

## 「テキストマイニング～文脈や語感のニュアンスの定量化」

栗田 昌孝 (BNPパリバ証券株式会社営業部 部長)

真壁 昭夫 (信州大学経済学部 教授)

テキストマイニングという解析法がある。形式化(定量化)されていないテキストデータ(通常 of 自然文)から鍵となる単語などに抽出し、その出現頻度や相関関係など解析することで、一定の知見や発想を得ようとするものだ。今回は東洋経済の四季報に記載されている各企業のアウトルックにテキストマイニングを行った。説明文に含まれるキーワードの発現回数から、文脈や語感のニュアンスを定量化しようとするものだ。そして「コンテキスト」という最も定性的なものの数値化に留まらず、実際に銘柄選択に利用可能かどうかにも運用バックテストも実施した。

### 1. 調査の目的

テキストマイニングという解析法がある。形式化(定量化)されていないテキストデータ(通常 of 自然文)から鍵となる単語などを抽出し、その出現頻度や相関関係など解析することで、一定の知見や発想を得るテキストデータ分析手法の総称のことだ。今回はテキストマイニングによって「コンテキスト(説明文脈)」という最も定性的なものを数値化して銘柄選択に利用可能かどうかを調べるために「東洋経済新報社発行の会社四季報(以下、四季報)」の企業アウトルック説明文を分析した。

説明文中に含まれる語句(キーワード)の発現回数から、文脈や語感のニュアンスを定量化した後、株価リターンとの関連性を検証することで、情報の効率性や市場参加者の行動様式を推定しようというのが本稿の試みである。

### 2. 先行研究

テキストマイニングの先行研究は高橋・津田(2004)によるアナリストレポートに対する研究がある。それによると、アナリストレポートのタイトルに「利益拡大」「予想上方サプライズ」というポジティブな語句が含まれたケースでは、レポート発表後のリスク調整リターンが有意にポジティブであり、逆に「利益下方サプライズ」「業績下方修正」など悪いニュースの後では、リターンは有意にネガティブに推移したという。また同時に、アナリストレポートの大型株への偏在性が非常に顕著であることやレポートが発表される前から、株価のドリフト現象が発生していることについても言及している。

### 3. データと検証結果

四季報の場合、アナリストレポートのように「大型株へのティルト・バイアス」が掛かっていない。むしろ全銘柄に等量のコメン

トが寄せられているので、実際の検証上「特定のグループ(例えば、大型株)ばかりを重複して調べているだけ」というバイアス回避しやすい。また、情報の入手方法も比較的簡単な上、発行頻度も年4回とアナリストレポートの発行数に比べ格段に少ないので、検証によって“有力な証拠”が発見された場合、それらは利用する側にとって扱いやすい“フレンドリーな知見情報”になる可能性が高いと思われる。

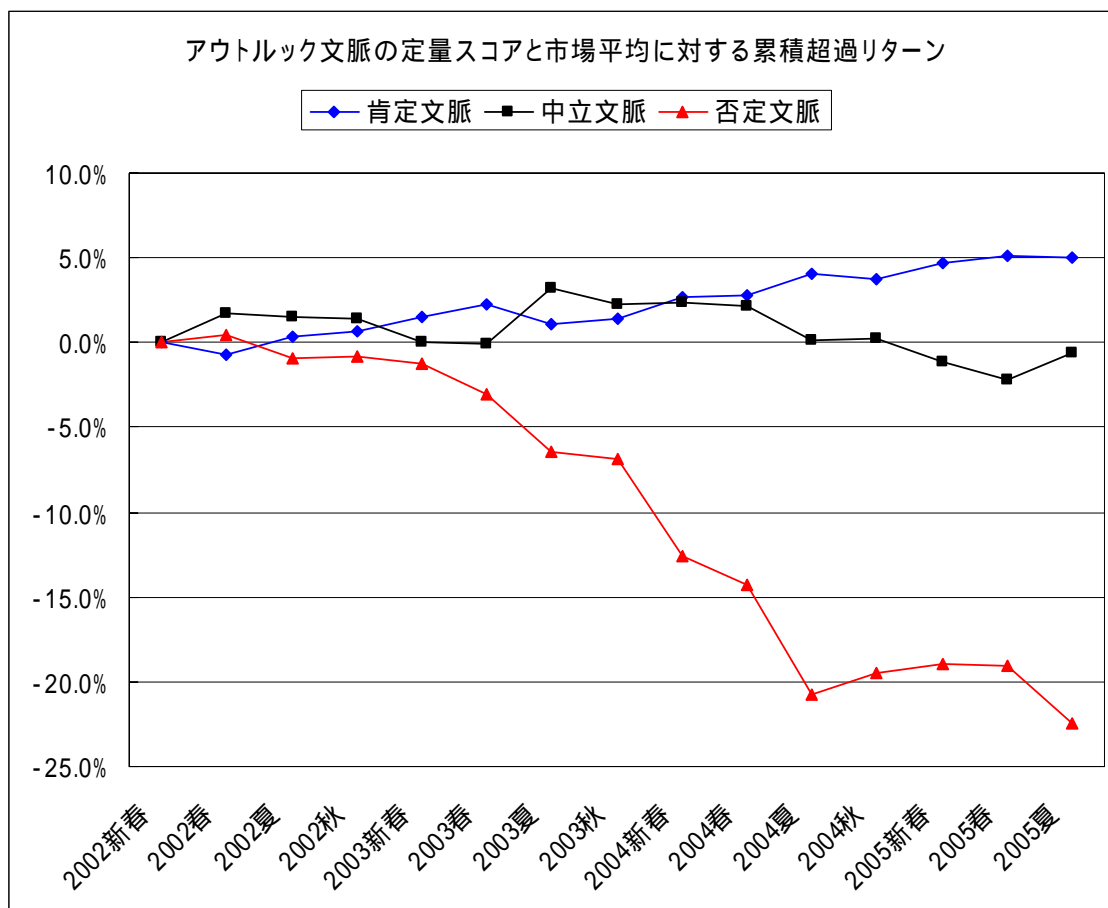
今回の検証において採用した8つの「肯定語と否定語のキーワード・ペア」はそれぞれ(以下で括弧内が否定語)、増益(減益)、大幅増益(大幅減益)、増収増益(減収減益)、強含み(弱含み)、増額(減額)、上方修正(下方修正)、回復(低迷)、改善(悪化)である。そして、各銘柄の説明文中で肯定語の採用がある毎に1ポイント加算、逆に否定語の使用に対しては1ポイント減じるという肯否の“カウント”を実施した。なお、ここではゼロを中立としており、カウントを開始するときの持ち点もゼロとした。そして、このようにして作成したスコアを「アウトルック文脈スコア」と名づけた。

2002年新春号から2005年夏号まで15季分について分析した結果が図1である。毎季ごとにスコアデータが有効な銘柄全体を「アウトルック文脈スコア」の正負によって、肯定文脈(スコアが正)、中立文脈(スコアがゼロ)、否定文脈(スコアが負)の3つに分類して等ウェイトのポートフォリオを作成した後、時系列方向のパフォーマンスを示している。なおベンチマークは期間中のサイズ効果の影響を排除するため有効銘柄全体の「単純平均リターン(等金額ウェイトポートフォリオのリターン)」を採用した。リバランスは四季報発行月(3, 6, 9, 12月)の最終週の週末終値で実施したと仮定している。また表1で

は3つのポートフォリオの組入銘柄数と回転率を記した。  
 否定文脈では相対的に銘柄数が少なく相対パフォーマンスが悪い。更に、否定文脈の悪いパフォーマンスに寄与した銘柄の寄与度を調べてみても特定の銘柄に依存しているわけではなかった。どうやら文脈スコアは「選んではいけない」

の判別に貢献するようだ。なかなか「売り」を示唆する情報が少ない中で、これは利用価値があるかもしれない。後は文脈スコアの“持ち味”が独自のキャラクターかどうかの判定であるが、これは講演にて紹介しよう。

**【図1】文脈ニュアンス別ポートフォリオの累積相対パフォーマンス**



**【表1】文脈ニュアンス別ポートフォリオの銘柄数と回転率**

		平均	最大	最小
肯定文脈	銘柄数	1,529	1,862	1,026
	回転率	36.4%	46.9%	29.5%
中立文脈	銘柄数	813	1,098	545
	回転率	63.8%	68.4%	58.7%
否定文脈	銘柄数	420	699	299
	回転率	69.9%	80.2%	48.3%