第２章　景気ウォッチャー調査の個票と各種経済指標の関係に関する検証の結果

２．１　検証手順の概観

内閣府が作成･公表した 2010 年 1 月から 2014 年 12 月の景気ウォッチャー調査の個票と各種経済指標の関係変化などを、ビッグデータ業務で用いるテキスト分析や回帰分析などの手法によって調査・分析する。これにより、東日本大震災経験後の我が国において経済構造変化や景気循環変化が起こったかどうか、並びに景況感の変化を示す統計数値（DI 値等）が景気ウォッチャー調査のコメントによって推計が可能かどうか等についての検証を試みる。

景気ウォッチャー調査のコメントによる景況感の変化を示す統計数値の推計については、具体的には、景気ウォッチャー調査のコメントについては類似の意味をもつことばを統一化した上で数値化し、景気ウォッチャー調査の現状判断 DI、先行き判断 DI、地域別支出総合指数とその構成要素（地域別民間住宅総合指数、地域別公共投資総合指数等）等との関係性を検証する。検証方法は単純な回帰式など簡便な方法のみならず、日本電気株式会社（以下、ＮＥＣ）の中央研究所において開発した、特徴表現抽出などの自然言語処理技術を用いたテキスト分析、多種多様なデータから複数の規則性を容易に導出できる異種混合学習技術を活用したアプローチを試みる。

分析作業を２つに大別すると、テキストマイニング[1] とデータマイニング[2]に分けられる。各々の概要について、次節以降で述べる。

２．１－１　テキストマイニング

テキストマイニングでは、景気ウォッチャー調査の「追加説明及び具体的状況の説明」及び「景気の先行きに対する判断理由」の欄に記載されたテキスト情報（以降、コメントと呼ぶ）から景気指標との相関性が高い単語や特徴表現を有意表現として抽出する。

有意表現を抽出するテキストマイニングのプロセスを図２．１－１に示す。本プロセスは 7 ステップからなり、それぞれの処理については２．２－１～２．２－７の各項にて詳述する。以下では、各ステップの概要をまとめる。

---

[1] 通常の文書からなるデータを単語や文節で区切り、それらの出現の頻度や共出現（共起ともいう）の相関、出現傾向、時系列などを解析することで有用な情報を取り出す、テキストデータの分析方法

[2] 統計学、人工知能等の様々なデータ解析技法を適用し、大量のデータの中から法則・パターンを見つけ出す分析方法

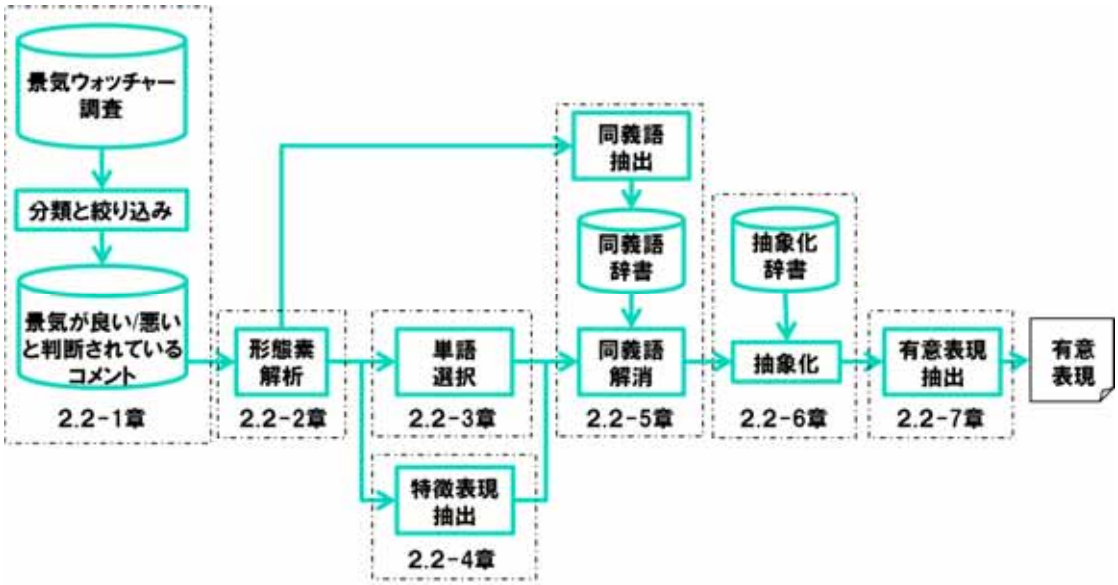

図２．１－１　テキストマイニングのプロセスと本資料中での説明箇所との対応関係

①分析対象の絞り込み：テキストマイニングでは、景気指標との相関性が高い有意表現を特定するために、景気ウォッチャー調査の中から、景気が良くなっている・やや良くなっている（及び良くなる・やや良くなる）と景気が悪くなっている・やや悪くなっている（及び悪くなる・やや悪くなる）と判断されている報告のみを用いて分析を行う。そこで、本ステップでは、景気ウォッチャー調査のコメントから景気が変わらないと記載しているものを除外し、残ったコメントを景気の良し悪しの２つに分ける作業を行う。

②コメントからの単語の抽出：コメントから、形態素解析[3]エンジンを利用して単語を抽出する。また、動詞などの活用によって表現が異なる単語を同一視するために、抽出した単語を原形（活用していない状態）に変換する。

③単語選択：有意表現の抽出には、助詞等の付属語は景気指標と相関性があるとは考えづらいため削除しておき、名詞や動詞等の自立語のみを選択して用いる。

④コメントから特徴表現の抽出：単語単位ではなく景気動向に関連する特徴的な表現（複合名詞やフレーズなど）を抽出する。

⑤同義語を統一化する処理：複数の同義判定方式を用いて、コメントの中から同義語を

[3] 自然言語で書かれた文を形態素（言語として意味を持つ最小単位）に分割し、形態素の付加情報（品詞、原形など）を判別する。

特定し、同義語辞書を作成する。抽出した同義語を元に、②で抽出した単語と④で抽出した特徴表現において同義のものを同じ表現に置き換える。

⑥単語と特徴表現を抽象化する処理：②で抽出した単語や④で抽出した特徴表現には、「ゲームセンター」や「パチンコ」など個々には異なるものであるが概念的には「娯楽施設」としてまとめられるもの、また「競争が激しい」と「激戦区」など「競争激化」という同じ状態を表している表現などが含まれる。本資料では、こうした抽象概念への変換を抽象化と呼び、人手で作成した抽象化辞書を用いて、⑤を終えた単語や特徴表現を抽象化する。

⑦有意表現の抽出：⑥までのステップで得られた単語や特徴表現が、①で分類した景気が良いとするコメントと景気が悪いとするコメントのいずれかに偏って存在することを景気指標と相関性のある単語とみなし、その度合いをカイ二乗値として算出する。

２．１－２　データマイニング

データマイニングのプロセスを図２．１－２に示す。まず前述のテキストマイニングプロセスで景気ウォッチャーの個票内の単語のスコア（カイ二乗値）を算出する。次に、カイ二乗値の上位 400 語を抽出する。抽出結果は別紙に記載する。次に、各地域・月ごとに、単語が含まれる件数および割合を計算する。さらに、DI 値の値および過去の DI 値の値を組み合わせて、分析データを作成する。

次に、分析データを時期によって学習期間と評価期間に分けて、学習データと評価データを作成する。次に学習データを学習システムに投入し、予測モデルを作成する。次に、評価データを予測モデルに投入し、評価期間の予測結果を得る。この予測結果を実際の値と比較して評価を実施する。

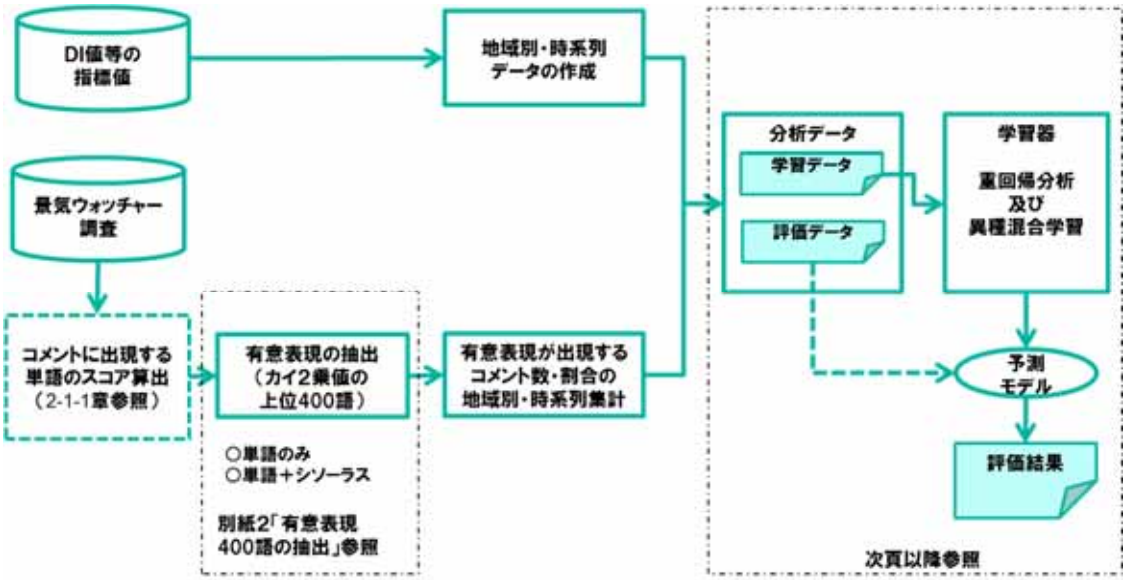

図２．１－２　データマイニングのプロセス