第４章　今後の分析業務に際しての課題と推進の方向性

第３章で詳述した今回分析は、いわゆるビッグデータによる経済・景気分析の第一段階として位置付けられることができる。

ビッグデータによる分析、貢献の対象は多岐にわたり、今回の試みは潜在的には非常に有望、かつ将来性のある分野である。経済・景気の分析、特に現状判断は一般に政府機関が公表する統計をベースとして行うことが多いが、国民経済計算を作成する際に使用する統計を例にとっても生産側のデータと需要側のデータに偏っており、販売側のデータは殆ど使用していない。すなわち、それぞれの統計やデータが固有に持つ限界、バイアスを他の統計データが補完、あるいはチェックすることを通じより客観性、網羅性を高める体制が十分であるとは必ずしも言い切れない。また、データの速報性の点においても改善は進んでいるもののそこには依然として限界が存在する。

以上のような既存の統計データのもつ欠点、弱点を補うという１点に着目しただけでも、ビッグデータによる経済・景気分析の重要性、有望性が理解できるものと思われる。

本章の以下各項においては、今後この分析業務をさらに発展、推進させていく際の課題や将来の方向性について記述する。

４－１　有識者ヒアリング、検証結果、検討委員会から得られた分析業務の課題

第２章、第３章から得られた知見や評価、さらには検討委員会における委員からの指摘を踏まえると、今回分析業務、より広くは景気・経済に関する分析をビッグデータ[1]を使用して行うことの課題は、大要以下のようにまとめることができよう。

４－１－１　データの収集、取得、利用に対するアベイラビリティの向上

ビッグデータを利用して経済や景気の動向を分析するには、当然のことながら最初に以下の点に関する構想・方針の策定が不可欠である。

① ビッグデータを用いて経済、あるいは景気のどの事象や項目に関する知見、情報、データを抽出・獲得したいのか（あるいはできるのか）についての目的／期待の

[1]一般に「ビッグデータ」というとデータそれ自体よりも「データと技術とビジネス（あるいは公共面での便益など）を統合した一連の価値創造活動、あるいはその手法」を示すことが多いが、本章では「世間一般に存在する大量のデータ」のことを指すこととする。

設定

② ①の目的／期待を達成するためにはどのようなデータを利用するのがよいのかについての仮説設定、ならびにデータの種類・内容等に関する調査

③ ②の作業の結果、候補として定まった「対象データ」をどのようにして収集・取得・利用するかについての手段／方法の確認、案出

上記のうち、①と②については有識者ヒアリングにおいて数々の知見や情報を頂戴し、今回分析業務、並びに将来的に分析業務を拡充させる際のターゲットについては概ね誤りなく定められたものとみられる。

しかし、③については分析主体（例えば内閣府）による具体的なリソースの投入や調達等といった実行段階の話であるため、学術・研究分野を中心にヒアリングを実施した結果だけでは必ずしも具体策のフィージビリティ確認が十分でない可能性がある。

特に、

・データの収集や取得が有料である
・データの利用は無料であったとしても、知的財産権、個人情報保護等の観点からその取扱いに際し特別な配慮が必要
・データの利用目的に制限がかかる（例えば商用目的の利用は不可）、あるいはデータ／分析結果の公開に制限がかかる

といったケースは十分考えられ、これらの点に関しては必ずしも明示的にルール化、慣習化されたものが確立されていないため、分析主体による個別の対処が必要となる。

内閣府をはじめとする政府機関によるビッグデータの分析・利用は現在までのところ事例があまりないと一般には認識されており、このため政府機関による（特に有料の）ビッグデータの利用に関する有用性、費用対効果に関する一般からの認識や理解を、今回分析業務等をきっかけとして高めることは今後重要度を増してくると考えられる。

また、政府機関がビッグデータ、あるいはその集計結果である二次データを有料で利用する場合においては、過去事例の不足、ビッグデータ利用による成果獲得の確実性[2]、データ価格の妥当性の観点から予算枠の確保等の場面で難航する可能性もある。経済、景気の分析目的でデータを入手、利用する際の価格については、既に学術・研究目的での利用実績がある程度進んでいる現状を踏まえ、データ出所・提供元との相談・交渉を粘り強く行い利用目的、費用対効果の観点等に照らして妥当かつ適正な水準を確保すること、さらにはデータの取り扱いに関する当事者間、一般の安心感を確保する枠組みを

[2] ビッグデータの分析は100%期待通りの成果が出るとは限らないという性質上、調査研究、研究開発に対する支出という考え方が基本になる可能性が高い。

作っていくことが、分析主体による将来的なデータ利用の自由度、可用性を広げることにつながるものと考えられる。

４－１－２　ビッグデータ使用による景気・経済の分析にかかる知見・知識・ノウハウのさらなる蓄積

金融・株式市場においては、商用を目的とした金融・為替・株式市場関連のデータ並びに分析・予測サービスの提供は国内外ともに普及が進みつつある。
しかし、ビッグデータを利用することによって景気や経済を分析するという分野の研究はあまり進んでいるとはいえないのが現状である。

経済や景気の分析を行う人材や部門は政府、学術、民間の各分野にそれぞれ存在する一方で、ビッグデータの分析業務を行う人材や部門は学術と民間の両部門が比較的多いのに対し、政府部門のそれは現状手薄いように見受けられる。さらに、「経済・景気」と「ビッグデータ」の両方が重なる人材や部門は圧倒的に学術分野に偏重しているというのが有識者ヒアリングから得られた結果であった。

また、同分野の研究・分析は直接的には商用目的に結びつきにくいという性格上、民間企業部門による同分野の研究に対するインセンティブが不足していることも影響しているものとみられる。

以上のような現状を踏まえると、ビッグデータ使用による景気・経済の分析にかかる知見・知識・ノウハウはかなり学術分野に偏在している可能性がある。これをさらに蓄積・拡充、そして学術分野以外にも広く展開させるような経済・利益上のインセンティブが少なくとも短期的には見えにくいと判断されるならば、この状況から脱却するための取り組みを政府部門、民間部門を巻き込んだ形で意識的に行う必要がある。

４－１－３　ビッグデータの分析技術や手順等の可視化（ホワイトボックス化）

ビッグデータの分析にかかる技術は時とともに進歩を続けており、このため分析の技術や解析の手順などについて一般の理解を得ることはますます難しくなる傾向にある。

しかしながら、政府機関が各種統計の結果を公表するに際してはその概要や作成手順、結果の解説等についても公開しており、これは説明責任上も当然のプラクティスである。今後、もし政府機関がビッグデータを利用した経済・景気の分析を景気判断等において利用、あるいは援用するということを（実験的にではなく）定例的に行うことになれば、

それに付随する各種情報の公開が求められることはほぼ確実とみられる。
特に、ビッグデータをどのように利用し、どのような技術を使って分析しているかについて一般に説明することはかなり難易度が高い可能性があるため、ともすると分析部分についてブラックボックス化してしまう懸念も存在する。

一方で、ビッグデータの分析ツールやアルゴリズム等はそれを開発、実用化している企業、研究者等にとっては知的財産権の根幹に関わる部分である。従い、もし政府機関が経済・景気の分析に資するビッグデータの分析を業務委託するなどの手段を採用した場合には、分析の概要や手法等を一般にわかりやすく説明するための工夫が必須であるとともに、業務委託先である企業等とは情報開示などにかかる適切な取り決め、ルールを構築することが求められることになる。

一般に、ビッグデータの利用や分析においては、個人情報保護の観点やデータの商用利用が問題視される傾向にある。さらに、政府機関がビッグデータを有料で利用する際には、その費用対効果等に関する説明責任も求められる可能性がある。これらの課題をひとつひとつクリアにしていき、ビッグデータの分析が（ブラックボックス化ではなく）ホワイトボックス化されていくことが、技術の進歩に呼応した形で健全かつ効果的なアウトプットを生み出し、ひいては国民生活の向上に寄与するものとみられる。

４－１－４　検討委員会における主な指摘事項

第３章における分析結果を踏まえ、検討委員会において委員から頂いた主な指摘事項は以下の通りであった。
今後の分析業務においては、これらの事項を念頭に置いて分析の枠組み、手法、精度の改善を図ることが望まれる。

・予測式（推計式）作成の際に、オーバーフィットを回避するためにデータを地域別に分けるなどの工夫をしていたが、根本的には 400 という有意表現の数が多すぎることがオーバーフィットの原因だと思う。

・データが少なく、学習期間が 36 カ月ということではオーバーフィットが起こるのはある意味仕方がないと思う。この条件下で、外挿の予測期間のレベル感としては下の方に出ているというバイアスはあったが、前月比のトレンドは比較的うまくとらえていたと思う。

・３－２において、2 値・3 値・5 値の分析の結果数値が少しずつ違うという点は非常

に面白いと思ったが、多分その理由は正解(率)の良し悪し、すなわち予測精度が違っていることにあると思われる。その辺りを明らかにしていくと精度を改善する余地がある。

・5 段階等に回答が区切られた「オーダーデータ」と回答テキストのキーワードとの関係（紐づけ）をどのように作っていくかについては、一般に、予測（目的）変数が 5 段階とした場合、いくつかの手法がある。①5 段階がそれぞれ別々の、独立した目的変数だと考えて判別する方法と、②段階 5 に入るか入らないかを予測して、次に段階 4 に入るかどうかを予測する方法。そのとき、(1)5 段階それぞれに対して最も確信度が高かったものに分類していくという多数決的なやり方もあるし、もしくは(2)段階は数値の連続値だと考えて、例えば３．いくつ(3.x)という値を回帰式によって予測するようにしておき、それがどの段階に最も近いかで分類する方法もある。なお、目的にもよるがプラス５とかマイナス５というように一番外れたところを予測したい際には別々に予測した方がよいことが多く、一方、前月との差を見たい場合などの「連続性をみたい」場合には数値的に予測する方がよかったりする。そこは目的による。

・キーワードの分析に関しては、もう少しハイブリッド的な分析ができるとよいという印象。単語の語数を拾うことはもちろん大事だが、どういう文脈で出てきているかについて、例えばある単語の後に否定的な言葉がついているとこれは関係無いということになるので、単語とその前後の文章を機械にまず書きださせて、それを人間が見て、こういう場合には使えない、こういう場合には使えるというようなことをチェックしながら高度な方法論が開発できる余地があるのではないかと思う。

・キーワードの分析は積み上げが大事なので、それぞれの単語に統計的有意水準（1%とか 5%とか）を付けてもらったほうが後々積み上げていく分析のときに役立つと思う。

・ビッグデータ分析においてデータ量を増やす場合には、いろいろな人から意見をとってきて単語の情報を増やすという空間的な増やし方と、時系列的な量を増やすという２通りの方法がある。今回の分析は大まかにみて全ての時期を同じように分析をかけているので、時系列な量のバイアスを完全に無視した方法ということになる。今回結果では、抽出単語の中に「中国」や「東日本」が出てきているが、「東日本」という単語が景気が悪いという意味を示す理由は時期によるものである。３年という短い期間だが、三つの期間くらいに分けて、時系列方向の統計的有意性も検証しないと的中精度はかなり落ちてしまう。
すなわち、単語には短期的に影響を与える単語と長期的に影響を与える単語の２種類

がある。例えば「タイの洪水」の場合に株価や景気を調べたら、その時は「タイの洪水」が景気を下げた単語として出てきても、「タイ」という単語は長期的に見ると景気を悪くする単語ではない。そこは時系列で分けて、時系列的な統計的有意精度も見ていく必要があると思う。常にいい、あるいは悪いといわれている語でないと予測には非常に使いづらいわけで、そこをチェックする。時期を分けるというのは時系列的な時期で分ける必要があるわけではなく、例えば「東日本」の語が頻繁に出てくるのならば、最初の 100 単語がでてきた時期のデータまでを使うとか、逆に頻度の低い単語ならば、1 年間で出てきた個数は統計的に有意なものはとれないので、そうした場合は例えば 2 年ごとに、時系列的に安定しているかどうかをチェックする。頻度をとるのは、時系列的な安定性を測る上でよく使われる方法である。

・特徴表現の抽出結果は実感にかなったピッタリしたものが出てきていると思うが、有意表現の抽出結果は逆に悪化しているような気がする（データの絶対数が圧倒的に足りないことが大きな原因だった可能性がある）。

４－２　今後の分析業務推進の方向性と将来像

４－２－１　短期的な視野（当面の推進策の可能性）

３－２、及び３－３における分析業務の結果を踏まえ、景気の現状をより早く把握（ナウキャスト）することを短期的に実現するための可能性を考察する。

４－２－１－１　分析業務から得られた評価、知見

・特徴表現抽出について
－景気ウォッチャー調査のコメントテキストからは、景気動向と関連する特徴表現を比較的うまく抽出できている
－ただし、景気ウォッチャー調査のコメントテキストをブログやツイートなど他のデータ群の文章にそのまま当てはめることには留意が必要

・景気ウォッチャー調査のコメントテキストと景気判断との関係性について
－上記の適切に収集されたテキストデータは景気判断の説明力を有する
－予測に強く効くキーワードの数は必ずしも多くない（数十個程度）
－ただし十分なサンプル数が確保できないと、オーバーフィッティングの問題が発生する

・景気判断の予測可能性について
－Twitter やインターネット上のニュース、経済専門情報サイトのテキスト等について、ある程度の数のキーワードの発生頻度を随時追跡しておくことによって、

ネット上での景気判断を随時予測できる可能性（検証作業は別途必要）

４－２－１－２　４－２－１－１を踏まえた当面の推進策（作業項目）の大まかなイメージ

※具体的なデータ入手先、対象データ項目、並びに具体的な作業項目やフローの特定などについては別途検討が必要

(1) 景気ウォッチャー調査のコメントテキストから景気判断の予測に強く効くキーワードを特定、抽出（数十個）

(2) (1)の月別出現回数データを収集

(3) Twitter やニュースサイト、経済専門情報サイトなどのテキストにおける、(1)のキーワードの月別出現回数データを収集(収集期間は要検討)

(4)(2)と(3)を比較し、(3)の中から説明変数として採用するキーワードを抽出

(5)(4)のデータと景気ウォッチャー現状判断 DI 値との相関関係を推計、評価
(学習期間、予測期間の設定も要検討)

(6)(5)で得られた推計式について、今後入手する Twitter やニュースサイト、経済専門情報サイトなどのテキスト量を外挿し予測値(当該月の現状判断 DI 値)とする

以上はあくまで可能性のある推進策のイメージのひとつであり、実際には Twitter やニュースサイト、経済専門情報サイトなどのテキストにおけるキーワードの出現回数データの中身などによっては方針を変更することもありえる。

４－２－２　中期的な視野（今後１～２年程度）

一方、今後１年あるいは２年程度の期間をかければ検討可能かもしれない事項としては、以下の２点が考えられる。

４－２－２－１　景気指標推定の速報性向上に寄与するビッグデータの収集並びに分析の本格始動

景気の現状をより早く把握することを目的として、上記４－２－１に示した業務を（短期的な視野に基づく簡易な形態ではなく）本格的な態勢で実施し、実用化するための準備を今後１～２年間で進めていくことが極めて重要と考えられる。

具体的には、ある程度（例えば過去 10 年ぐらい）の期間にわたる Twitter やネットニュース、経済専門情報サイトなどのテキストデータを収集／取得し、今回分析業務で実施したような景気に関連する重要なキーワードの抽出と、それらと景気指標との関係性を推計し検証していくことであり、このための予算や要員、実施体制など各種リソー

ス面での準備も並行して検討していくことが有効とみられる。

４－２－２－２　ビッグデータ景気動向指数(DI)（仮称）のテスト作成

景気判断を目的としたビッグデータの分析対象は決してひとつに限られるわけではない。単一の指標、分析結果に頼ることなく、複数の分析結果をもとに既存の公式統計の弱点（特に速報性）を補完し、より多面的な判断基準を提供するツールとしてビッグデータ分析は極めて有効と考えられる。

内閣府が毎月作成、公表している景気動向指数(DI)は、採用系列のうち改善している指標の割合のことで、景気の各経済部門への波及の度合いを表す。

これと同様の考え方に基づき、ナウキャスティング（速報性）を重視したビッグデータ分析結果のデータ系列を例えば５つ程度採用し、各指標の改善／悪化の割合を計測するアイデアをここでは提示する。この指標の最大の長所はデータ（指標）の公表頻度が月次よりも多い（例えば週次）ことであり、速報性の点では既存の公表統計と比較して格段に改善することが期待される。

現時点では候補５つを明示することはできないが、例えば４－２－１に挙げたテキストデータをもとにした景気現状判断指標や、東大物価指数のうち売上高に関する指標は採用系列の候補として有力とみられる。採用系列を公募し、１年間程度の共同研究をベースに採用系列としての信頼性等を検証の上テスト公表していくという方式も考えられる。

４－２－３　長期的な視野（将来像）

一方、より長期的な観点から将来の姿を考えると、以下のようなトレンド到来の可能性を念頭に置き、それに対応したビッグデータの分析について検討をしておくことが望まれる。

なお、これらはいずれも主要な進歩のドライバーは技術であり、技術の進歩に伴いデータ量の増加や種類の多様化が進み、それによってビッグデータ分析とその利活用が進展していくという前提（ストーリー）に立っていることに注意を要する。すなわち、実際には技術以外にも規制、政策、経済情勢など各種のパラメーターが影響を与えることが考えられる。

４－２－３－１　ビッグデータの「個人化」「ミクロ化」

情報通信技術の急速な進歩や、センシングデバイスなどによる計測技術の発展により、経済活動を行う様々な場面で、人間（個人、あるいは個人の集合体である群）の行動を

計測し記録する技術の観測範囲および精度が飛躍的に向上している。
一例をあげれば、現在はまだ発展途上の技術ではあるものの、将来的に音声の分析が多くの場所でできるようになると、人間の行動に関する情報の把握度合いがより多様化、精緻化することが考えられる。すなわち、あらゆる場面における個人の経済活動がデータの形で把握でき、それが時系列データとして蓄積されていくと、特に個人消費を中心としたビッグデータ分析の精度が格段に向上する可能性がある。

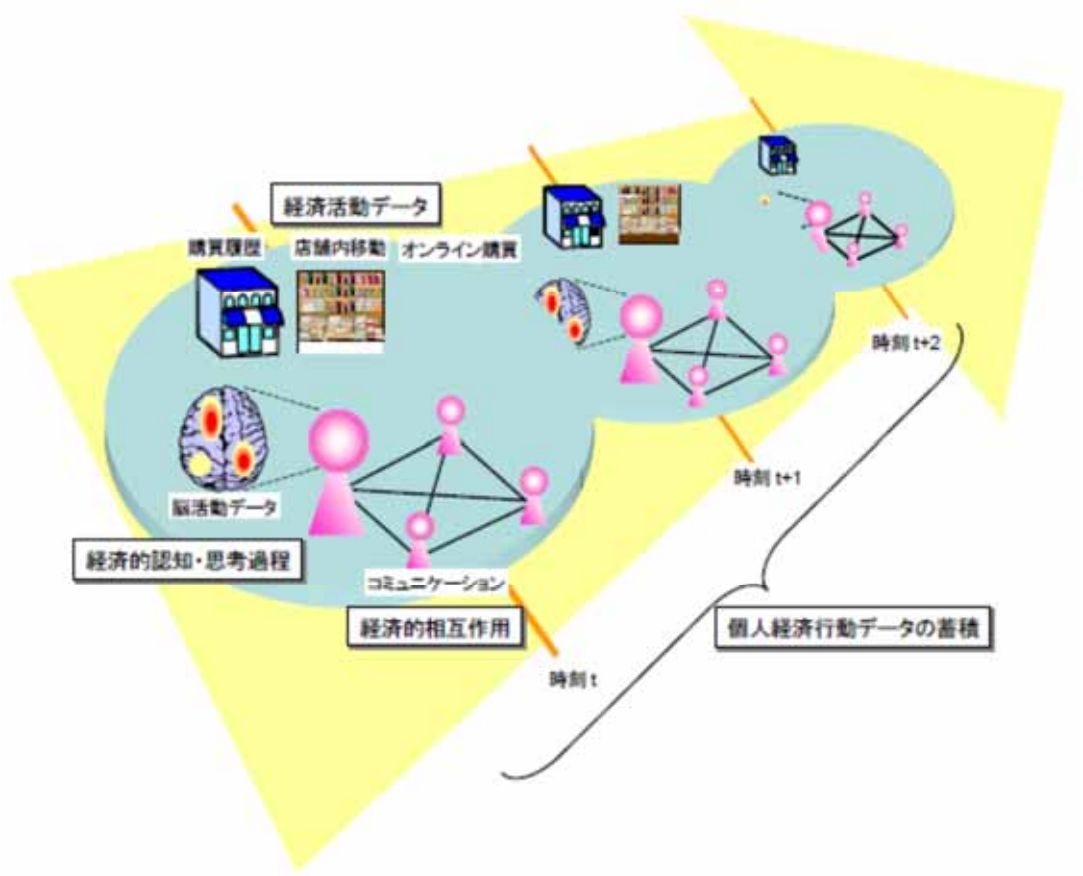

図 4－2－1　個人経済行動の観測・モデリングの詳細化のイメージ
（出所）「分野横断型科学技術アカデミック・ロードマップ報告書」p.132
横断型基幹科学技術研究団体連合 平成 21 年 3 月

４－２－３－２　ビッグデータの「マクロ化」「リアルタイム化」

一方で、消費者レベルの経済活動にとどまらず、企業部門、政府部門を含めた各種のデータが集約されていくと、マクロレベルでの経済活動の把握がビッグデータを通じて可能になっていくものとみられる。

特に、企業間のモノやサービス、金融取引に関するデータについては、その取引頻度が非常に細かいレベルまで把握することが可能になりつつある。これがひいてはビッグデータのリアルタイム化を推し進めることにつながるものとみられる。

もっとも、景気判断や経済活動の動きを把握する目的からすれば、日次ベースよりも細かいデータを利用する可能性は、現在既に利用可能な為替レートや株価等以外については当面は小さいとみられるが、実態経済の場面においてリアルタイム化が進むというという認識は持っておいた方がよいものとみられる。

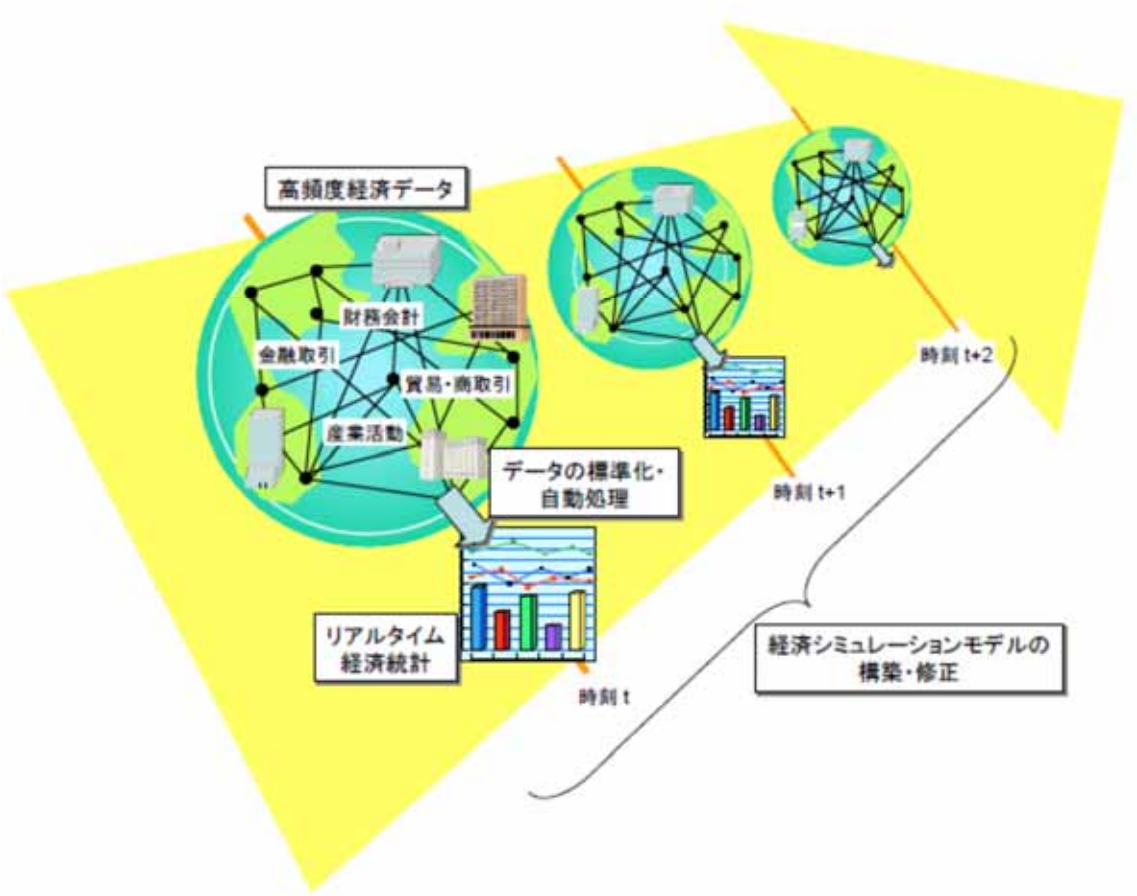

図 4−2−2　経済現象観測の大規模化・リアルタイム化のイメージ
（出所）「分野横断型科学技術アカデミック・ロードマップ報告書」p.134
横断型基幹科学技術研究団体連合 平成 21 年 3 月

４－２－４　検討委員会における主な指摘事項

今後の分析業務推進の方向性と将来像については、検討委員会においても委員から多くの指摘、コメントを頂いた。

上記４－２－１～４－２－３の視野に加え、以下の指摘事項も踏まえた推進策の検討と実行が望まれる。

・短期、中期についてこの順番でやっていくのは基本的に賛成。ただし、短期的視野（４－２－１）の中で、景気ウォッチャーに出てくるコメントの単語とツィッター、ブログに出てくるコメントの単語とではかなり傾向が違う、あるいは全くひっかからない可能性もある。それに対してはいくつか技術的な方法があると思う。
例えば少し古いが Google が出している Google コーパスというデータがある。約 200 億文（2007 年時点）の日本語のウェブサイトにあるテキストを集めており、その中で例えば「消費税」という語ならば、「消費税」と隣り合って出てくる単語、もしくは近くに出てきそうな（関係のありそうな）単語をリストアップしたデータがある[3]。それを使うことによって、景気ウォッチャー調査で出てきたある単語 A に対し、単語 A が直接なくても、A と一緒に出てきそうな単語 A ’ や A ”が Twitter やブログテキストに出たら、スコアを 0.5 くらい(半分くらい)与えるようなことをして、一緒に出てきそうな単語を数える方法が良く使われる。そのような工夫が必要になると思う。

・今回分析では、「たばこ」や「さかな」等の個別の単語を説明変数に使っているが、個別の単語の分析ではなく、単語をグループ化して分析するという視野が中期的には必要になると思う。例えば数十種類の単語をグループ化して「生鮮食料品」とし、「生鮮食料品」の合計回数を説明変数とする。説明変数を減らしてオーバーフィットを防ぐという方法が有望ではないかと思う。
ビッグデータ景気動向指数の目的にもよると思うが、例えば景気指標が上がったとき、タバコという単語が増えたから指標が上がったということをユーザーがみて嬉しいかといえば、そんなことはないと思う。さかなという単語が上がったときに、消費税という文脈の中で生鮮食料品がどうか、などの視点でとらえた方がビッグデータ景気動向指数のテキスト分析の良さがでると思う。

・ビッグデータ景気動向指数を実用化するためのアイデア（問題点と発展の方向性）が２点ある。
① 問題点:一次データを内閣府自身が集めて蓄えるのは非常に大変だと思う。少なくとも中期的な目標の時点までは民間かアカデミックなところが解析した、回数が出ている二次データを利用した方が楽だと思う。ただし、例えば民間の二次データを使ってそれをユーザーに公表する際に、自身が一次データから分析した場合と比較して、データが操作された、バイアスのかかったものではないかという懸念、危険性はあることには注意が必要。一次データから二次データ

[3] Google 日本語コーパスに関連する情報は以下のサイトに詳しい。
http://www.arg.ne.jp/node/3365 (Google、日本語 N-gram データを公開)
http://www.gsk.or.jp/catalog/ (言語資源協会における言語資源カタログ- Web 日本語 N グラム第 1 版を参照)

の生成は比較的単純な計算なので、アルゴリズムを作ってしまえば自動的にできる。しかし、泥臭いところ、すなわちツィッターなりブログの膨大なテキストをインターネット上から拾ってきて、一次データを貯めるところが一番大変な作業。

② (ビッグデータ景気動向指数を)実装した場合の有望な点として、たとえ予測がある程度難しいとしても、過去の月次で出ている DI 値を週次、日次ベースで補う指標（過去のデータの間を埋められるような指標）を作るだけでも非常に価値があると思う。

ただしその際に技術的に問題となるのは、もともとは月一回しかなかったものを日や週単位でどうやってデータを類推していくかという、その方法。これについては、2段階の作業が必要になると思う。

a.最初に、DI 値を推測するために別の細かい、テキストデータではなく数値データ（例えば株価等）を使って DI 値との関連性を見る。これはある程度できると思う。

b.次に、数値データによって推定された細かい（階段を埋めた）データを Twitter やブログのテキストデータで推測していく。

この方法を取れば、最終的にはビッグデータによってより細かい頻度の DI 値を指標として作ることができるのではないかと思う。間を埋めるための他の数値指標を使うことを、今後 1-2 年でやっていくことを考慮する方がよいと思う。

・消費税、増税、駆け込みなどのキーワードを手作業でピックアップした資料を記者ブリーフィングなどで用意されているとのことなので、これに少し加えるだけで情報量が非常に増えると思う。

例えば、「消費税」と、あるポジティブな文脈、あるいはネガティブな文脈で一緒に出てきそうな単語は何か、を示すだけで、例えば最初は消費税が駆け込みだけと一緒にでてきたのが、その後は安売りなどの語にフォーカスがシフトした、などの事象が見える可能性がある。一緒に出てくる単語の頻度を時系列的にみせるだけでも、実は一般的な着目の変化を図示化できる可能性があるかもしれない。

そして、そのフォーカスがシフトしたことによって、景気動向が具体的にどう変わっていったのかがわかると予測により結びつくと思うが、その前段階（フォーカスシフト）だけであってもビッグデータの価値はあるのではないかと思う。これだけピックアップできているワードがあるのであれば、その可能性は高いと思う。

・既存の経済統計とこのビッグデータ景気動向指数のようなものは、両方に「バイアス」があると思う。どっちにどういうバイアスがあるかということを比較、相対的に整理しておくことが良いと思う。（物価指数については多少そういう分析があったが。物価指数以外の景気動向についても同じような議論があって良い）

・東大物価指数で典型的に出てきたが、月次より短い頻度で分析ができるようになると、新しい概念が必要になる。具体的には、前年同日比というのがあったが、主に曜日の調整した後での前年同日での概念に対応する言葉がまだない。従ってそういった新しい概念設定が必要になってくると思う。

・もう少し手軽にできることとしては、毎月の景気ウォッチャー調査で、今は手作業でやっていると思うが「ビッグデータ分析で出た今月の頻出語句」のようなコーナーを作り、今月はこういう言葉が出ているというような付録をつけるだけでも有用。世の中に対しそういうことができるんだなという認知をしてもらう、将来的にこのような研究が増えるような土台を作るのも良いと思う。

・短期的にも、景気ウォッチャー調査だけではなかなかわからない、なぜ景気が良くなっているのか、悪くなっているのかをテキストから推察することができると思う。
単語を（トピック別に）紐付ける = 分類するという方法において、機械的な分類の仕方でよく使われるのが LDA(Latent Dirichlet Allocation)だが、これを使うと文章の中にどういう成分が入っているのかが分かってくる。
すなわち、もともとの文章の中から単語の出現頻度によって、景気が良いという文章、悪いという文章がそれぞれどうなっているのか、が分かってくる。次に LDA をかけることによって、何で景気が悪くなっていたのかというひとつひとつの抽出項目を取り出すことができる。次に抽出した項目の時系列変化を見ると、今景気が悪くなっているのは何が原因なのかというのが分かってくる。
それが分かってくると「タイの洪水」が仮に景気が悪い原因として出てきた時には、これは長期的なものではないので長期的に危惧するものではないという判断ができる。一方、長期的な分類として出てくる単語が抽出成分が強く出てくると、なかなか景気を回復させることは厳しい、という判断もできる。

４－３　分析業務の強化のためのキャパシティビルディング

４－３－１　データ収集、取得

ビッグデータの分析にはまず「データ」の収集、取得（利用権を含む）が必須である。経済・景気の分析に役立つ可能性があるデータの種類については第２章（有識者ヒアリングの実施結果）で記載した通りだが、これらの中には有償、無償の両方があり、さらに一次データ（生データ）、二次データ（一次データを加工して統計量などのデータにしたもの。これらの多くは有償）の２種類がある。

これらのうちどのデータを収集、取得するかは分析業務の方針に関わる重要な事項で

ある。しかし、データが有償でそのデータ価格が高いことが分析業務実施のボトルネック（しかも最大の障害）になることは十分に考えられる。

４－１－１でも記述した通り、上記の場面においてデータ価格、及びデータの取り扱いに付帯する各種条件の交渉は勿論必須である。しかし、今回分析業務等から得られた知見や情報に鑑みると、例えば Twitter のテキストデータを加工した二次データは複数の企業が取り扱っている模様であるし、加工データのタイプもバリエーションがある。分析目的に合致したデータ種類の選択肢と、それらの費用対効果を徹底的に検討していくことによって、データ価格の問題をクリアし自部門内のデータベースとして蓄積、活用を図っていく道を探っていくことが大切である。

また、４－２－２でも記述した通り、景気分析に資する可能性のあるビッグテータの種類は必ずしもひとつとは限らないし、分析の対象を１種類のデータに固定・集中する必要もない。景気分析をより総合的、マクロ的な観点から行うには、自部門にデータベースとして保有、活用するビッグデータの種類やポートフォリオ管理についても構想立案が必要になってくるものと考えられる。

４－３－２　分析のための要員

ビッグデータの分析を実施するには、基本的に

・自組織内で行う

・自組織以外の外部にアウトソースする

のどちらかとなる。自部門で分析手法の設計をしつつも、分析業務の作業はアウトソースするという上記両者のいわば中間的な形態は勿論考えられるが、いずれにせよ自部門の中でビッグデータの分析に関する知見が極めて薄いと、たとえ業務をアウトソースするとしてもアウトソース先の選定や評価において困難に直面する可能性がある。

従い、少なくとも自組織内にビッグデータの分析に関する知見を持つ人員を最低１名、業務の継続的遂行を考慮するならば組織全体の規模にもよるが２～３名の人員を確保するようなスタンスが望まれることになる。

しかし、ビッグデータの分析業務を始めようとしたときに最初から分析要員が自組織内に在籍していることは現実的な想定ではない。従い、実際の場面においては

・分析要員を中途採用などの方式で採用する

・民間企業等からの出向の形で分析要員を確保する

などの方策が考えられる。ただし、後者においては出向期間が解けると自組織内に知見・ノウハウが残らなくなってしまうため、後者の場合は出向者の確保と同時に自部門内で分析要員を養成することを並行して行うことが現実的な選択肢のひとつになるか

もしれない。

有識者ヒアリングにおいては、ビッグデータ分析（自然言語処理）のためのプログラミング言語の習得には半年程度を要し、アカデミアと共同研究するのが一番技術の習得が早い（水野貴之氏）という指摘を頂いている。

もっとも、今回分析業務においてはテキストマイニング（自然言語処理）とデータマイニング（相関等の分析）という２種類の分析を段階別に実施している。これがもしビッグデータを利用した景気分析手法（手順）のプロトタイプのひとつだとすれば、この両方の技術をひとりで、しかも短期間でゼロから習得し、実務に使えるようにするには上記指摘の半年程度の期間ではまず困難であろうとみられる。

自前での要員育成を目指すには、より中長期的視点での対処が必要不可欠であろう。

４－３－３　有識者等とのネットワーク、産学官協力

４－１－２で示した通り、ビッグデータ使用による景気・経済の分析にかかる知見・知識・ノウハウはかなり学術分野に偏在している可能性がある。従って、今回分析業務における有識者ヒアリングや検討委員会の機会をきっかけに、今後知見を有する多方面の有識者と意見交換、さらには共同研究の可能性を探るなどの手段によって自部門の知見・知識・ノウハウをさらに高めていくことは重要である。

さらに、民間企業部門にも多数とみられるいわゆる「データサイエンティスト[4]」との関係強化も知見や具体的分析手法に関する知識を広げる上で非常に有効であろう。景気や経済をビッグデータを利用して分析することに興味、関心を持つ人物や組織から成るフォーラムは現在のところまだ構築されていないように見受けられるが、そのような産官学各分野が結集した意見交換、協力の場が形成されれば、最新動向をはじめとした知見、知識の向上にも大いに役立つと考えられる。

（以上）

[4] データの分析手法に詳しく、分析ツールの操作に長け、ビッグデータから主にビジネスに活用できる情報を引き出す目利きとして活躍する人物の総称。数学、統計学、計算機科学、情報工学、パターン認識、機械学習、データマイニング、データベース、可視化などの分野との関わりが深く、さらに金融、経済など特定の分野への造詣が深いことが多い。