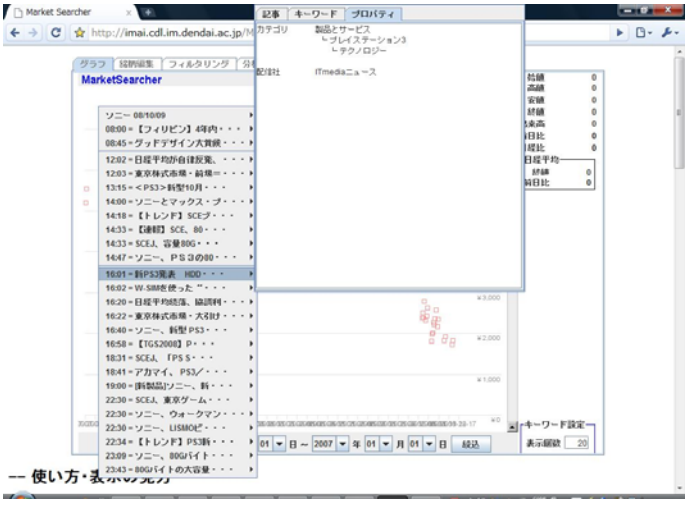

図 7: カテゴリ表示例

を、表示しているプロット中に含まれる記事全てを対象として行う（図 10）。フィルタリングを行った後に分析をすることで、例えば、日経平均より＋５%以上変動があった銘柄に対する記事に重要な語を抽出することができる。また、記事を東証営業日前、営業時間中、営業時間後、全体にわけることで、市 場に影響を与える記事の偏りを探る手がかりにする狙いがある。

5 まとめと今後の課題

テキストと株価の関係を分析するための基盤環境として、テキスト・株価・企業名情報の統合データベースの構築と、それを利用した分析ツール MarketSearcher の紹介を行った。本システムはまだ構築途上であり、今後もさらに開発を進めていく予定である。あわせて、本

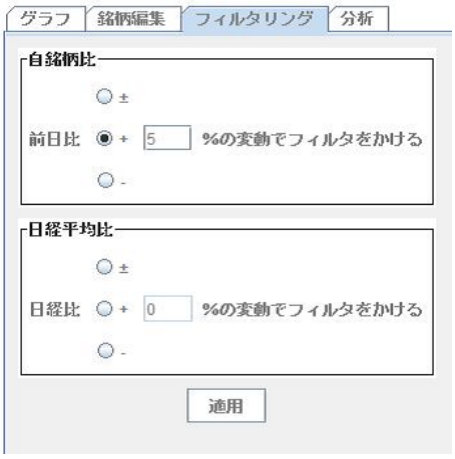

図 8: フィルタリング機能

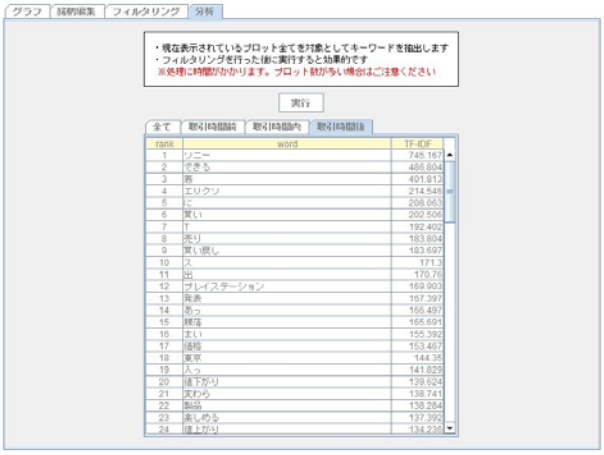

図 10: 分析実行例

ツールを利用したテキスト分析についても今後進めていく予定である。

参考文献

[1] 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)

[2] 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)

[3] 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)

[4] 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)

[5] Kabukun: 初心者・中級者の個人投資家向け株式投資情報サイト. http://www.kabukun.net/modules/xfsection/.

[6] 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)

[7] Hiroshi Nakagawa, Automatic Term Recognition based on Statistics of Compound Nouns, Terminology 6(2), pp. 195–210 (2000)

[8] Masayuki Asahara and Yuji Matsumoto, Extended Models and Tools for High-performance Part-of-speech Tagger, Proceedings of the 18th International Conference on Computational Linguistics, pp. 21–27 (2000)

[9] 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)

[10] Wikipedia, the free encyclopedia. http://www.wikipedia.org/

[11] PanRolling 相場データ集・国内相場版 (2007)

図 2: MarketSearcher 実行例

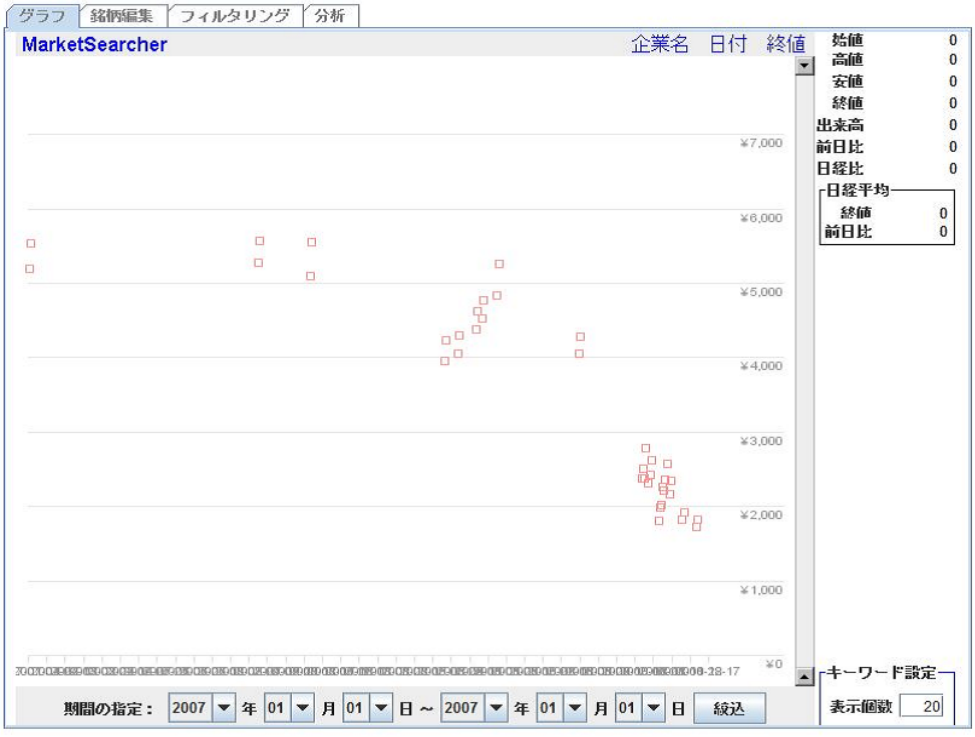

図 9: フィルタリング結果の例