ニュース記事分析によるトピック・銘柄の関係知識獲得

Knowledge Acquisition of Relations between Topics and Stock Brands from News Articles

牧野　恭子　　櫻井　茂明　　松本　茂

Kyoko Makino, Shigeaki Sakurai, Shigeru Matsumoto

東芝ソリューション株式会社　IT 技術研究所
Advanced IT Laboratory, Toshiba Solutions Corporation

Abstract: In this study, we propose a method that acquires a topic dictionary from financial news articles. The topic dictionary is the knowledge related to the stock brands, and is composed of topics (e.g. influenza), groups of stock brands (e.g. pharmaceutical companies and / or spinning companies), and the strength of relations between them. The strength of a relation is updated by referring to the number of news headlines of the topic and the volume of transaction of the stock, and is used in order to strengthen the relation by calculating the weighted sum of news headlines. This study shows the correlation coefficient between the weighted sum of transactions and the volume is higher than the one between the non-weighted sum and the volume of transactions. The topic dictionary is expected to help catching the influence which newest topics give to the volume of transactions.

1. はじめに

「新型インフルエンザが日本上陸」「インフルエンザの流行開始」などの世間のできごとで、特定銘柄の株価・出来高が大きな影響を受けることがある。インフルエンザ流行の場合、「インフルエンザ関連銘柄」として知られる薬品会社、マスクを製造する紡績会社などの株価や出来高に大きな変動が起こる。こうした現象に関する知識のある投資家は、インフルエンザに関するニュースを入手すると、ニュースには明記されていないが関連性の強い銘柄を想起して取引を行う。同様の行動をとる投資家が多ければ、ニュース配信直後に関連銘柄の株価や出来高に大きな変化が起こる。このように、株取引においては、銘柄の名前が記載されていない社会ニュースなどから関連銘柄を想起できる知識の有効性が高い。

上記のような関連銘柄の知識は、各社の事業内容から得ることもできるが、日々配信される経済ニュース、特に株式市場の状況を伝えるニュースから得ることもできる。例えば、経済ニュースには、「北海道や沖縄でインフルエンザ流行の兆しが報じられたことからA薬品やB紡績などインフルエンザ関連銘柄が買われた。」のように、銘柄(A 薬品、B 紡績)を「インフルエンザ関連銘柄」とグループ定義し、「インフルエンザ流行の兆しが報じられたことから」と株価変動材料を推測するものがある。このようなニュースから知識を抽出し管理することで、以後の同様の社会事象発生時に関連銘柄を想起した有益な取引の判断支援をすることが期待できる。

ニュースと関連銘柄の知識の参照を支援する試みとして、例えば[1]では、構築中のテキスト・株価・企業名情報の統合データベース及び分析ツールを紹介している。また、経済変動要因(知識)を、SVM、ナイーブベイズ、クラスタリングなどで新聞記事から機械抽出・学習する試みもなされている[2-7]。その他にも、人手で整備した知識や、新聞記事をはじめとするテキストから機械抽出した知識を、経済市場予測に活用する研究は数多くなされている。しかし、これらの研究では、日々の予測精度を参照した知識の随時更新は言及されていない。

本報告では、経済ニュースから、話題(トピック)と、その影響が波及する複数銘柄との関係を知識として抽出し、辞書化する技術を述べる。トピック（例「インフルエンザ」）と、その影響を受ける銘柄(例「インフルエンザ」関連商品を扱う薬品会社、紡績会社など)の関連を抽出し、トピックに関するニュース件数と各銘柄の出来高を参照して関連の強さである影響度を日々評価更新する。影響度の導入により、知識を更新すると同時に、ニュース件数と出来高の関連を強化し、最新の話題が企業に及ぼす影響の把握を支援することを試みる。また、影響度の随時更新による社会状況変化(例えば震災)への追従効果を述べる。

2. ニュース記事分析による、話題と関連銘柄の関係知識の獲得と更新

2.1 経済ニュースに記載される知識

本報告では、前述のような経済ニュースから抽出される情報を、以下のように定義する。

- トピック：「インフルエンザ」などの話題
- グルーピング知識：「インフルエンザの関連銘柄はA薬品、B紡績」のように銘柄をトピック名でグルーピングする知識
- トピック定義語：グルーピング知識を定義する「関連銘柄」「特需」などの表現
- 材料知識：「(インフルエンザ)流行の兆し」のような株価変動の材料の知識

経済ニュースとして、Yahoo!ファイナンスの「経済総合」「市況・概況」「日本株」「産業」の4ジャンルで配信されるニュースを収集し、分析した。2011/1/11(火)～2011/1/14(金)の平日4日間に配信された計4,037件のニュースの分析結果を図1に示す。

図1に示すとおり、経済ニュース(平日1日あたり約1,000件)のうち、3.6%(144件)にはグルーピング知識が含まれる。また、21.8%(882件)には材料知識が含まれる。グルーピング知識は「インフルエンザ関連銘柄」「インフルエンザ特需」のように、前述のトピック定義語(関連銘柄、特需など)とその直前の名詞で定義されることが多い。また、材料知識は「が報じられたことから」「手掛かり材料」「材料視」「嫌気」「好感」などの手掛かり語と、その手掛かり語と係り受けする名詞句で説明されることが多い。この傾向を踏まえると、トピック定義語・材料手掛かり語を利用し、経済ニュースからグルーピング知識や材料知識を抽出できると期待できる。

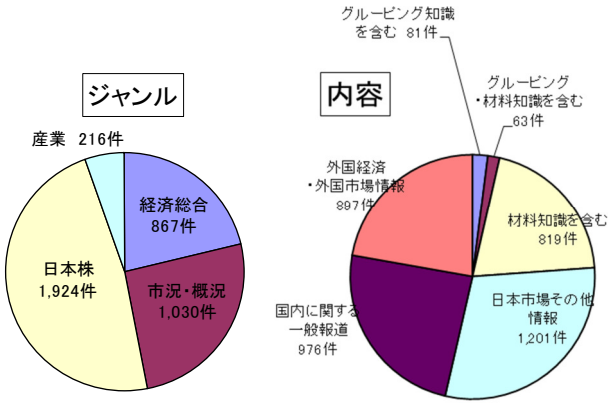

図1　Yahoo!ファイナンスの経済ニュース配信状況分析結果(2011/1/11～2011/1/14)

2.2 トピック辞書の定義

本報告で提案するトピック辞書は、表1のように、第1階層から第3階層までの最大4階層からなるシソーラスに、トピックと銘柄の関係の強さを示す数値情報「影響度」を付与し、2.1で述べたグルーピング知識とその関連情報を表現するものである。

表1　トピック辞書の例

第1階層：トピック	第2階層：サブトピック	オプション：材料知識	第3階層：銘柄	確信度
インフルエンザ	インフルエンザ薬	流行の兆し	A薬品	1.67
	マスク	流行の兆し	B紡績	1.53
			C社	1.21
…	…	…	…	…
為替	円高		D社	0.85
為替	円安		E社	1.33
…	…	…	…	…

※　以降の記載では、「トピック」と「サブトピック」を合わせて「トピック」として扱う。

第1階層・第2階層・第3階層からなるシソーラス部分が、2.1で述べた「グルーピング知識」に相当する。「影響度」は、例えば、トピック名の含まれるニュースの配信件数と、各銘柄の出来高から、銘柄の株取引に対するトピックの影響の強さを算出したもので、日々更新される。トピックと銘柄の各組み合わせに影響度の情報が付与されていることで影響の大きさを知ることができ、日々のニュース配信件数・出来高で影響度を随時更新することで、トピックのニュース配信内容の変化やトピックと銘柄の関係の変化のような社会状況変化に、トピック辞書に表現された知識が迅速に対応できる。グルーピング知識に、オプションとして、2.1で述べた「材料知識」を加えて、株取引への影響発生をより詳細な知識で説明することも可能である。

このような要素で構成し構築するトピック辞書は、以下の特長を持つことが期待できる。

- 最大4階層のシソーラスと数値情報に構造を限定することで、一般のシソーラス／オントロジーの構築と比較し、情報源からの用語抽出を限定的な処理で容易に行うことができる。
- 先行研究と比較して短周期で更新(学習)を行うことで、状況変化に迅速に対応する。
- 情報源(テキスト情報)とは異なる評価指標(出

来高)を利用した影響度の付与により、影響の強さという有効な情報が追加される。

トピック辞書とその活用例を図 2 に示す。影響度が大きいほど、トピックのニュースが銘柄の出来高に与える影響が大きいことを示している。

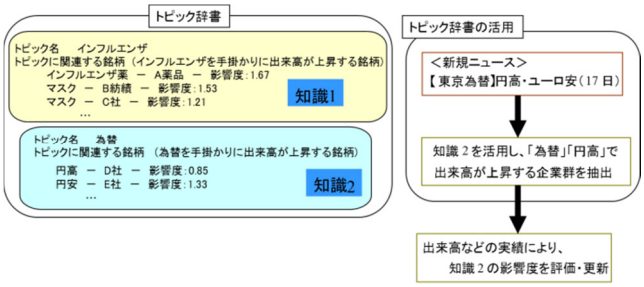

図 2　トピック辞書と活用例

3. 経済ニュースからのトピック辞書抽出

経済ニュースからのトピック辞書抽出実験を行った。具体的には、経済ニュースに記載された「インフルエンザ」と「A 薬品」「B 紡績」などのトピック名と銘柄の組み合わせについて、試作アルゴリズムによる抽出結果と、同じニュースから人手で抽出した内容の比較を行った。

なお、以降の実験では、形態素解析器としてフリーソフトウェア「茶筌」(chasen-2.4.2, http://chasen-legacy.sourceforge.jp/)を用いる。形態素解析器の辞書は、NAIST-Japanese-dic と、企業名や企業名略称、株式用語など計 7,317 語からなるユーザ辞書を用いている。

トピック辞書の抽出、すなわちトピックと銘柄の抽出と組み合わせ作成は、図 3 の手順で行った。

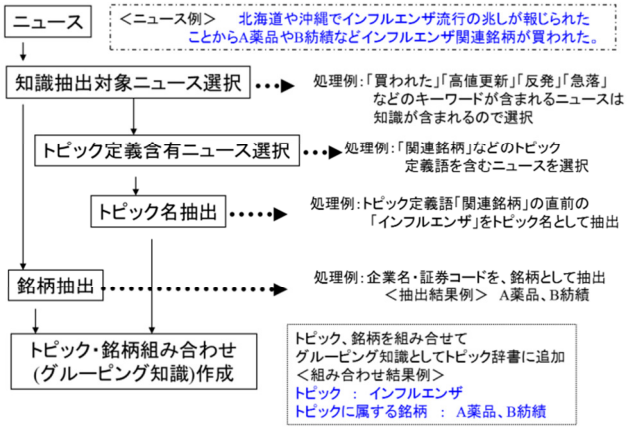

図 3　トピック辞書抽出手順

図 3 中の「トピック定義含有ニュース選択」「トピック名抽出」の処理では、2.1 で述べたトピック定義語を参照している。「トピック定義語」として、「関連銘柄」「特需」「関連株」「株」など、経済ニュースでトピック名の後に付与される表現 25 種と、「低位株」「材料株」など約 100 種の除外表現を用いる。また、「トピック・銘柄組み合わせ作成」処理は文単位で行い、トピック名と銘柄(一つ以上)が揃った段階で、トピック名と企業の組み合わせを出力する。その後、次の文節・文以降に対して、新規トピックとそのトピックと組み合わせる銘柄の抽出を試みる。

トピック辞書抽出と精度評価に用いるデータは、2009/11/30～2011/8/2 に Yahoo!ファイナンスのジャンル「経済総合」「市況・概況」「日本株」「産業」で配信されたニュース(見出し・本文)のうち表現「東芝」を含む 4,017 件である。この 4,017 件を図 3 の「知識抽出対象ニュース」相当とし、これらに対して「トピック定義含有ニュース選択」以降の処理を行った。得られたトピックと銘柄の組み合わせのうち、銘柄が「東芝」であるものを評価対象とした。なお、ここで言う「東芝」は、「東芝テック」などの東芝グループ別会社を除いている。

同じデータから人手で抽出したトピック辞書(トピック名と「東芝」のセット)と比較すると、機械抽出の精度は以下のようになった。

A)　人手で抽出したトピック辞書のトピック名と「東芝」の組み合わせ(正解)　：　1,256 件
B)　機械抽出されたトピック名と「東芝」の組み合わせ(評価対象)　：　1,392 件
C)　B)のうち、A)と同じニュースから同じトピック名で抽出されたもの(正解数)　：　525 件
D)　適合率　：　525 / 1,392 = 37.7%
E)　再現率　：　525 / 1,256 = 41.8%

上記の結果は、適合率・再現率とも、実用に十分とは言えない値となった。

機械抽出されたトピック名のうち、不正解であったもの(867 件)は、主に次のような原因で誤抽出されている。

- トピック名と前後して記載されている企業名称・略称を企業名と認識できず、少し離れて記載された「東芝」をトピック名と組み合せた。
- 一つの文に、「東芝、X 社、自動車株が」のように、個別企業と、個別企業とは無関係のトピック名(自動車)が記載され、「東芝」と無関係のトピック名を組み合せた。
- 除外表現の設定が不充分で、トピック名として不適切なもの(トピックの定義ではない一般用語の「公募株」など)を「東芝」と組み合せた。

また、人手で作成したトピック辞書(正解)のうち、機械抽出できなかったもの(731 件)は、主に次のような理由で抽出対象外となった。

- ニュース中で、トピック名と「東芝」の記載が別の文に分かれていた。
- 処理に用いたトピック定義語が使われずにトピック(ハイテクなど)が記載されていた。

機械抽出の適合率が 37.7%にとどまった原因のうち、「除外表現の設定が不充分で、トピック名として不適切なもの(公募株など)を『東芝』と組み合せた」は、除外表現を見直すことで改善できる。「トピック名と前後して記載されている企業名称・略称を企業名と認識できず、少し離れて記載された『東芝』をトピック名と組み合せた」は、形態素解析辞書の固有名詞登録追加により多少の改善が期待できる。企業名を高精度で認識するためには、さらに、新語獲得・判定の機械化が必要となる。「一つの文に、『東芝、X 社、自動車株が』のように、個別企業と、個別企業とは無関係のトピック名が記載され、『東芝』と無関係のトピック名を組み合せた」は、係り受け解析結果で文節区切りのみ利用するのではなく、係り受け情報も活用することで、ある程度の改善が期待できる。

再現率が 41.8%にとどまった原因のうち、「処理に用いたトピック定義語が使われずにトピック(ハイテクなど)が記載されていた」は、トピック名抽出を図 3 に例示した処理に限定せず、係り受け解析の結果も活用してトピック定義語のないトピック名も抽出することで、改善が期待できる。また、固有名詞(組織)と共起頻度の高い表現をトピック名候補として抽出する、などの手法も考えられる。

「ニュース中で、トピック名と『東芝』の記載位置が別の文に分かれていた」については、複数の文のつながりの解析、代名詞の参照の解析をはじめとする高度処理の適用により、改善の可能性がある。

4. トピック辞書の効果

4.1 個別銘柄のトピック辞書効果検証

4.1.1 実験内容

3.と同様に経済ニュースから人手で抽出した東芝に関するトピック辞書(230 トピック)に対して、出来高による影響度更新を行った。

トピック辞書抽出に使った経済ニュースは、2009/11/30～2011/11/2 の Yahoo!ファイナンスのジャンル「経済総合」「市況・概況」「日本株」「産業」で配信されたニュース(見出し・本文)のうち、表現「東芝」を含む合計 4,700 件である。このデータから東芝とトピック名の組み合わせ(東芝に関するトピック辞書)を人手で抽出し、評価対象とした。なお、ここで言う「東芝」は、3.と同様に「東芝テック」などの東芝グループ別会社を除いている。

トピック辞書利用の効果検証は、インターネットのニュース配信サイト excite, goo, Infoseek, livedoor, Yahoo!から収集したニュース見出しを用いて行った。効果検証に用いた各サイトのニュース件数を表 2 に示す。ニュース件数と出来高の相関が強い場合に、トピック辞書利用効果が高いと判断する。

表 2　トピック辞書利用効果評価実験データの構成
(ニュース見出し、2010/8/28～2011/11/11)

サイト名	ニュース件数	一日あたりのニュース件数
excite ニュース	420,775	954
goo ニュース	323,367	733
Infoseek ニュース	740,759	1,680
livedoor ニュース	717,996	1,628
Yahoo!ニュース	778,710	1,765

4.1.2 グルーピング知識の影響度更新

ニュース件数と出来高によるグルーピング知識の影響度更新は、以下のタイミングで実行する。

① 各ニュース配信時に、ニュース見出しから、「東芝」(「東芝テック」などの別会社を除く)と、東芝が関連付けられたトピック名を抽出する。
② 株取引のある日の 15:00 を集計ポイントとし、前集計ポイントと該当集計ポイントの間に発生したニュースで、「東芝」もしくは東芝に関連するトピック名が抽出されたニュース件数を集計する。
③ ニュース件数集計後、翌営業日の証券市場開始前に、影響度を更新する。

なお、ニュース見出しからトピック名を抽出する際は、経済ニュースで該当トピックが銘柄と関連づけられた日時以降のニュース見出しのみを対象とした。グルーピング知識、すなわち各トピックと銘柄の間の影響度更新は、銘柄と関連付けられているトピックがニュース配信された日に、以下の情報を参照し実行する。

- 該当トピックの配信件数
- 該当トピックのニュース配信件数に影響度を乗じた結果(補正ニュース件数)
- 補正ニュース件数と出来高それぞれの、直前 5 営業日の平均と比較した変化率(補正ニュース件数変化率、出来高変化率)

翌日の出来高の変動予測は、該当銘柄に関連するトピックの補正ニュース件数を手掛かりとすることを想定している。

影響度の初期値を 1 とし、値を 0～5 の間におさめるような調整式を設定して影響度更新を実験した。あるトピックと該当銘柄に関する影響度の調整方法は以下のとおりである。なお、以下に記載する閾値や調整係数の値は、小規模実験で良好な結果が得られた値を仮に採用したものである。

(ア) 銘柄名とトピックをともに含むニュースが 3 件以上ある場合は、影響度を増加させる
- 出来高変化率が 1.1 を超える場合は、(出来高変化率 – 1.1) * 0.5 を影響度に加える。
- 出来高変化率が 1.1 以下の場合は、0.05 を影響度に加える。

(イ) (ア)に該当しないが、トピックを含むニュースが 31 件以上ある場合は、出来高変化率・補正ニュース件数変化率に応じて影響度を増減させる
- 出来高変化率が 1.1 を超え、かつ、補正ニュース件数変化率と異なる場合は、影響度を更新する。(影響度に加える値：(出来高変化率 - 補正ニュース件数変化率) * 0.5)
- 出来高変化率が 1.0 未満の場合は、影響度を減少させる。(減少分：(1.0 - 出来高変化率) * 0.5)

(ウ) 影響度を 0～5 におさめるための例外処理
- 影響度の計算結果が 0 を下回った場合は、影響度を 0 とする。
- 前日の影響度が 4.95 未満であり、さらに新たな影響度が 5.0 を超える場合は、上限規制処理として、影響度を 5.0 とする。
- 前日の影響度が 5.0 以上で、さらに影響度を増加させる判断となった場合は、上限規制処理として、影響度に 0.05 を加える。

4.1.3 トピックの影響度更新結果

4.1.2 で述べたグルーピング知識の影響度調整の結果、東芝に関しては、表 3 のようなトピックと影響度の組み合わせが得られた。

表 3 は見出しの日において影響度が大きいトピック上位 15 種を挙げたものである。東芝が裸眼３Ｄテレビを発売した後の 2011/3/1(表 3 の上部左側) では、「3D」「3Dテレビ」などの影響度が大きい、すなわち東芝の出来高に関連の強いトピックと評価されている。震災直後の 2011/3/17(表 3 の上部右側)には、震災で影響を受けたライフラインのうち、東芝が関連する「電力」などの影響度が大きな値に調整されている。震災の約 1 ヶ月半後の 2011/4/28(表 3 の下部左側)には、「電力」がニュースで取り上げられても東芝の出来高に影響がなくなりつつあり、影響度も小さく調整されている。地上アナログ放送が終了した直後の 2011/7/29(表 3 の下部右端)には、「テレビ」の影響度がやや小さくなるが、電力不足対応で注目されるようになった「蓄電池」が大きな影響度を得るようになっている。表 3 より、経済ニュースからの知識(トピック)獲得及び出来高とニュース配信件数による影響度調整は、世間のできごとの影響の把握と知識化に有効であると言える。

表 3　東芝に関する主なトピックと影響度

No.	2011/3/1(震災直前)		2011/3/17(震災直後)	
1	３Ｄ	3.878	テレビ	5.000
2	テレビ	3.260	ｉＰｈｏｎｅ	4.237
3	３Ｄテレビ	3.206	３Ｄ	3.623
4	半導体	2.512	電力	3.474
5	液晶テレビ	2.471	原子力	3.474
6	ＬＥＤ	1.977	３Ｄテレビ	3.256
7	スマートフォン	1.876	鉄道	3.213
8	メモリー	1.823	原発	2.974
9	ＬＳＩ	1.760	液晶テレビ	2.521
10	システムＬＳＩ	1.660	半導体	2.512
11	ｉＰｈｏｎｅ	1.580	ＬＥＤ	2.202
12	原発	1.462	メモリー	1.823
13	発電	1.332	ＬＳＩ	1.760
14	ノートＰＣ	1.300	システムＬＳＩ	1.660
15	原子力	1.282	スマートフォン	1.566

No.	2011/4/28 (震災後 1 ヶ月半)		2011/7/29 (震災後 4 ヶ月半)	
1	テレビ	4.641	テレビ	3.866
2	３Ｄテレビ	3.256	半導体	3.326
3	液晶テレビ	2.898	３Ｄテレビ	3.256
4	３Ｄ	2.853	液晶テレビ	3.098
5	半導体	2.712	ｉＰｈｏｎｅ	2.242
6	鉄道	2.390	ＬＥＤ	2.147
7	ｉＰｈｏｎｅ	2.222	メモリー	2.069
8	ＬＥＤ	2.162	ＨＤＤ	1.939
9	原子力	1.868	ノートＰＣ	1.898
10	メモリー	1.823	ＬＳＩ	1.810
11	ＬＳＩ	1.810	蓄電池	1.677
12	システムＬＳＩ	1.660	システムＬＳＩ	1.660
13	ＨＤＤ	1.520	ＮＡＮＤ	1.543
14	ノートＰＣ	1.450	中小型液晶	1.518
15	ＮＡＮＤ	1.342	液晶パネル	1.518

4.1.4 影響度導入によるニュース件数と出来高の相関係数の変化

トピック辞書の有無による、ニュース件数と出来高の相関係数の変化を評価した。

具体的には、以下の三者それぞれと東芝の東証 1 部における出来高との相関係数を比較し評価した。

A. 「東芝」のみ :「東芝」を見出しに明示的に含むニュース件数

B. 「東芝」＋230 トピック(影響度なし) :「東芝」もしくは東芝と関連付けられたトピック名を見出しに明示的に含むニュース件数

C.「東芝」＋230 トピック(影響度補正あり):「東芝」を見出しに明示的に含むニュース件数に、東芝と関連付けられたトピック名を見出しに明示的に含むニュース件数をそのトピックの影響度で重み付けをした数を加えたニュース件数

株取引のある日の 15:00 を集計ポイントとしたニュース件数と、東芝の東証 1 部の出来高の相関係数を、評価値として求めた。結果を表 4 に示す。

表 4　東芝に関するトピック辞書・影響度導入効果

相関係数	対象とするニュース		
評価期間	A.「東芝」のみ	B.「東芝」＋230 トピック(影響度なし)	C.「東芝」＋230 トピック(影響度補正あり)
全期間	0.036	0.037	0.242*
通常期	0.024	0.011	0.364*
震災直後	－0.128	0.055	0.229
復興期	0.108	－0.259*	0.003

※　全期間：2010/8/31-2011/11/11
　　通常期：2010/8/31-2011/3/10
　　震災直後：2011/3/11-2011/4/27
　　復興期：2011/5/12-2011/11/11

※　数値の後に「*」が付与されたセルは、p 値<0.05 で無相関が棄却された「対象とするニュース」と「評価期間」の組合せ

表 4 より、明示的に「東芝」と書かれたニュースのみの件数と出来高の相関係数(A.「東芝」のみ)は -0.128～0.108 であり、ほぼ無相関である。経済ニュースから獲得した 230 種のトピック名も東芝関連とみなしニュース配信件数を単純集計する(B.「東芝」＋230 トピック(影響度なし))と、相関係数は-0.259～0.055 となり、ニュース配信件数が増えても出来高に影響がないか、出来高は逆に減少する傾向となる。しかし、影響度を導入し、ニュース配信件数を補正する(C.「東芝」＋230 トピック(影響度補正あり))ことで、相関係数は 0.003～0.364 となり、やや相関があるとみなせる状態に改善する。したがって、この銘柄･評価期間に関して言えば､影響度導入による、ニュース件数補正の効果があるといえる。特に、震災直後は､『A.「東芝」のみ』の相関係数が負に転じているのに対して､『B. 「東芝」＋230 トピック(影響度なし)』と『C.「東芝」＋230 トピック(影響度補正あり)』は、正の相関を保っている。復興期のみ、影響度導入により相関係数が『A.「東芝」のみ』を下回る結果となっている。大震災のような大きな変動の影響が薄れた時期に、大震災で大きく調整が入った影響度をリセットするようなしくみを加えることも検討する必要がある。

無相関の検定を行った結果を表 4 のセル中に「*」で示す。無相関の検定結果を見ると、『B. 「東芝」＋230 トピック(影響度なし)』は復興期以外で無相関が棄却されないが､『C.「東芝」＋230 トピック(影響度補正あり)』の影響度導入によって、全期間と通常期について、ニュース件数と出来高が無相関であることが、p 値<0.05 で棄却されている。影響度導入により、この期間のニュース件数と出来高に相関がある状態に改善できたといえる。したがって、初期のトピック辞書を高精度で構築し、影響度を導入することで、最新の話題が関連銘柄の株取引に及ぼす影響を把握することが容易になると考えられる。

4.2 複数銘柄のトピック辞書の効果検証

4.2.1 実験内容と結果

アセットアライブ株式会社によって提供されている、2011/10/25 現在の市場テーマ・キーワードと関連株の組み合わせ情報を、
http://www.asset-alive.com/thema/から収集し、理想的なトピック辞書とみなす。この理想的トピック辞書に掲載されている銘柄のうち､2010/8/30～2011/12/30 の全平日に株取引があり､4.1 で述べる影響度調整が 1 回以上実施される、すなわち関連ニュースの件数が規定値以上の日がある 741 銘柄を対象とし、出来高による影響度更新の効果を検証した｡効果検証は、4.1 と同様に､各銘柄のニュース件数と出来高の相関係数により行った。ニュース見出しは、表 2 と同じニュースソースから収集した 2010/8/30～2011/12/30 のデータである。

初期のトピック辞書の全トピック数は 956 種であった。このうち、2010/8/30～2011/12/30 にニュース見出し中に 1 回以上現れたトピックは 707 種、いずれかの銘柄で影響度調整が 1 回以上行われたトピックは 374 種であった。

4.1 と同様に、銘柄名(正式名称、略称、ブランド名、”＜＞”で囲まれた銘柄コード)のみを含むニュースの件数など 3 種のニュース件数と、各銘柄の出来高の相関係数を求めた。741 銘柄の集計結果を相関係数の分布として図 4 に示す。

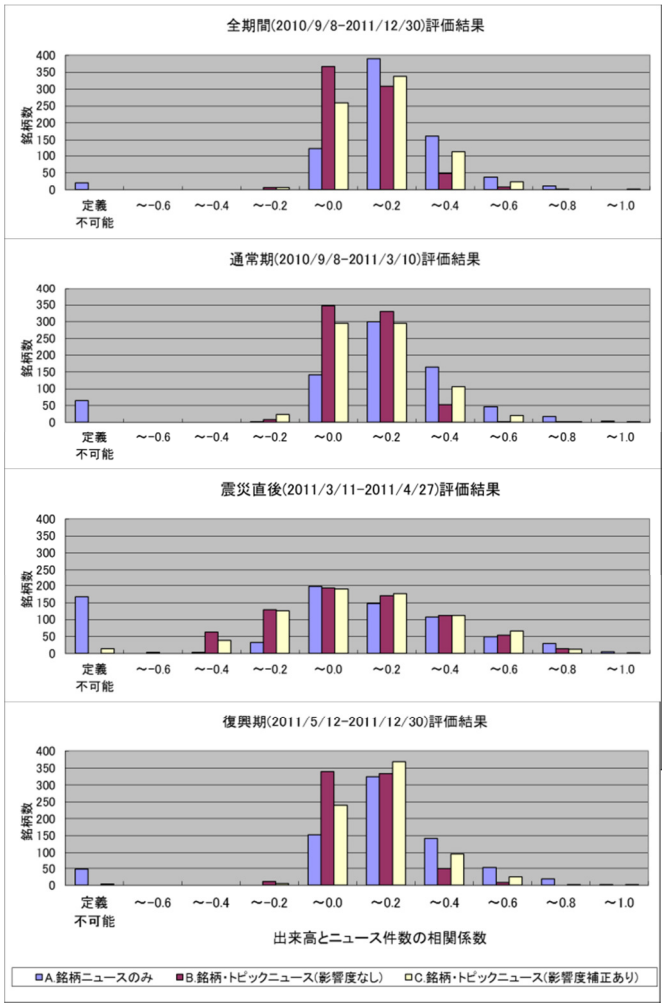

図 4　741 銘柄に対する影響度導入の評価結果

4.2.2 影響度導入効果の全般的考察

図 4 より、トピック辞書を導入することで、銘柄に関するニュースがなく相関係数を定義できない銘柄数(グラフの左端に位置づけられる銘柄数)が激減している。つまり、より多くの銘柄をニュースと対応づけられるようになっている。また、震災直後については、影響度を導入することで、相関係数が 0.2 を超える銘柄の数(グラフの右側に位置づけられる銘柄数)がやや多くなっており、影響度の導入は世間の状況の変化に柔軟に対応する効果が期待できる。

震災直後以外は、トピック辞書の導入、特に影響度による補正で、出来高との相関係数が大きくなる銘柄もある反面、出来高との相関係数が小さくなる銘柄も発生し、全体傾向としてはグラフに示した銘柄数の分布状況に大きな変動はない。相関係数の分布で全体評価を行った結果では、効果の有無を判断し難い状況となっている。

4.2.3 影響度導入の銘柄別考察

影響度導入の効果を銘柄別にみると、トピック辞書と影響度を導入したニュース件数補正により、「全期間」「通常期」「震災直後」「復興期」の 4 種の期間全てで出来高との相関係数が改善したものは 27 銘柄であった。このうち、トピック辞書と影響度の導入で相関係数が 0.5 以上改善する場合のある銘柄は 9 銘柄であった。9 銘柄中 2 銘柄は多彩な商品を扱う商事会社であり、他の 2 銘柄は証券業者、残り 5 銘柄が所属するトピックは、例えば、「オンラインゲーム」「クラウド」「スマートフォン」「電力」であった。社会で大きな話題となった各種トピック(該当銘柄の事業表現)を適切に取り入れることが効果的であったと考えられる。

4.2.4 影響度導入のトピック別考察

影響度導入の効果をトピック別にみると、トピック辞書と影響度の導入により所属銘柄全てでニュース件数と出来高との相関係数が改善したトピックと該当期間は、トピック 59 種で延べ 93 期間であった。該当するトピックは、例えば「蓄電池」「電力」「冷凍食品」「家電量販店」「警備」である。また、所属する全銘柄・4 種の期間全てで相関係数が悪化したトピックは 24 種であり、例えば「リチウムイオン電池」「海洋掘削」「ハイブリッド車」「エアコン」「キャラクター」「海外旅行」であった。

各銘柄の所属トピック数は 2～110 である。1 銘柄は、平均 8.6 グループ、ニュースのあるグループに限定すれば平均 6.6 グループ、影響度調整が行われたグループに限定すれば平均 4.4 グループに所属している。全銘柄で相関係数の改善効果が見られた 59 種のトピックに所属する銘柄は、同時に他のトピックにも所属している。相関係数の改善は、59 種のトピックのいずれかのみの効果とは断言できない。しかし、所属銘柄全てで改善が見られることから、59 種のトピックは、評価期間において所属する銘柄の話題性を推測する手掛かりとなった可能性が高いと言える。

同様に、相関係数が悪化したトピック 24 種に所属する銘柄に関しても、相関係数の悪化が該当トピックのみが原因とは断言できない。しかし、24 種のトピックは、所属銘柄に影響のない場面でも使われることの多い表現や、「海外旅行」の所属銘柄が 1 銘柄

であるなど、トピックに対する所属銘柄が最新の状況と比較して少なすぎるものが多い。影響度の導入により一般的な表現の影響を徐々に排除することが期待できるが、トピック辞書構築の際は、トピックの表現や所属銘柄を適切にするよう十分な注意が必要と考えられる。

本報告では検討の対象外としたが、一銘柄が同日に複数のトピックで話題となった場合の各トピックの影響判断や、トピックとして扱わなかった「決算」などの大きな株取引変動要因となる話題の影響排除も、今後検討したいと考えている。

4.2.4 トピック辞書の利用効果

4.2.1～4.2.3 でトピック辞書の効果をニュース件数と銘柄の出来高で評価したが、時系列に基づいたルール発見法への適用[8]においても、741 銘柄に関するトピック辞書(影響度なし)を活用することで、適合率は若干低下したものの再現率が大きく向上し、トピック辞書の効果が確認できた。

5. まとめと今後の課題

トピック（例「インフルエンザ」）と、その影響下にある対象物(例「インフルエンザ」関連銘柄である薬品会社、紡績会社など)の関連を保持し、さらに出来高を参照して関連の強さである影響度を評価更新するトピック辞書と、その利用効果を述べた。

経済ニュースからの銘柄「東芝」に関するトピック辞書抽出では、人手で抽出した正解と比較し、適合率 37.7%、再現率 41.8%であった。

銘柄「東芝」に関して人手で作成したトピック辞書に影響度を導入すると、ニュース件数と出来高の相関係数が、-0.128～0.108(ほとんど相関がない)から 0.003～0.364 程度(ほとんど相関がない～やや相関がある)に向上することが確認できた。特に、大震災直後の変動への対応効果が見られた。さらに、741 銘柄に対して影響度導入効果を評価したところ、事業内容が多彩な商事会社や震災後に注目されたトピック「蓄電池」「電力」などで変動への対応効果が見られた。

精度の高いトピック辞書を作成し、影響度を導入して学習により評価・更新すれば、最新の話題が企業に及ぼす影響を把握することが容易になる。精度向上のためには、前章までに述べた課題の他に、以下の課題解決も必要と考えている。

- トピック辞書構築に関する課題
  - 出現頻度などに基づいてトピック名の良否を判定する基準の設定
  - トピック定義語などの手掛かり語の評価と保守
  - トピック名と一般ニュースの記載表現が異なる場合への対応として、共起頻度が高い表現の保持(例えば、トピック「防衛」で、経済ニュースで共起する表現「情勢緊迫化」も保持する、など)
  - トピック辞書構築時の情報源拡大(各企業の公開情報参照など)
- トピック辞書更新に関する課題
  - 経済ニュースに記載される「参入」「撤退」などのイベントに対応したエントリー追加・削除の導入
  - 本報告では固定とした影響度更新に使う各種パラメータの、評価・設定機能検討

参考文献

[1] 吉田稔, 廣川敬真, 浦信将, 山田剛一, 増田英孝, 中川裕志: 株価情報とニュース記事の統合的検索・分析システム, 第 2 回 人工知能学会ファイナンスにおける人工知能応用研究会, SIG-FIN-002-10, pp. 59-64, (2009)

[2] 酒井浩之, 増山繁: 経済新聞記事内容の個々の企業におけるインパクトの判定, 情報処理学会研究報告. 自然言語処理研究会報告 2006(94), pp. 43-50, (2006)

[3] 坂地泰紀, 酒井浩之, 増山繁: 景気動向を示す根拠表現の抽出と分析, 情報処理学会研究報告. 自然言語処理研究会報告 2007(76), pp. 151-156, (2007)

[4] 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, http://www.iser.osaka-u.ac.jp/rcbe/6thworkshop/presentationHP.pdf/SatoruTakahashi070210.pdf, (2012 年 10 月 3 日現在)

[5] 張へい, 松原茂樹: 株価データに基づく新聞記事の評価, 第 22 回人工知能学会全国大会, 1E2-4, (2008)

[6] 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, 情報処理学会研究報告. 情報学基礎研究会報告 2001(20), pp. 137-144, (2011)

[7] 松井藤五郎, 石田智也, 中嶋啓浩, 和泉潔, 吉田稔, 中川裕志: ニュース記事クラスタリングによる取引高予測の試み, 2011 年度人工知能学会全国大会, 2H1-OS18-7, (2011)

[8] 櫻井茂明, 牧野恭子, 松本茂: 学習データの拡充による評価主体検知性能の改善, 2012 年度人工知能学会全国大会, 2B1-R-3-1, (2012)