## 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。

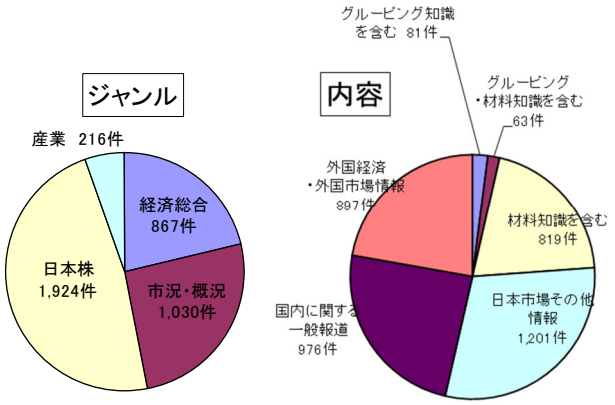


**図1　Yahoo!ファイナンスの経済ニュース配信状況分析結果(2011/1/11～2011/1/14)**

### 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。

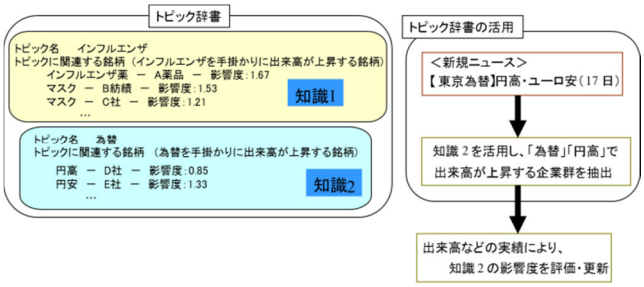


**図 2　トピック辞書と活用例**

## 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。

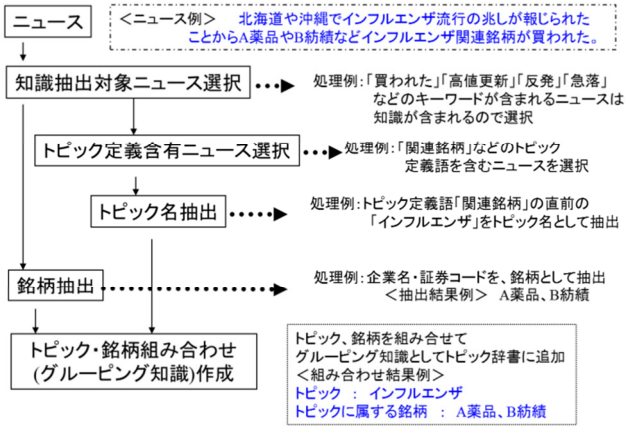


**図 3　トピック辞書抽出手順**

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。