E-021

アナリストレポートからの資産運用知識の学習システム

An Efficient Learning System for knowledge of Asset Management from Analyst Report

高橋　悟[†‡]　　津田　和彦[‡]
Satoru Takahashi　　Kazuhiko Tsuda

1．はじめに

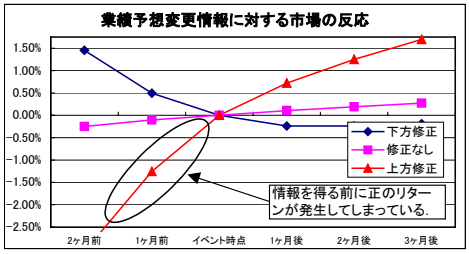
近年，資産運用を取り巻く環境は大きく変化している．特にインターネットをはじめとする情報通信技術の急速な発展により，資産運用に関する膨大な量の情報がタイムラグなしに流通するようになっている．これらの膨大な量の情報をすべて人手で処理するのは不可能であるが，その中には資産運用に有効な情報が含まれている．その情報をより効率的かつ迅速に利用することが，資産運用における重要な課題となっている．

本研究では，資産運用に有効な知識を習得するためのシステムを提案する．資産運用における重要な情報源として，セルサイドアナリストが公表する企業業績評価レポート（以下アナリストレポート）があげられる．アナリストレポートに対してテキストマイニングを用いることにより，資産運用に有効な情報の機械的な抽出を行う．テキストマイニングによりキーワード・係り受け関係・キーワードに対するウエイトを抽出し，株価リターンを教師信号とする自動学習知識ベースを提案する．

2．数値データと株価の関係

資産運用において，企業の業績予想などの数値データは株価に与えるインパクトが大きく，重要な情報として扱われる[1],[2]．そこで，数値データである業績予想値と株価リターンとの間にどのような関係があるか，イベントスタディーを用いて分析を行う．業績予想データをデータベースから取得した日を基準日とし，業績の上方修正，下方修正があったグループに分け，その前後の月次株価リターンについて分析した(図 1).

<図 1>　予想と市場の反応



○分析方法：業績予想値が上方修正，下方修正されたかによってグルーピングを行う．ベンチマークリターンを東証1部の単純平均リターンとし，各グループの対ベンチマーク超過リターンを計測し，その累計値を算出する．

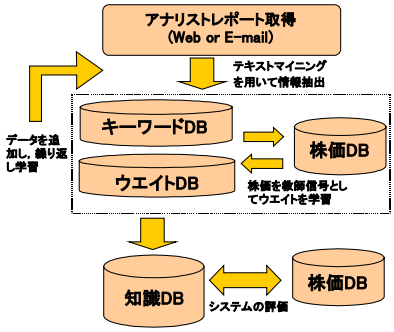
図 1 から，情報発表後から業績が良いグループの株価は上昇し，業績が悪いグループの株価は下落しており，業績予想値が株価に大きな影響を与えていることがわかる．しかし，情報が利用可能となる前から市場の株価は情報を織り込んだ形で反応していることがわかる．この理由として，業績予想値などの数値データは，発表から加工を通して投資家に届くために，利用可能となるまでにタイムラグが存在することが挙げられる．また，業績予想値が公表されるアナリストレポートの中には，業績に対するアナリストの見解などの数値では表現できないテキスト情報が多く含まれている．例えば，「特許取得」や「リストラ」などは株価に与えるインパクトは大きいが，数値データとしては評価できない．市場ではこれらの情報がすばやく株価に反映され，図 1 のような株価パターンを形成しているものと考えらえられ，いかに情報を効率的かつ迅速に利用するかが重要である[3].

このテキスト情報は，人手で処理するには膨大な量とパターンを有するため，テキストマイニングを用いて，情報の効率的利用を行うシステムを提案する．

3．学習システムの構成

はじめにテキストデータをどのように資産運用に適用していくのか，学習システム全体のデザインについて述べる．現在の資産運用では，テキストデータのシステマティックな利用は行われていない．その理由として，テキストデータからの情報抽出と，その評価方法が確立されていないことが大きな要因としてあげられる．本研究で提案するシステムの特徴として，テキストデータからの知識ベース構築方法とシステムの評価の 2 点が上げられる．学習システムの概念図を図 2 に示す．

<図 2>　学習システムの概念図



3.1 知識ベースの構築方法

まずテキストマイニングを用いてテキストデータからキーワードの抽出を行う．その際，表 1 で示すように，キーワードの表記ゆれを調整する．

<表 1>　キーワードの表記のゆれの調整

キーワード	ゆれの調整	カテゴリー
予想が強気	予想下方サプライズ	予想情報
予想より悪い	予想下方サプライズ	予想情報
予想と同水準	予想サプライズなし	予想情報
予想はリーズナブル	予想サプライズなし	予想情報
予想はやや強気	予想上方サプライズ	予想情報
予想は増額修正	予想上方サプライズ	予想情報
⋮	⋮	⋮

†三井アセット信託銀行 Mitsui Asset Trust and Banking Co, Limited
‡筑波大学 Tsukuba University

各キーワードに対して株価データを教師信号とし，ウエイトを学習する．その学習を繰り返すことにより，知識ベースを構築する．

3.2 学習システムの評価について

学習システムの評価は，評価データを用いて行う．学習データを基に構築した知識ベースのルールが，どの程度の評価データに適合するか分析を行う．さらにその結果を知識ベースの構築に還元し，知識ベースの改善と拡大を図る．評価方法は，株価リターンを教師信号として行う．テキストからの情報抽出と，その後の株価推移を比較し，システムの有効性の検証を行う．

4．キーワードと株価の分析

本研究では，学習システムの構築に向けて，キーワードと株価の関係について分析を行った．アナリストレポートの表題からキーワードの抽出を行い，キーワード出現の有無とその後の株価リターンについて有意な差が存在するか分析を行った．以下に分析結果を述べる．

4.1 分析対象データ

分析期間は 2000 年 4 月 1 日～2000 年 4 月 30 日とし，対象銘柄は東証 33 業種分類における電気業種を中心に分析する [1]．4 月を分析期間としたのは，3 月末決算予想値がもっとも多く発表される期間であるためである．また，電機業種を選択したのは，業績予想値に敏感に反応する業種の代表だからである．当該期間中に，銘柄数 447，レポート数 866 のデータを取得した．その中から，表題に情報のない (例えば，「東京朝会速報 April 2, 2002」等の)アナリストレポートは削除した．その結果，銘柄数は 153，レポート数は 416 となった．

4.2 キーワード抽出と表記のゆれの修正

キーワードは形態素解析を用いて抽出した．単純な形態素解析の結果では，すべての品詞を含めると 367 個の形態素を抽出した．その中から助詞などの重要性の低いものを除き，かつ，表 1 で示した表記のゆれの修正を行った．その結果，分析対象キーワード数は 93 個であった．

4.3 教師信号について

教師信号として，アナリストレポート発信後の株価リターンを用いる．本研究ではキーワードの持つ情報の短期・中期での効果を検証するため，アナリストレポート提供後 5 営業日と 20 営業日の対 TOPIX 超過リターンを用いた．また銘柄 i の対 TOPIX 超過リターンは以下のように計算される．

$$\text{対TOPIX超過リターン}_i = P_{i,t+1} / P_{i,t} - TOPIX_{t+1} / TOPIX_t$$

where

$P_{i,t}$:銘柄 i の t 時点の価格

$TOPIX_t$:*TOPIX* の t 時点の価格

リターンとして超過リターンを用いるのは，個別銘柄のリターンは市場全体の動きに大きく影響を受けるため，その影響を排除して分析を行う必要があるからである．

[1] ひとつのアナリストレポートの中にはいくつかの銘柄が対象として書かれていることが多い．その場合，業種毎に分類されて書かれているが，各証券会社により業種分類の定義が異なっているため，取得したデータの中には東証 33 業種における電機機器以外の銘柄も多く含まれている．

4.4 分析結果

各キーワードをアナリストレポートから抽出し，キーワード毎にキーワードの有無でデータを 2 分類した．その後グループ毎の平均リターン格差が有意であるかどうか検定を行った．検定は，グループの分散が未知であるとし，Welch 検定を行った(表 2).

＜表 2＞　Welch 検定の結果の抜粋

項目		グループ再編	投資
銘柄数	キーワードあり	17	50
	キーワードなし	397	364
	キーワードありの銘柄の割合	4.11%	12.08%
5営業日リターンの平均	キーワードあり	0.76%	−2.82%
	キーワードなし	−1.08%	−0.76%
	リターン格差	1.84%	−2.06%
5営業日リターン格差のWelch検定	t値	2.76	−2.64
	検定結果	キーワードあり	キーワードなし
20営業日リターンの平均	キーワードあり	6.203%	−2.658%
	キーワードなし	−2.471%	−2.040%
	リターン格差	8.67%	−0.62%
20営業日リターン格差のWelch検定	t値	4.40	−0.53
	検定結果	キーワードあり	差がない

表 2 より，「グループ再編」のキーワードは，短期・中期ともにキーワードありの方が，リターンが高いことがわかる．また，「投資」というキーワードの短期リターン格差は，含まれていないものと比較して有意に低いことがわかる．しかし，「投資」のキーワードは「設備投資」などを連想させ，この結果は直感と異なっている．そこで，「投資」のキーワードを抽出したアナリストレポートの内容を直接検証してみる．表題は，「株式戦略:日米ハイテク株の動向と株式投資戦略」であり，内容としては，「テクノロジー株の株価が上昇し，ターゲットプライスに近づいたため，ウエイトを引き下げるべきだ」というものであった．内容を実際に読むと否定的であることがわかるが，表題だけからはそのことは読み取れない．よって，内容からも情報を抽出し，知識データベースを拡大する必要があることがわかる．

5．おわりに

本研究では，アナリストレポートから情報を学習し資産運用に応用するシステムを提案した．また基礎分析として，アナリストレポートの表題に対して，キーワードと株価リターンの関係について分析を行った．その結果，表題についてもある程度の情報が存在するが，より有効な知識ベースを構築するためには，アナリストレポートの内容についても学習する必要があることが判明した．今後アナリストレポートの内容について分析を行う予定である．

参考文献

[1] Barberis,N., Shleifer,A., and Vishny,R.: A model of investor sentiment, Journal of Finance 52,pp. 2003-2049.

[2] Bernard,L.,V.: Stock Price Reactions to Earnings Announcements, Advances in Behavioral Finance, Russell Sage Foundation.

[3] Wuthrich,B., Permunetilleke,D., Leung,S., Cho,V., Zhang,J., and Lam,W.: Daily Prediction of Major Stock Indices from Textual WWW Data, KDD'98, pp. 364-368